Linear Correlation in LM's Compositional Generalization and Hallucination

Anonymous ACL submission

Abstract

The generalization of language models (LMs) is undergoing active debates, contrasting their potential for general intelligence with their struggles with basic knowledge composition (e.g., reverse/transition curse). This paper uncovers the phenomenon of *linear correlations* in LMs during knowledge composition. For explanation, there exists a linear transformation between certain related knowledge that maps the next token prediction logits from one prompt to another, e.g., "*X lives in the city of*"→ "*X lives in the country of*" for every given *X*. This mirrors the linearity in human knowledge composition, such as *Paris*→ *France*. Our findings indicate that the linear transformation is 1) resilient to large-scale fine-tuning, 2) generalizing updated knowledge when aligned with real-world relationships, 3) but causing hallucinations when it deviates. Empirical results suggest that linear correlation can serve as a potential identifier of LM's generalization. Finally, we show such linear correlations can be learned with a single feed-forward network and pre-trained vocabulary representations, indicating LM generalization heavily relies on the latter.[1]

1 Introduction

What knowledge do language models (LMs) learn beyond memorizing the training data? The generalization ability of LMs is undergoing an active debate. Optimists claim that LMs might have the capability in entirely novel tasks with their emergent behavior (Wei et al., 2022) by scaling-up parameters, while pessimists argue that LMs struggle with composing simple knowledge (Peng et al., 2024a; Thomm et al., 2024), such as reverse or transition curses claiming that LMs cannot even simply compose knowledge by reversing or transiting (Berglund et al., 2024; Zhu et al., 2024).

While macroscopically investigating how skills emerge in language models remains challenging, we can gain microscopical insight from the generalization behavior on the smallest learning unit, next token prediction (NTP). We unveil an interesting linear correlation between logits of related NTPs, such as *City*→*Country*, from the **source knowledge** like logits of $F_{City}(X) = \text{NTP}(\text{"}X \textit{ lives in the city of}\text{"})$ to the **target knowledge** like logits of $F_{Country}(X) = \text{NTP}(\text{"}X \textit{ lives in the country of}\text{"})$. Between logits in knowledge subdomains (e.g., $\{Paris, Shanghai, \cdots\}$ for $F_{City}(X)$), we can fit a linear transformation (W, b) that well approximates $F_{Country}(X) = W \cdot F_{City}(X) + b$ for any X as the input. To fit the transformation, we sample numerous output logits from prompts with arbitrary inputs Xs as shown in Figure 1. Then, (W, b) is fitted with partial logit pairs and tested on the rest. The Pearson correlation coefficients for evaluation reflects the inherent relations of knowledge in the real world, with high correlations in cases like *City*→*Country* and low correlations in cases like *City*→*Gender*.

Examining W, we find that its weights mirror the linearity in the knowledge composition of humans. In the *City*→*Country* case, the W assigns high weights to real-world (City, Country) pairs such as *Paris*→*France*. In other words, probability $P(F_{Country}(X) = France)$ is correlated with $P(F_{City}(X) = Paris)$. However, there also exists counterfactual weights learned in W, for instance, the weight fit in W for (*Indianapolis*, *India*) is much higher than the correct (*Indianapolis*, *USA*). We say W is *precise* when W assigns high weights for the correct knowledge pairs. W's precision is generally low for knowledge pairs with low correlations, but a high linear correlation also does not guarantee high precision. This motivates us to explore the connection between 1) such linear correlations, 2) W's precision, and 3) LM's compositional generalization. Importantly, if the same W

[1]The code will be released for reproduction.

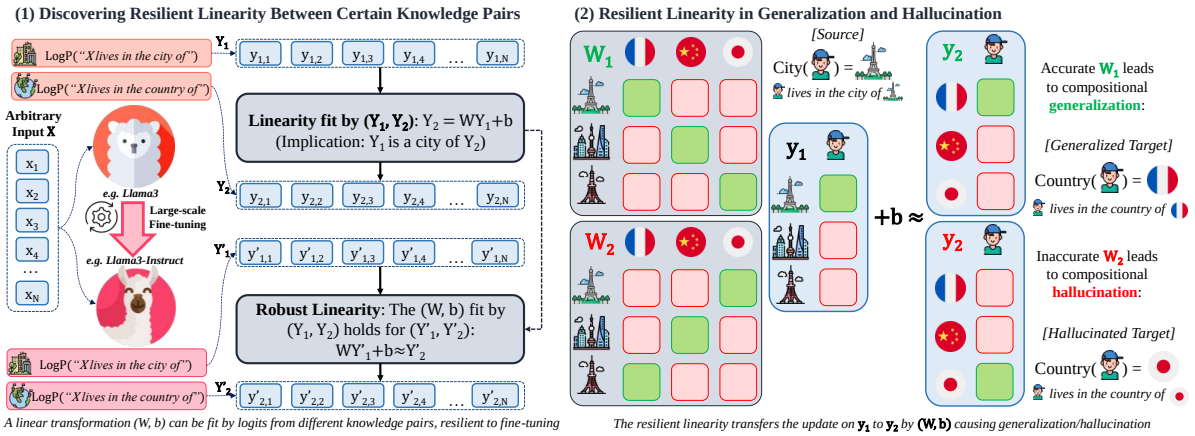


Figure 1: Demonstration of our main discoveries. 1) We can fit a linear transformation between the output of source and target knowledge prompts, which is resilient against fine-tuning. 2) Updating the source knowledge will generalize to the target one via resilient linearity, causing compositional generalization/hallucination.

and b also fit the parameter updates after gradient propagation, then learning source knowledge will simultaneously update the target knowledge.

We begin with one-step parameter updates, fine-tune the LM with a piece of source knowledge, and then check the gradients on the source and target knowledge. When the linear correlation between the source and target knowledge is high, we find W capable of estimating the gradients on the target knowledge based on the source gradient. We then extend the comparison to LMs before and after large-scale post-training, which shows W fitted before post-training to retain the estimation ability for the LM after post-training. Thus, W between highly correlated knowledge is found resilient against gradient propagation, which consistently plays an important role in generalization.

To assess the role of linear correlation in LM generalization, we test source–target knowledge pairs with varying correlation intensity and W precision. Generalization succeeds only when both are high, suggesting LMs struggle with non-linear generalization, limiting the effectiveness of simple fine-tuning (Cohen et al., 2024). High correlation with low W precision can cause compositional hallucinations (e.g., $P(City) = \textit{Indianapolis} \rightarrow P(Country) = \textit{India}$). These correlations, observable before fine-tuning, help diagnose potential faults in LM knowledge composition.

Finally, we explore the linear correlation's origin and hypothesize that vocabulary representations are key. Even when we remove the LM's complex internals (position embeddings, self-attention, etc.) and use only a mean-pooling layer plus a single feedforward network, the model still learns to compose knowledge from few paired texts (e.g., $F_{City} = \textit{Paris}$ paired with $F_{Country} = \textit{France}$). The simplified archecture shows similar generalization performance as the original Transformer. However, altering lexical mappings (e.g., *Paris*→*Japan*) disrupts this ability, underscoring the critical role of vocabulary representations. Our contributions are as follows,

- We unveil the linear correlation between the LM's output logits for related knowledge.
- We find such linear correlation existing between gradients and resilient against training, which connects it to compositional generalization and hallucination of LMs.
- We attribute the cause of linear correlation between NTPs to vocabulary representations.

2 Related Works

2.1 Language Model Interpretation

Language models (LMs) (Achiam et al., 2023; Team et al., 2024; Groeneveld et al., 2024; Dubey et al., 2024) are gaining widespread attention across various fields due to their strong performance on a variety of tasks, like reasoning and knowledge retrieval. However, the black-box nature of (neural) LMs hinders human's understanding of their working mechanism. Various methods have been developed to interpret LM behavior by analyzing its parameters and intermediate representations. Several works suggest that LMs store knowledge inside the feedforward layers (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022), which are used in a key-value matching manner to map inputs into related knowledge (Geva et al., 2022). Some parameters are also found to perform certain relational transformations for the LM (Todd

et al., 2024; Zhang et al., 2024), known as the task representations (Lampinen and McClelland, 2020). For certain subsets of relations, LMs have been unexpectedly found to encode knowledge in a linear manner (Hernandez et al., 2024), suggesting a potential role of linearity in their understanding of relational structures. However, it remains unknown how the LM understands the transformation between relations. Our work shows the linearity between the output from several relation pairs given the same input.

2.2 Model Generalization

The power of modern deep neural networks lies in their remarkable ability to generalize effectively to unseen inputs. However, the exact mechanisms through which these models achieve generalization remain poorly understood. For instance, in the context of knowledge editing, numerous research studies have observed that standard fine-tuning methods for updating knowledge often struggle to meet critical objectives simultaneously . On one hand, they fail to prevent unintended modifications to unrelated knowledge. On the other hand, they frequently fall short of ensuring that logical deductions based on the updated knowledge are properly incorporated (Cohen et al., 2024; Zhong et al., 2023). Previous research has proposed various metrics and methods to measure and predict generalization in deep neural networks. However, these approaches don't cover the perspective of correlation in model generalization (Yu et al., 2022; Garg et al., 2022; Kang et al., 2024).

2.3 Hallucination Detection

Hallucination remains one of the most significant challenges in the deployment of language models (LMs) (Zhang et al., 2023; Huang et al., 2024). Numerous studies have explored approaches to predict and mitigate this issue. For instance, some prior works utilize trained classifiers to identify hallucinations (Jiang et al., 2024; Quevedo et al., 2024; Chen et al., 2024). Another method involves detecting hallucinations by clustering semantically similar responses and calculating entropy across these clusters (Farquhar et al., 2024). Additionally, the MIND framework has been proposed to exploit the internal states of LMs during inference, enabling real-time hallucination detection (Su et al., 2024). Moreover, formal methods guided by iterative prompting have been employed to dehallucinate LM outputs (Jha et al., 2023). RAG has also been used to detect and correct hallucinations in LM (Mishra et al., 2024). Our study presents an innovative approach to predicting hallucinations, different from existing methodologies, by leveraging the correlation.

3 Discovering Linear Correlation

3.1 Preliminary and Motivation

Next Token Prediction. Neural language models have been scaled up to numerous parameters but can still be understood as a mapping function among vocabulary representations $V \in \mathbb{R}^{\#\text{Vocab} \times d}$. We denote the embedding of the word X as $V_X \in \mathbb{R}^d$. For an input word sequence, such as "*X lives in the city of*", the embeddings of the involved words will be processed with other components in the LM $\theta_{\neg V}$ (positional embedding, self-attention networks, etc.) to encode the input context as $C = F([V_X, \cdots, V_{of}]) \in \mathbb{R}^d$. Most[2], if not all, LMs tie the input and output vocabulary embeddings together (Press and Wolf, 2017) to use the dot product $C \cdot V_Y$ as the logit of Y for the next token prediction. Finally, the vocabulary-wise dot products are normalized by a softmax layer to represent the probability of a certain token (Y for example).[3]

$$P_{\theta_{\neg V}}(Y|[V_X, V_{lives}, \cdots, V_{of}]) = \frac{e^{C \cdot V_Y}}{\sum_{Z \in \text{Vocab}} e^{C \cdot V_Z}} \quad (1)$$

For a subset of all possible sequences that follow the template "*X lives in the city of*" and takes arbitrary X as the input, we can view template representations $[V_{lives}, \cdots, V_{of}]$ as constant to map a variable X (V_X) with the *City* relation.

$$P_{\theta_{\neg V}}(Y|[V_X, V_{lives}, \cdots, V_{of}]) = P_{\theta_{\neg V}, [V_{\text{lives}}, \cdots, V_{\text{of}}]}(Y|V_X) \quad (2)$$

Here, the encoding function $F_{\text{City}} = F(\cdot|[V_{lives}, \cdots, V_{of}])$ (subscript *City* denotes the semantics of constant representations) affects the final probabilistic distribution by mapping V_X to C near vocabulary embeddings of cities, such as $V_{Paris}, V_{Shanghai}, V_{Tokyo}$.

Motivation: Linearity in Relation. Some knowledge like $F_{\text{CityToCountry}}$ ("*X is a city in the country of*") are found linear (Hernandez et al., 2024) between vocabulary representations, which

[2] In Appendix E, we empirically show our conclusion also holds for an parameter untied LM - Mistral (Jiang et al., 2023)

[3] We omit the discussion of potential bias terms, multiple token input for simplification and reading fluency.

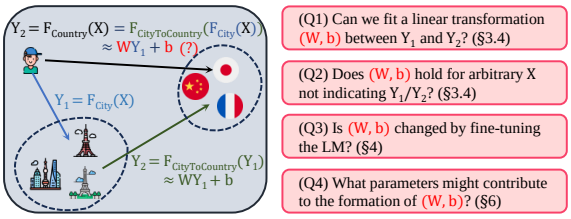


Figure 2: Our hypothesis and questions about how LMs compose knowledge by learning (W, b).

means F can be well approximated by (W, b) s.t. $C = WV + b$. While not all mappings have such an interesting property, this phenomenon indicates the potential for LMs to compose knowledge in their parameters.

Knowledge Composition. There exists compositional relations between knowledge such as F_{Country} ("*X lives in the country of*") can be composed by other relations as $F_{\text{CityToCountry}}(F_{\text{City}})$ since one's residential city **(source knowledge)** indicates one's residential country **(target knowledge)**. Suppose the LM applies $F_{\text{City}}(V_X)$ to map V_X close to a city embedding like V_{Paris}), then may the LM learn (W, b) inside parameters and perform $F_{\text{Country}}(V_X) = F_{\text{CityToCountry}}(F_{\text{City}}(V_X)) = WV_{Paris} + b = V_{France}$? While the hypothesis can be made for non-linear relations in the composition as well, we emphasize the linearity as it corresponds to the key-value matching (Geva et al., 2021) behavior of Transformers. The linear transformation can be simply performed by a feedforward network activated by self-attention.

Motivated by the potential role of linearity in compositional knowledge, we conduct experiments to validate the hypothesis that LMs learn such linear transformation inside the parameters for composition. The roadmap of our exploration is presented in Figure 2, with questions we will answer in the following sections. We will demonstrate that

- Such (W, b) exists for logits prompted from certain related knowledge pairs, which is applicable to arbitrary inputs, not necessarily indicating a known output (§ 3.4).
- Such linearity stays resilient against large-scale fine-tuning, which guarantees the LM's generalization to compositional knowledge (§ 4).
- Such linearity can be highly attributed to the vocabulary representations. (§ 6).

