Cell-Based Representation of Relational 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

Understanding a discourse requires tracking entities and the relations that hold between them. While Large Language Models (LLMs) perform well on relational reasoning, the mechanism by which they bind entities, relations, and attributes remains unclear. We study discourse-level relational binding and show that LLMs encode it via a Cell-based Binding Representation (CBR): a low-dimensional linear subspace in which each "cell" corresponds to an entity–relation index pair, and bound attributes are retrieved from the corresponding cell during inference. Using controlled multi-sentence data annotated with entity and relation indices, we identify the CBR subspace by decoding these indices from attribute-token activations with Partial Least Squares regression. Across domains and two model families, the indices are linearly decodable and form a grid-like geometry in the projected space. We further find that context-specific CBR representations are related by translation vectors in activation space, enabling cross-context transfer. Finally, activation patching shows that manipulating this subspace systematically changes relational predictions and that perturbing it disrupts performance, providing causal evidence that LLMs rely on CBR for relational binding. Code and data are available at https://github.com/cl-tohoku/CBR-Subsapce.

1 Introduction

A core requirement for language comprehension is to keep track of entities and the relations between them as a discourse unfolds (Webber, 1979; Van Dijk et al., 1983; Zwaan and Radvansky, 1998). It is believed that comprehenders achieve this via a fundamental "binding" operation that, on some representational level, "binds together" the internal representations of entities among which a discourse relation holds (Treisman, 1996). For example, a reader may bind their internal representation of the *table* in Figure 1 (a) to that of *Australia* since the *manufactured in* relation holds between these two entities. Recent work has found evidence that Large Language Models (LLMs) are able to track entities across discourse and has started to uncover mechanisms supporting relational binding (Feng and Steinhardt, 2023; Kim and Schuster, 2023; Feng et al., 2024; Dai et al., 2024; Gur-Arieh et al., 2025). However, this line of research has primarily focused on very short texts involving only a small number of entities (see related work in §2), leaving discourse-level relational binding in LLMs largely unexplored. Here, we extend the scope of analysis towards discourse-level relational structures spanning multiple sentences and involving multiple entity–relation bindings, and show that relational binding in LLMs can be understood in terms of what we call **Cell-based Binding Representation (CBR)**.

A CBR consists of cells that are arranged in a more or less grid-like pattern in a linear subspace of activation space, with each cell corresponding to an entity-relation pair that the model will decode to its bound entity (called attribute) during inference. We assume relational triples of the form (e, r, a), where e denotes an entity, r a relation, and a an attribute. The ordering reflects that the attribute a is contextualized and bound to the entity e under relation r (e.g., (*table*, *manufactured in*, *Australia*) in Figure 1 (a)), consistent with the autoregressive encoding of LLMs in which later tokens are represented conditioned on the preceding context. As we will show, a CBR abstracts discourse-level relational structure into discrete entity indices ei and relation indices ri, enabling attributes to be represented as bound to specific $[ei, ri]$ pairs.

Furthermore, these indices are linearly decodable from model activations, revealing a low-dimensional and interpretable relational binding subspace organized along two dominant directions corresponding to entity indices ei and relation in-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47464–47524

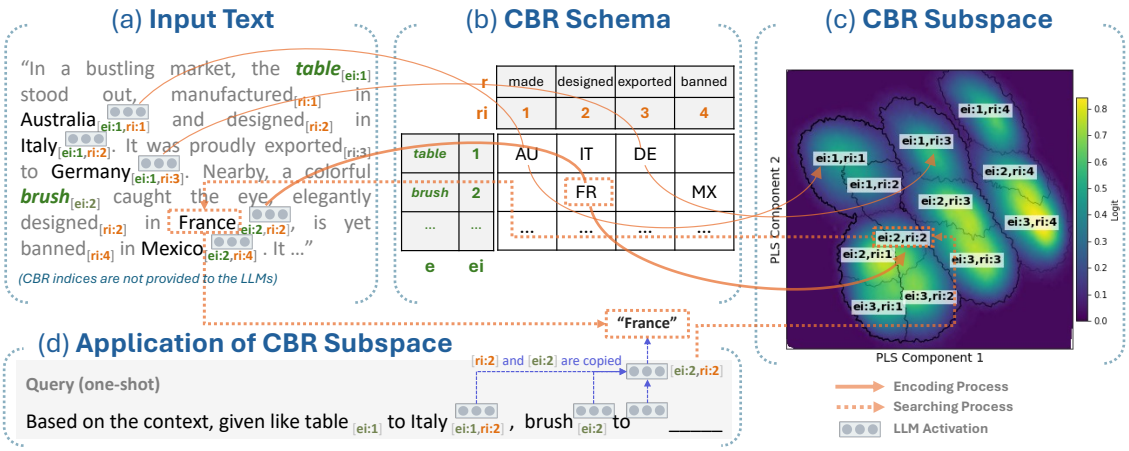


Figure 1: Overview of our Cell-based Binding Representation (CBR): (a) discourse annotated with entity and relation indices; (b) CBR schema, called Indexed Relational Schema, that binds entity index ei and relation index ri to their corresponding attributes; (c) visualization of the cells corresponding to each entity-relation index pair; and (d) cell–based retrieval, in which the model projects the query *brush* and hidden relation *designed in* onto the cell $[ei : 2, ri : 2]$ to retrieve the answer *France*.

dices ri. We further show that context-specific CBR representations are related through translation vectors in activation space, enabling cross-context transfer. Finally, through causal interventions using activation patching, we demonstrate that manipulating activations within this subspace systematically changes relational predictions, providing evidence that LLMs actively use this cell-based representational mechanism to bind and retrieve relational information over discourse.

2 Background: Relational Binding in Linguistics and LLMs

In linguistics, relational binding is a core feature of semantic formalisms such as Discourse Representation Theory (DRT Heim, 1982; Kamp, 2013). For example, the text shown in Figure 1 (a) can be represented in DRT as follows: $[x, y; \textit{table}(x), \textit{brush}(y), \textit{manufactured\_in}(x, \textit{Australia}), ..., \textit{designed\_in}(y, \textit{France}), ...]$. Here, entity x and *Australia* are bound to form a *manufactured_in* relation, while y is relationally bound to *France* through the predicate *designed_in*.

Recent work has found evidence that LLMs are able to track entities across discourse and has started to uncover mechanisms supporting relational binding. Kim and Schuster (2023) provide behavioral evidence that LLMs can learn to track entity states through sequences of state-changing operations, but also show that this capacity is not consistently present in text-only pretrained models and degrades as settings become more complex. Feng and Steinhardt (2023) identify a Binding ID mechanism which binds entity and attribute representations, and Feng et al. (2024) build on this by extracting explicit logical propositions from internal activations using propositional probes in a binding subspace. Dai et al. (2024) refine the Binding ID picture by identifying an Ordering ID that causally controls binding, while Gur-Arieh et al. (2025) show that models rely on a mixture of positional, lexical, and reflexive mechanisms.

Despite these advances, prior work has important limitations for understanding *discourse-level relational binding*. Much of the evidence and analysis is derived from very short contexts, typically involving a small number of entity-attribute pairs. This leaves open whether the same mechanisms scale to multi-sentence discourse, where relational structure is richer and where models must maintain multiple relations per entity and integrate information across sentences. Moreover, prior accounts focus on binding as a one-dimensional phenomenon (e.g., entity-to-attribute or order-based binding), whereas discourse-level semantics requires bindings indexed not just by *which entity*, but also by *which relation* holds for that entity.

Our work builds directly on these mechanistic insights, especially the idea that binding information is encoded in a low-dimensional and interpretable subspace and can be tested via causal interventions, but extends them to discourse-level relational structures spanning multiple sentences. Concretely,

we show that relational binding in LLMs can be understood in terms of a Cell-based Binding Representation (CBR): a grid-like organization of activation space in which each "cell" corresponds to an entity-relation pair $[ei, ri]$ that can be decoded to its bound attribute during inference.

3 Cell-based Binding Representation

This section presents our main results, proceeding in three steps. First, we identify and visualize the CBR subspace (§3.1). Second, we verify its causal relevance (§3.2). Finally, we analyze the generality and robustness of the CBR subspace across domains, templates, and model families (§3.3).

3.1 Identifying the CBR Subspace

Motivated by prior work on linear representations of relational binding in LLMs (Feng and Steinhardt, 2023; Feng et al., 2024; Dai et al., 2024), we hypothesize that LLMs encode discourse-level relational binding in a low-dimensional linear subspace of activation space and hence devise a linear probing method for identifying this subspace. Following Dai et al. (2024), we assume that this subspace encodes the indices of entities and relation types according to their order in the discourse.

We formalize entity and relation indices as what we term an Indexed Relational Scheme (IRS). In contrast to lexicalized discourse representations such as DRT, which encode relations using explicit predicate symbols (e.g., *manufactured_in*$(x, Australia)$), the IRS represents relational structure in a *delexicalized* form using discrete indices. Specifically, an IRS abstracts discourse-level bindings into *entity indices* ei and *relation indices* ri, corresponding to the order in which entities and relation types are introduced, and associates each attribute token with a specific index pair $[ei, ri]$. Formally, an IRS is a set of indexed triples $\{(ei, ri, a_{[ei,ri]})\}$, where $a_{[ei,ri]} \in A$ is an attribute bound to the entity-relation cell (ei, ri). For example, in Figure 1 (b), *Australia* is represented as bound to $[ei : 1,\ ri : 1]$ and *France* to $[ei : 2,\ ri : 2]$. This delexicalized indexing scheme provides a compact and testable target for probing how LLM activations encode relational bindings over discourse.

Data and models. We automatically generate multi-sentence discourses containing multiple entities, multiple relation types, and multiple attributes per entity, while ensuring that each attribute is uniquely bound to a specific *(entity, relation)* pair. Each discourse introduces entities and relation types in a controlled order, yielding a ground-truth IRS annotation for every attribute token, i.e., its entity index ei and relation index ri.

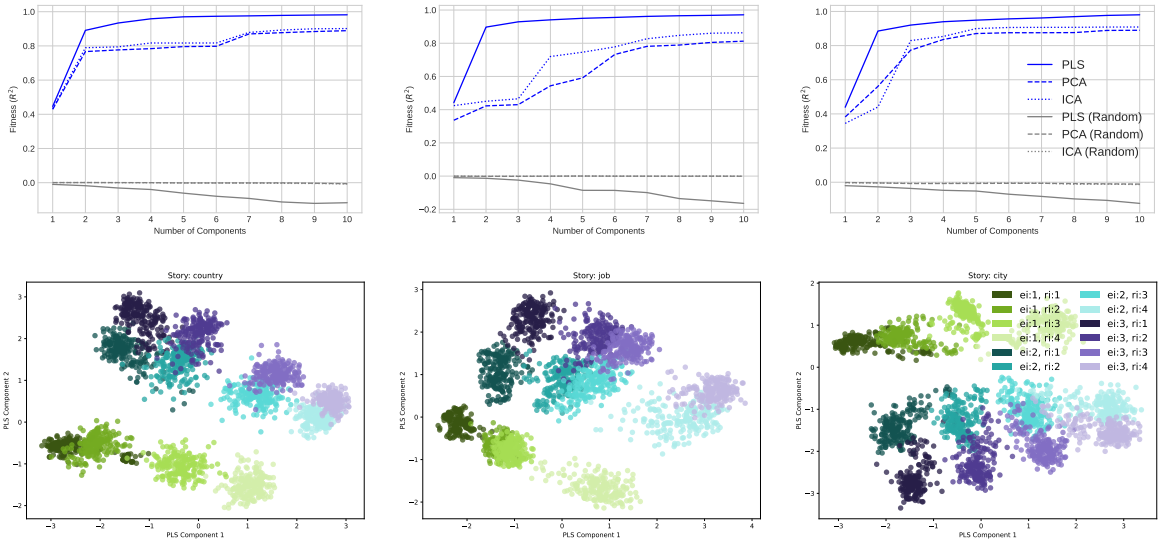
To test robustness across semantic domains and surface forms, we construct five discourse contexts that vary in entity, relation, and attribute inventories (e.g., countries, cities, occupations, and objects). We additionally use multiple templates and naturalistic paraphrased variants for each context to reduce reliance on superficial, repetitive positional cues and to better capture real-world discourse patterns. Appendix §A.2 provides full details on templates, sampling procedures, and dataset statistics. Overall, our experiments cover five domains [1] and two model families.

Method. To identify the CBR subspace, we fit a Partial Least Squares (PLS; Wold et al., 2001) regression model to map activations onto entity and relation indices ei and ri. For each attribute token in a discourse, we collect its activation vector $\mathbf{h}_i \in \mathbb{R}^d$ from a middle layer (layer 15 in our case), where d denotes the dimensionality of the activation. Each activation vector is paired with its index label $\mathbf{y}_i \in \mathbb{R}^2$ (i.e., $[ei, ri]$). Stacking all activations yields the matrix H and stacking all labels yields Y, and fitting a PLS yields a projection matrix W_{CBR} that projects the model activations H onto the directions in which H and Y maximally covary. In other words, PLS identifies the low-dimensional directions in activation space that are most predictive of the entity and relation indices, thereby giving us a candidate CBR subspace.

Results. We vary the number of components of PLS models and evaluate using goodness of fit. As shown for Llama3-8B-Instruct across three domains in Figure 2 (top) and for all models and domains in (§A.4) and (§A.5), PLS models achieve near-perfect fits with a small number of components, i.e., both entity and relation indices can be linearly decoded from low-dimensional subspace of activation space. We further evaluate decoding performance under monotonic transformations of the indices, details are provided in Appendix (§A.9)

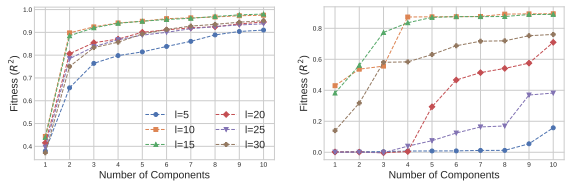
Projecting the activations of attribute tokens (e.g., "Australia" in Figure 1 (a)) onto the top two PLS components, we obtain the visualization

[1] In this paper, the terms "context" and "domain" are used interchangeably.



(a) Domain: *country* (b) Domain: *job* (c) Domain: *city*

Figure 2: **Top:** PLS goodness-of-fit when predicting entity and relation indices $[ei, ri]$ from Llama3-8B-Instruct attribute activations across three domains. For comparison, we also fit a Principal Component Analysis regression (PCA), Independent Component Analysis (ICA) regression and include random-label controls. **Bottom:** Visualization of attribute activations projected onto the top two PLS components, showing a grid-like distribution organized by entity index (ei) and relation index (ri).



(a) PLS (b) PCA

Figure 3: Layer-wise and component-wise analysis of CBR signal strength. Predictive performance of CBR indices across layers using PLS and PCA.

shown in Figure 2 (bottom) for Llama3-8B-Instruct and Figure 13 (§A.7) for Qwen3-8B. The plots reveal two dominant and interpretable directions: one that separates the points by ei and another that separates them by ri. Attributes associated with the same entity cluster along the ei direction, while those participating in the same relation align along the ri direction. These patterns suggest that LLMs encode relational binding structure in a low-dimensional linear subspace, supporting our hypothesis of a CBR subspace that jointly represents entity and relation indices.

