Representational Analysis of Binding in Language Models

Qin Dai[1] **Benjamin Heinzerling**[2,1] **Kentaro Inui**[3,1,2]
[1]Tohoku University [2]RIKEN AIP [3]MBZUAI
qin.dai.b8@tohoku.ac.jp benjamin.heinzerling@riken.jp
kentaro.inui@mbzuai.ac.ae

Abstract

Entity tracking is essential for complex reasoning. To perform in-context entity tracking, language models (LMs) must bind an entity to its attribute (e.g., bind a container to its content) to recall attribute for a given entity. For example, given a context mentioning "The coffee is in Box Z, the stone is in Box M, the map is in Box H", to infer "Box Z contains the coffee" from the context, LMs must bind "Box Z" to "coffee". To explain the binding behaviour of LMs, Feng and Steinhardt (2023) introduce a Binding ID mechanism and state that LMs use a abstract concept called Binding ID (BI) to internally mark entity-attribute pairs. However, they have not captured Ordering ID (OI), namely ordering index of entity, from entity activations that directly determines the binding behaviour. In this work, we provide a novel view of the BI mechanism by localizing OI and proving the causality between OI and binding behaviour. Specifically, we discover the OI subspace and reveal causal effect of OI on binding that when editing activations along the OI encoding direction, LMs tend to bind a given entity to other attributes (e.g., "stone" for "Box Z") accordingly. The code and datasets used in this paper are available at https://github.com/cl-tohoku/OI-Subspace.

1 Introduction

The ability of a model to track and maintain information associated with an entity in a context is essential for complex reasoning (1; 2; 3; 4; 5). To recall attribute information for a given entity in a context, a model must bind entities to their attributes (6). For example, given a Sample 1, a model must bind the entities (e.g., "Box Z", "Box M" and "Box H") to their corresponding attributes (e.g., "coffee", "stone" and "map") so as to recall (or answer) such as what is in "Box Z". Binding has also been studied as a fundamental problem in Psychology (7).

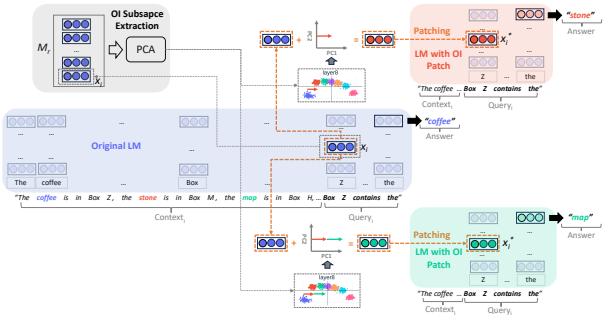

Figure 1: Our main finding on Ordering ID (OI) subspace intervention. Patching entity (e.g., "Z") representations along OI direction (i.e., PC1) in activation space yields corresponding changes in model output.

To uncover how Language Models (LMs) realize binding in term of internal representation, an existing research (6) introduces the Binding ID mechanism that LMs apply an abstract concept called Binding ID (BI) to bind and mark Entity-Attribute (EA) pairs (e.g., "Box Z" and "coffee" in Sample 1, where BI is denoted as a numbered square). However, they have not captured the Ordering ID (OI) information from the entity (or attribute) activations that causally affects binding behaviour and thus BI information as well. Here, OI is defined as the input order (or ordering index) of entities and attributes, no matter they are bound by a relation (e.g., "is_in" in Sample 1) or not, such as the indexing number in Sample 1 and Sample 2. We can observe that in a 1E-to-1A bound context, such as in Sample 1, BI and OI are interchangeable.

(1) **Context**: The coffee$_{\boxed{0}}$ is in Box Z$_{\boxed{0}}$, the stone$_{\boxed{1}}$ is in Box M$_{\boxed{1}}$, the map$_{\boxed{2}}$ is in Box H$_{\boxed{2}}$. **Query**: Box Z$_{\boxed{0}}$ contains the

(2) **Non-related Context**: The coffee$_{\boxed{0}}$ and Box Z$_{\boxed{0}}$ are scattered around, the stone$_{\boxed{1}}$ is here and Box M$_{\boxed{1}}$ is there, the map$_{\boxed{2}}$ and Box H$_{\boxed{2}}$ are in different place. **Query**: Box Z$_{\boxed{0}}$ contains the