3.2 Method and Evaluation

We search for the potential linear transformation between pairs of source and target knowledge. Continuing with the $(F_{\text{City}}, F_{\text{Country}})$ example, the transformation will be established between $C_{\text{City},X} = F_{\text{City}}(V_X)$ and $C_{\text{Country},X} = F_{\text{Country}}(V_X)$. We then decode the two representations by the LM head to produce logits $\text{LogP}_{\text{City},X}$ and $\text{LogP}_{\text{Country},X}$ both in shape $\mathbb{R}^{\#\text{Vocab}}$.

$$\text{LogP}_{\text{City},X} = C_{\text{City},X} \cdot V; \text{LogP}_{\text{Country},X} = C_{\text{Country},X} \cdot V \quad (3)$$

As the dot product with V is linear, the potential linearity holds after the transformation. We can calculate $W \in \mathbb{R}^{\#\text{Vocab} \times \#\text{Vocab}}$ and $b \in \mathbb{R}^{\#\text{Vocab}}$ for the transformation between logits. We learn (W, b) for logit transformation (rather than hidden state) to improve the interpretability of the fitted W. For example, a high weight in $W_{(\text{France,Paris})}$ indicates a correct understanding of knowledge composition.

In practice, only a subdomain[4] D of the LM's large vocabulary is meaningful for the predicted logits, such as $D_{\text{City}} = \{$*Paris*, *Shanghai*, *Tokyo*, $\cdots\}$ for $\text{LogP}_{\text{City}}$ and $D_{\text{Country}} = \{$*France*, *China*, *Japan*, $\cdots\}$ for $\text{LogP}_{\text{Country}}$. Thus, we are more interested in the submatrix of W for these meaningful words. Our main experiments will focus on those values in W representing the linear transformation $W_{(D_{\text{City}}, D_{\text{Country}})}$ between such output subdomains. The specific procedure to build such subdomains is presented in Appendix D.

Based on the prior discussion above, we propose a method to search for the linear transformation. We first build a comprehensive input set by enumerating a large number of words in the LM's vocabulary. While some words might indicate clear answers for certain knowledge (e.g., *Obama* as X for F_{Country}), most of them do not (e.g., *Lit* as X for F_{Country}). We feed all inputs to different prompts and collect the output logits such as $\text{LogP}_{\text{City}}$ and $\text{LogP}_{\text{Country}}$. For each logit, we only keep dimensions for words falling inside the corresponding output vocabulary domain such as D_{City} and D_{Country}. By collecting numerous (10K in our experiments) logit pairs, we fit the linearity transformation (W, b) with half of those pairs $(\text{LogP}_{\text{City, X}}, \text{LogP}_{\text{Country, X}}), \forall X \in$ Train and then evaluate the transformation on other half of pairs $(\text{LogP}_{\text{City, X}}, \text{LogP}_{\text{Country, X}}), \forall X \in$ Test.

Evaluation. With (W, b), we make predictions on the test pairs, $\text{LogP}_{\text{Country, X}} = W \cdot \text{LogP}_{\text{City, X}} + b, \forall X \in$ Test. We compare the predictions with the

[4] General subdomain size is $\sim$ 100, listed in Appendix B.

test references using the correlation metric, Pearson correlation, to evaluate how similar the logits are distributed. The evaluation is applied by both instance-wise (averaged over instance-wise logits on $x_1, x_2, \cdots \in X$) and label-wise (averaged over label-wise logits across instances on $d_1, d_2, \cdots \in D$). Our main content focuses on the label-wise Pearson correlation as we find that the global bias b plays an important role in the instance-wise predictions as shown in Appendix C. The label-wise evaluation eliminates the effect of b, which concentrates on the logit correlation matrix W. Another advantage of instance-wise correlation is that the metric is calculated based on distributions with the same dimensions. Besides, the correlation weights on different labels also reflects how well each label is approximated by the linear transformation.

3.3 Experiment Setup

While numerous compositional knowledge pairs exist in natural language, we focus on large families of knowledge composition that share a commonality. Specifically, we include four large families, attribute, cross-language, simile, and math. We include 111 prompts in our experiments to cover broad knowledge fields as listed in Appendix B.

Attribute. Updating one attribute of a subject will affect other attributes as well. The *City→Country* example illustrated before shows such a compositional relation in the spatial attribute. Another example is: $F_{\text{CEO}} \rightarrow F_{\text{Company}}$

Cross-language. Knowledge update is expected to be propagated to other languages, like the *English → French* example: $F_{\text{City}} \rightarrow F_{\text{Ville}}$.

Simile. Simile builds equivalence among the attributes between objects. Thus, updating a simile to a subject will result in updating the corresponding attribute. An example is $F_{\text{SameColorAsFruit}} \rightarrow F_{\text{Color}}$.

Math. Numbers have denser compositional relations with each other, such as "*X+1=*2"→"*X+2=*3". We involve the four basic arithmetic operations in experiments to explore the knowledge composition in math. An example is $F_{\text{X+1}} \rightarrow F_{\text{X+2}}$.

For each family, we include the results on $10 \sim 20$ knowledge prompts in the main content to save the length and place the others in Appendix E. Table 1 showcases some examples of prompts and domains.

Family	Prompt	Domain Examples
Attribute	"X lives in the city of" "X lives in the country of"	Paris, Vienna France, Austria
X-Lang.	"X vit dans la ville de" "X lebt in der Stadt von"	Paris, Vienne Paris, Wien
Simile	"X has the same color as" "X's color is"	Apple, Banana Red, Yellow
Math	"X+1=" "X*2="	1, 2, 3, 4, 5 2, 4, 6, 8, 10

Table 1: Examples of prompts and domains in different families of knowledge composition.

We include different LLaMA-3 (Dubey et al., 2024) models in our experiments with parameter numbers of 1B, 3B, 8B, and 70B. We include the before and after post-training LMs for the evaluation of linear correlation's resilience against fine-tuning. The variance in the model scale allows us to explore the generality and scaling law of the linear correlation inside different models. We include LMs from the same family to ensure consistency in tokenization and training data, allowing for a more controlled and convenient discussion. Results on other LMs for broader generality are also included in Appendix E.

3.4 Experiment Results

Figure 3 presents the main findings on linear correlations between NTP logits, with full results available in Appendix E.

Attribute Strong correlations emerge among semantically related attributes (e.g., *city*, *country*, *continent*) and within thematic clusters such as spatial, language, job, and family-related attributes. In contrast, unrelated attributes (e.g., gender vs. continent) show weak correlation, suggesting that LMs can disentangle unrelated factors and mitigate bias. However, some expected relationships (e.g., *Language→Continent*, *CEO→Company*) show weak correlation, indicating gaps in knowledge structuring.

Cross-language Moderate cross-lingual correlations exist, but same-concept alignment across languages (e.g., English-Chinese) is weaker than within-language semantic links. This likely reflects English dominance in LLaMA-3's training, further examined using Aya in Appendix G.

Simile Simile analysis (Table 8) shows moderate correlation between objects and their attributes, suggesting LMs can connect figurative expressions with underlying semantics.

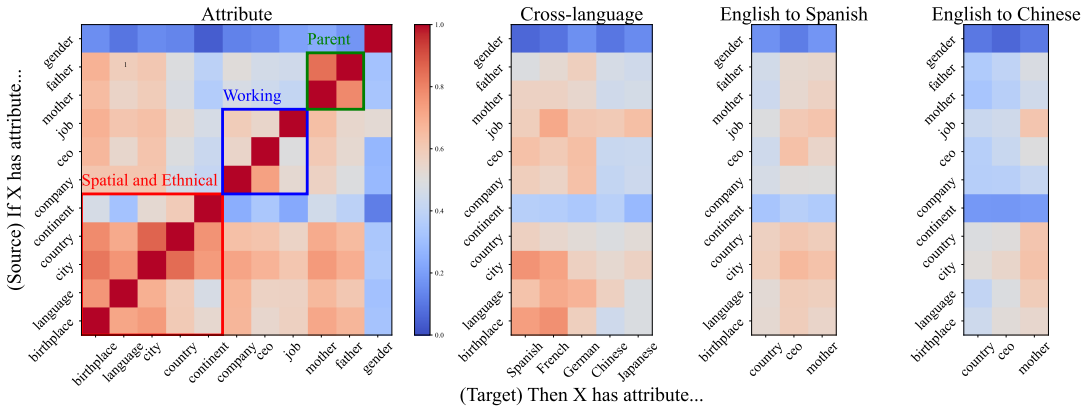


Figure 3: The linear correlation between NTP logits of `llama-3-8b`.

Relation Pair	Hit@Top-N					
	Influenced Target			Influencing Source		
	1	3	5	1	3	5
City→Country	0.42	0.45	0.48	0.67	0.74	0.78
CEO→Company	0.09	0.09	0.14	0.05	0.05	0.08
City_{en}→City_{es}	0.91	0.91	0.92	0.67	0.74	0.78
City_{en}→City_{zh}	0.10	0.13	0.16	0.09	0.11	0.15
Fruit→Color	0.25	0.38	0.47	0.38	0.50	0.54
Food→Taste	0.28	0.50	0.62	0.14	0.36	0.43
X+1→X+2	0.00	0.50	0.60	0.10	0.30	0.50
X+1→X*2	0.10	0.40	0.50	0.10	0.30	0.70

Table 2: The precision of composition built up in W.

If City =	Then Country =
Shanghai	**China**, Italia, Albania, USSR, Korea
NYC	**USA**, USSR, UAE, China, CCCP
Oslo	CCCP, **Norway**, Kosovo, Israel, Oman
Seattle	Uruguay, Serbia, Kosovo, Romania, Slovenia
Indianapolis	India, Indonesia, France, Iraq, Netherlands
If $X+1=$	**Then $X+2=$**
1	1, **2**, 4, 6, 3
2	2, **3**, 4, 5, 7
3	3, 6, 5, **4**, 7
4	4, 0, 2, 1, 10
5	5, **6**, 8, 7, 9

Table 3: Cases of top-influenced tokens pairs in target knowledge.

Math Strong correlations appear among outputs of the same math operator (Figure 7), but further analysis reveals that such correlation may not reflect precise computation.

3.5 W can Reflect Real-world Knowledge

The weight matrix W can reflect compositional relations between source and target domains. Thus, we check whether the W's weights reflect real-world knowledge. Specifically, for each token in the source (target) domain, we check whether the top-influenced (influencing) outputs (inputs), i,e. have the highest weights, are consistent with the real world. We use Hit@Top-N ($N = 1, 3, 5$) metric to evaluate whether there is a correct influenced (influencing) token with a top weight. In experiments that require closed reference, we test subset of knowledge pairs with clear causal relations (e.g., *City→Country* rather than *Mother→Father*). The experiment scale is relatively small due to the sparsity of knowledge composition references.

We analyze the W precision of 2 cases from each family with the results presented in Table 2. We find the LM have a relatively precise understanding of the correlation between certain highly correlated attributes like *City→Country*. In transformation matrix W, 42% cities learn the top-1 weight with their influenced countries and 67% countries have a correct top-1 influencing city. For less correlated *CEO→Company* attributes, W is also imprecise, suggesting the failure to reflect the real-world causal relation. This phenomenon is also observed in the cross-language family for the strongly correlated *English→Spanish* and the weakly correlated *English→Chinese*. However, a strong correlation does not necessarily guarantee a precise W as shown in the math cases.

In Table 3, we showcase some top-influenced tokens in the attribute and math correlations to visualize how W reflects real-world correlations. In the *City→Country* case, some cities like *Shanghai* and *NYC* are matched with the correct countries while some others like *Oslo*, *Seattle*, and *Indonesia* are not. The *Indonesia→India* case indicates a bias introduced by superficial similarity into the weights in W. The math cases show the correlation is dominated by identical mapping. While the LM tries to model a correct correlation as many secondly influenced numbers are correct, the domination of identical mapping hinders the precision of W to reflect real-world correlation. More cases in Appendix H further support our observation and extend it to non-causal correlations like parent names.

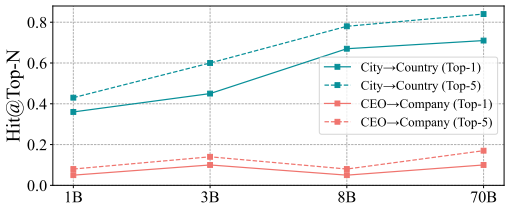


Figure 4: The scaling-up of the precision of W.

Relation Pair	Logit Correlation	Grad. Correlation
City→Country	0.89	0.79
CEO→Company	0.55	0.47
City_{en}→City_{es}	0.70	0.79
City_{en}→City_{zh}	0.58	0.46
Fruit→Color	0.48	0.46
Food→Taste	0.47	0.47
X+1→X+2	0.93	0.87
X+1→X*2	0.73	0.66

Table 4: Gradient correlation between relations.

3.6 Is W More Accurate in Larger LMs?

Our discovery indicates that W reflects real-world correlations between knowledge. We check whether the weights of W are more in line with the real world knowledge for larger LMs. Thus, we plot the Top-N metric of correlations in LLaMA-3 of different model sizes in Figure 4. In the *City→Country* case, we can view a clear scaling-up of W's precision, showing that larger LMs also better organize their knowledge. However, *CEO→Company* is shown to be a hard causal relation, whose W's precision is not successfully scaled up by a larger model size.

4 Resilient Correlation against Training

4.1 Gradient Correlation

As many weights in W reflect the real-world correlation, we hypothesize that they are resilient against gradient propagation because they capture inherent patterns that resist change. Thus, we check whether the gradients on related knowledge prompts are also linearly correlated. We choose to train `llama-3.2-3b`[5] with a common setup for large LMs (AdamW (Loshchilov and Hutter, 2019) with 5×10^{-6} learning rate).

The gradient correlation results are presented in Table 4, demonstrating a correlation between the gradients on different NTP logits. Specifically, with the gradient ∇LogP on a logit, we can estimate the gradient on a correlated logit by $W \cdot \nabla\text{LogP}$. If W is a precise one, the learned knowledge will also be correctly synchronized by knowledge composition caused by W, such as Shanghai→China. Thus, the correlation between gradients indicates a potential mechanism behind how LMs compose learned knowledge.

[5]We select 3B LM for efficiency, which shows a similar correlation behavior as the 8B LM in Appendix E.

Corr.	Prec.	Relation Pair	Generalization (Random)
High	High	City→Country	53.70% (0.78%)
		Country→Continent	50.93% (20.00%)
		City_{en}→City_{es}	39.10% (0.41%)
High	Low	X+1→X+2	0.00% (9.09%)
		X+1→X*2	8.18% (9.09%)
Low	Low	Fruit→Color	11.60% (6.67%)
		Food→Taste	19.44% (10.00%)
		CEO→Company	4.34% (1.00%)
		Language→Continent	23.65% (20.00%)
		City_{en}→City_{zh}	2.49% (0.41%)
		City_{en}→City_{ja}	4.60% (0.41%)

Table 5: The ratio of successful generalization in relations with different linear correlation and W precision.

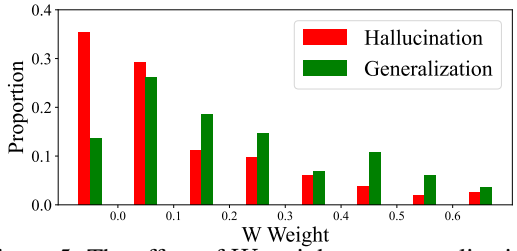


Figure 5: The effect of W weights on generalization.