In addition, to identify where the CBR signal is most prominent, we conducted layer-wise and component-wise analyses using both PLS and PCA in (§A.8). An example analysis for the *city* domain is shown in Figure 3. We observe that middle layers (approximately Layers 10–20) exhibit the strongest signal, with Layer 15 achieving peak or near-peak performance across contexts, while earlier and later layers show weaker alignment. In the PLS analysis, performance improves with dimensionality and stabilizes at around 2-5 components; fewer components reduce accuracy, whereas additional components yield only marginal gains. We therefore use Layer 15 and 2-5 components as empirically supported main experimental settings for the following analyses.

3.2 Causal Effect of the CBR Subspace

Effect on attribute prediction. To understand if and how LLMs use the CBR subspace for relational binding, we perform causal interventions via activation patching (Vig et al., 2020; Geiger et al., 2020, 2021; Wang et al., 2022; Stolfo et al., 2023; Heinzerling and Inui, 2024; Hanna et al., 2024), using a one-shot query setting as shown below (1). We opt for one-shot queries since alternatives such as providing preceding context (2) and using a direct query (3) would allow the model to rely on superficial cues and alternative strategies such as context matching via induction heads (Olsson et al., 2022). In contrast, the one-shot query format requires the model to rely on relational structure rather than surface form, as the correct answer must be inferred from the underlying entity–relation bindings.

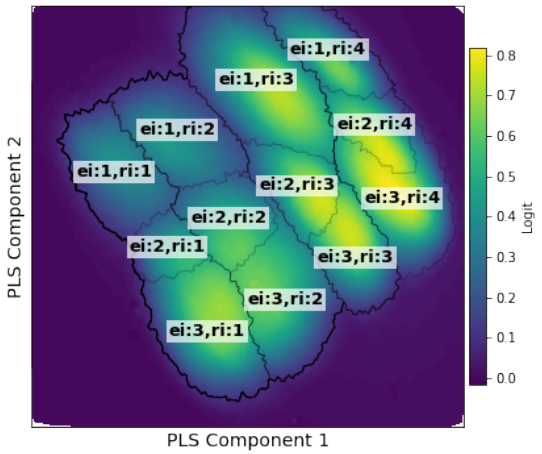


Figure 4: Logit landscape of attribute predictions resulting from causal interventions in the CBR subspace. Each "cell" corresponds to an entity–relation index pair $[ei, ri]$, with boundaries marking where the predicted attribute switches. See detailed explanation in §3.2.

(1) **Query** (one-shot): Based on the context, given like Sean to Perm, Jose to?

(2) **Context**: Sean, who hails from Phoenix, currently resides in Perm. ... Meanwhile, Jose was born in Austin and is now living in Berlin. ...

(3) **Query** (direct): Based on the context, Jose is now living in?

We now causally intervene on the CBR subspace by patching activations along the top two PLS directions, which we hypothesize to encode entity and relation indices. Concretely, we uniformly sample 10^4 points (denoted p_j) from the range defined by the minimum and maximum values of the learned CBR subspace obtained through the projection matrix W_{CBR} (see §3.1). The embedding of each point is denoted as $\mathbf{h}_{p_j}$. For each attribute instance in 50 randomly selected samples (e.g., "Berlin" in Sample 2), we gradually move its activation (denoted as $\mathbf{h}_{ei,ri}$) towards one of the sampled target points according to Equation 2, where α is a hyperparameter and $\mathbf{h}^*_{ei,ri}$ denotes the updated activation of an attribute, effectively sweeping the activation of attribute across the CBR plane. At each step, we compute the logit score of the corresponding attribute predicted by the LLM. For instance, given a Context 2 and a Query 1 for "Berlin", we patch its activation in the Context and observe the logit score of the corresponding attribute "Berlin". The resulting logit landscape is shown in Figure 4. Additional results for other datasets and corresponding results for Qwen3-8B are reported in §A.20.

$$\mathbf{s}_{ei,ri\to p_j} = \mathbf{h}_{p_j} - W_{\text{CBR}}\mathbf{h}_{ei,ri}, \quad (1)$$

$$\mathbf{h}^*_{ei,ri} = \mathbf{h}_{ei,ri} + \alpha W^T_{\text{CBR}}\mathbf{s}_{ei,ri\to p_j} \quad (2)$$

The logit landscape shows that the CBR subspace is partitioned into "cells" arranged in a grid-like pattern, with each cell corresponding to a specific entity–relation index pair $[ei, ri]$. Within each cell, the attribute bound to that particular index achieves the highest logit score, and the logit value decreases smoothly as the patched activation moves away from center of the cell. Also see the cross-section plots along ei and ri directions in Appendix A.21. Taken together, these results show that the identified CBR subspace is causally relevant for relation binding: Attributes are predicted based on the corresponding cell. In other words, relational binding in LLMs is causally mediated by the geometry of the CBR subspace.

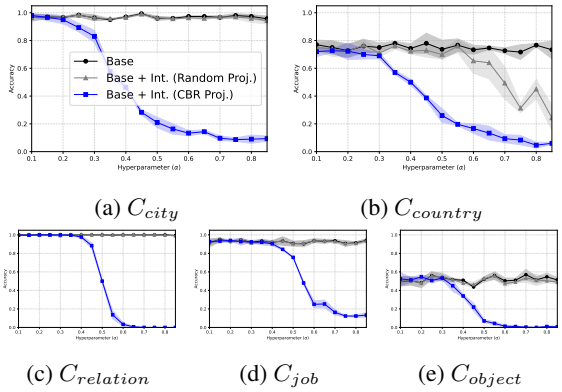
Control: Intervention along random directions. To test whether LLMs rely on the CBR subspace for relational binding, we also perturb attribute activations along the CBR directions using Equation 4 and compare the results with perturbations along a random subspace defined by a randomly generated projection matrix in Equation 3. An example of the CBR subspace distribution after the perturbation (i.e., $\mathbf{h}^*_{ei,ri}$) is visualized in Figure 71 (§A.22). We use the same one-shot query setting to query each attribute in the context.

$$\mathbf{h}^*_{ei,ri} = \mathbf{h}_{ei,ri} + \alpha W^T_{\text{rand}}(W_{\text{CBR}}\mathbf{h}_{ei,ri}), \quad (3)$$

$$\mathbf{h}^*_{ei,ri} = \mathbf{h}_{ei,ri} + \alpha W^T_{\text{CBR}}(W_{\text{CBR}}\mathbf{h}_{ei,ri}) \quad (4)$$

If the LLM indeed uses the CBR subspace to make predictions, perturbing activations along this subspace should degrade performance. Figure 5 shows the results as a function of perturbation strength. The corresponding results for Qwen3-8B are reported in Appendix (§A.22). We observe that as the perturbation weight increases, the accuracy of attribute predictions decreases significantly. In contrast, perturbation along the random subspace has little or no effect on accuracy. These results provide strong evidence that LLMs utilize the CBR subspace when predicting attributes: disruptions in this subspace directly impair model performance, whereas unrelated directions do not.

CBR Subspace based Mechanism To understand how LLMs use the CBR subspace to retrieve



(a) C_{city} (b) $C_{country}$

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 5: Effect of perturbing activations along the CBR subspace versus a random subspace on Llama3-8B-Instruct. The X-axis shows the perturbation weight α in Equation 3 and 4, and the Y-axis shows the attribute prediction accuracy. Perturbations along the CBR subspace (i.e., blue line) lead to a significant drop in accuracy, while perturbations along a random subspace (i.e., grey line) have minimal effect. This indicates that LLMs rely on the CBR subspace to make relationally bound predictions.

the correct attribute given a context (e.g., Sample 2) and a query (e.g., Query 1), we propose a high-level mechanism illustrated in Figure 1 (d). When answering a one-shot relational query, the model appears to perform two parallel operations: (i) it extracts relation index information (e.g., $ri:2$) from the attribute exemplar provided in the one-shot part, and (ii) it extracts entity index information (e.g., $ei:2$) from the query entity itself. These two indices together define a point corresponding to a cell in the CBR subspace, which is encoded into the activation of the last token. The model then uses this representation to locate the answer attribute in the context whose indices match this combination.

To test this mechanism, we perform several Activation Patching (AP) interventions that steer model activations along specific CBR subspace directions: (a) Relation-index steering that shifts ri in the one-shot attribute from $ri:j$ to $ri:j+1$, as shown in Figure 72 (§A.23), (b) Entity-index steering that shifts ei in the query entity from $ei:j$ to $ei:j+1$, as shown in Figure 73 (§A.23), (c) Last-token steering that modifies ei or ri at the final token before prediction, as shown in Figure 74 (§A.23). The AP is applied using Equation 5 and 6, where $\mathbf{s}_{j\to j+1}$ is the steering vector that shifts ri (or ei) from j to $j+1$, $\mathbf{h}_j$ denotes the activation of a target token from middle layers [2] whose ri (or ei) is j, α is a hyperparameter, and W_{CBR} is learned from the corresponding tokens, following the method described in Section 3.1.

[2]We select Layers 10-20, and set $\mathbf{s}_{j\to j+1} \in \mathbb{R}^5$.

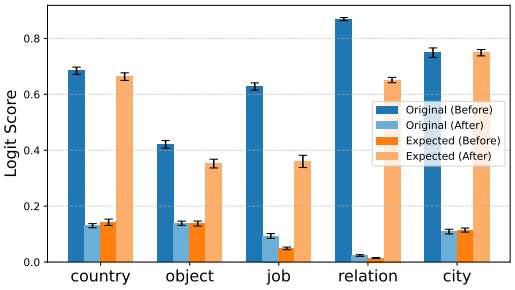


Figure 6: Activation patching via Relation-index (i.e., ri) steering on the activation of the attribute token (e.g., "Italy") in query part across five contexts on Llama3-8B-Instruct. We show the change in logit scores for the original answer and the expected answer before and after activation patching, which are denoted as "Original (Before)", "Original (After)", "Expected (Before)" and "Expected (After)" respectively.

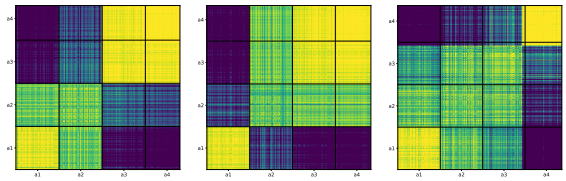
$$\mathbf{s}_{j\to j+1} = \frac{1}{n}\sum_{k=1}^{n}\left(W_{\text{CBR}}\mathbf{h}^k_{j+1} - W_{\text{CBR}}\mathbf{h}^k_j\right), \quad (5)$$

$$\mathbf{h}^*_j = \mathbf{h}_j + \alpha W^T_{\text{CBR}}\mathbf{s}_{j\to j+1} \quad (6)$$

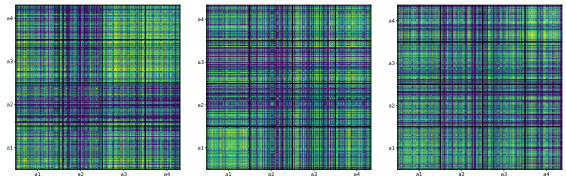
Following Wang et al. (2022), we evaluate the effect of steering by measuring the change in logit scores for both the original correct answer and the expected answer after intervention. As shown in Figure 6, steering along the CBR subspace direction consistently suppresses the logit of the original answer and increases the logit of the expected answer, precisely in line with the intended index manipulation. The results for (b) Entity-index steering and (c) Last-token steering are reported in Appendix (§A.23). The corresponding results for Qwen3-8B are presented in Appendix (§A.24). These results demonstrate that LLMs rely on CBR-subspace representations to retrieve attributes, and that targeted activation patching along CBR directions can systematically alter the model's predictions.

3.3 Generality of the CBR Subspace

Does the CBR subspace encode semantics or is it merely positional? To examine whether the CBR subspace captures semantic information in addition to indices, we construct an additional dataset in which relations exhibit controlled patterns of semantic similarity. As shown in Table 1, the four relations are grouped so that: (i) the first and last



(a) 2to2 (b) 1to3 (c) 3to1



(d) 2to2 (Rand.) (e) 1to3 (Rand.) (f) 3to1 (Rand.)

Figure 7: Cosine similarity heatmaps of attribute representations projected into the CBR subspace (above) and a random subspace (below).

two relations share similar meaning (denoted as "2to2"), (ii) the last three relations share similar meaning (denoted as "1to3"), and (iii) the first three relations share similar meaning (denoted as "3to1"). For instance, in the "1to3" pattern, the last three attributes of "Bob": "editor", "guard" and "student" are bound to semantically similar relations. This design allows us to test whether semantic similarity among relations is reflected in the geometry of the CBR subspace.

Using the projection matrix W_{CBR} learned from the main dataset (§A.2), we project attribute (e.g., "$a1$" in Table 1) activations into the CBR subspace and compute pairwise cosine similarities between projected representations. We compare these results with a control condition where activations are projected using a random matrix W_{rand} of the same dimensionality. The resulting cosine similarity heatmaps for Llama3-8B-Instruct are shown in Figure 7. Additional results for other datasets are reported in Appendix (§A.11), and corresponding results for Qwen3-8B are presented in Appendix (§A.10).

The CBR-projected representations clearly reproduce the intended semantic similarity structure: relations designed to be semantically close form coherent similarity blocks, while semantically distant relations remain well separated. In contrast, the random projection produces a comparatively less clear and meaningful structure. These results suggest that the CBR subspace embeds semantic information such that attributes bound to semantically similar relations yield similar representations in the CBR subspace, and thus similar indices.

Pattern	**Discourse**
2to2	Paul currently works as a writer and serves as a doctor … Meanwhile, Bob currently works as a coach$_{a1}$ and serves as an editor$_{a2}$, thriving in his creative endeavors. Yet, he hates being a guard$_{a3}$ and dislikes being a student$_{a4}$, feeling constrained in those roles. …
1to3	Paul currently works as a writer, … Meanwhile, Bob is a coach$_{a1}$, but he too feels the weight of frustration, particularly when he thinks about being an editor$_{a2}$. He hates the idea of being a guard$_{a3}$ and finds no joy in the thought of being a student$_{a4}$. …
3to1	Paul currently works as a writer, … Meanwhile, Bob works as a coach$_{a1}$ and serves as an editor$_{a2}$. He takes on the role of a guard$_{a3}$, but he dislikes being a student$_{a4}$. …

Table 1: Samples of the dataset for semantic information analysis in the CBR subspace, where "$a1$" denotes the first attribute for a given entity (e.g., "Bob"), and so on.

Does the CBR subspace generalize across domains? As described in §3.1, we learn an CBR projection matrix $W_{\mathbf{CBR}}$ independently for each context. Here, we analyze the generality of these $W_{\mathbf{CBR}}$: specifically, whether a $W_{\mathbf{CBR}}$ learned from one context can effectively recover index information in another. Therefore, we further hypothesize that the CBR subspace may exhibit shared structure across contexts. In particular, we consider two possibilities.

Hypothesis 1 (Global CBR subspace). The CBR subspace is global across contexts. Under this hypothesis, a single context-independent projection matrix W extracts CBR indices from all contexts, such that $Wh_{c_1} \approx Wh_{c_2}$, where $h_c \in \mathbb{R}^d$ denotes the activation vector under context c, and $W \in \mathbb{R}^{k\times d}$ projects activations into the CBR subspace.