Since binding is the foundational skill that underlies entity tracking (6), in this work, we take the entity tracking task (8; 9) as a benchmark to analyze the LM's binding behaviour. Based on the analysis of internal representation on this task, we localize the OI information from the activiations and provide a novel view of the BI mechanism. Specifically, we apply Principle Component Analysis (PCA) as well as other dimension reduction methods such as Partial Least Squares to analyze the activations of LMs, and which are empirically proven to be effective. We discover that LMs encode (or store) the OI information into a low-rank subspace (called OI subspace hereafter), and the discovered OI subspace can causally affect binding behaviour and thus BI information as well. That is, we find that by causally intervening along the OI encoding Principle Component (PC), LMs swap the binding and infer a new attribute for a given entity accordingly. For example, as shown in Figure 1, by patching activations along the direction (i.e., PC1), we can make the LMs to infer "Box Z contains the stone" and "Box Z contains the map" instead of "Box Z contains the coffee". Therefore, our findings extend the previous BI based understanding of binding in LMs (6) by revealing the causality between OI and binding.

In addition, we find that such OI subspace that determines binding is prevalent across multiple LM families such as Llama2 (10) (and Llama3 (11)), Qwen1.5 (12) and Pythia (13), and the code fine-tuned LM Float-7B (9). Please see our paper (14) for more details.

2 Finding OI Subspace

In this section we describe our Principle Component Analysis (PCA) based method to localize the OI subspace in activations of LMs. As shown in Figure 1, given a LM (e.g., Llama2), and a collection of texts which describes a set of EA pairs related by a relation such as "is_in" in Sample 1, we extract the activation of entity token (e.g., "Z") in query (denoted as $\mathbf{x}_i$) from a certain layer [1)]. We then construct a activation matrix $M_r \in R^{n\times d}$ for a relation r, where n denotes the number of entities and d denotes the dimension of the activation. The row i of M_r is the activation of an entity token (i.e., $\mathbf{x}_i$).

PCA has been applied for identifying various subspace (or direction) such as the subspace encoding language bias (15), truth value of assertions (16) and sentiment (17).

1) The layer is determined by a development set

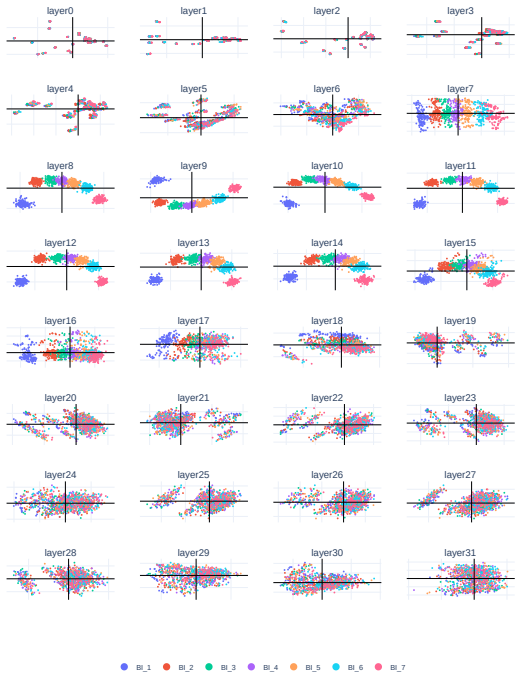

Figure 2: Layer-wise OI subspace visualization on Llama2-7B, where "BI" primarily denotes OI.

Inspired by these studies, we choose PCA as our first attempt to localize OI subspace. Specifically, the PCA of a activation matrix is $M_r = U_r \Sigma_r V_r^T$, where the columns of $V_r \in R^{d\times d}$ are principle directions of M_r. We takes first c columns of V_r as the OI direction, denoted as $B_r \in R^{d\times c}$.

We adopt a subset of the entity tracking dataset (8; 9), which contains $n = 1000$ samples, to create layer (l) wise activation matrix M_r^l. We then use the M_r^l to extract the layer-wise OI subspace projection matrix $B_r^l \in R^{d\times 2}$ to visualize the activations. Figure 2 shows the embedding visualization on Llama2-7B, where each point represents the activation of an entity projected via the B_r^l, and the colors represent OIs. From which, we can observe that middle layers, such as layer 8, have a clearly visible direction along which OI increases, while the others have tangled distribution.