4.2 Correlation after Large-scale Post-training

We extend our investigation from single updates to large-scale post-training of LMs, testing whether the fitted linear transformation (W, b) from a pre-trained LM (e.g., `llama-3-8b`) still applies to its post-trained counterpart (e.g., `llama-3-8b-instruct`). Comparing correlation matrices before and after post-training (Figures 3 and 8), we observe that the linear correlation remains robust despite extensive optimization, demonstrating that W is resilient to large-scale post-training. This highlights the persistent role of linear correlation in LM generalization. Moreover, as detailed in Appendix F, this correlation resilience is even more pronounced in larger LMs.

5 Correlation is a Double-edged Sword

The potential role of the linear correlation in knowledge composition inspires us to investigate how W implicates the generalization of LMs. We anticipate the resilient correlation to be a two-edged sword, which propagates knowledge with a precise W but also exacerbates hallucination with a imprecise W. For validation, we continue to fine-tune the `llama-3.2-3b` model.

City	Reference	Generalized	W_{ref}	W_{gen}	W_{max}
Shanghai	China	China	0.50	0.50	0.50
NYC	USA	USA	0.58	0.58	0.58
Copenhagen	Denmark	Denmark	0.47	0.47	0.47
Karnataka	India	India	0.34	0.34	0.56
Indianapolis	USA	India	−0.05	0.15	0.17
Dresden	Germany	Israel	0.04	0.13	0.15
Canberra	Australia	Canada	0.04	0.10	0.10
Helsinki	Finland	Sweden	0.42	0.11	0.42

Table 6: Generalization cases in *City→Country*.

We examine how generalization depends on correlation and the precision of W. Table 5 compares relation pairs with varying levels of correlation and precision (excluding the rare case of low correlation and high precision). Results show significant generalization only occurs when both are high. We test more low-correlation pairs to confirm their poor generalization, suggesting linear correlation predicts generalization. When correlation is high but W is poor, the LM hallucinates as expected. In the *X+1→X+2* case, learning "*X+1=*N" generalizes strongly to "*X+2=*N" as in Figure 3.

We analyze the role of W weights in both generalized and hallucinated cases of *City→Country*. As shown in Figure 5, higher W weights on ground-truth pairs generally promote successful generalization by enabling more effective gradient propagation, consistent with the gradient correlation trends in Table 4. However, high W weight alone does not guarantee generalization. Case studies in Table 6 reveal, 1) Correct generalizations typically align with top W weights; 2) Exceptions arise when target entities like *India* have high prior probability, enabling generalization even with lower W, whereas low-prior entities like *Finland* fail to generalize despite strong gradient correlations.

The hallucinated cases can also be divided into two categories. 1) Wrong W weight, a major reason of compositional hallucination. The fifth to seventh cases show low ground-truth W weights, consequently leading to unsuccessful generalization. These cases also show a relatively low maximal weight in W, which is potentially an indicator of imprecise W weights. 2) Low prior probability. The last case shows a high W weight between *Helsinki* and *Finland* but the prior probability of *Finland* is much lower than *Sweden*, which results in a compositional hallucination. This is a mirror case of the *Karnataka→India* generalization.

6 What Causes the Correlation?

To explore the source of linear correlations, we hypothesize that vocabulary representations—beyond pre-training data or architecture—drive this behavior, as similar patterns persist across various LM architectures (Appendix E). To test this, we conduct an ablation study replacing LLaMA-3's intermediate layers with mean pooling and a basic feed-forward network (Figure 6). The model is trained on 1,024 paired texts (e.g., "X lives in Shanghai" / "X lives in China") for 1,000 epochs to capture compositional relations. It is then evaluated on 128 unseen subjects (e.g., *"Z lives in Shanghai"*) over 2,000 epochs to see if it can infer related knowledge (e.g., *"Z lives in China"*).

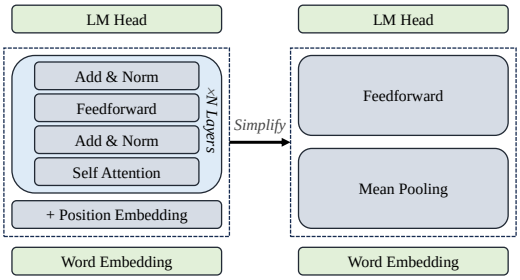


Figure 6: We replace the deep intermediate layers of LMs with an initialized shallow bag-of-word network.

Mapping	Generalization
(City→Country)	
Shanghai, Tokyo, Paris→China, Japan, France	97.66%
Shanghai, Tokyo, Paris→Japan, France, China	22.66%
S, T, P→C, J, F	36.72%
(Country→Continent)	
China, France, Canada→Asia, Europe, North	78.12%
(CEO→Company)	
Elon, Andy, Tim→Tesla, Amazon, Apple	58.59%
(+1→+2)	
1, 2, 3→3, 4, 5	9.38%

Table 7: Generalization effects of vocabulary mappings.

Several test results are presented in Table 7, showing a consistent generalization performance as the initial deep Transformer model. When we switch the correspondence between cities and countries or keep only the first letter, the generalization behavior disappears, which highly attributes generalization to the vocabulary representations.

7 Conclusion

This work reveals a new perspective on how LMs generalize by knowledge composition. We detect linear correlations between related NTP logits, which are resilient to training. Such correlations are found to propagate updates on knowledge to one another, leading to compositional generalization and hallucination. We attribute the correlation to vocabulary representations with an ablation study. Future topics include further investigating the formation of such linear correlation and utilizing it for generalizable learning.

Limitations

As a pioneering study, our work focuses on uncovering the phenomenon of linear correlations in language models but leaves several key aspects for future research:

- **Theoretical Explanation** We do not provide a formal theory explaining why resilient linear correlations emerge. Future work can explore the underlying model architectures, optimization dynamics, and linguistic structures that drive this phenomenon.
- **Data Distribution Effects** Our study does not systematically analyze how training data influences the formation of these correlations. Investigating which data properties contribute to their emergence could provide deeper insights.
- **Identifying Correlated Knowledge Pairs** While we observe linear correlations in specific cases (e.g., city–country), we do not establish a general method to predict what knowledge pairs exhibit this property. Future work can develop theoretical or empirical criteria for identifying such relationships.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2024. Hallucination detection: Robustly discerning reliable answers in large language models. *Preprint*, arXiv:2407.04121.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. *arXiv preprint arXiv:2005.02433*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. 2022. Leveraging unlabeled data to predict out-of-distribution performance. *Preprint*, arXiv:2201.04234.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 30–45. Association for Computational Linguistics.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Evan Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15789–15809. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152. IEEE.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. On large language models' hallucination with regard to known facts. *Preprint*, arXiv:2403.20009.

Katie Kang, Amrith Setlur, Dibya Ghosh, Jacob Steinhardt, Claire Tomlin, Sergey Levine, and Aviral Kumar. 2024. What do learning dynamics reveal about generalization in llm reasoning? *Preprint*, arXiv:2411.07681.

Andrew K. Lampinen and James L. McClelland. 2020. Transforming task representations to perform novel tasks. *Proc. Natl. Acad. Sci. USA*, 117(52):32970–32981.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

Binghui Peng, Srini Narayanan, and Christos H. Papadimitriou. 2024a. On limitations of the transformer architecture. *CoRR*, abs/2402.08164.

Letian Peng, Chenyang An, and Jingbo Shang. 2024b. Correlation and navigation in the vocabulary key representation space of language models. *CoRR*, abs/2410.02284.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 157–163. Association for Computational Linguistics.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 2024. Detecting hallucinations in large language model generation: A token probability approach. *Preprint*, arXiv:2405.19648.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *Preprint*, arXiv:2403.06448.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. 2024. Limits of transformer language models on learning to compose algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15894–15939. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy

Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. 2022. Predicting out-of-distribution error with the projection norm. *Preprint*, arXiv:2202.05834.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6228–6247. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael I. Jordan, Jiantao Jiao, Yuandong Tian, and Stuart Russell. 2024. Towards a theoretical understanding of the 'reversal curse' via training dynamics. *CoRR*, abs/2405.04669.

A Results for Main Content

In Table 8, Figures 7 and 8, we illustrate the experiment results for the main content because of the length limitation. Table 8 demonstrates the correlation between simile objects and attributes. Figure 7 shows a high correlation between math calculation results. Figure 8 presents the linear correlation between logits from knowledge before and after large-scale post-training, which is compared with the results in Figure 3 to conclude a resilient linear correlation against fine-tuning. The cross-tuning results for simile and math families are presented in Table 9 and Figure 9, which validate a resilient correlation against post-training for highly correlated knowledge pairs. Note that the concepts in *Object* (apple, t-shirt, laptop, chair, washing machine, etc.) for simile relations do not directly indicate attributes, so they are not used for evaluation when reference is required.

Due to content limitations, we focus on describing the phenomenon rather than fully explaining its origins. We hope our findings serve as a foundation for further research into the mechanisms and implications of linear correlations in LMs.

B Prompts and Setups

Table 10 shows the statistics of the prompts used in our experiments. Tables 11, 12, 13 further list all the specific prompts used in our experiments. The domain size of most prompts is around 100 expect for some domains with limited valid outputs like *Continent* and *Color*.

C Instance-wise Correlation

Figure 10 shows the instance-wise Pearson correlation evaluation results on different knowledge pairs. We use attribute correlation as an example to show that the target knowledge of each instance can be well approximated by a linear transformation on the source knowledge. In the main content, we demonstrate the label-wise correlation because we find the bias term b to dominate the prediction on many knowledge pairs that are poorly linear correlated (especially in gradient). Some target knowledge is predictable with only the prior probability from bias even without any linear indicator. Thus, the label-wise correlation is a more challenging metric by eliminating the effect of b with a better reflection of how the source knowledge influences the target knowledge.

D Subdomain Building Procedure

To build the subdomains, we do not simply collect the top predictions from the next token predictions because many predictions are introduced by the frequency and similarity bias (e.g., stop words like *the*) in the next token representation space (Demeter et al., 2020; Peng et al., 2024b). Instead, we enumerate the common answers by `gpt-4o` (Achiam et al., 2023) and search engines. Then we keep the first tokens of the tokenization for these answers which are not subwords. For example,*China* will be represented by *China*, *South Korean* will be represented by *South*, and *Brunei* will be dropped because it is tokenized into [*Br*, *unei*]. We exclude subwords because they cannot identify complete semantics without tokens after them. The discussion for subword cases is included in Appendix J.

E Whole Attribute Results and Extra Discussion

From Figure 11 to Figure 19, we present the whole correlation matrices inside all kinds of LMs for different prompts. We can observe the existence of correlation behavior among different LMs. While the correlation in different LMs behaves differently, some common pairs like *City→Country* hold for all different LMs. Also, models from the same LLaMA-3 family tend to behave in a similar way. We can also observe many spurious correlations such as *Hobby→Mother*, which generally have low causal relations in the real world. Larger LMs tend to be better at disentangling such kind of spurious correlations as the smallest `GPT2-Medium` model shows a much stronger correlation. In Figures 18 and 19, Table 14, we illustrate that the 3B model has a similar correlation behavior as the 8B one.

F More Resilient Correlation in Larger LMs

In Figure 20, we find the linear correlation is more resilient against fine-tuning by plotting the correlation before and after post-training in 1B, 3B, 8B LLaMA-3 LMs as we find more strong correlations in larger LMs. In Figure 21, we also plot the correlation matrix between logits from `mistral-7b-v0.3` before and after post-training, which supports the existence of resilient linear correlation in LMs with vocabulary representation untied.

Table 8: Correlation between gradients on simile objects and attributes.

Relation Pair	Fruit-Color	Food-Taste	Gem-Color	Name-Country	Animal-Size
Correlation	48.42	46.68	27.46	67.35	59.59
Relation Pair	Object-Genre	Object-Heat	Object-Size	Object-Price	Object-Color
Correlation	77.68	73.11	71.41	72.87	70.87

Figure 7: The linear correlation between NTP logits of `llama-3-8b` in math operations.

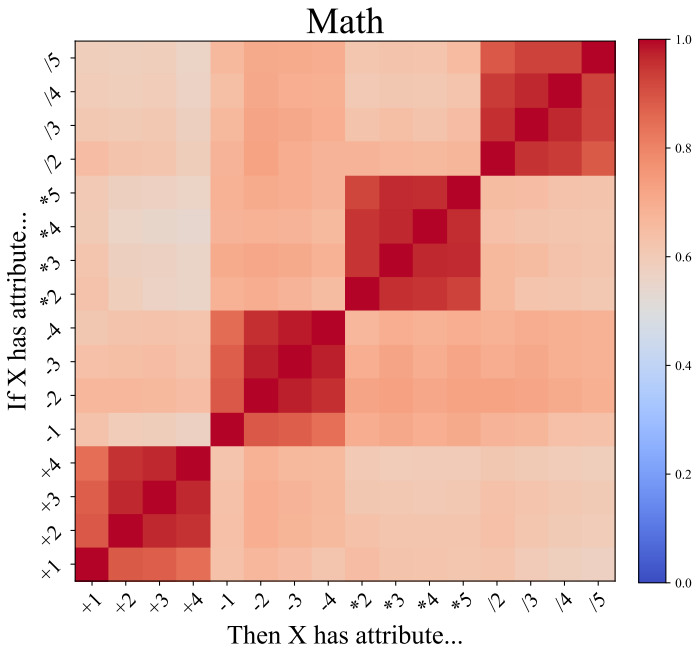


Figure 8: The linear correlation between NTP logits of `llama-3-8b` **before and after large-scale post-training**.