Hypothesis 2 (Context-dependent subspace with consistent structure). The exact alignment of the CBR subspace may vary across contexts, but the mappings between these context-specific subspaces follow a consistent second-order structure. In particular, activations across contexts may be related by a translation in activation space:

$$W_{c_1} h_{c_1}^{(ei,ri)} \approx W_{c_1}\left(h_{c_2}^{(ei,ri)} + \Delta_{c_2\to c_1}^{(ei,ri)}\right), \quad (7)$$

where $h_c^{(ei,ri)} \in \mathbb{R}^d$ denotes the activation vector under context c with CBR indices (ei, ri), $W_{c_1} \in \mathbb{R}^{k\times d}$ is the projection matrix learned in context c_1, and $\Delta_{c_2\to c_1} \in \mathbb{R}^d$ is a context-dependent translation vector aligning activations from context

(a) w/o Δ (b) w/ Δ

Figure 8: Each cell shows the R^2 fitness score obtained from Llama3-8B-Instruct. The projection matrix learned from one context (column) is used to predict the index information of another context (row).

c_2 (e.g., *country*) with those from context c_1 (e.g., *relation*).

To evaluate this, we compute cross-context fitness scores. For each source (or trained) context, we apply its $W_{\mathbf{CBR}}$ to the activations from a different target context and measure how well the projected representations predict the index information of target context. The resulting scores are shown in Figure 8a for Llama3-8B-Instruct and Figure 25 (§A.12) for Qwen3-8B.

We observe that the cross-context R^2 scores vary across contexts, suggesting that the first hypothesis is less likely, while the second hypothesis is more plausible. To further examine the second hypothesis, we learn a translation vector using Equation 8, defined as the mean difference between activation vectors under two contexts for the same (ei, ri) pair. The results after applying this translation via Equation 7 are shown in Figure 8b. The improved R^2 scores support the **Hypothesis 2**, suggesting that while position of the CBR subspace varies across contexts, the mappings between these context-specific subspaces exhibit a consistent second-order structure. Further ablation analyses on $\Delta_{c_2 \to c_1}$ are presented in Appendix (§A.13), and a discussion of the underlying reason is provided in Appendix (§A.14).

$$\Delta_{c_2 \to c_1}^{(ei,ri)} = \mathbb{E}\left[h_{c_1}^{(ei,ri)} - h_{c_2}^{(ei,ri)}\right] \quad (8)$$

How stable is the CBR subspace? To examine the stability of the CBR subspace, we evaluate its behavior under controlled discourse perturbations. By permuting the order of subsequent mentions while keeping the original introduction fixed, we test two hypotheses: (1) ri encodes introduction order as structural discourse information, or (2) ri merely reflects surface-level token order.

To this end, we introduce two types of perturbations. First, we ablate certain attributes and relations associated with the second and third entities while leaving the relations of the first entity intact, as illustrated in Figure 9 (a). This manipulation preserves the original entity and relation indices but alters surface-level content. Second, we shuffle the order of attributes for the second and third entities while keeping the original indices unchanged.

Representation Stability. In both cases, the resulting CBR subspace visualizations preserve a similar geometric organization: although individual points exhibit some positional changes relative to the base input, the distribution continues to align with the directions corresponding to ei and ri (Figure 9 (b)). These results indicates that the overall structure of the CBR subspace is robust such permutations.

Stability of CBR Index Prediction. We assess whether projection matrices learned under these permuted conditions generalize across datasets. Using the CBR projection matrix learned from the ablated or shuffled dataset to predict index information in the original dataset (and vice versa), we observe high predictive performance (R^2 is around 0.8) in Figure 9 (c). This further confirms that the CBR subspace is highly consistent and thus its overall structure is preserved through ablation and shuffling. Additional results for other datasets are reported in Appendix (§A.15), and corresponding results for Qwen3-8B are presented in Appendix A.16. Overall, these results reconfirm our conclusions about the stability of CBR subspace.

Stability of the CBR-Based Mechanism. Our activation steering analysis, introduced in (§3.2), continues to causally manipulate model outputs under the perturbed conditions in the same manner as in the base setting (Appendix A.25). This provides complementary evidence that the index-based mechanism remains functionally intact despite perturbation of the surface geometry.

Prevalence of CBR Across Diverse Patterns. To rule out the possibility that the observed CBR structure is a pattern-specific artifact, we construct 13 distinct discourse patterns with progressively increasing structural complexity and variation, such as pattern 5: (e_1r_1, e_1r_2, e_2r_1, e_2r_3, e_3r_2, e_3r_4) and pattern 8: (e_1r_1, e_1r_2, e_2r_1, e_2r_3, e_2r_4, e_3r_1, e_3r_2, e_3r_4), where e_ir_j denotes the attribute binds the i-th entity through the j-th relation , as detailed in the Appendix (§A.19). Across all patterns, the CBR index prediction scores (measured by R^2)

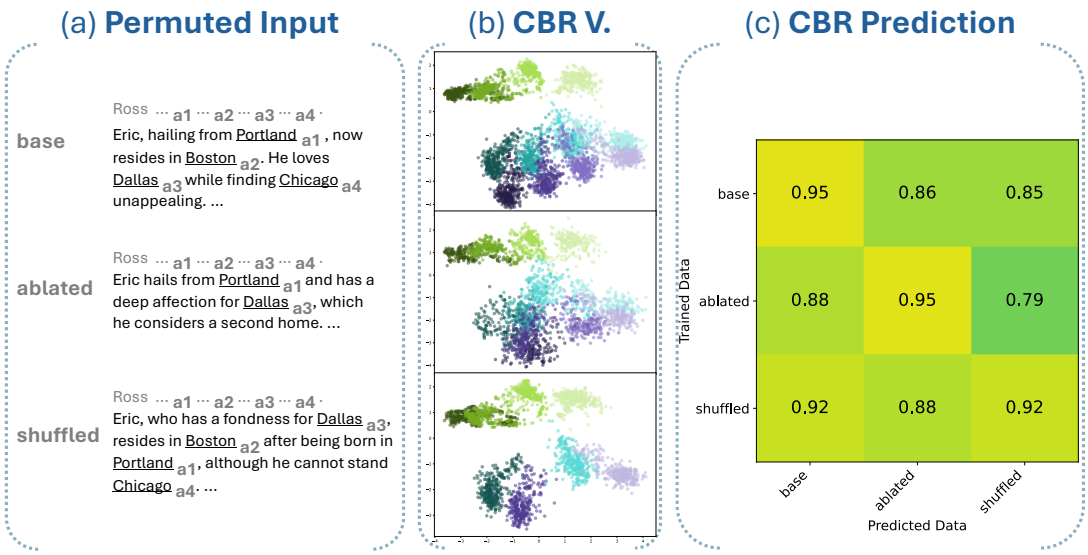


Figure 9: (a) Samples of Ablated and Shuffled Dataset from C_{city}. The ablated sample removes the attribute "a2" (i.e., "Boston") and "a4" (i.e., "Chicago"), while the shuffled sample alters the order of "a1" (i.e., "Portland") and "a3" (i.e., "Dallas"). (b) Visualization of the corresponding CBR subspace. (c) Cross-dataset R^2 scores for CBR index prediction on Llama3-8B-Instruct.

remain stable and consistent at around 0.95 across contexts. The observed CBR index–based structure is unlikely to be an artifact of a specific repetitive pattern. In addition, Appendix A.18 shows that the CBR structure persists under reference-distance manipulations, indicating that the indexing scheme captures more than superficial cues.

How to apply the CBR Subspace? Our analysis suggests that the CBR subspace provides a useful representation for monitoring relational bindings in LLMs. We compare our method with the Hessian-based approach of Feng et al. (2024), which proposes propositional probes for monitoring latent world states in LLMs and identifying binding entities in discourses such as: "*The nurse lives in Singapore. The CEO lives in Canada. The person living in Singapore is male. The person living in Canada is female.*" In this setting, the correct binding is *CEO–female* rather than the stereotypical association *CEO–male*. Using the same setup, our CBR-based method predicts the correct bound entity index ei from the activation of the attribute token (e.g., "*female*") and achieves accuracies of 0.95 (pro) and 0.94 (anti), which are higher than the reported mean accuracies of the Hessian-based baseline. These results indicate that CBR could effectively capture binding information and monitor internal world states in LLMs. Details of the evaluation are provided in the Appendix (§A.28).

The CBR structure may also benefit downstream NLP tasks such as Relation Extraction (RE). The CBR representation could support more stable RE and enable new approaches for interpreting and controlling LLM reasoning. An analysis of CBR indices on a commonly used real-world document-level RE dataset is provided in the Appendix (§A.29). The results indicate that the CBR signal could be identified in the real-world dataset when the target samples contextually resemble training examples.

4 Conclusions

In this work, we investigated how LLMs internally represent relational binding in discourse. To this end, we proposed the Cell-based Binding Representation (CBR) framework and applied PLS to identify a low-dimensional subspace that linearly encodes entity and relation indices. Visualizations reveal two interpretable directions corresponding to entity and relation indices. Causal interventions show that manipulating the CBR subspace reliably alters LLMs predictions, supporting a CBR subspace based mechanism in which models retrieve attributes by combining the index information embedded in activations of LLMs. These findings suggest that discourse-level relational binding in LLMs is supported by a structural, compositional, and low-dimensional representation, which extends the Linear Representation Hypothesis.

Limitations

Several limitations remain in this work. First, our analysis operates at the level of model activations, linear subspaces and CBR related head detection[3], and we do not perform fine-grained, head-level or neuron-level circuit analysis to localize the exact transformer components that implement the CBR-based mechanism. Second, while we empirically demonstrate the existence of a CBR subspace, we do not investigate how this subspace is generated during the model's forward computation. In particular, this work does not address how LLMs select entities that carry index information, how the model resolves situations where multiple entities are present but only a subset participates in relational structures, or how to assign index information when the role of an entity shifts (e.g., from entity to attribute) as discourse unfolds.

In addition, although we use GPT-4o-mini to generate more naturalistic narrative discourse and apply various permutations to simulate the feature of real world discourse, our datasets are still synthetic and controlled in nature. They represent human-like but machine-generated text, and thus may not fully capture the complexity and variability of the real-world discourse. As a result, we do not investigate the extent to which the CBR schema generalizes to or breaks down in naturally occurring human text.

Finally, although we propose the CBR subspace based mechanism for explaining relational binding, LLMs may also rely on other mechanisms, such as surface-level cues or statistical regularities, to encode binding information. This work does not analyze the potential interactions between these mechanisms or their relative contributions to binding behavior.

Ethical Statement

LLMs (i.e., Llama3-8B-Instruct, Qwen3-8B, GPT-4o-mini) are applied according their intended research purposes and licenses. In addition, we construct synthetic datasets by following the framework of existing benchmarks, with entities and attributes sampled from diverse sets of single-token names and other concepts. Because our data are automatically generated and do not involve human annotation, this work does not introduce risks associated with annotation bias or the disclosure of personal or sensitive information. The datasets and code are publicly available to ensure the reproducibility of our experiments.

[3]We discuss this in Appendix (§A.30).

Acknowledgements

This work was supported by AMED Grant Number JP25wm0625405 and Japan Science and Technology Agency under Grant No. JST BOOST JPMJBY24F9.

References

Meta AI. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xin Wei Chia, Swee Liang Wong, and Jonathan Pan. 2025. Probing latent subspaces in llm for ai security: Identifying and manipulating adversarial states. *arXiv preprint arXiv:2503.09066*.

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)*, pages 8493–8502.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. 2024. Representational analysis of binding in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17468–17493.

Ahmed Oumar El-Shangiti, Tatsuya Hiraoka, Hilal AlQuabeh, Benjamin Heinzerling, and Kentaro Inui. 2025. The geometry of numerical reasoning: Language models compare numeric properties in linear subspaces. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–561.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2024. Monitoring latent world states in language models with propositional probes. *arXiv preprint arXiv:2406.19501*.

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

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the third blackboxnlp workshop on analyzing and interpreting neural networks for NLP*, pages 163–173.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

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

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. 2025. Mixing mechanisms: How language models retrieve bound entities in-context. *arXiv preprint arXiv:2510.06182*.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2024. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, and 1 others. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Irene Roswitha Heim. 1982. *The semantics of definite and indefinite noun phrases*. University of Massachusetts Amherst.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195.

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

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Hans Kamp. 2013. A theory of truth and semantic representation. In *Meaning and the Dynamics of Interpretation*, pages 329–369. Brill.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of llms: Stages of inference? *arXiv preprint arXiv:2406.19384*.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

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

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

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Adam Paszke, Sam Gross, Francisco Massa, and 1 others. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Baturay Saglam, Paul Kassianik, Blaine Nelson, Sajana Weerawardhena, Yaron Singer, and Amin Karbasi. 2025. Large language models encode semantics in low-dimensional linear subspaces. *arXiv preprint arXiv:2507.09709*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 8472–8487.

Matthieu Tehenan, Christian Bolivar Moya, Tenghai Long, and Guang Lin. 2025. Linear spatial world models emerge in large language models. *arXiv preprint arXiv:2506.02996*.

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

Anne Treisman. 1996. The binding problem. *Current opinion in neurobiology*, 6(2):171–178.

Teun Adrianus Van Dijk, Walter Kintsch, and 1 others. 1983. Strategies of discourse comprehension.

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

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

Bonnie Lynn Webber, editor. 1979. *A Formal Approach to Discourse Anaphora*, 1 edition. Routledge, London.

Svante Wold, Michael Sjöström, and Lennart Eriksson. 2001. Pls-regression: a basic tool of chemometrics chemometr. *Intell. Lab*, 58(2):109–130.

Thomas Wolf, Lysandre Debut, Victor Sanh, and 1 others. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2024. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36.

An Yang and Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. 2024. Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution. *arXiv preprint arXiv:2410.00153*.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162.

A Appendix

A.1 Related Work

Linear Representation Recent research on the Linear Representation Hypothesis (LRH) shows that language models encode a wide range of semantic and structural concepts as linear directions or low-dimensional subspaces. Prior work has found linear representations for Othello board states (Li et al., 2022; Nanda et al., 2023), truth values of propositions (Marks and Tegmark, 2023), sentiment (Tigges et al., 2023), semantic concepts (Zhao et al., 2024; Saglam et al., 2025), spatial information (Tehenan et al., 2025), safe and jailbroken states (Chia et al., 2025) and numeric attributes such as elevation, population, and dates (Gurnee and Tegmark, 2023; Heinzerling and Inui, 2024; El-Shangiti et al., 2025).

In addition, several studies have demonstrated that various forms of binding exhibit linear structure in LLM representations (Feng and Steinhardt, 2023; Dai et al., 2024; Feng et al., 2024; Gur-Arieh et al., 2025). For example, given an input such as "Alice lives in Laos.", Feng et al. (2024) typically analyzes the binding subspace between the entity "Alice" and the attribute "Laos". Gur-Arieh et al. (2025) reveal several mixing mechanisms such as positional, lexical and reflexive mechanism. Such analyses are limited to pairwise bindings between an entity and an attribute, and therefore fail to capture the full binding structure among entities, relations, and attributes. Moreover, these approaches primarily focus on entity–attribute pair binding within individual sentences and do not examine discourse-level relational structures, which are fundamental to discourse understanding for LLMs. To address these gaps, we show that LLMs encode discourse-level relational structure via a Cell-based Binding Representation (CBR). We analyze relational binding at the level of entire discourses, capturing structures that involve three or more interacting elements. In addition, we introduce IRS as a systematic annotation scheme for probing how LLMs bind entities, relations, and attributes into coherent discourse-level representations.

Learned Knowledge Prior work has attempted to localize and edit factual relations, such as "capital of", that language models acquire during pretraining and store in their parameters (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022; Geva et al., 2023; Hernandez et al., 2023). These studies primarily examine how static knowledge is encoded in model weights. In contrast, our work focuses on in-context representations of relations and investigates how relational structure is encoded in model activations during discourse processing.