We also observe similar pattern of distribution on Llama3-8B, Float-7B and other LM families such as Qwen1.5 and Pythia. This indicates that LMs use the middle layers to encode OI information, and the finding is prevalent across multiple LM families. This finding is also consistent with the "stages of inference hypothesis" (18) stating that the function of early layers is to perform detokenization, middle layers do feature engineering, and late

Context	Query	Answer for # Step 1	2	3	4	5	6
The coffee is in Box Z, the stone is in Box M, the map is in Box H, the coat is in Box L, the string is in Box T, the watch is in Box E, the meat is in Box F.	Box Z contains the	stone	map	map	string	watch	meat
The letter is in Box Q, the boot is in Box C, the fan is in Box N, the crown is in Box R, the guitar is in Box E, the bag is in Box D, the watch is in Box K.	Box Q contains the	boot	fan	crown	guitar	watch	watch
The cross is in Box Z, the ice is in Box D, the ring is in Box F, the plane is in Box Q, the clock is in Box X, the paper is in Box I, the engine is in Box K.	Box Z contains the	ice	ring	ring	clock	paper	engine

Table 1: Attributes inferred by Llama2-7B as a result of directed activation patching along OI-PC in the OI subspace on the dataset of "r: is_in", where color denotes the BI.

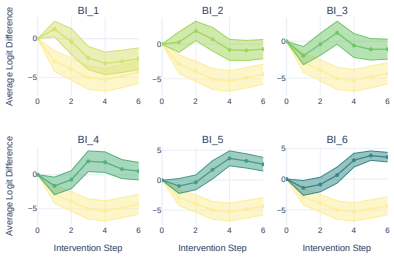

Figure 3: Logit Difference (LD) for OI-PC based intervention across datasets on Llama2-7B, where x axis denotes the number of intervention steps on e_0, y axis does the LD, BI_i represents each target attribute and the light yellow bottom line indicates the LD of original attribute (i.e., a_0). Here, $l = 8$, $v = 2.5$, and $\alpha = 3.0$.

Figure 4: Logit flip for OI-PC based intervention across datasets on Llama2-7B, where x axis denotes the number of intervention steps on e_0, y axis does the proportion of each inferred attribute in model output.

layers map the representations from the middle layers into the output embedding space for next-token prediction. According to the hypothesis, we would expect to find the ordering feature most prominently represented in middle layers, which is exactly what the visualization shows. We call this dimension that represents OI as OI Principle Component (OI-PC). In the following section, we apply causal intervention on the OI-PC to analyze how OI-PC affect the model output.

3 Causal Interventions on OI-PC

In order to test if OIs are not only encoded in the OI subspace, but that these representations can be steered so as to swap the binding and change LM's output, in this section, we perform interventions to analyze the causality. That is, we want to find out if making interventions along OI-PC leads to a change in LM's binding computation.

Activation Patching (AP) (19) has been recently proposed to causally intervene computational graph of a LM so as to interpret the function of a target computational node (or edge). Different with the common AP setup, we realize AP by directly editing activations along a particular direction (i.e., along OI-PC), similar to the activation editing method of (20; 21; 22).

3.1 Setting

Dataset To explore the internal representation that enables binding, we adopt the entity tracking dataset (8; 9). The dataset consists of English sentence describing a set of objects (here called attributes) located in a set of boxes with difference labels (here called entities), and the task is to infer what is contained by a given box. For instance, when a LM is presented with "The coffee is in Box Z, the stone is in Box M, the map is in Box H, ... Box Z contains the", the LM should infer the next token as "coffee". Each sample involves 7 EA pairs.

Metrics We apply two evaluation metrics: logit difference (23) and logit flip (24). The logit difference metric calculates difference in logits of a target token between original and intervened setting. The "logit flip" accuracy metric represents the proportion of candidate tokens in model output after a causal intervention.

3.2 Results: Direct Editing OI Subspace

We intervene via the Equation 1, where $\mathbf{x}_{0,l}$ is the original activation of e_0 (i.e., the leftmost entity) in layer l, $\mathbf{x}^*_{0,l}$ is the intervened activation, B_r is the OI subspace projection matrix mentioned in Section (§2), α is a hyper-parameter to scale the effect of intervention and β ($0 \leqslant \beta \leqslant 6$) denotes

the number of steps.

$$\mathbf{x}^*_{0,l} = \mathbf{x}_{0,l} + \alpha B_r^T (B_r \mathbf{x}_{0,l} + \beta v) \quad (1)$$

Table 1 lists several examples under the OI subspace intervention on the entity tracking dataset (8; 9). We can see that when adding 1 step along OI-PC, the model selects "stone" for entity "Z" instead of its original attribute "coffee". Similarly, when the step is doubled, the model will select attribute "map" for the entity, and so on. This indicates that changing the value along OI-PC can induce the swap of attribute.