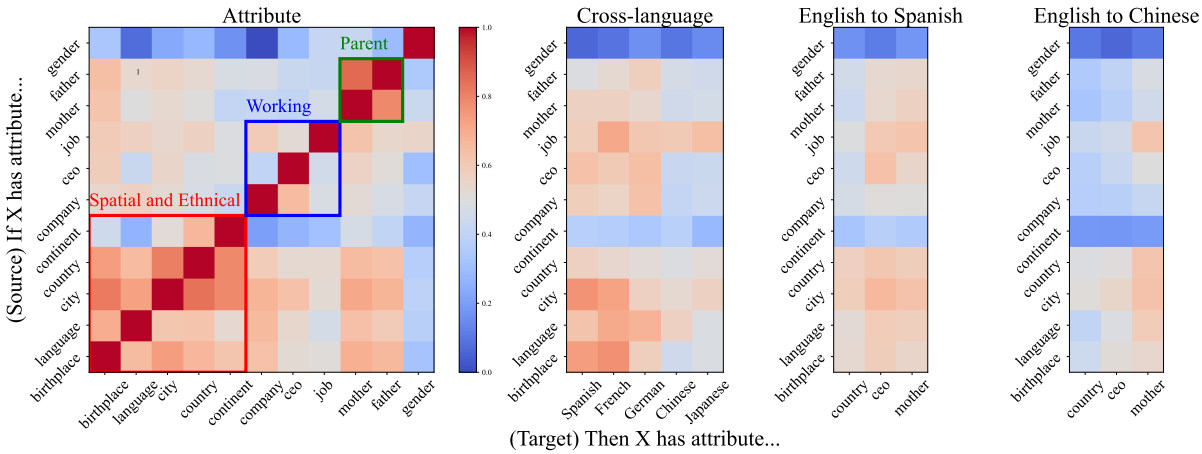
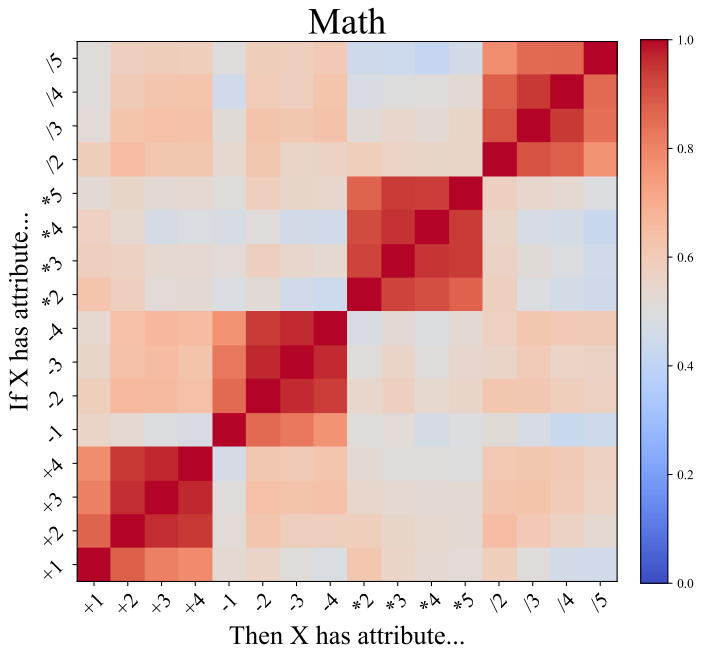


Table 9: Correlation between logits on simile objects and attributes **before and after large-scale post-training**.

Relation Pair	Fruit-Color	Food-Taste	Gem-Color	Name-Country	Animal-Size
Correlation	44.11	37.06	33.66	67.30	49.65
Relation Pair	Object-Genre	Object-Heat	Object-Size	Object-Price	Object-Color
Correlation	72.03	63.75	66.13	71.09	66.27

Figure 9: The linear correlation between NTP logits in math operations **before and after large-scale post-training**.



Template	Domain Size
Attribute	23
Cross-language	$11 \times 5 = 55$
Simile	17
Math	$4 \times 4 = 16$
Total	111

Table 10: The statistics of prompts in different families.

	Knowledge	Template	Domain Size
Attribute	birthplace	"*{} was born in the city of*"	242
	city	"*{} lives in the city of*"	242
	country	"*{} lives in the country of*"	128
	continent	"*{} lives in the continent of*"	6
	language	"*{} speaks the language of*"	217
	company	"*{} works for the company of*"	100
	landmark	"*{} lives near the landmark of*"	100
	ceo	"*{} works for the CEO called*"	101
	mother	"*{}'s mother's name is*"	100
	father	"*{}'s father's name is*"	100
	job	"*{}'s job is*"	105
	personality	"*{}'s personality is*"	100
	pet	"*{}'s pet is*"	100
	sport	"*{}'s favorite sport is*"	102
	food	"*{}'s favorite food is*"	104
	drink	"*{}'s favorite drink is*"	102
	gender	"*{}'s gender is*"	3
	vehicle	"*{}'s preferred mode of transportation is*"	51
	color	"*{}'s favorite color is*"	15
	music	"*{}'s favorite music genre is*"	100
	hobby	"*{}'s favorite hobby is*"	101
	flower	"*{}'s favorite flower is*"	97
	vacation	"*{}'s favorite vacation spot is*"	101

Table 11: Templates used in our experiments (Part 1: Attribute).

G Multilingual LM

Figure G demonstrates the cross-lingual correlation of the multilingual LM, `aya-expanse-8b`, which outperforms LLaMA-3 in multilingual tasks but still lags behind in English (Üstün et al., 2024). The results show Aya to have a stronger cross-lingual correlation between knowledge pairs, especially in Chinese and Japanese. On Latin language, Aya's advantage becomes smaller because these languages share quite a lot entity names with English and LLaMA-3 can benefit from its English ability to complement the weakness in multilingual ability.

H Extra Case Study

We provide extra cases for analysis in this section. In Table 15, we provide massive cases on the influencing cities in the *City→Country* knowledge composition, which shows that the LM establishes correlation between many (City, Country) pairs such as (*Edinburgh*, *Scotland*), (*Islamabad*, *Pakistan*), and (*Afghanistan*, *Kabul*). Tables 16 and 17 showcase the correlation between knowledge pairs that do not have a clear reference. Taking parent correlation as an example, Table 16 shows correlation of parent names from the same ethnicity like (*Chen*, *Mei*) and (*Santiago*, *Sofia*).

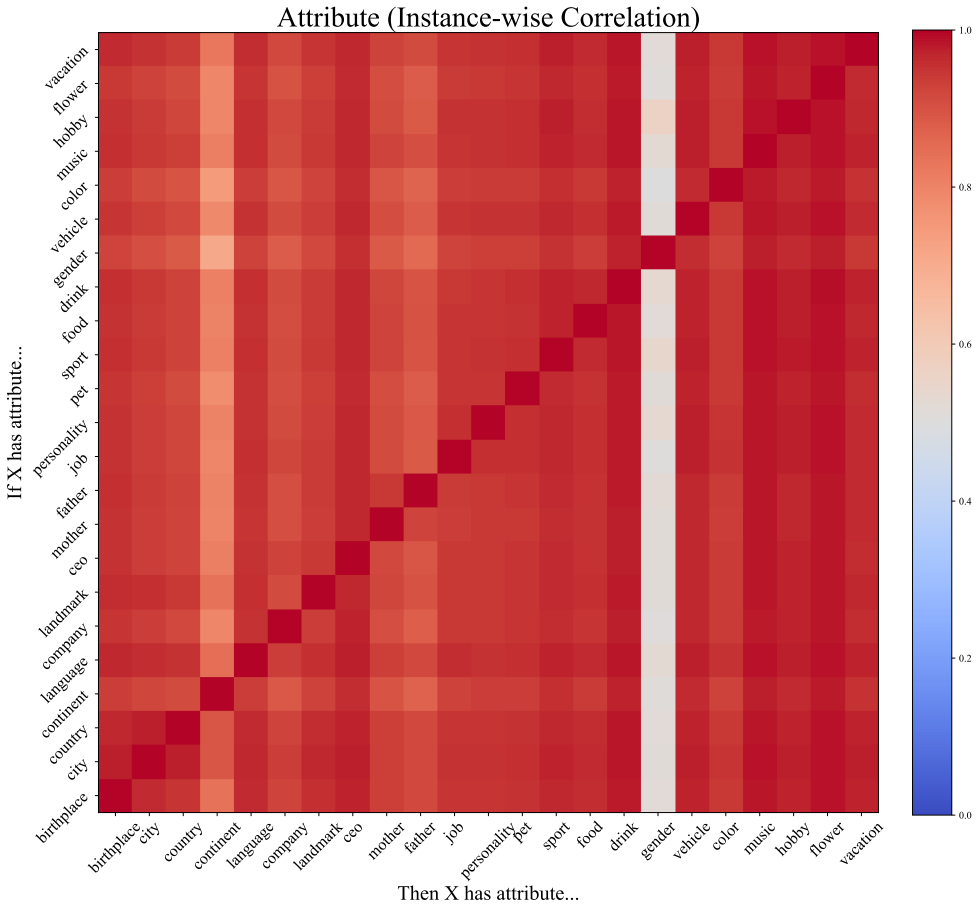


Figure 10: The instance-wise correlation between NTP logits of `llama3-8b` (attribute as an example).

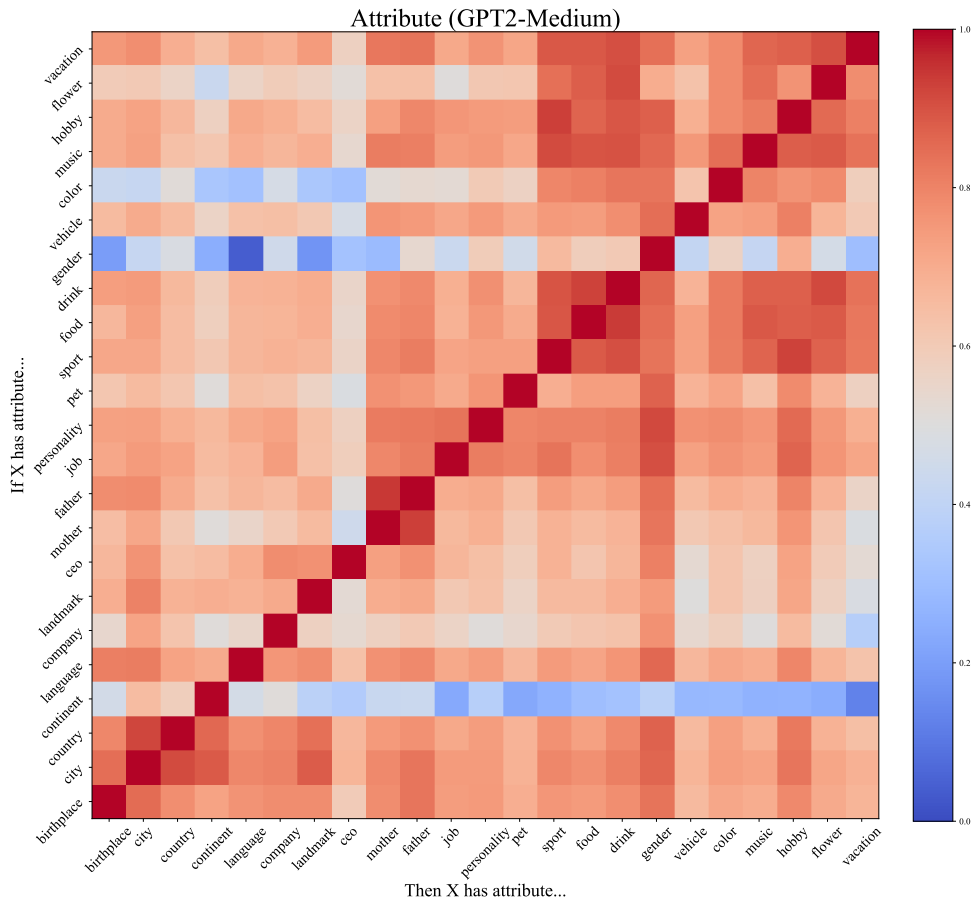


Figure 11: The attribute correlation between NTP logits of `gpt2-medium`.

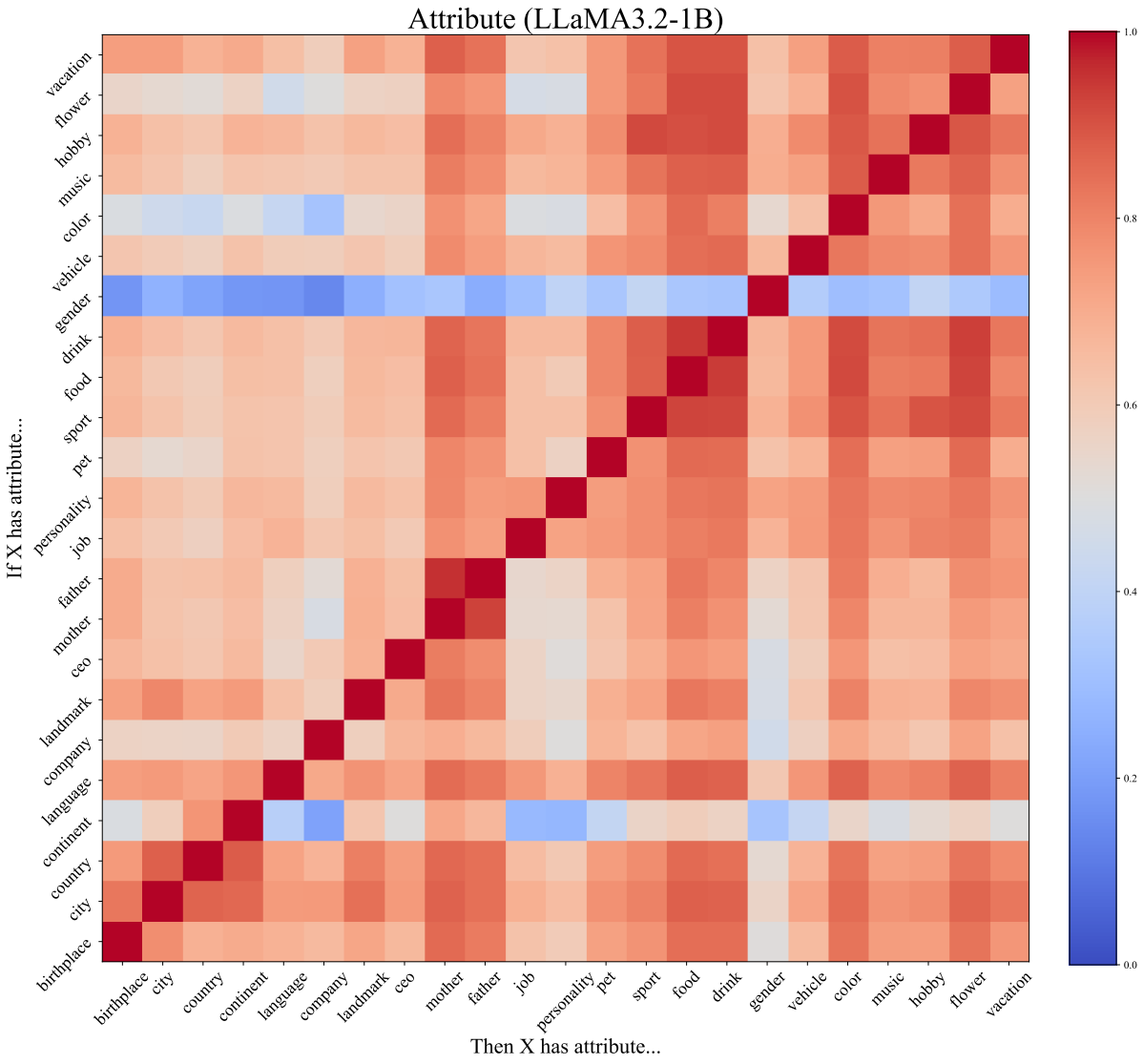


Figure 12: The attribute correlation between NTP logits of `llama-3.2-1b`.