Mechanistic Analysis Substantial progress has been made in uncovering internal circuits that implement specific computations in language models (Elhage et al., 2021; Wang et al., 2022; Wu et al., 2024). Recent work has begun to analyze mechanisms for entity tracking and binding, including the identification of circuits for entity tracking (Prakash et al., 2024), the proposal of a Binding ID (Feng and Steinhardt, 2023) and binding subspace based high-level mechanism (Feng et al., 2024) to explain binding. While these studies provide valuable insights into sentence-level or pairwise binding, they do not capture the structured index information over extended discourse, especially entity–relation bindings. Our work complements this line of research by introducing CBR subspace based mechanism and empirically analyzing how both entity and relation indices are represented and combined in model activations to support discourse-level relational binding.

A.2 Datasets and Language Models for CBR Subspace Analysis

Datasets To analyze the CBR subspace, we construct a controlled dataset that systematically varies entity and relation indices while keeping semantic content simple and interpretable. Following the data curation method (Feng et al., 2024), we begin by defining a closed world consisting of five contexts: $C_{relation}$, C_{object}, C_{city}, C_{job} and $C_{country}$, and each context (e.g., $C_{country}$) includes four relation types (e.g., "manufactured in", "designed in", "exported to" and "banned in"), reflecting distinct event roles, as shown in the first row of Table 5 (§A.3). Entity and attribute for a relation type are sampled from five sets of one-token words: S_{name}, S_{object}, S_{city}, S_{job} and $S_{country}$, as shown in Table 4 (§A.3). Here, S_{name}, for instance, denotes a set of $name$ tokens such as "Eric".

Taking $C_{country}$ as an example, we first sample three entities (e.g., "boot") from S_{object} and twelve attributes (e.g., "Mexico") from $S_{country}$ as shown in Table 2 and organize them into a structured table (i.e., Table Template Input in Table 3) that specifies which attribute is bound to which entity and relation. These tables are then instantiated into discourse templates to generate concise descriptions of the relational structure (i.e., the Discourse Template Input in Table 3). The CBR subspace visualization and causal intervention results for this template input are shown in Appendix A.26 and Appendix A.27, respectively. Finally, to produce naturalistic inputs suitable for analyzing LLM activations, we prompt GPT-4o-mini to rewrite each structured description into a short, coherent narrative (i.e., Story Input in Table 3) while preserving the underlying relational bindings. This pipeline ensures that surface variation does not obscure the latent entity–relation indices, enabling precise analysis of how these indices are encoded in model representations. Each context contains $1,000$ naturalistic narratives. Additional samples for Table Template Input, Discourse Template Input and Story Input are shown in Table 5, 6 and 7 in Appendix (§A.3) respectively.

	Entity	**Attribute**
$C_{country}$	boot, radio, chair (from S_{object})	Mexico, Jordan, Turkey, India, Japan, Brazil, France, Canada, Sweden, Argentina, Australia, Spain (from $S_{country}$)
$C_{relation}$	Eric, Ian, Dan (from S_{name})	Tara, Brad, Jack, Nick, ... (from S_{name})
$C_{...}$	...	...

Table 2: Samples of Entity and Attribute.

Language Models We adopt Llama3-8B-Instruct (AI, 2024) and Qwen3-8B (Yang and Team, 2025) for CBR subspace analysis. Llama3-8B-Instruct is a 32-layer Transformer with a hidden dimension of 4096, while Qwen3-8B consists of 36 Transformer layers with a hidden size of 4096.

	Table Template Input
$C_{country}$	| Product | Manufactured in | Designed in | Exported to | Banned in | | boot | Mexico | Jordan | Turkey | India | | radio | Japan | Brazil | France | Canada | | chair | Sweden | Argentina | Australia | Spain |
$C_{relation}$	| Name | Spouse | Child | Teacher | Boss | | Eric | Tara | Brad | Jack | Nick |...
$C_{...}$	...
	Discourse Template Input
$C_{country}$	The boot is manufactured in Mexico and designed in Jordan, and it is exported to Turkey, but it is banned in India . The radio is manufactured in Japan and designed in Brazil, and it is exported to France, but it is banned in Canada . The chair is manufactured in Sweden and designed in Argentina, and it is exported to Australia, but it is banned in Spain .
$C_{relation}$	Eric is married to Tara and has a child named Brad, he was taught by Jack and works under Nick . ...
$C_{...}$	...
	Story Input
$C_{country}$	In a bustling market, a unique boot caught everyone's attention, as it was manufactured in **Mexico** and designed in Jordan. It had made its way to Turkey for export, but unfortunately, it faced a ban in India. Meanwhile, a sleek radio, manufactured in Japan and designed in Brazil, was shipped to France, yet it couldn't be sold in Canada due to restrictions. Lastly, a stylish chair, crafted in Sweden and designed in Argentina, found its way to Australia, but it was banned in Spain. Each item told a tale of international trade and the complexities of global regulations.
$C_{relation}$	Eric lives happily with his spouse Tara and they have a child named Brad. He fondly remembers being taught by Jack and appreciates working under Nick. Meanwhile, ...
$C_{...}$	...

Table 3: Samples of Dataset.

	$\|S\|$	**Sample**
S_{name}	47	"Ray", "Eric", "Leo", "Ross", "James", "Matt", "Brad", "Jeff", "Todd", ...
S_{object}	51	"window", "glass", "door", "paper", "book", "toy", "mirror", "ball", "clock", ...
S_{city}	20	"Atlanta", "Seattle", "Phoenix", "London", "Hamilton", "Boston", "Kansas", "Toronto", "Miami", ...
S_{job}	16	"writer", "student", "driver", "artist", "editor", "actor", "athlete", "guard", "chef", ...
$S_{country}$	23	"Georgia", "India", "Japan", "Spain", "Italy", "Australia", "China", "Russia", "Egypt", ...

Table 4: Samples of One-token Words.

A.3 Additional Data Samples

	Table Template Input
$C_{country}$	| Product | Manufactured in | Designed in | Exported to | Banned in | | ring | Russia | France | Iraq | Singapore | | plant | China | India | Australia | Jordan | | jar | Sweden | Iran | Pakistan | Georgia |
C_{city}	| Name | Birthplace | Lived City | Loved City| Disliked City | | Brad | Paris | Houston | London | Detroit | | Gary | Austin | Berlin | Chicago | Portland | | James | Hamilton | Split | Atlanta | Dallas |
$C_{relation}$	| Name | Spouse | Child | Teacher | Boss | | Brad | Ava | Jack | Paul | Rob | | Gary | Kim | Joe | Fred | Leo | | Mike | Tara | Tom | Lee | Jake |
C_{job}	| Name | Current Job | Dream Job | Previous Job | Disliked Job | | Sean | guard | builder | student | chef | | Luke | coach | actor | judge | artist | | Sam | teacher | writer | driver | manager |
C_{object}	| Name | Created Object | Bought Object | Sold Object| Favorite Object | | Nick | lamp | brush | mat | table | | Jay | stamp | jar | shirt | belt | | Gary | ball | book | phone | fork |

Table 5: Other Samples for Table Template Input.

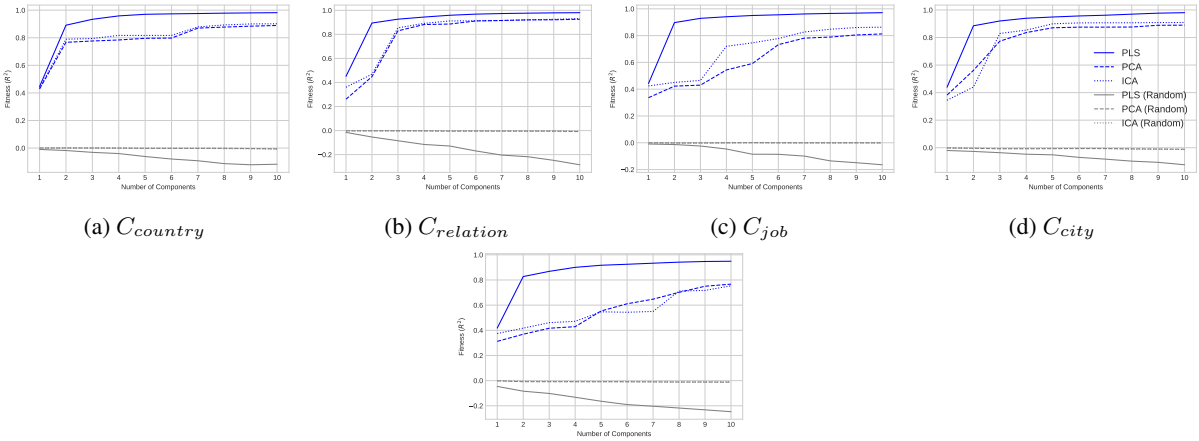
	Discourse Template Input
$C_{country}$	The ring is manufactured in Russia and designed in France, and it is exported to Iraq, but it is banned in Singapore . The plant is manufactured in China and designed in India, and it is exported to Australia, but it is banned in Jordan . The jar is manufactured in Sweden and designed in Iran, and it is exported to Pakistan, but it is banned in Georgia .
C_{city}	Brad was born in Paris and currently lives in Houston, he loves London and dislike Detroit . Gary was born in Austin and currently lives in Berlin, he loves Chicago and dislike Portland . James was born in Hamilton and currently lives in Split, he loves Atlanta and dislike Dallas .
$C_{relation}$	Brad is married to Ava and has a child named Jack, he was taught by Paul and works under Rob . Gary is married to Kim and has a child named Joe, he was taught by Fred and works under Leo . Mike is married to Tara and has a child named Tom, he was taught by Lee and works under Jake .
C_{job}	Sean currently works as a guard and dreams of becoming a builder, he previously worked as a student, and he dislikes being a chef . Luke currently works as a coach and dreams of becoming a actor, he previously worked as a judge, and he dislikes being a artist . Sam currently works as a teacher and dreams of becoming a writer, he previously worked as a driver, and he dislike s being a manager .
C_{object}	Nick created the lamp and also bought the brush, he sold the mat, and his favorite object is the table . Jay created the stamp and also bought the jar, he sold the shirt, and his favorite object is the belt . Gary created the ball and also bought the book, he sold the phone, and his favorite object is the fork .

Table 6: Other Samples for Discourse Template Input.

	Story Input
$C_{country}$	In a bustling marketplace, a unique ring caught the eye of many. This exquisite piece was manufactured in Russia and designed in France. It found its way to Iraq, although it faced a ban in Singapore. Nearby, a vibrant plant stood tall, manufactured in China and designed in India. It was exported to Australia, yet it was banned in Jordan. Lastly, a beautifully crafted jar sparkled under the sun, manufactured in Sweden and designed in Iran. It was destined for Pakistan but faced a ban in Georgia. Each product had its own journey, marked by both success and restriction.
C_{city}	Brad, born in Paris, now lives in Houston. He has a fondness for London but harbors a dislike for Detroit. Meanwhile, Gary hails from Austin and currently resides in Berlin. He adores Chicago, although he has an aversion to Portland. Lastly, James was born in Hamilton and makes his home in Split. He is passionate about Atlanta, yet he dislikes Dallas. Each of these men carries a unique relationship with the cities that have shaped their lives.
$C_{relation}$	Brad is happily married to Ava, and together they have a son named Jack. Throughout his education, he learned from Paul and currently works under the supervision of Rob. Similarly, Gary shares a life with Kim, and they are proud parents of a boy named Joe. His guiding mentor was Fred, and he reports to Leo at work. Lastly, Mike enjoys a loving marriage with Tara, and their child is Tom. He was instructed by Lee and is employed under Jake's leadership. Each family thrives, supported by their mentors and bosses.
C_{job}	Sean is currently a guard, aspiring to be a builder someday. He once held the position of a student but found no joy in being a chef. Luke, on the other hand, works as a coach and dreams of becoming an actor. His previous role was as a judge, yet he has no fondness for being an artist. Lastly, Sam is a teacher who hopes to transition into a writer. Before this, he worked as a driver, and he particularly disliked being a manager. Each of them navigates their careers, longing for something more fulfilling.
C_{object}	Nick was an inventor who created a lamp. He found a great deal and bought a brush. In his entrepreneurial spirit, he sold a mat. His favorite object, however, was the table. Jay was an artist who created a stamp and decided to buy a jar for his projects. He sold a shirt that he no longer needed, but his favorite object remained the belt. Lastly, Gary was a playful spirit who created a ball. He bought a book to inspire his creativity, sold a phone he no longer used, and cherished the fork as his favorite object.

Table 7: Other Samples for Story Input.

A.4 Identification of CBR Subspace on Llama3-8B-Instruct

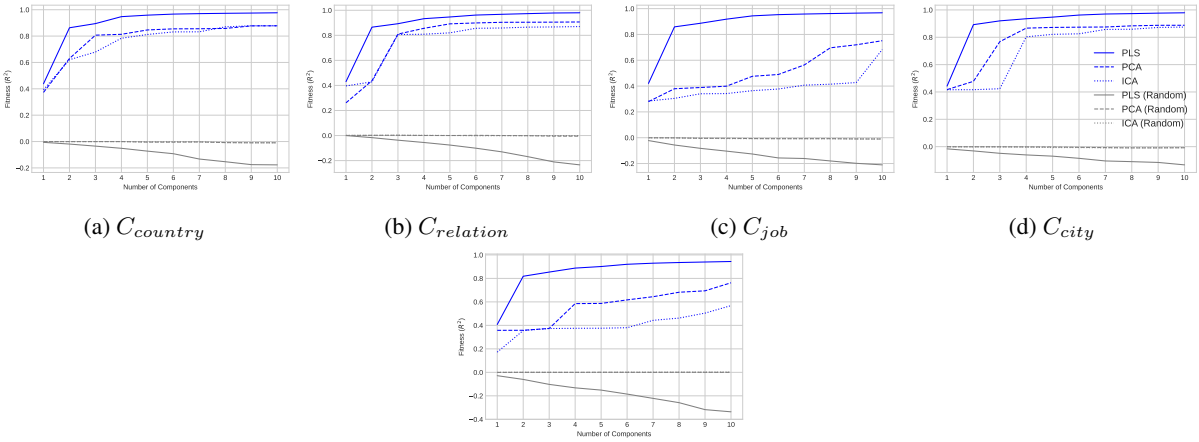


(a) $C_{country}$ (b) $C_{relation}$ (c) C_{job} (d) C_{city}

(e) C_{object}

Figure 10: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct using linear subspace methods. Each subplot shows how accurately entity and relation indices can be predicted as the dimensionality of the projected subspace increases for a given discourse context. The Y-axis indicates the fitness score (R^2), and the X-axis shows the number of components used for PLS, ICA and PCA projections. "(Random)" refers to model trained on random indices, serving as baseline controls.