Besides the qualitative analysis, we also conduct quantitative analysis for the causality between the OI subspace based AP and the binding behaviour of LMs. We plot mean-aggregated effect of the OI-PC based AP in Figure 3. Figure 3 indicates how the Logit Difference (LD) of each attribute changes as the step increases. We can observe that as the number of steps increases, LD of the original attribute decreases. In contrast, LD of other attributes gradually increase until a certain point and then gradually decrease. Given a candidate attribute, its LD peak roughly corresponds to the number of steps that is equal to its BI. For instance, when adding 3 steps, the points of BI_3 (i.e., attributes of BI= 3) achieve the highest LD score. This indicates that by adjusting the value along the OI-PC, we can adjust BI information and thus increase the logit score of the corresponding attribute.

Similarly, Figure 4 illustrates the relation between the number of steps and the logit flip, which gauges the percentage of the predicted attributes under an intervention. Figure 4 shows that as the step increases, the proportion bar becomes darker, it means that the model promotes the proportion of the corresponding attribute in its inference. For instance, when adding 3 step on the subspace, the a_3 (i.e., BI_3) becomes the major of the answers. This proves that the OI-PC based interventions can causally affect BI information as well as the computation of Binding in a LM.

OI Subspace and Other Information: the independence of OI subspace from positional information (i.e., *postion_ids*) is studied in Appendix (§A.1), and its relationship with the existence of a binding relation (e.g., "is in") is analyzed in Appendix (§A.2).

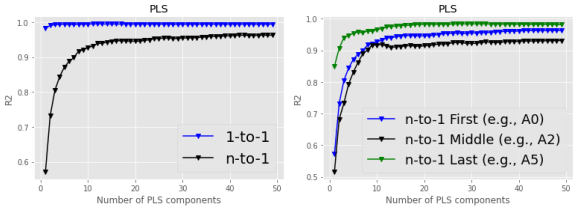

Figure 5: PLS components and R2 score

Input (1-to-1)
"A_0 is in E_0, A_1 is in E_1, A_2 is in E_2, A_3 is in E_3, A_4 is in E_4, A_5 is in E_5, A_6 is in E_6."
Input (n-to-1)
"A_0 is in $\mathbf{E_0}$, A_1 is in E_1, A_2 is in $\mathbf{E_0}$, A_3 is in E_2, A_4 is in E_3, A_5 is in $\mathbf{E_0}$, A_6 is in E_4."

Table 2: A n-to-1 sample where entity "E_0" has 3 attributes

3.3 OI Subspace for n-to-1 Setting

Section (§2) reveals that LMs encode OI into OI-PC under 1-to-1 setting, that is, one entity only has one attribute. In this session, we analyze how a LM encodes OI under n-to-1 setting, where one entity possesses multiple attributes. To do so, we create an alternative dataset as shown in Table 2, and analyze it via Partial Least Squares (PLS) (25). PLS aims to learn low-dimensional representation of the activation of an entity (e.g., E_0) that keeps (or predicts) the OIs of its corresponding multiple attributes (e.g., OI= 0, OI= 2 and OI= 5). Figure 5 (left) shows that compared to 1-to-1, in the n-to-1 setting, the LM has a high R2 score when PLS components are more than 20, indicating that a LM encodes OI via relatively high-rank subspace while dealing with multiple attributes. In addition, Figure 5 (right) shows the R2 score for each attribute, indicating that the encoding capacity of OI varies with the order of each bound attribute in the n-to-1 setting.

4 Conclusion and Future Work

In this work, we study the in-context binding, a fundamental skill underlying many complex reasoning and natural language understanding tasks. We provide a novel view of the Binding ID mechanism (6) that there exists a subspace in the activation of LMs that primarily encodes the ordering information and which is used as the prototype of BIs to causally determine binding. Our future work includes: 1. the analysis of OI subspace in a more realistic setting; 2. OI subspace based mechanistic analysis.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR20D2 and JSPS KAKENHI Grant Number 21K17814.

References

[1]Lauri Karttunen. Discourse referents. In Notes from the linguistic underground, pp. 363–385. Brill, 1976.

[2]Irene Heim. File change semantics and the familiarity theory of definiteness. Semantics Critical Concepts in Linguistics, pp. 108–135, 1983.

[3]Mante S Nieuwland and Jos JA Van Berkum. When peanuts fall in love: N400 evidence for the power of discourse. Journal of cognitive neuroscience, Vol. 18, No. 7, pp. 1098–1111, 2006.

[4]Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. Computational Linguistics, Vol. 34, No. 1, pp. 1–34, 2008.

[5]Hans Kamp, Josef Van Genabith, and Uwe Reyle. Discourse representation theory. In Handbook of Philosophical Logic: Volume 15, pp. 125–394. Springer, 2010.