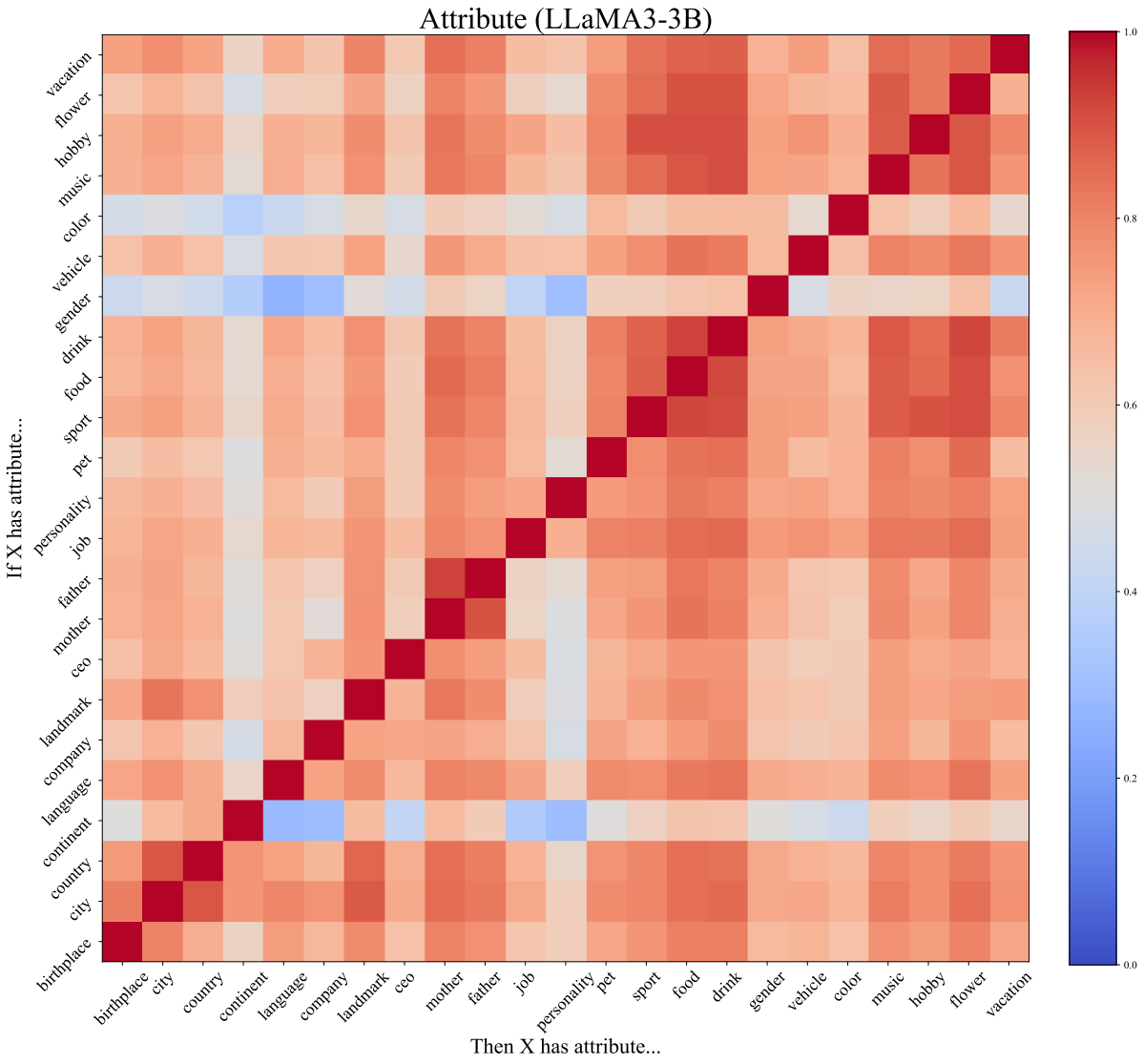


Figure 13: The attribute correlation between NTP logits of `llama-3.2-3b`.

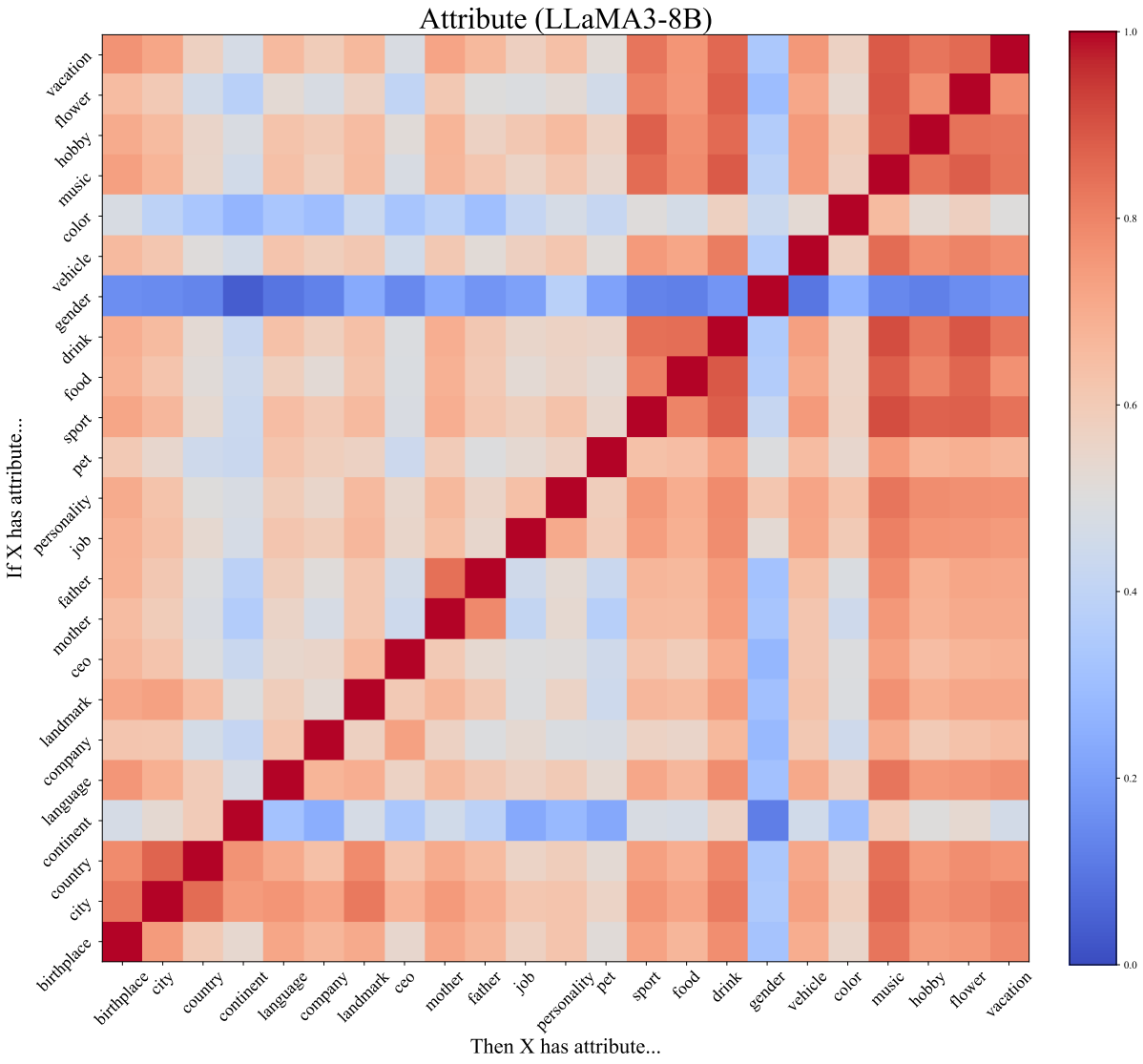


Figure 14: The attribute correlation between NTP logits of `llama-3-8b`.

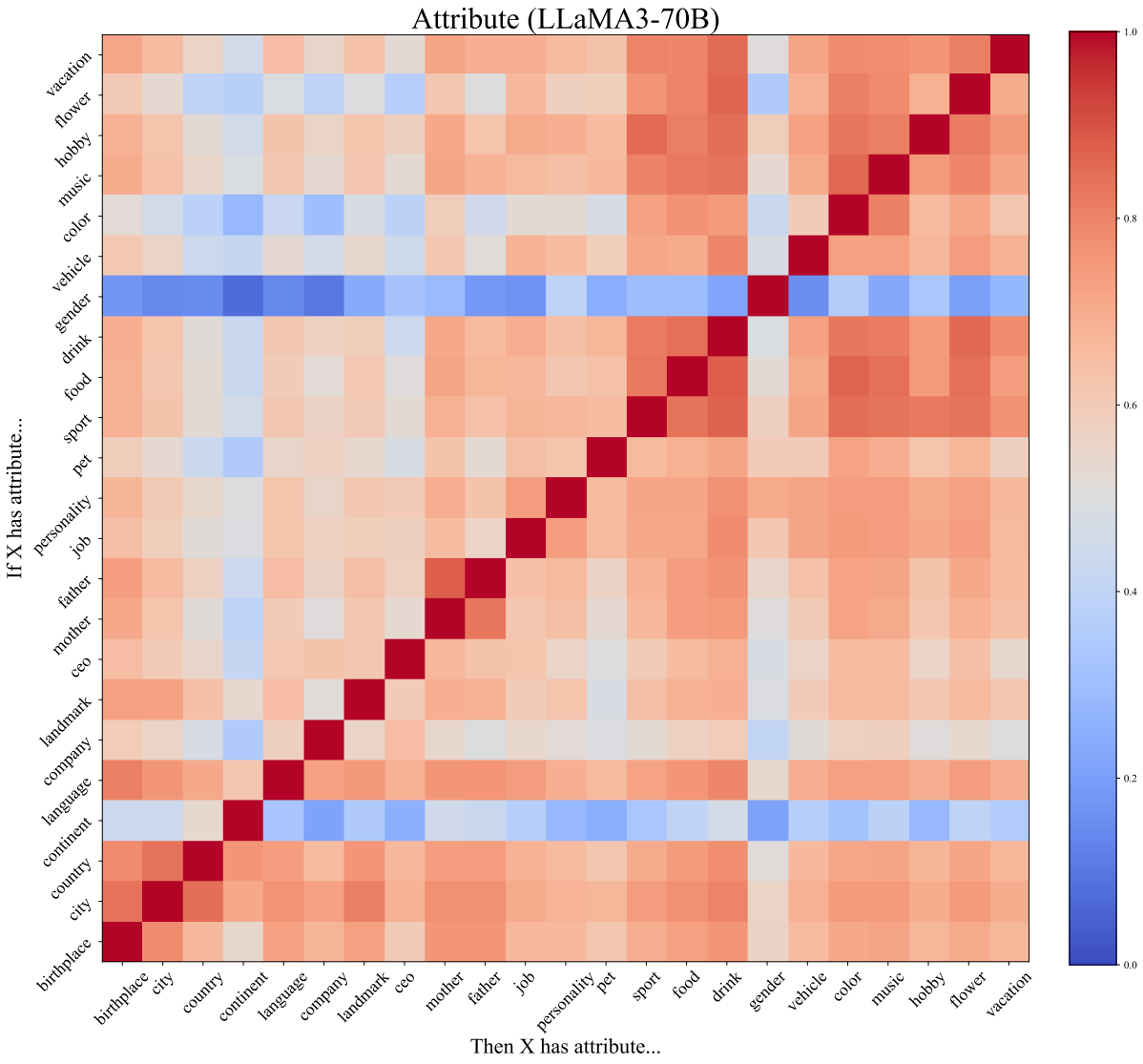


Figure 15: The attribute correlation between NTP logits of `llama-3-70b`.

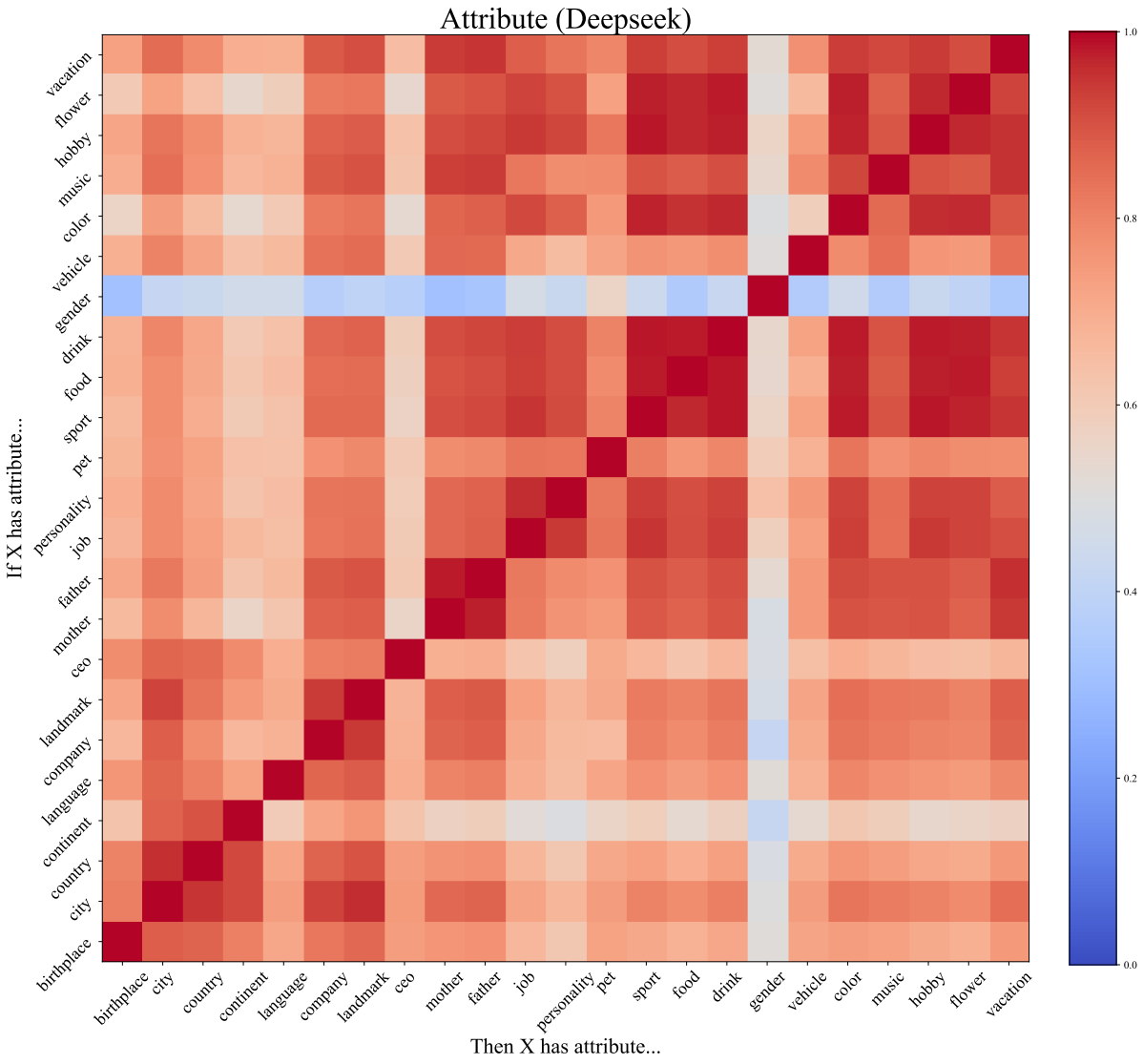


Figure 16: The attribute correlation between NTP logits of `deepseek-r1-distll-qwen-7B`.