A.5 Identification of CBR Subspace on Qwen3-8B



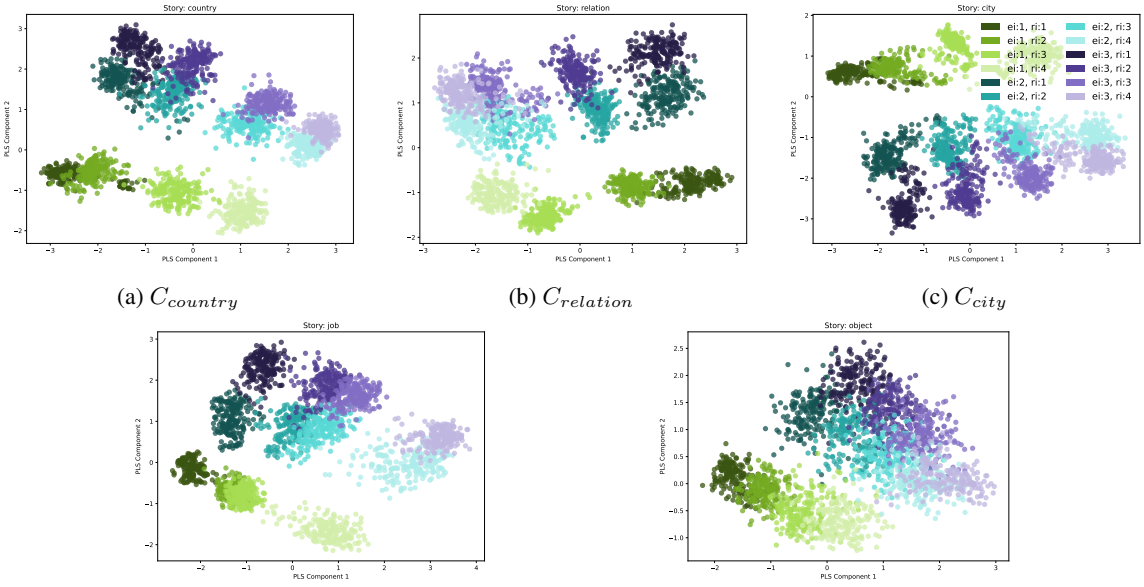
(a) $C_{country}$ (b) $C_{relation}$ (c) C_{job} (d) C_{city}

(e) C_{object}

Figure 11: Decoding performance of $[ei, ri]$ from activations of Qwen3-8B using linear subspace methods. Qwen3-8B shows consistency with Llama3-8B-Instruct, indicating that the CBR subspace emerges across model families and may reflect a general property of activation in LLMs.

Figure 11 shows the decoding performance to predict $[ei, ri]$ from the activation of Qwen3-8B. The results are consistent with Llama3-8B-Instruct, indicating that the CBR subspace is a general phenomenon shared across LLMs from different families. In addition, detailed layer-wise prediction results are provided in the Appendix. A.8.

A.6 Visualization of CBR Subspace on Llama3-8B-Instruct

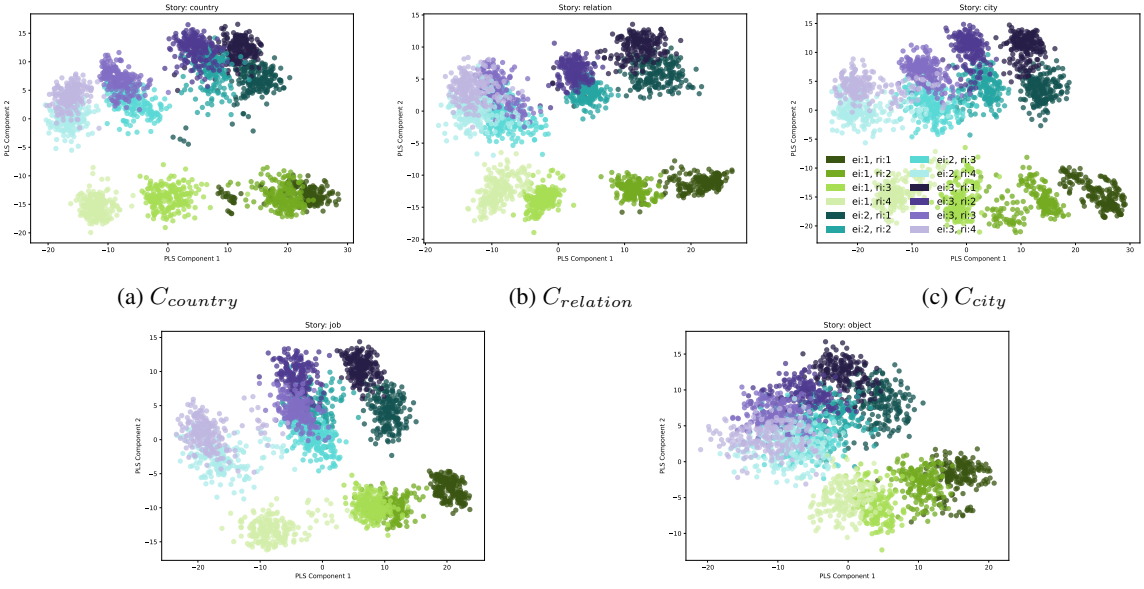


(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

Figure 12: Visualization of the CBR subspace from Llama3-8B-Instruct. Each point represents the projected activation of an attribute token. Colors indicate groups of attributes sharing the same $[ei, ri]$. The structure reveals clear distribution along both the ei and ri increasing (or decreasing) directions.

A.7 Visualization of CBR Subspace on Qwen3-8B



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

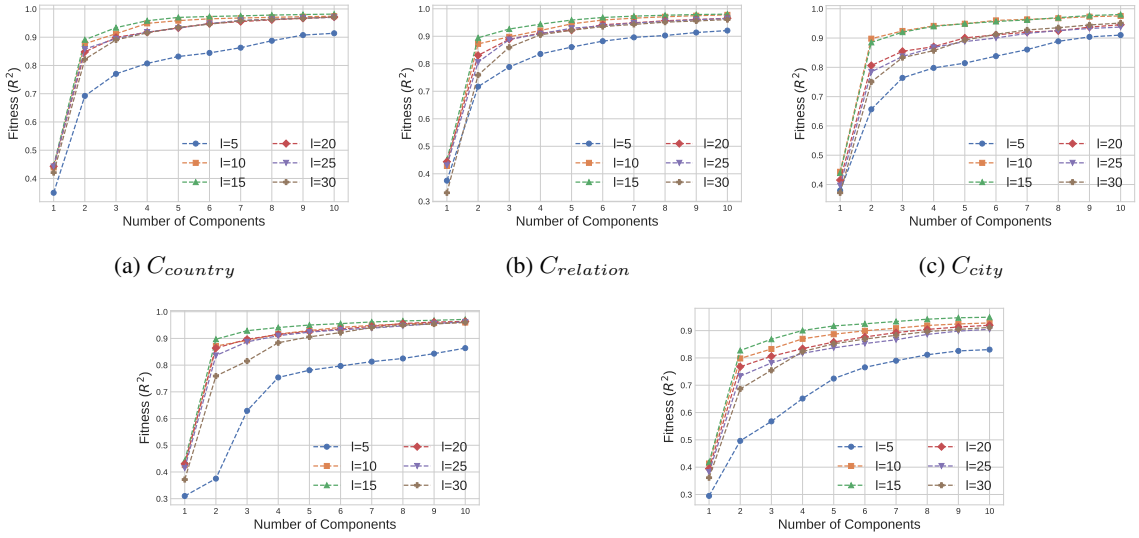
(d) C_{job} (e) C_{object}

Figure 13: Visualization of the CBR subspace from Qwen3-8B.

The CBR subspace visualization for Qwen3-8B, shown in Figure 13, is similarly organized along the entity and relation index, indicating that the CBR subspace is a general geometric property common to both LLM families.

A.8 Layer-wise Identification of CBR Subspace

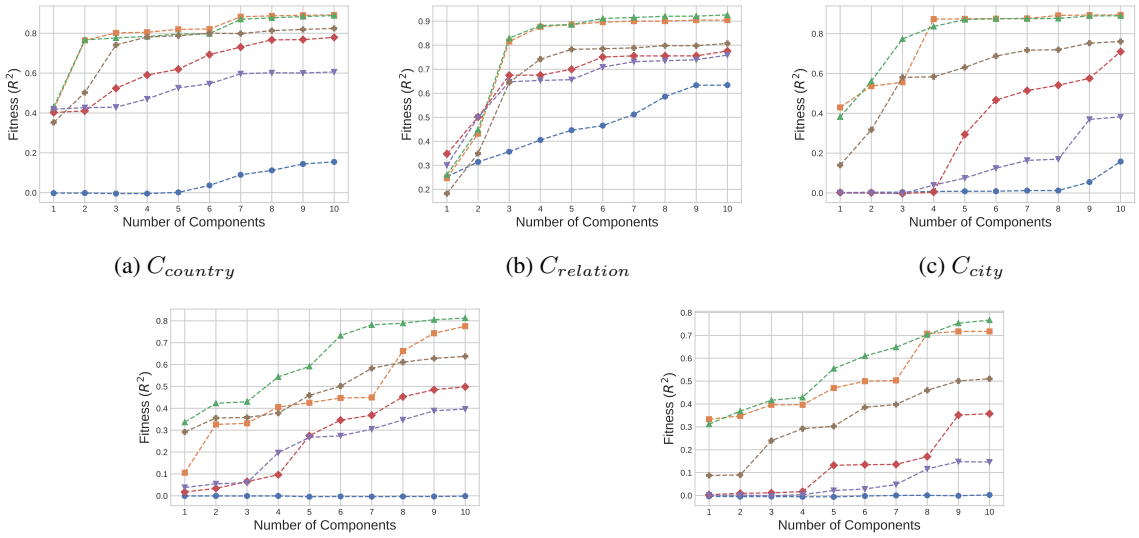
We further apply linear subspace based methods (i.e., PLS and PCA regression) to predict entity and relation indices in a layer-wise setting. The results, shown in Figure 14, 15, 16 and 17, indicate that the R^2 score peaks in the middle layers (e.g., $l = 15$) and is lower in both earlier layers (e.g., $l = 5$) and later layers (e.g., $l = 25$). This pattern is consistent with the "stages of inference hypothesis" (Lad et al., 2024), which states that intermediate layers are primarily responsible for feature engineering (e.g., the feature of the discourse relational structure in our case). In addition, the same trend is observed for both Llama3-8B-Instruct and Qwen3-8B, suggesting that this layer-wise organization of the CBR subspace is consistent across model families.



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

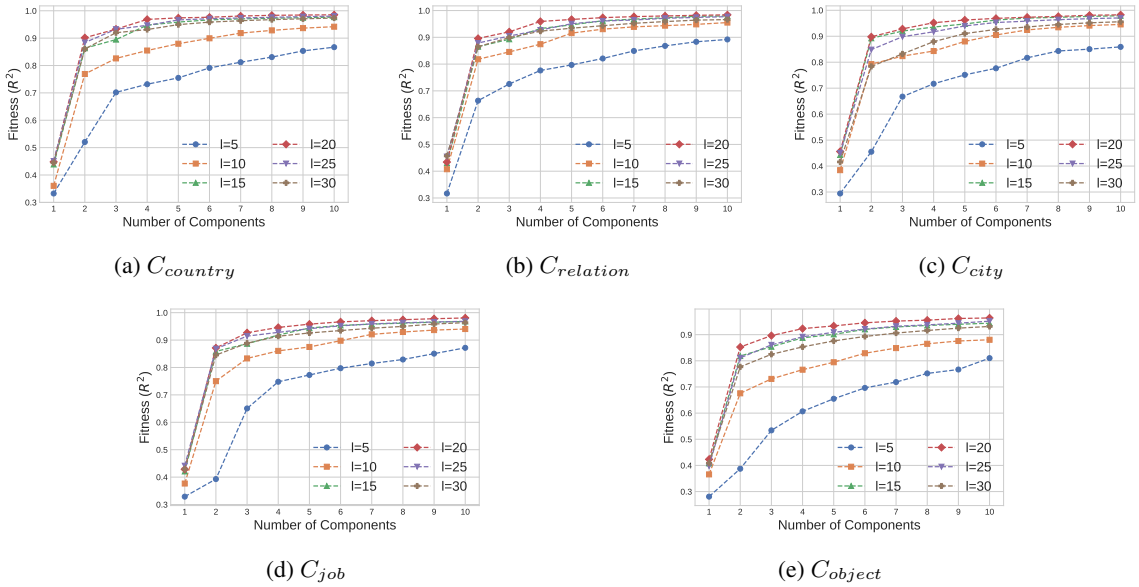
Figure 14: Decoding layer-wise performance of $[ei, ri]$ from activations of Llama3-8B-Instruct using PLS. Performance peaks in the middle layers, while both lower and higher layers show reduced fitness.



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

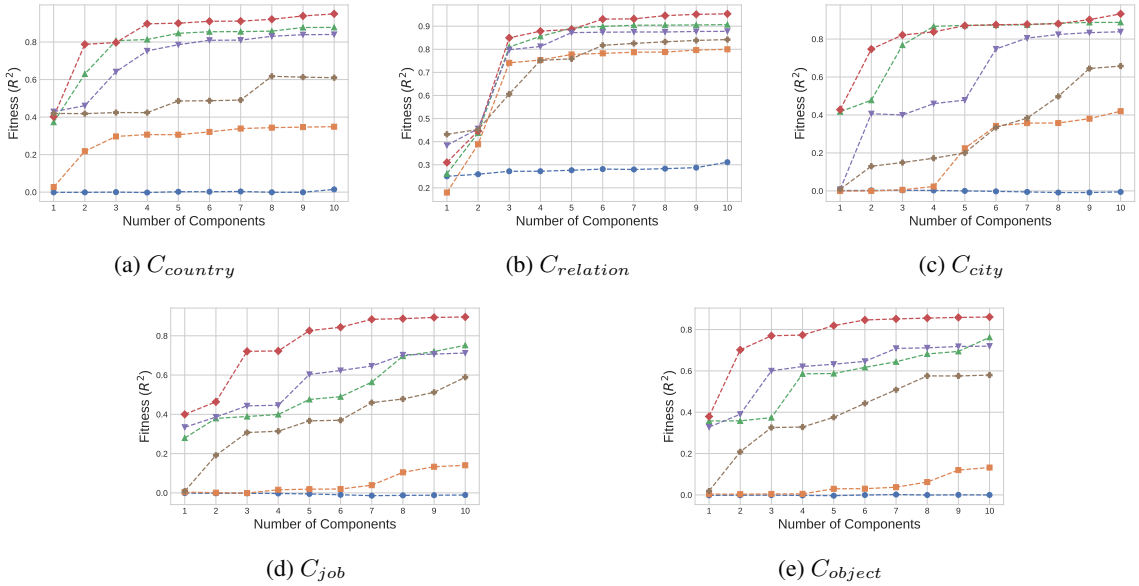
Figure 15: Decoding layer-wise performance of $[ei, ri]$ from activations of Llama3-8B-Instruct using PCA regression.



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

Figure 16: Decoding layer-wise performance of $[ei, ri]$ from activations of Qwen3-8B using PLS.



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

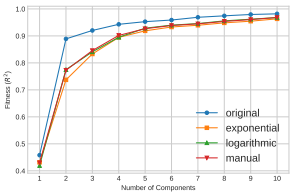
Figure 17: Decoding layer-wise performance of $[ei, ri]$ from activations of Qwen3-8B using PCA regression.

A.9 Monotonic Encoding of CBR Indices

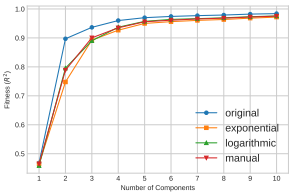
To analyze the monotonic property of the CBR index, we consider not only the standard integer indices (i.e., $[1, 2, 3, 4]$), but also several alternative monotonic sequences. Specifically, we construct exponential sequences (i.e., $[1, 3, 9, 27]$), logarithmic sequences (i.e., $[1, 4.64, 21.54, 100]$), and manually selected monotonic values (e.g., $[1, 5, 30, 100]$) to represent the indices of ei and ri. For example, under the manually selected sequence, the CBR index of the second entity or relation is assigned the value 5 instead of 2.