[6]Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? arXiv preprint arXiv:2310.17191, 2023.

[7]Anne Treisman. The binding problem. Current opinion in neurobiology, Vol. 6, No. 2, pp. 171–178, 1996.

[8]Najoung Kim and Sebastian Schuster. Entity tracking in language models. arXiv preprint arXiv:2305.02363, 2023.

[9]Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. arXiv preprint arXiv:2402.14811, 2024.

[10]Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[11]AI@Meta. Llama 3 model card. 2024.

[12]Jinze Bai, Bai, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

[13]Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning, pp. 2397–2430. PMLR, 2023.

[14]Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. arXiv preprint arXiv:2409.05448, 2024.

[15]Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. A simple and effective method to eliminate the self language bias in multilingual representations. arXiv preprint arXiv:2109.04727, 2021.

[16]Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv preprint arXiv:2310.06824, 2023.

[17]Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. arXiv preprint arXiv:2310.15154, 2023.

[18]Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference? arXiv preprint arXiv:2406.19384, 2024.

[19]Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems, Vol. 33, pp. 12388–12401, 2020.

[20]Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa, and Kentaro Inui. Tracing and manipulating intermediate values in neural math problem solvers. arXiv preprint arXiv:2301.06758, 2023.

[21]Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. arXiv preprint arXiv:2403.10381, 2024.

[22]Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear, 2024.

[23]Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. arXiv preprint arXiv:2211.00593, 2022.

[24]Atticus Geiger, Wu, et al. Inducing causal structure for interpretable neural networks. In International Conference on Machine Learning, pp. 7324–7338. PMLR, 2022.

[25]Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. Analytica chimica acta, Vol. 185, pp. 1–17, 1986.

A Appendix

A.1 OI Subspace and Position

To prove the independence between OI subspace and Positional Information (PI), which is namely the *postion_ids* of input tokens, we create the following alternative dataset. The dataset is created by adding Filler Words (FW) with various length, such as "OK", "I see that" and "There is no particular reason", in front of the entity tracking dataset (8; 9), as shown in Table 3. Since the length (i.e., the number of tokens) of FW directly changes the PI of its following entities and attributes without affecting their OIs, we take the length as the measure of intervention on PI and apply Spearman's rank correlation ρ to calculate the correlation between the length (denoted as PI) and the OI-PC value. Figure 6 shows ρ between PI and OI-PC as well as between OI and OI-PC. We can observe that OI-PC has high ρ with OI but almost zero ρ with PI, indicating that the discovered OI-PC is highly correlated with OI information but independent with PI. Therefore, the OI-PC does not simply encode absolute token position.

Figure 6: Spearman's rank correlation between OI-PC and PI (or OI), where "PC*i*" denotes the *i*-th PC of the OI subspace and "PI" is the length of FW.

Input (original)
"The apple is in Box E, the bell is in Box F, ..."
Input (with filler words)
"*I will find out that* the apple is in Box E, the bell is in Box F, ..."

Table 3: An example of the dataset with filler words "*I will find out that*".

Input (original)
"The *apple* is in *Box E*, the *bell* is in *Box F*, ..."
Input (Non-related)
"I see *apple*, somewhere else there is *Box E*, the *bell* and *Box F* are scattered around, ..."

Table 4: An example of the dataset with non-related expression.

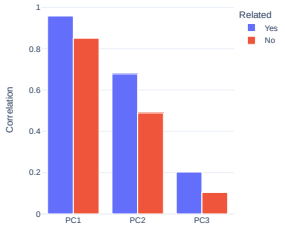

Figure 7: OI-PC based correlation between attributes and their corresponding entities, where "PC*i*" denotes the *i*-th PC of the OI subspace, "Yes" and "No" represent the related (i.e., original) and non-related dataset respectively.

A.2 OI Subspace and Relatedness

In order to uncover the relationship between OI-PC and the relatedness, which namely means the existence of a binding relation, we create an alternative dataset by converting relational expression into non-related one, as shown in Table 4. We can observe that non-related expression could make a target EA pair (e.g., "Box E'' and "apple") semantically unrelated but retain their OI (e.g., the OI of "apple'' and "bell" are still 0 and 1 respectively). We select Spearman's rank correlation ρ as the correlation metric and compare the ρ of the non-related dataset with the related one in the Figure 7.

We can observe that ρ of non-related dataset is slightly lower than the related (i.e., original) one, indicating that the OI-PC might contain limited relational information so that removing it can marginally decrease the ρ. However, there is still strong correlation between the non-related (or non-bound) entity attribute pair, indicating that the OI-PC primarily encodes the OI information but not the relatedness.