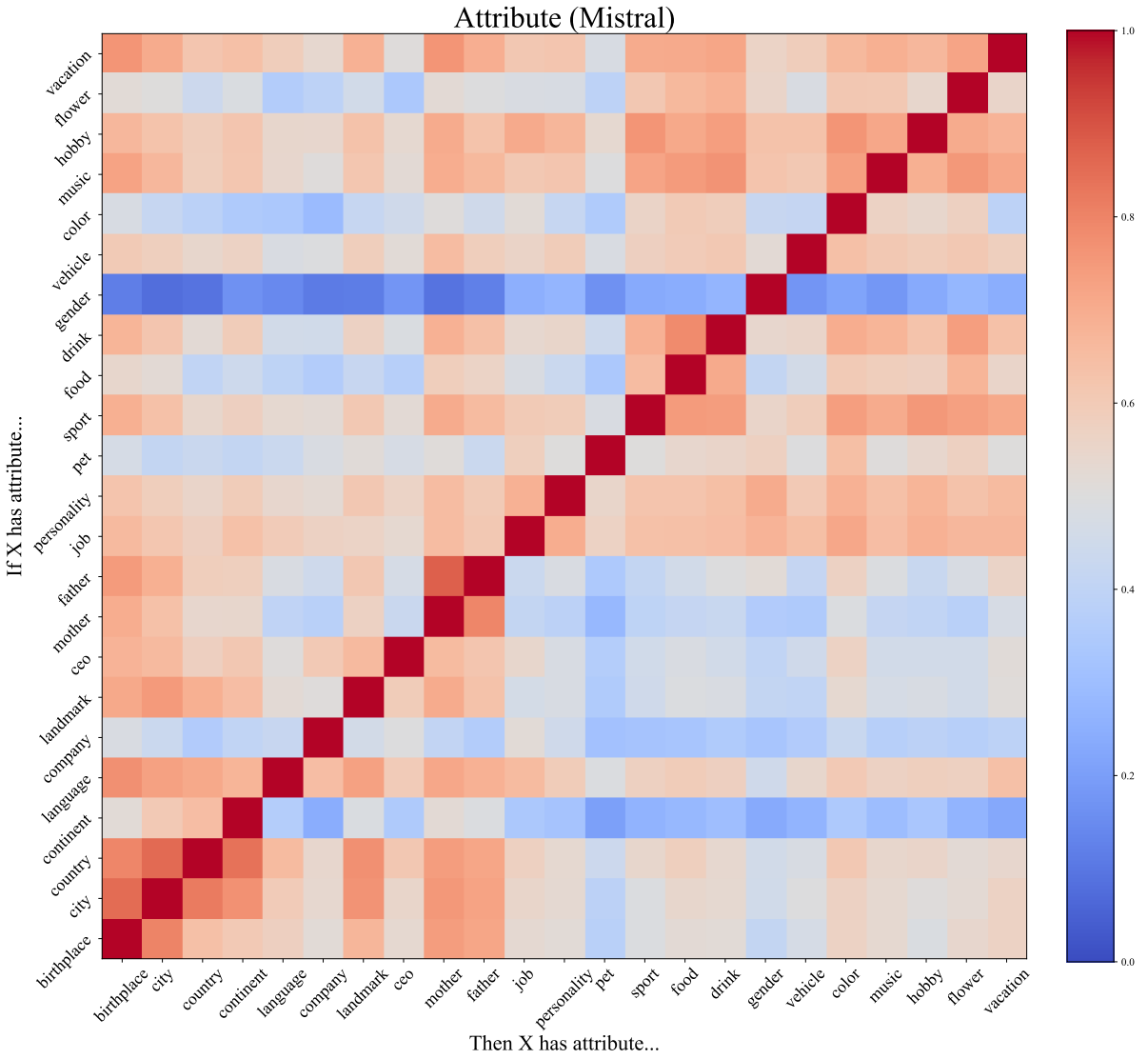


Figure 17: The attribute correlation between NTP logits of `mistral-7b-v0.3`.

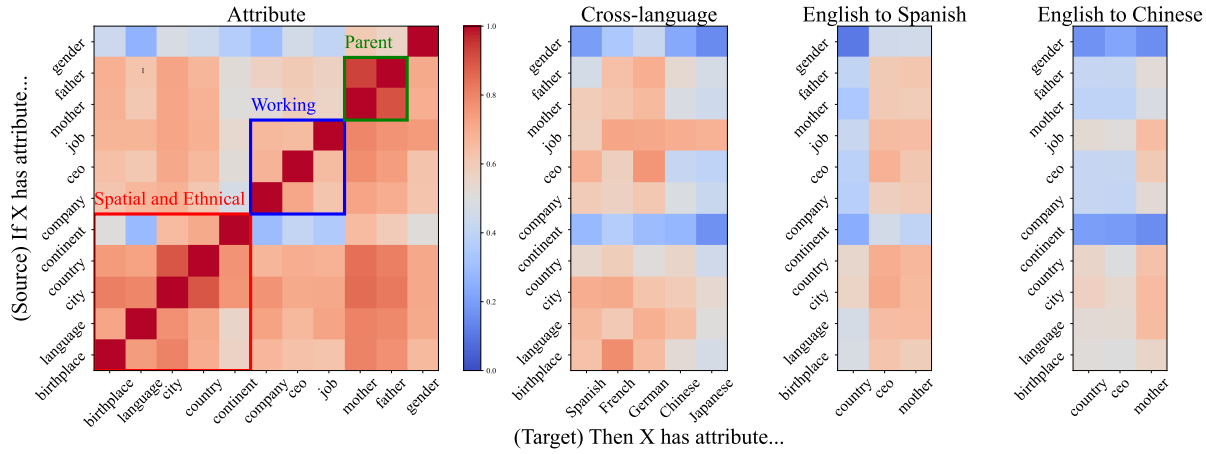


Figure 18: The linear correlation between NTP logits of `llama-3.2-3b`.

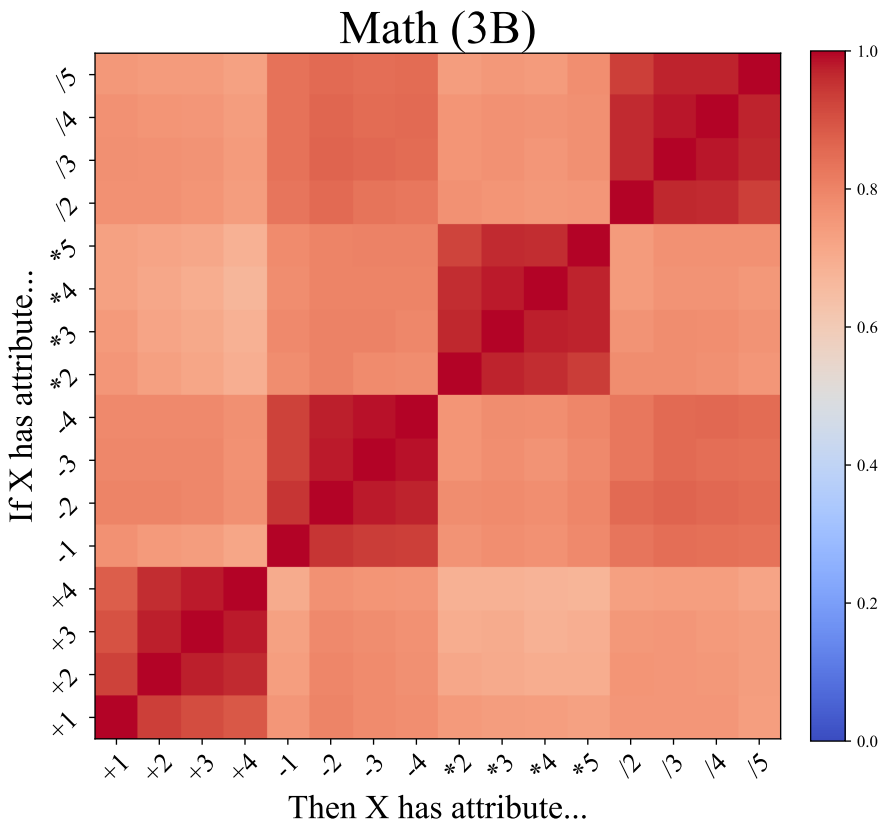


Figure 19: The linear correlation between NTP logits of `llama-3.2-3b` before and after large-scale post-training.

Figure 20: The correlation becomes more resilient in larger LMs.

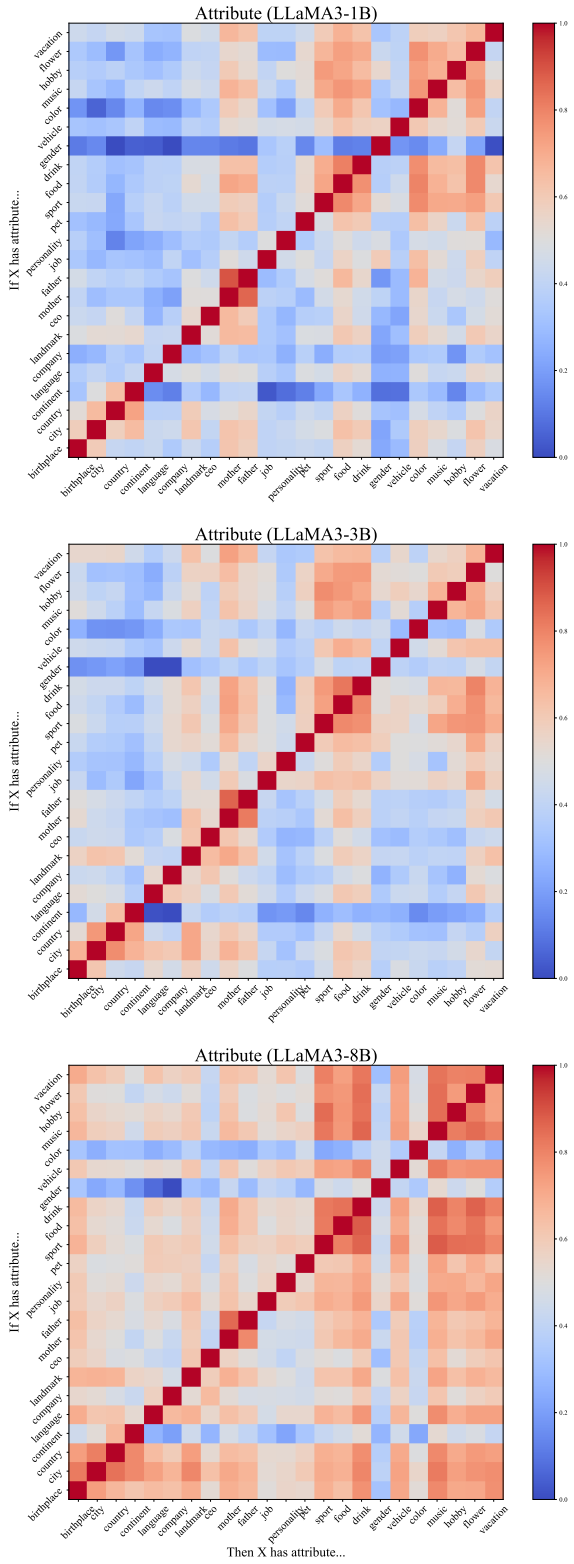
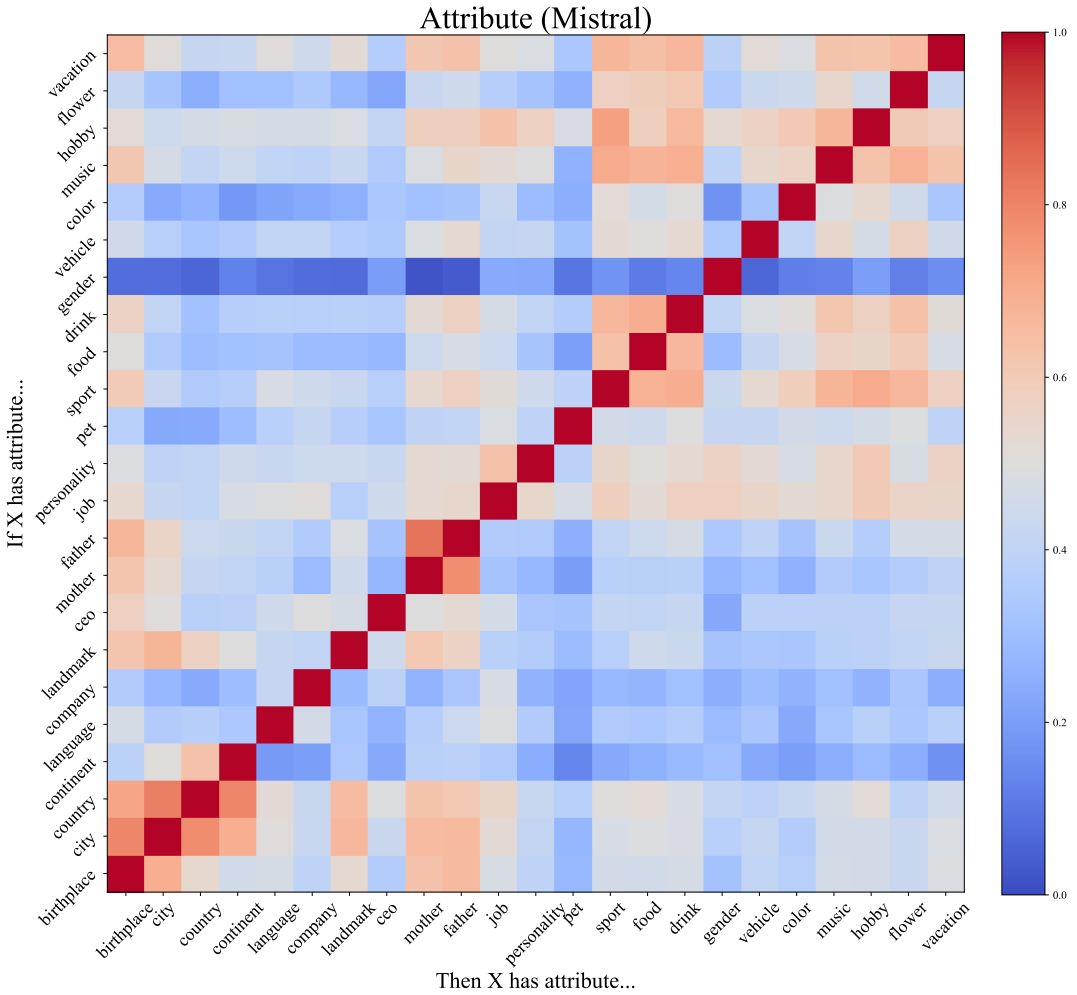


Figure 21: The correlation between logits from `mistral-7b-v0.3` before and after post-training.