We then apply PLS to predict these indices from the hidden activations. The decoding performance is shown in Figure 18 and 19. The results show that PLS can successfully predict not only the original CBR indices $[1, 2, 3, \ldots]$, but also arbitrary monotonic transformations of these indices, including exponential sequences and irregularly spaced monotonic values. Performance is slightly higher when using the original integer indices.

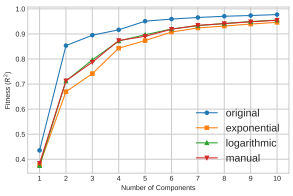
This observation suggests that the activations encode the *ordinal structure* of CBR cells rather than their exact numerical indices. In particular, the representations appear to organize CBR cells along a latent direction in activation space that preserves their ordering, allowing linear probes to recover any monotonic transformation of the indices. These findings indicate that CBR indices are encoded through an *order-preserving embedding* in the activation space of LLMs.



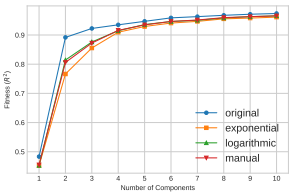
(a) C_{city}



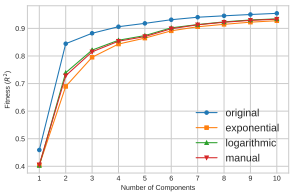
(b) $C_{country}$



(c) $C_{relation}$

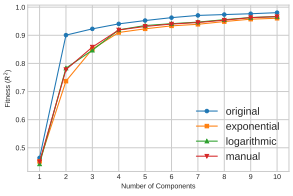


(d) C_{job}

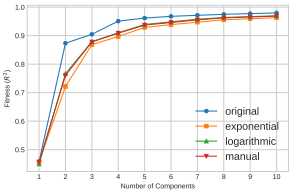


(e) C_{object}

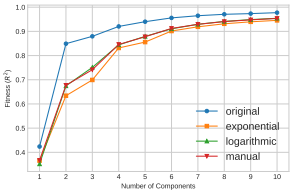
Figure 18: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct under different monotonic index sequences. "original" denotes the original integer index.



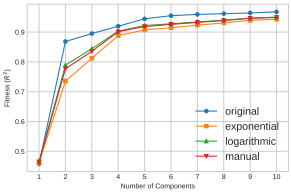
(a) C_{city}



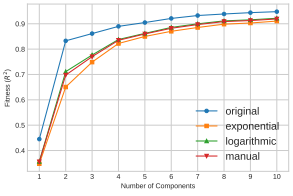
(b) $C_{country}$



(c) $C_{relation}$



(d) C_{job}

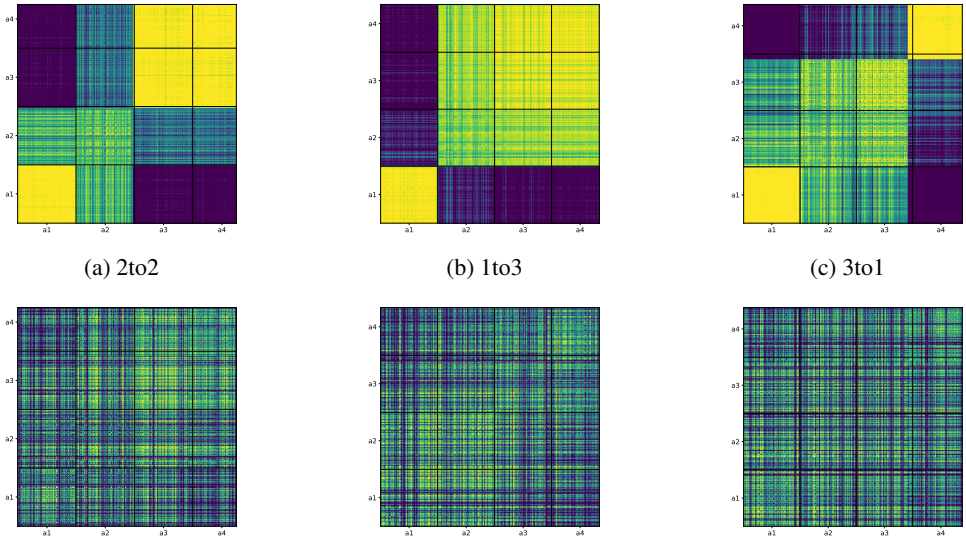


(e) C_{object}

Figure 19: Decoding performance of different monotonic $[ei, ri]$ from activations of Qwen3-8B.

A.10 CBR Subspace and Semantic Information on Qwen3-8B

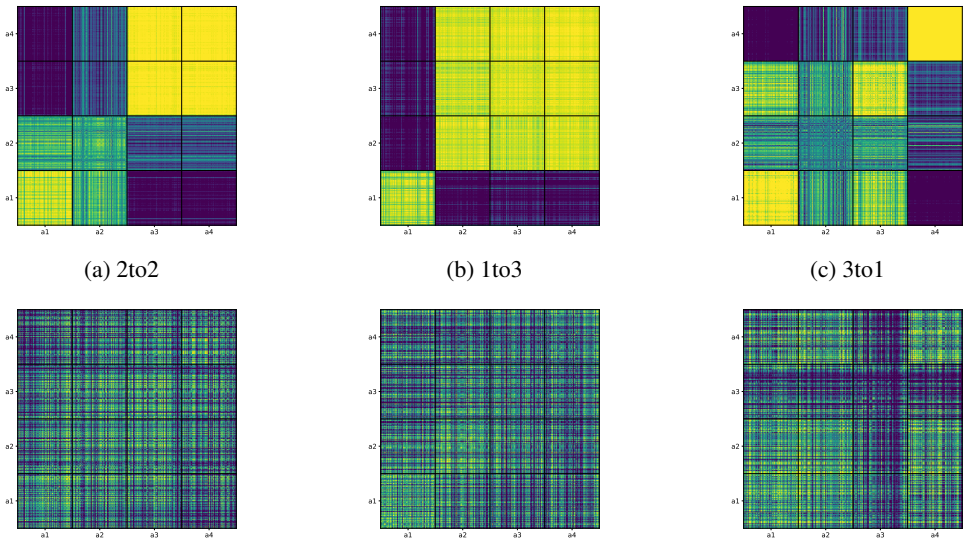
The heatmap in Figure 20, 21 and 22 illustrates CBR subspace embedding similarity among attributes, showing that the CBR subspace in Qwen3-8B also captures semantic similarity patterns.



(a) 2to2 (b) 1to3 (c) 3to1

(d) 2to2 (Rand.) (e) 1to3 (Rand.) (f) 3to1 (Rand.)

Figure 20: Cosine similarity heatmaps of attribute representations projected into the CBR subspace (above) and a random subspace (below) from Qwen3-8B on C_{job}. The CBR projection recovers the intended semantic similarity structure among relations, whereas the random projection does not clearly capture it.



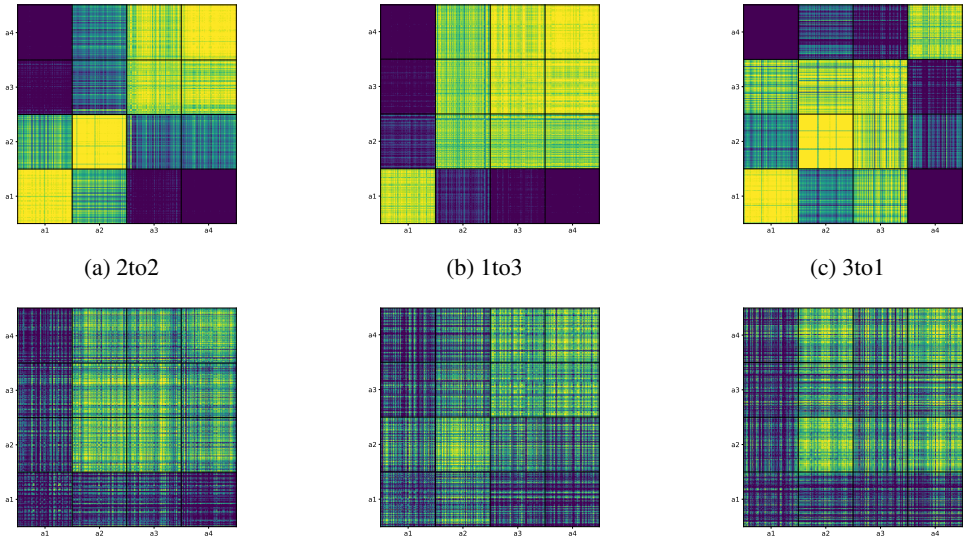
(a) 2to2 (b) 1to3 (c) 3to1

(d) 2to2 (Rand.) (e) 1to3 (Rand.) (f) 3to1 (Rand.)

Figure 21: Cosine similarity heatmaps of attribute representations from Qwen3-8B on C_{city}.

A.11 CBR Subspace and Semantic Information across Contexts

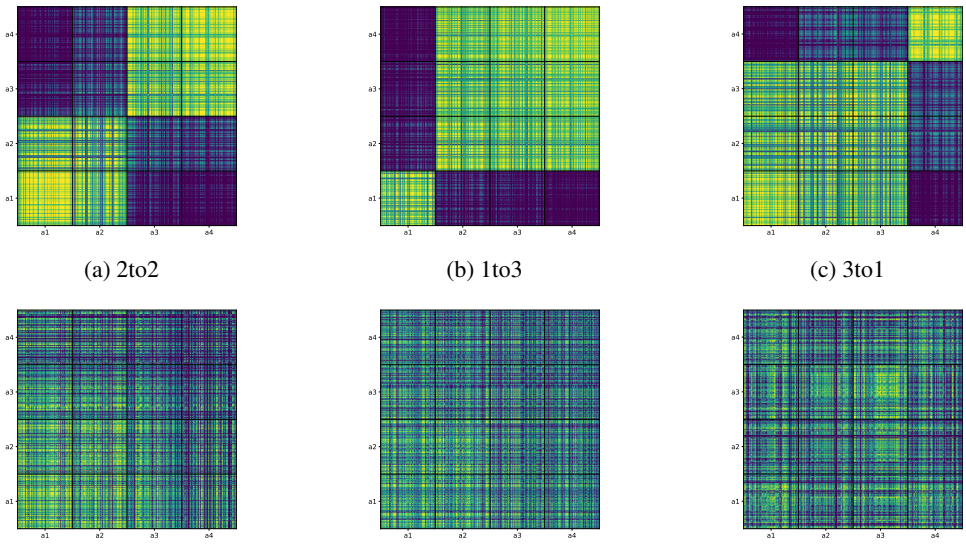
The heatmap in Figure 23 and 24 illustrates CBR subspace embedding similarity among attributes in C_{city} and $C_{country}$ respectively. The consistency of these pattern across different contexts indicates that the CBR subspace in Llama3-8B-Instruct can capture semantic similarity pattern.



(a) 2to2 (b) 1to3 (c) 3to1

(d) 2to2 (Rand.) (e) 1to3 (Rand.) (f) 3to1 (Rand.)

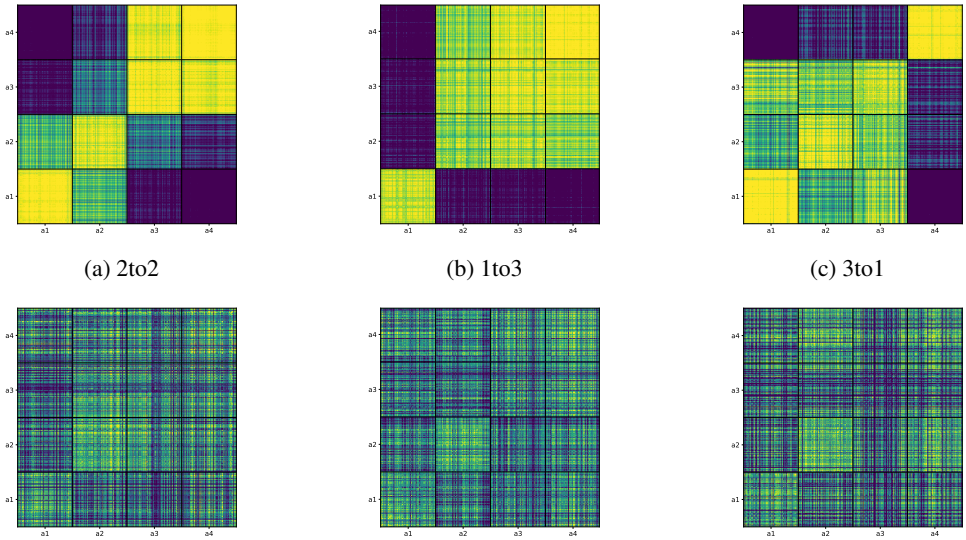
Figure 22: Cosine similarity heatmaps of attribute representations from Qwen3-8B on $C_{country}$.



(a) 2to2 (b) 1to3 (c) 3to1

(d) 2to2 (Rand.) (e) 1to3 (Rand.) (f) 3to1 (Rand.)

Figure 23: Cosine similarity heatmaps of attribute representations from Llama3-8B-Instruct on C_{city}.



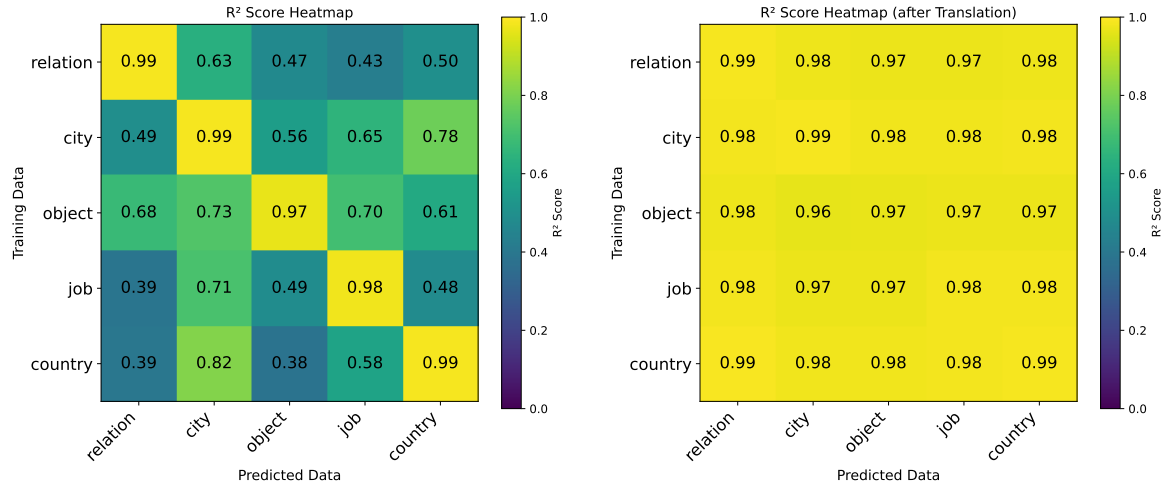
(a) 2to2 (b) 1to3 (c) 3to1

(d) 2to2 (Rand.) (e) 1to3 (Rand.) (f) 3to1 (Rand.)

Figure 24: Cosine similarity heatmaps of attribute representations from Llama3-8B-Instruct on $C_{country}$.