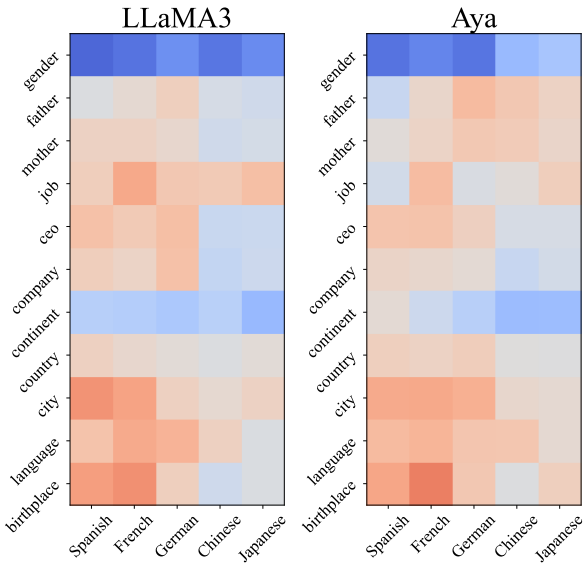


Figure 22: The comparison between Aya and LLaMA in cross-lingual correlation.

	Knowledge	Template	Domain Size
Spanish	birthplace	"*{} nació en la ciudad de*"	242
	city	"*{} vive en la ciudad de*"	242
	country	"*{} vive en el país de*"	128
	continent	"*{} vive en el continente de*"	6
	language	"*{} habla el idioma de*"	217
	company	"*{} trabaja para la empresa de*"	100
	ceo	"*{} trabaja para el CEO llamado*"	101
	job	"*El trabajo de {} es*"	105
	mother	"*El nombre de la madre de {} es*"	100
	father	"*{} el nombre del padre es*"	100
	gender	"*El género de {} es*"	3
French	birthplace	"*{} est né dans la ville de*"	242
	city	"*{} vit dans la ville de*"	242
	country	"*{} vit dans le pays de*"	128
	continent	"*{} vit sur le continent de*"	6
	language	"*{} parle la langue de*"	217
	company	"*{} travaille pour l'entreprise de*"	100
	ceo	"*{} travaille pour le PDG appelé*"	101
	job	"*{} travaille comme*"	105
	mother	"*Le nom de la mère de {} est*"	100
	father	"*Le nom du père de {} est*"	100
	gender	"*{} est de sexe*"	3
German	birthplace	"*{} wurde in der Stadt geboren*"	242
	city	"*{} lebt in der Stadt*"	242
	country	"*{} lebt im Land*"	128
	continent	"*{} lebt auf dem Kontinent*"	6
	language	"*{} spricht die Sprache von*"	217
	company	"*{} arbeitet für das Unternehmen von*"	100
	ceo	"*{} arbeitet für den CEO namens*"	101
	job	"*Der Beruf von {} ist*"	105
	mother	"*Der Name von {}'s Mutter ist*"	100
	father	"*Der Name von {}'s Vater ist*"	100
	gender	"*Das Geschlecht von {} ist*"	3
Chinese	birthplace	"*{}*所出生的城市是"	242
	city	"*{}*所居住的城市是"	242
	country	"*{}*所居住的国家是"	128
	continent	"*{}*所居住的大陆是"	6
	language	"*{}*说的语言是"	217
	company	"*{}*工作的公司是"	100
	ceo	"*{}*工作的公司的*CEO*是"	101
	job	"*{}*的工作是"	105
	mother	"*{}*的母亲的名字是"	100
	father	"*{}*的父亲的名字是"	100
	gender	"*{}*的性别是"	3
Japanese	birthplace	"*{}*が生まれた都市は"	242
	city	"*{}*が住んでいる都市は"	242
	country	"*{}*が住んでいる国は"	128
	continent	"*{}*が住んでいる大陸は"	6
	language	"*{}*が話している言語は"	217
	company	"*{}*が働いている会社は"	100
	ceo	"*{}*が働いている会社の*CEO*は"	101
	job	"*{}*の仕事は"	105
	mother	"*{}*の母の名前は"	100
	father	"*{}*の父の名前は"	100
	gender	"*{}*の性別は"	3

Table 12: Templates used in our experiments (Part 2: Cross Language).

	Knowledge	Template	Domain Size
Simile	object_color	"*The color of {} is the same as*"	85
	object_price	"*The size of {} is the same as*"	85
	object_heat	"*The heat of {} is the same as*"	85
	object_genre	"*The genre of {} is the same as*"	85
	object_size	"*The size of {} is the same as*"	85
	simile_color	"*The color of {} is*"	15
	simile_price	"*The size of {} is*"	2
	simile_heat	"*The heat of {} is*"	4
	simile_genre	"*The genre of {} is*"	22
	simile_size	"*The size of {} is*"	3
	simile_taste	"*The taste of {} is*"	3
	name_country	"*{} lives in the same country as*"	128
	gem_color	"*The color of {} is the same as the gem called*"	50
	animal_size	"*The size of {} is the same as the animal called*"	100
	food_taste	"*{} has the same taste as the food:*"	95
	fruit_color	"*{} X has the same color as the fruit:*"	99
Math	X+N	"*{}+N=*"	11
	X-N	"*{}-N=*"	11
	X*N	"*{}*N=*"	11
	X/N	"*{}/N=*"	11

Table 13: Templates used in our experiments (Part 3: Simile and Math).

Relation Pair	Fruit-Color	Food-Taste	Gem-Color	Name-Country	Animal-Size
Correlation	48.37	46.95	50.48	78.83	69.43
Relation Pair	Object-Genre	Object-Heat	Object-Size	Object-Price	Object-Color
Correlation	81.92	76.48	84.23	84.23	81.08

Table 14: Correlation between logits of `llama-3.2-3b` on simile objects and attributes.

Country	Influencing Cities
Sweden	**Stockholm**, Brisbane, Johannesburg, Cardiff, Chicago, Hyderabad, Aleppo, Lima, Rochester, Salem
Cuba	**Havana**, Chicago, Columbus, stockholm, Rochester, Hyderabad, Scarborough, Johannesburg, singapore, Hamburg
Switzerland	Columbus, Stuttgart, Cardiff, Leicester, Chicago, Brisbane, Saras, stockholm, vegas, Bethlehem
Ghana	Winnipeg, Nairobi, Johannesburg, Leicester, Atlanta, Tulsa, Maharashtra, Greenville, Brisbane, Lima
Poland	**Warsaw**, Cardiff, Liverpool, Maharashtra, stockholm, Amsterdam, Atlanta, Kashmir, Perth, Aleppo
Turkey	**Istanbul**, Chicago, Toronto, Maharashtra, stockholm, Johannesburg, Cardiff, Lima, Columbus, Ankara
Sudan	Nairobi, stockholm, Lima, Tulsa, Johannesburg, Maharashtra, Winnipeg, Hyderabad, Wilmington, Kashmir
Romania	Cardiff, Rochester, Johannesburg, **Budapest**, Seattle, Rajasthan, Hyderabad, Chicago, Kyoto, Lima
Samoa	Maharashtra, Leicester, Winnipeg, Chicago, Honolulu, Brisbane, Nairobi, Hyderabad, Lima, Cardiff
Iceland	Cardiff, Leicester, Chicago, Amsterdam, Wilmington, Islamabad, Winnipeg, Kyoto, Hyderabad, stockholm
Nigeria	Winnipeg, Nairobi, Maharashtra, Lagos, Johannesburg, Stuttgart, Leicester, Abu, Chicago, Tulsa
Iraq	Chicago, Hyderabad, Wilmington, Lima, Baghdad, stockholm, Kashmir, Tulsa, Belfast, singapore
Laos	Bangkok, Leicester, Chicago, Kashmir, Tulsa, stockholm, Winnipeg, Lima, Rajasthan, Johannesburg
USSR	**Moscow**, NYC, Midlands, stockholm, Chicago, Cardiff, Maharashtra, Pyongyang, Boulder, Columbus
Kosovo	Kashmir, Seattle, Leicester, stockholm, Tulsa, Belfast, Mosul, vegas, Rochester, Buenos
China	**Beijing**, **Shanghai**, Hyderabad, Brisbane, Columbus, stockholm, Maharashtra, Amsterdam, Leicester, Hamburg
Guatemala	Greenville, Tulsa, Leicester, Buenos, Johannesburg, Kashmir, Wilmington, Lima, Chicago, Rochester
Tunisia	Johannesburg, stockholm, Hamburg, Columbus, Leicester, Tulsa, Stuttgart, Winnipeg, Cardiff, Maharashtra
Denmark	**Copenhagen**, Cardiff, Leicester, Brisbane, Hyderabad, Atlanta, Saras, Chicago, Hamburg, Salem
Nicaragua	Nairobi, Bangkok, Rochester, Leicester, Amsterdam, Kerala, Maharashtra, Belfast, Winnipeg, Chicago
Türkiye	Maharashtra, München, Seattle, **İstanbul**, stockholm, Jakarta, Istanbul, Toronto, Milwaukee, Kyoto
Bosnia	Hyderabad, Islamabad, Belfast, Johannesburg, Jakarta, Cardiff, Rochester, Kashmir, Leicester, Lima
Netherlands	**Amsterdam**, Cardiff, Midlands, Columbus, Karachi, stockholm, Nottingham, Maharashtra, Saras, Wilmington
Malaysia	Leicester, **Kuala**, Cardiff, Hamburg, Maharashtra, Baltimore, Chicago, Columbus, Johannesburg, Hyderabad
Venezuela	Wilmington, vegas, Cardiff, Maharashtra, Rochester, Brisbane, stockholm, Buenos, Lima, Tulsa
Sri	Leicester, Atlanta, Kashmir, Rajasthan, Nairobi, Cardiff, stockholm, Lima, Maharashtra, Islamabad
Ireland	**Dublin**, Cardiff, Belfast, Leicester, Tehran, Johannesburg, Stuttgart, Aleppo, Bethlehem, Hyderabad
Liberia	Leicester, Winnipeg, Nairobi, Johannesburg, Chicago, Kerala, Rochester, Maharashtra, Atlanta, Greenville
Afghanistan	**Kabul**, Cardiff, Islamabad, stockholm, Tulsa, Chicago, Maharashtra, Kashmir, Rajasthan, Leicester
America	Columbus, **Chicago**, Belfast, Sofia, Hyderabad, **Seattle**, Cardiff, Johannesburg, Maharashtra, Moscow
Austria	Cardiff, **Vienna**, Hamburg, Hyderabad, Leicester, Bethlehem, Stuttgart, stockholm, Columbus, Rajasthan
Scotland	Cardiff, Glasgow, **Edinburgh**, Stuttgart, stockholm, Belfast, Leicester, Columbus, Maharashtra, Lima
Libya	Chicago, stockholm, Columbus, Leicester, Aleppo, Cardiff, Mosul, Lima, Wilmington, Johannesburg
Uruguay	Buenos, Seattle, Hyderabad, Maharashtra, Hamburg, Johannesburg, Wilmington, Leicester, Columbus, Cardiff
Bangladesh	Winnipeg, Cardiff, Leicester, Maharashtra, Tulsa, Atlanta, Chicago, Bangalore, Islamabad, Kashmir
Bahrain	Leicester, Chicago, Brisbane, Kashmir, Lima, Riyadh, Dubai, Wilmington, Atlanta, Saras
Pakistan	**Islamabad**, Cardiff, Jakarta, Karachi, Tulsa, Leicester, Winnipeg, Atlanta, Maharashtra, Wilmington
Fiji	Lima, Leicester, Fargo, Kashmir, Brisbane, Winnipeg, Johannesburg, Cardiff, Tulsa, Edinburgh
Cambodia	Bangkok, Tulsa, Leicester, Cardiff, stockholm, Kashmir, Johannesburg, Wilmington, Kabul, Lima
Singapore	**singapore**, Chicago, Leicester, Brisbane, Hamburg, Columbus, Atlanta, Kashmir, Johannesburg, Cardiff
Macedonia	Leicester, Stuttgart, Winnipeg, Rochester, Kashmir, Johannesburg, Jakarta, Maharashtra, Budapest, Lima
Mongolia	Winnipeg, Chattanooga, Leicester, Lima, Cardiff, Kyoto, Maharashtra, Johannesburg, Rajasthan, Hamburg
Peru	**Lima**, Perth, Maharashtra, Winnipeg, Leicester, Chattanooga, Seattle, Hyderabad, Nairobi, Chicago
Myanmar	Bangkok, Cardiff, Tulsa, Leicester, Winnipeg, Kashmir, Maharashtra, Kyoto, Lima, Chicago
Trinidad	Leicester, Cardiff, Maharashtra, Brisbane, Rochester, Tulsa, Winnipeg, Abu, vegas, Johannesburg
Colombia	Maharashtra, Columbus, Lima, Seattle, Rochester, Wilmington, Johannesburg, Stuttgart, Amsterdam, Hyderabad
Maurit	Winnipeg, Leicester, Johannesburg, Edinburgh, Cardiff, Chicago, Stuttgart, stockholm, Moscow, Wilmington
Iran	**Tehran**, Cardiff, Lima, Kashmir, Hyderabad, Leicester, Aleppo, Chicago, Stuttgart, Hamburg
India	Indianapolis, Cardiff, Maharashtra, Chicago, Hyderabad, Leicester, Lima, Columbus, Winnipeg, stockholm
Spain	**Madrid**, Hyderabad, stockholm, Spokane, Cardiff, Amsterdam, Rome, Barcelona, Dallas, Johannesburg
Honduras	Wilmington, Winnipeg, Buenos, Hamburg, Nairobi, stockholm, Johannesburg, Amsterdam, Columbus, Lima
USA	**NYC**, Moscow, Columbus, Midlands, **Chicago**, Sofia, Karnataka, Karachi, Cardiff, Sevilla

Table 15: The most influencing cities of counties in the *City→Country* correlation.