A.12 Generality of CBR Subspace on Qwen3-8B

As shown in Figure 25, Qwen3-8B exhibits a similar tendency to Llama3-8B-Instruct, where the projection matrix is influenced by contextual variations. However, by applying the translation vectors, the projection matrix trained on one context can correctly predict the CBR indices in another context.



(a) w/o Δ (b) w/ Δ

Figure 25: Each cell shows the R^2 fitness score obtained from Qwen3-8B. The projection matrix learned from one context (column) is used to predict the index information of another context (row). Higher values indicate better cross-context generality.

A.13 Ablation Study on Translation Vectors

To further verify the effectiveness of the translation vector, we evaluate several variants that modify the translation vector, including randomizing its norm or direction. We also compare it with baseline methods, such as using a random vector or a transformation matrix $M_{c_1 \to c_2}$ learned by minimizing the mean squared error between h_{c_1} and $M_{c_1 \to c_2} h_{c_2}$. The results are shown in Figure 26 and Figure 27. We observe that these alternative methods perform worse than the translation vector, further demonstrating its effectiveness.

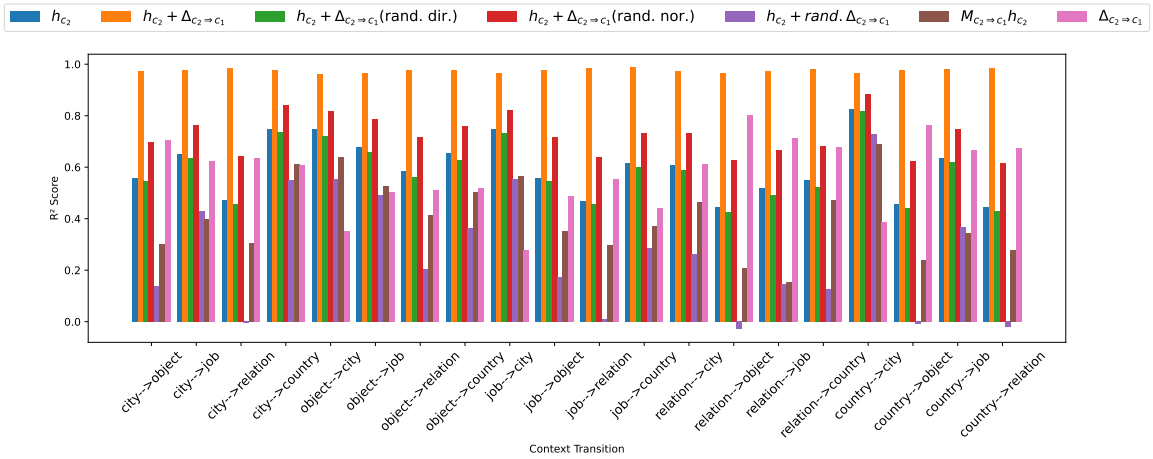


Figure 26: Performance comparison of the translation vector $\Delta_{c_1 \to c_2}$ and baseline methods for cross context CBR decoding from Llama3-8B-Instruct. The translation vector outperforms alternatives including random vectors (denoted as $rand\Delta_{c_1 \to c_2}$), randomized direction ($\Delta_{c_1 \to c_2}(rand.dir.)$) and norm ($\Delta_{c_1 \to c_2}(rand.nor.)$), a learned transformation matrix ($M_{c_1 \to c_2}$) and the $\Delta_{c_1 \to c_2}$ alone).

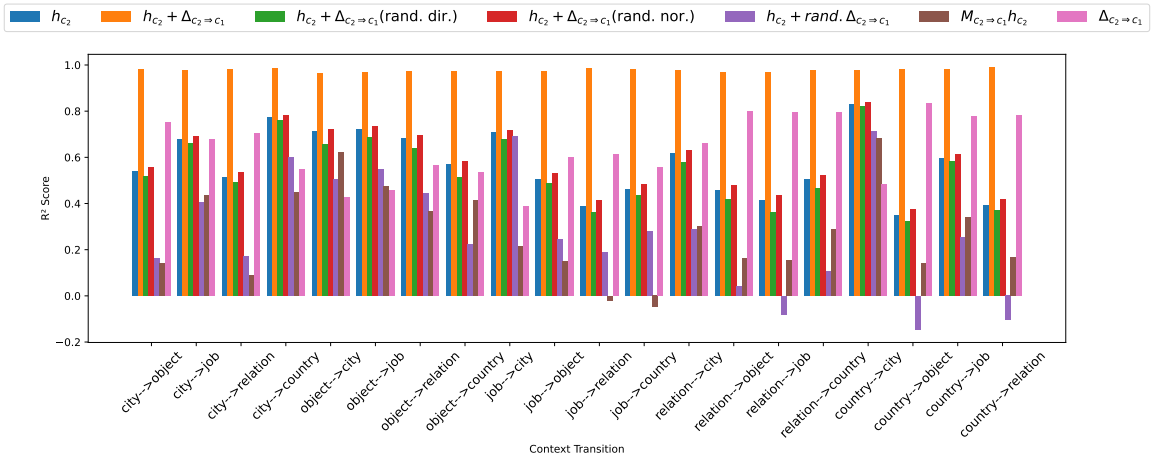


Figure 27: Performance comparison from Qwen3-8B.

One possible explanation of the effectiveness is that the translation vector partially contains CBR index information. However, this explanation is only partially supported by the results. As shown in Figures 26 and 27, in some context transformations, such as "object–>city" and "job–>city", the performance of using only the translation vector $\Delta_{c_2 \to c_1}$ is worse than directly using h_{c_2}.

The fact that $\Delta_{c_2 \to c_1}$ sometimes outperforms h_{c_2} suggests that the translation vector may contain index-related information, which could be explained by superposition theory (Elhage et al., 2022). The theory suggests that different features may be encoded in overlapping neuron directions. Under this view, the translation vector may capture a mixture of contextual and index-related components in the shared activation space.

A.14 Translation Vector under Superposition

To explain the effectiveness of translation vectors, we interpret the CBR representation through the lens of superposition theory (Elhage et al., 2022). In this framework, a hidden activation h can be expressed as a superposition of feature directions:

$$h = \sum_i f_i v_i,$$

where f_i denotes the activation strength of feature i, and v_i represents the direction (or subspace) encoding that feature. Importantly, multiple features may share neurons through superposition.

Assume that a hidden state in context c_1 contains both a CBR index feature and context-specific features:

$$h_{c_1} = f_{\text{index}} v_{\text{index}} + f_{c_1} v_c \tag{9}$$

where v_{index} denotes the direction encoding the CBR index and v_c represents the direction corresponding to contextual features. These directions are not necessarily orthogonal. Let M_{c_1} denote the projection matrix trained under context c_1 to decode the CBR indices. Applying this projection yields:

$$M_{c_1} h_{c_1} = M_{c_1} f_{\text{index}} v_{\text{index}} + M_{c_1} f_{c_1} v_c \approx \text{indices}_{\text{CBR}}. \tag{10}$$

Since v_{index} and v_c are generally not orthogonal, the contextual component satisfies $M_{c_1} f_{c_1} v_c \neq 0$. However, if the activation comes from another context c_2,

$$h_{c_2} = f_{\text{index}} v_{\text{index}} + f_{c_2} v_c, \tag{11}$$

then applying the same projection gives:

$$M_{c_1} h_{c_2} = M_{c_1} f_{\text{index}} v_{\text{index}} + M_{c_1} f_{c_2} v_c. \tag{12}$$

Because $f_{c_1} v_c \neq f_{c_2} v_c$, the contextual component changes. Compared to Equation 10, this leads to

$$M_{c_1} h_{c_2} \not\approx \text{indices}_{\text{CBR}}.$$

Nevertheless, if contexts c_1 and c_2 are similar, M_{c_1} tends to remain effective for context c_2. This is supported by the observations in Appendix (§A.29). Alternatively, their contextual features may be related by a translation vector Δ in the activation space:

$$f_{c_1} v_c \approx f_{c_2} v_c + \Delta_{c_2 \to c_1}.$$

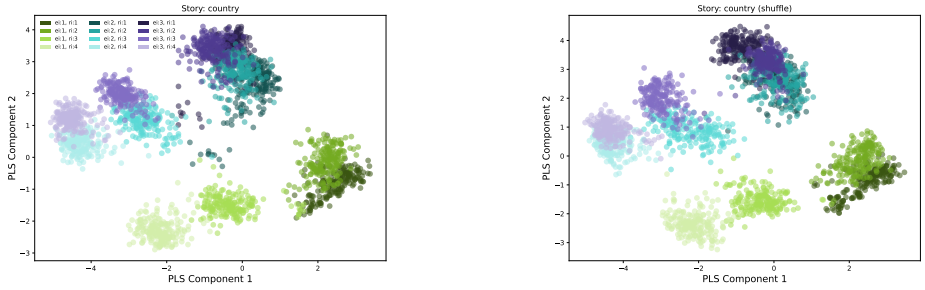
Consequently, applying this translation before decoding yields:

$$M_{c_1}(h_{c_2} + \Delta_{c_2 \to c_1}) = M_{c_1}(f_{\text{index}} v_{\text{index}} + f_{c_2} v_c + \Delta_{c_2 \to c_1}) \approx M_{c_1}(f_{\text{index}} v_{\text{index}} + f_{c_1} v_c) \approx \text{indices}_{\text{CBR}}.$$

This may explain why translation vectors can restore decoding performance across contexts: they compensate for context-dependent components in the superposed representation while preserving the shared CBR index direction.

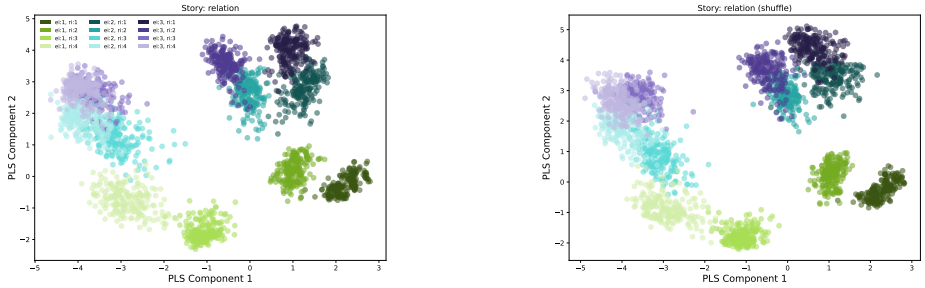
A.15 Consistency of CBR Subspace across Contexts

Section (§3.3) mentions that the overall structure of the CBR subspace remains stable under the permutation of shuffling and ablation, because the projected representations preserving their original geometric arrangement. We apply shuffling across all other discourse contexts in Figures 28, 29, 30, and 31. The visualizations after ablation are shown in Figures 32, 33, 34, and 35, where representations are visualized using the projection matrix learned from the permuted dataset. In all cases, the same pattern holds: the global geometry of the CBR subspace aligns with the ei and ri directions despite surface level variations. This cross-context stability confirms our earlier observations and further supports the conclusion that the CBR geometry primarily depends on the index structure rather than superficial variations.



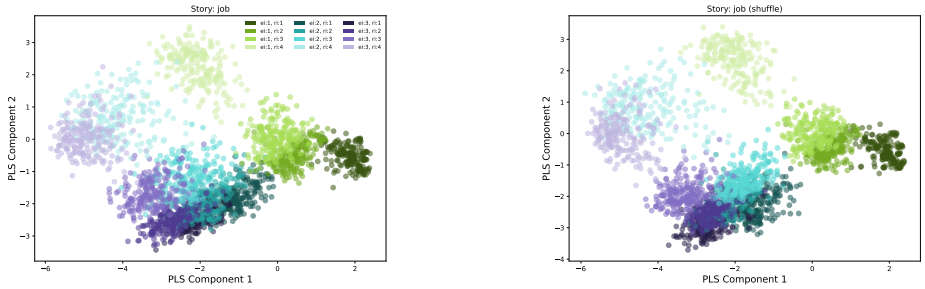
(a) base (b) shuffled

Figure 28: Visualization of the CBR subspace before and after **shuffling** the relation order of the second and third entities from Llama3-8B-Instruct on $C_{country}$.



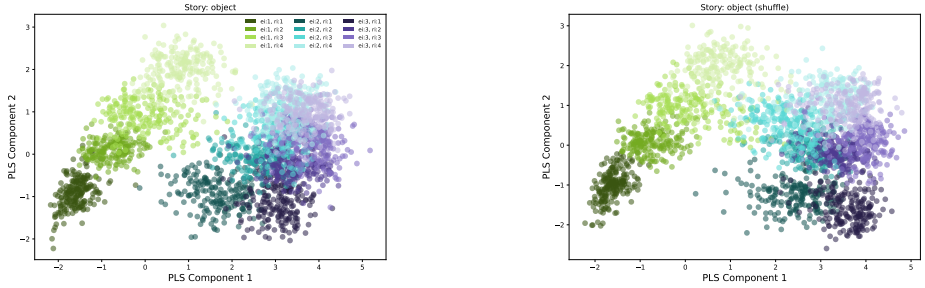
(a) base (b) shuffled

Figure 29: Visualization of the CBR subspace before and after **shuffling** the relation order of the second and third entities from Llama3-8B-Instruct on $C_{relation}$.



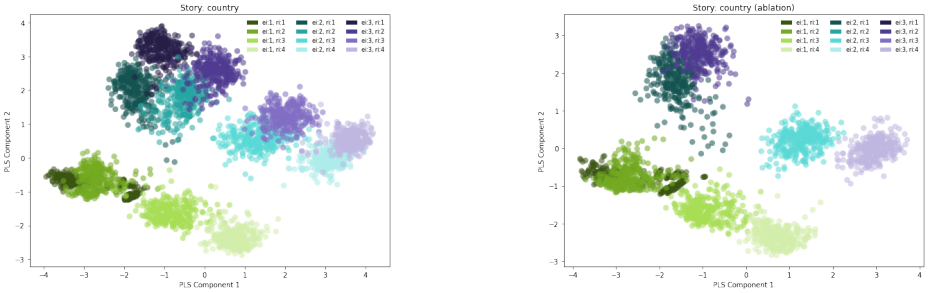
(a) base (b) shuffled

Figure 30: Visualization of the CBR subspace before and after **shuffling** from Llama3-8B-Instruct on C_{job}.



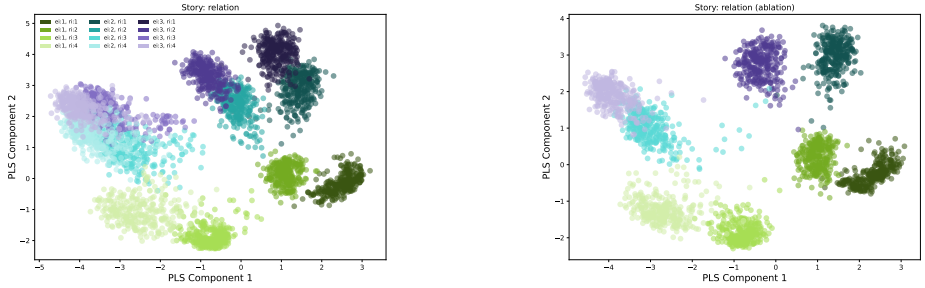
(a) base (b) shuffled

Figure 31: Visualization of the CBR subspace before and after **shuffling** from Llama3-8B-Instruct on C_{object}.