Father	Influencing Mothers
Omar	Olivia, Nora, Sara, Sofia, Naomi, Diana, Uma, Rosa, Eden, Jade
Victor	Victoria, Sofia, Maria, Savannah, Sophie, Uma, Sonia, Angela, Grace, Ivy
Andre	Angela, Sofia, Sophie, Savannah, Maria, Rebecca, Ivy, Clara, Chloe, Nina
Julio	Sofia, Chloe, Maria, Carmen, Rebecca, Ivy, Rosa, Olivia, Sonia, Savannah
Enrique	Carmen, Chloe, Rosa, Clara, Sofia, Emma, Maria, Rebecca, Fiona, Olivia
Amir	Sara, Sofia, Amelia, Eden, Mei, Nora, Uma, Bella, Victoria, Diana
Xavier	Sophie, Maria, Sonia, Olivia, Emma, Leah, Clara, Uma, Jasmine, Carmen
Javier	Carmen, Chloe, Sofia, Ivy, Maria, Jasmine, Olivia, Rosa, Fiona, Jennifer
Vlad	Elena, Sofia, Chloe, Mia, Nina, Angela, Diana, Naomi, Savannah, Clara
Roberto	Chloe, Sofia, Rosa, Carmen, Lucia, Olivia, Clara, Mei, Maria, Elena
Lars	Sophie, Clara, Maria, Nina, Ella, Sara, Harper, Savannah, Rebecca, Fiona
Min	Sonia, Mei, Angela, Eden, Clara, Chloe, Grace, Maria, Harper, Savannah
James	Grace, Fiona, Ella, Savannah, Emma, Angela, Chloe, Harper, Leah, Maria
Giovanni	Lucia, Fiona, Sofia, Savannah, Rosa, Diana, Bella, Chloe, Carmen, Mei
Ivan	Ivy, Elena, Sofia, Nina, Maria, Ada, Emma, Sophie, Savannah, Sakura
Diego	Chloe, Sofia, Maria, Rosa, Angela, Carmen, Savannah, Diana, Clara, Mei
Fernando	Maria, Rosa, Fiona, Savannah, Carmen, Angela, Sofia, Luna, Clara, Ada
Ethan	Elena, Leah, Jennifer, Emma, Jasmine, Chloe, Clara, Mei, Ada, Serena
Chen	Mei, Chloe, Grace, Nina, Eden, Harper, Sofia, Rebecca, Sakura, Sonia
Gabriel	Maria, Sophie, Eden, Leah, Sara, Grace, Chloe, Rebecca, Elena, Luna
Boris	Bella, Elena, Angela, Fiona, Nina, Ada, Sofia, Sophie, Nora, Leah
Jean	Sophie, Angela, Chloe, Maria, Naomi, Carmen, Savannah, Nina, Rebecca, Lucia
Dmitry	Sofia, Elena, Chloe, Diana, Nina, Savannah, Mia, Clara, Sakura, Ivy
Ahmed	Sara, Sofia, Sophie, Nora, Uma, Victoria, Eden, Sonia, Jennifer, Mei
Wei	Mei, Chloe, Grace, Rebecca, Mia, Sofia, Ada, Nina, Angela, Harper
Ibrahim	Sofia, Sara, Eden, Uma, Victoria, Nora, Bella, Ada, Sophie, Elena
Liam	Fiona, Emma, Mia, Chloe, Nora, Leah, Grace, Jasmine, Jade, Angela
Mustafa	Sara, Sofia, Nora, Victoria, Ada, Uma, Eden, Jade, Rosa, Elena
Jorge	Maria, Carmen, Rosa, Chloe, Sofia, Diana, Elena, Fiona, Angela, Nora
Leonardo	Clara, Sofia, Jennifer, Olivia, Chloe, Jasmine, Fiona, Rosa, Lucia, Diana
Luca	Fiona, Lucia, Sofia, Angela, Maria, Savannah, Emma, Clara, Sakura, Leah
Carlos	Carmen, Maria, Rosa, Olivia, Chloe, Sofia, Clara, Sakura, Savannah, Fiona
Pedro	Maria, Rosa, Carmen, Chloe, Olivia, Clara, Sakura, Sofia, Ivy, Ada
Michel	Sophie, Lucia, Nina, Maria, Leah, Eden, Elena, Sara, Sonia, Carmen
Kai	Mei, Maria, Nina, Angela, Chloe, Eden, Jade, Uma, Sakura, Ada
Benjamin	Leah, Eden, Bella, Rebecca, Sophie, Grace, Nina, Harper, Lucia, Victoria
Noah	Rebecca, Chloe, Nina, Nora, Eden, Naomi, Sara, Grace, Leah, Ada
Ali	Sara, Nora, Eden, Victoria, Uma, Sofia, Mei, Jade, Bella, Sonia
Levi	Chloe, Leah, Eden, Sara, Nina, Elena, Harper, Bella, Rosa, Rebecca
Antonio	Rosa, Maria, Angela, Lucia, Sofia, Chloe, Savannah, Olivia, Carmen, Fiona
Rafael	Sofia, Rosa, Carmen, Maria, Clara, Leah, Ivy, Chloe, Naomi, Lucia
Marco	Maria, Sofia, Jasmine, Lucia, Clara, Angela, Chloe, Mei, Rebecca, Carmen
Stefan	Elena, Fiona, Angela, Savannah, Clara, Sophie, Mei, Maria, Eden, Rebecca
Chung	Mei, Chloe, Grace, Maria, Angela, Sonia, Harper, Clara, Savannah, Mia
Abdul	Uma, Sara, Sofia, Nora, Jennifer, Ada, Rosa, Victoria, Eden, Bella
Muhammad	Sofia, Sara, Victoria, Mei, Emily, Jennifer, Nora, Uma, Eden, Naomi
Hugo	Maria, Sophie, Chloe, Clara, Fiona, Emma, Savannah, Angela, Carmen, Ivy
Axel	Sophie, Angela, Rebecca, Nina, Ada, Emma, Fiona, Ivy, Eden, Savannah
Lucas	Lucia, Maria, Clara, Fiona, Uma, Chloe, Harper, Savannah, Sophie, Jasmine
Mason	Harper, Leah, Jasmine, Chloe, Angela, Nina, Ada, Sofia, Ella, Emma
Hassan	Sara, Eden, Nora, Victoria, Bella, Sofia, Naomi, Savannah, Mei, Diana
Pablo	Maria, Chloe, Sofia, Rosa, Savannah, Rebecca, Carmen, Elena, Fiona, Luna
Raphael	Rebecca, Sophie, Elena, Leah, Rosa, Grace, Eden, Fiona, Clara, Sonia
Elijah	Elena, Eden, Rebecca, Chloe, Savannah, Ella, Leah, Emily, Grace, Uma
Louis	Sophie, Nina, Savannah, Grace, Rosa, Maria, Rebecca, Fiona, Leah, Sonia
Ricardo	Chloe, Carmen, Sofia, Rosa, Jennifer, Clara, Rebecca, Sakura, Mei, Olivia
Samuel	Sonia, Savannah, Leah, Eden, Rebecca, Sophie, Grace, Ada, Emma, Clara
William	Grace, Emma, Emily, Leah, Ada, Harper, Angela, Victoria, Fiona, Diana
Salman	Sonia, Sofia, Nora, Uma, Sara, Bella, Eden, Jennifer, Victoria, Leah
Oliver	Olivia, Sophie, Harper, Elena, Nina, Maria, Grace, Diana, Emma, Nora
Angelo	Angela, Sofia, Fiona, Clara, Chloe, Rosa, Carmen, Savannah, Lucia, Nina
Hans	Sophie, Rebecca, Angela, Savannah, Eden, Ella, Clara, Maria, Uma, Mei
Jamal	Sofia, Jasmine, Uma, Sara, Mei, Eden, Naomi, Victoria, Bella, Diana
Santiago	Sofia, Maria, Rosa, Carmen, Chloe, Savannah, Mei, Olivia, Ivy, Luna

Table 16: The most influencing fathers of mothers in the *Mother→Father* correlation.

	Attribute	Influencing Objects
Genre	toys	toy, puzzle, drum, shoes, sweater, electric, fridge, gloves, chair, jeans
	transport	headphones, pen, plate, drum, electric, car, couch, smartphone, rug, suitcase
	kitchen	drum, jeans, pen, plate, toy, backpack, rug, fridge, chair, grill
	furniture	drum, chair, fridge, electric, rug, camera, puzzle, shoes, sweater, plate
	decor	drum, rug, vase, pen, sweater, jeans, smartphone, backpack, washing, speaker
	accessories	drum, shoes, plate, laptop, electric, oven, gloves, curtains, jeans, chair
	sports	basketball, pen, drum, jeans, plate, skateboard, tennis, rug, charger, puzzle
	travel	pen, drum, water, yoga, suitcase, sunglasses, watch, plate, jeans, fridge
	art	drum, puzzle, pen, scarf, water, camera, couch, toy, chair, jeans
	fitness	yoga, puzzle, drum, pen, couch, electric, sweater, scarf, rug, camera
	outdoors	drum, plate, pen, fishing, electric, water, couch, camera, toy, puzzle
	bags	drum, fridge, sweater, gloves, jeans, backpack, pen, rug, electric, umbrella
	electronics	electric, drum, headphones, plate, toy, pen, laptop, jeans, sweater, couch
	clothing	drum, sweater, electric, shoes, skateboard, pen, jeans, camera, rug, fridge
	food	fridge, drum, pen, water, scarf, couch, plate, smartphone, sweater, speaker
	photography	camera, water, drum, puzzle, scarf, skateboard, yoga, headphones, rug, couch
	literature	book, iron, pen, drum, yoga, couch, water, speaker, scarf, fan
	appliances	electric, sweater, jeans, plate, shoes, fridge, drum, chair, oven, laptop
	home	electric, oven, drum, smartphone, pen, backpack, rug, jeans, fridge, puzzle
	music	guitar, drum, headphones, scarf, basketball, pen, toy, puzzle, suitcase, water
Heat	warm	hoodie, sweater, clock, lamp, drum, earrings, yoga, apple, tennis, oven
	hot	hoodie, puzzle, tennis, drum, oven, jeans, car, lamp, earrings, fan
	neutral	jeans, speaker, blanket, sofa, car, puzzle, earrings, hoodie, tennis, rug
	cold	hoodie, car, earrings, fan, lamp, curtains, couch, clock, puzzle, sweater
Size	large	smartphone, jeans, drum, puzzle, hoodie, umbrella, pencil, clock, car, backpack
	medium	hoodie, tripod, car, keyboard, drum, suitcase, smartphone, basketball, curtains, bottle
	small	smartphone, hoodie, car, drum, pencil, jeans, backpack, keyboard, puzzle, toy
Color	black	jeans, iron, fan, umbrella, hoodie, suitcase, puzzle, bowl, printer, electric
	green	backpack, plate, puzzle, jeans, couch, umbrella, drum, soap, car, sweater
	blue	jeans, electric, puzzle, plate, backpack, fishing, bottle, chair, car, umbrella
	beige	jeans, soap, hoodie, drum, puzzle, bottle, suitcase, oven, bed, speaker
	gold	puzzle, backpack, car, earrings, iron, bottle, drum, jeans, plate, fan
	natural	jeans, bottle, puzzle, earrings, car, plate, oven, yoga, suitcase, drum
	silver	bottle, jeans, puzzle, iron, drum, mirror, soap, electric, backpack, earrings
	orange	puzzle, car, drum, backpack, jeans, umbrella, bottle, electric, oven, plate
	red	car, drum, earrings, puzzle, microwave, pen, umbrella, bowl, electric, backpack
	gray	jeans, soap, mouse, puzzle, plate, sweater, umbrella, printer, bed, backpack
	brown	soap, iron, puzzle, sweater, umbrella, backpack, speaker, drum, hoodie, couch
	yellow	plate, yoga, car, backpack, umbrella, soap, drum, puzzle, sweater, fan
	purple	puzzle, drum, electric, hoodie, backpack, jeans, microwave, mouse, bottle, bowl
	white	plate, suitcase, fan, jeans, puzzle, backpack, soap, umbrella, sweater, drum
Price	high	smartphone, drum, air, car, hoodie, jeans, backpack, umbrella, puzzle, electric
	low	drum, jeans, backpack, smartphone, car, hoodie, air, umbrella, puzzle, electric

Table 17: The most influencing objects of attributes in the simile correlation.

I Low Dispersion in Label-wise Correlation

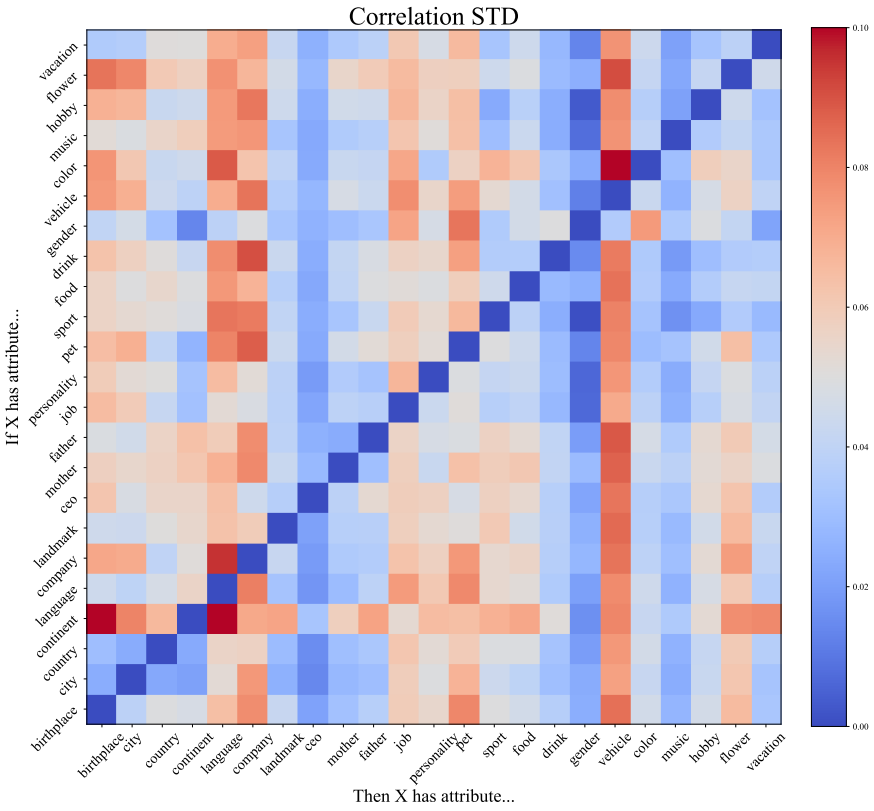
A potential concern on the correlation metric is whether the correlation reflects the majority property of different labels or some highly correlated cast bias into the evaluation. We plot the std of label-wise correlation distributions of `llama-3-8b` in Figures 23 (on the same model) and 24 (before and after post-training). The result shows the distributions to be concentrated with a std generally lower than 0.05, which addresses the misrepresentation concern.

J Subword Issue

Finally, we show the precision of W is highly affected by the semantics of the input and output tokens. We first categorize the tokens into 3 categories, 1) Subword, a token being part of a word, such as a prefix like *Br* in *Brunei*, 2) Word in a phrase, a token is a whole word but also a part of a phrase like *North* in *North America*, 3) Whole semantics, the rest of tokens with a full meaning in itself like *USA*.

The results in Table 18 show the semantic completeness to be an important factor in whether knowledge can be generalized. With higher semantic completeness (Whole Semantics $>$ Word in a Phrase $>$ Subword), the W's precision also rise as the token indicates a clearer entity. Consequently, it can be better updated by the generalization behavior caused by the linear correlation. The only precise mapping (and successful) generalization for "Word in a Phrase" is *Riyadh*$\rightarrow$*Saudi Arabia*, where the first token *Saudi* has a strong indication of the country.

Figure 23: The std of correlation distribution between logits.



Completeness	Correlation	Precision (Hit@Top-5)	Generalization
Whole Semantics	0.85	0.49	55.67%
Word in a Phrase	0.86	0.10	2.00%
Subword	0.87	0.00	0.00%

Table 18: The correlation and W precision of tokens with different levels of semantic completeness.

Figure 24: The std of correlation distribution between logits **before and after large-scale post-training**.