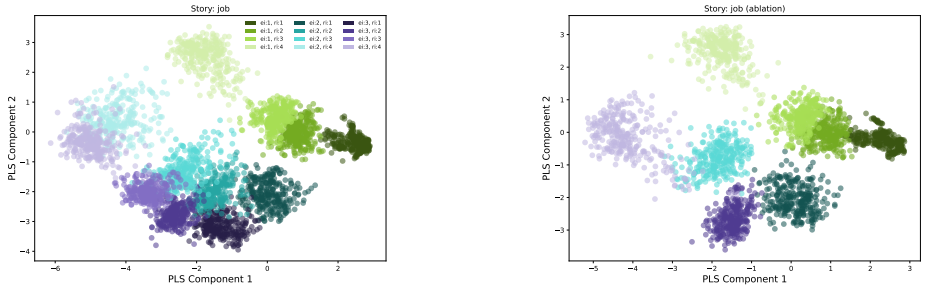
(a) base (b) ablated

Figure 32: Visualization of the CBR subspace before and after relation **ablation** from Llama3-8B-Instruct on $C_{country}$.



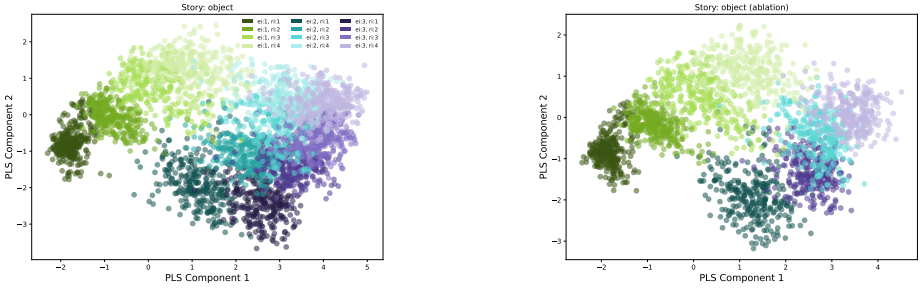
(a) base (b) ablated

Figure 33: Visualization of the CBR subspace before and after relation **ablation** from Llama3-8B-Instruct on $C_{relation}$.



(a) base (b) ablated

Figure 34: Visualization of the CBR subspace before and after relation **ablation** from Llama3-8B-Instruct on C_{job}.



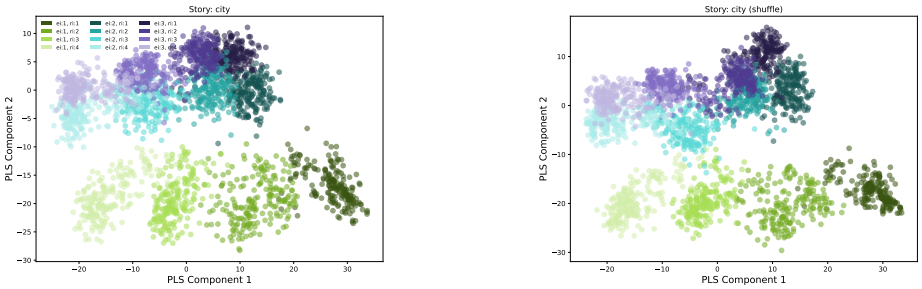
(a) base (b) ablated

Figure 35: Visualization of the CBR subspace before and after relation **ablation** from Llama3-8B-Instruct on C_{object}.

A.16 Consistency of CBR Subspace on Qwen3-8B

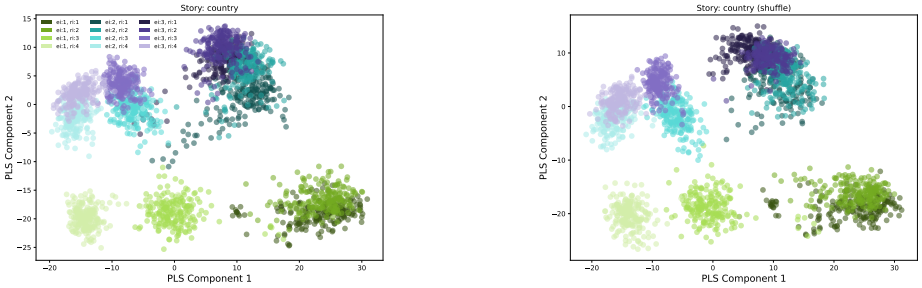
Figure 36, 37, 38, 39 and 40 visualize the CBR subspace before and after shuffling the relation order of the second and third entities in all the contexts. The stability of the distribution shows that the CBR subspace in Qwen3-8B primarily depends on index structure rather than superficial permutation.

Figure 41, 42, 43, 44 and 45 visualize the CBR subspace before and after relation ablation in all the contexts. The overall geometric structure aligns with the ei and ri directions despite surface level variations, indicating that removing relational content that does not affect the overall structure of CBR subspace in Qwen3-8B.



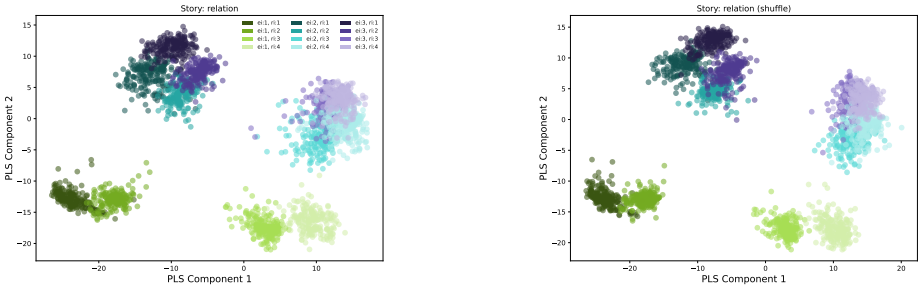
(a) base (b) shuffled

Figure 36: Visualization of the CBR subspace before and after **shuffling** from Qwen3-8B on C_{city}.



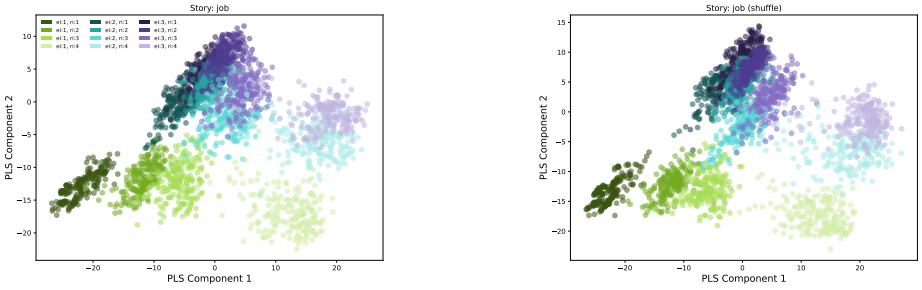
(a) base (b) shuffled

Figure 37: Visualization of the CBR subspace before and after **shuffling** from Qwen3-8B on $C_{country}$.



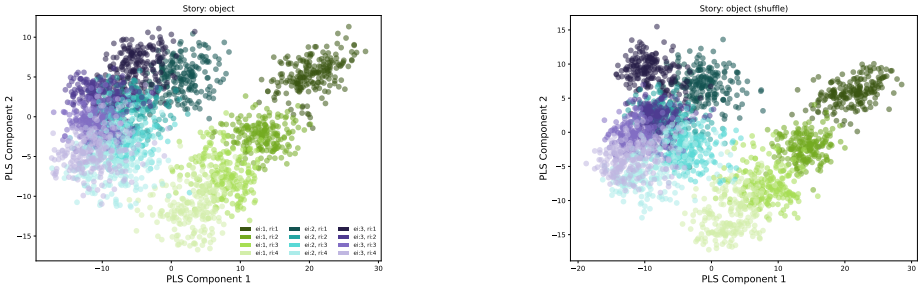
(a) base (b) shuffled

Figure 38: Visualization of the CBR subspace before and after **shuffling** from Qwen3-8B on $C_{relation}$.



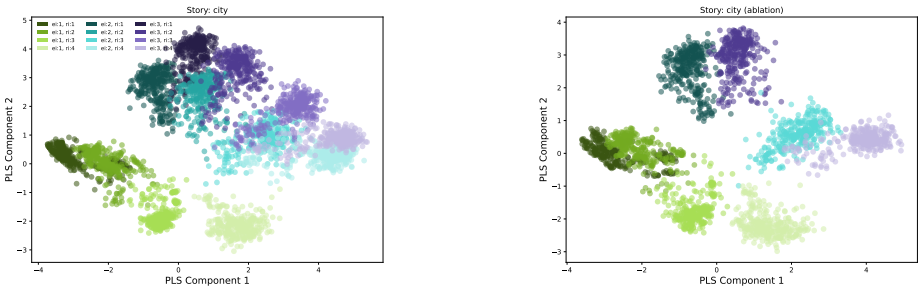
(a) base (b) shuffled

Figure 39: Visualization of the CBR subspace before and after **shuffling** from Qwen3-8B on C_{job}.



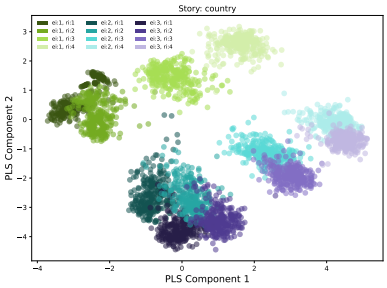
(a) base (b) shuffled

Figure 40: Visualization of the CBR subspace before and after **shuffling** from Qwen3-8B on C_{object}.

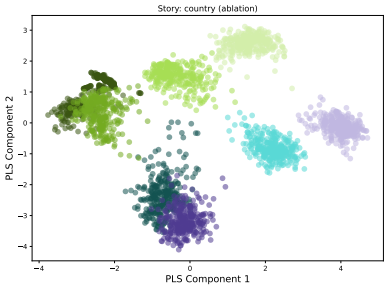


(a) base (b) ablated

Figure 41: Visualization of the CBR subspace before and after relation **ablation** from Qwen3-8B on C_{city}.

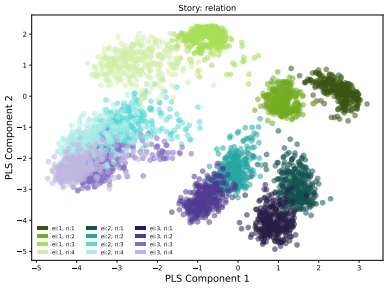


(a) base

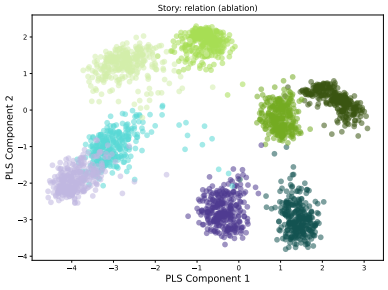


(b) ablated

Figure 42: Visualization of the CBR subspace before and after relation **ablation** from Qwen3-8B on $C_{country}$.

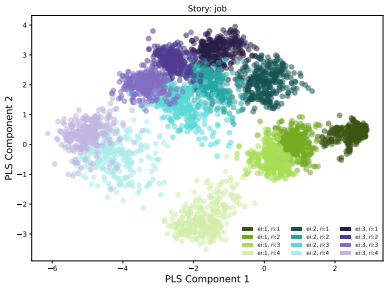


(a) base

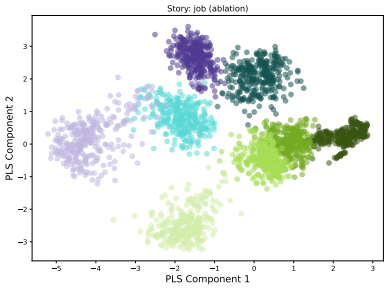


(b) ablated

Figure 43: Visualization of the CBR subspace before and after relation **ablation** from Qwen3-8B on $C_{relation}$.

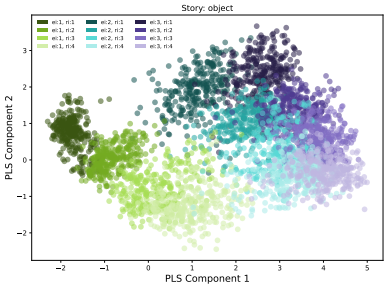


(a) base

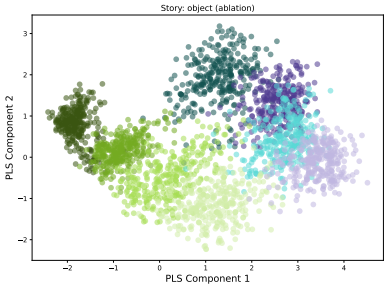


(b) ablated

Figure 44: Visualization of the CBR subspace before and after relation **ablation** from Qwen3-8B on C_{job}.



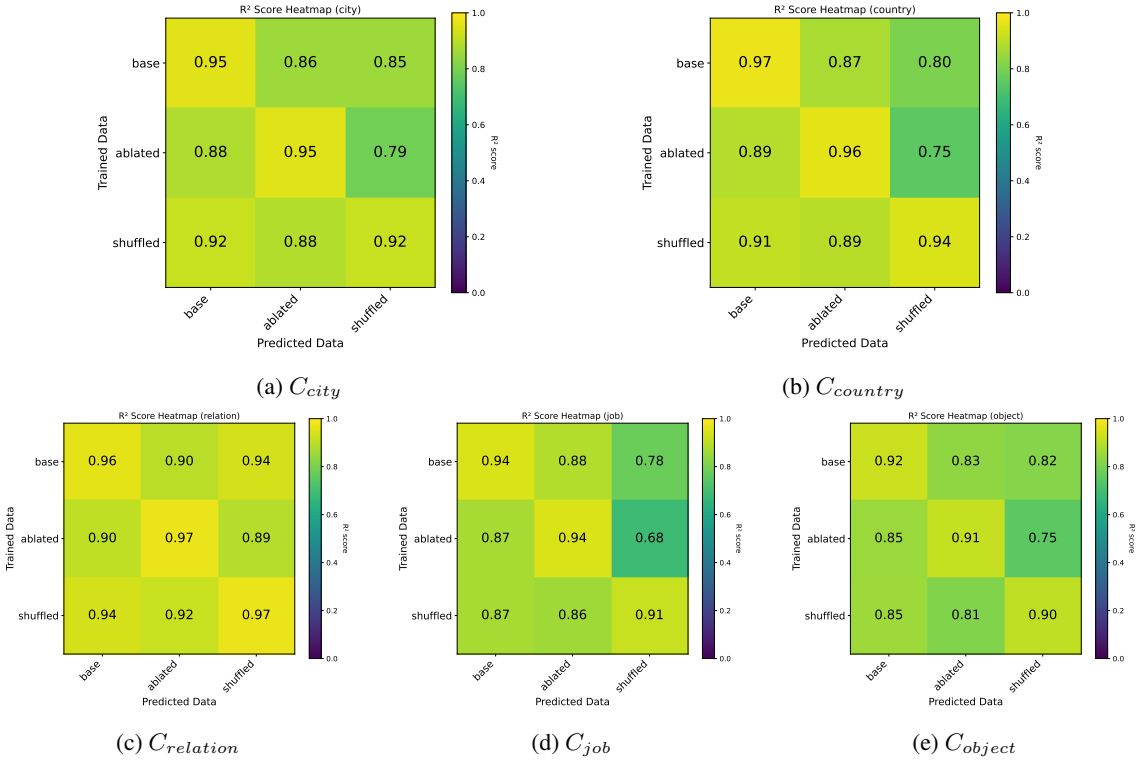
(a) base



(b) ablated

Figure 45: Visualization of the CBR subspace before and after relation **ablation** from Qwen3-8B on C_{object}.

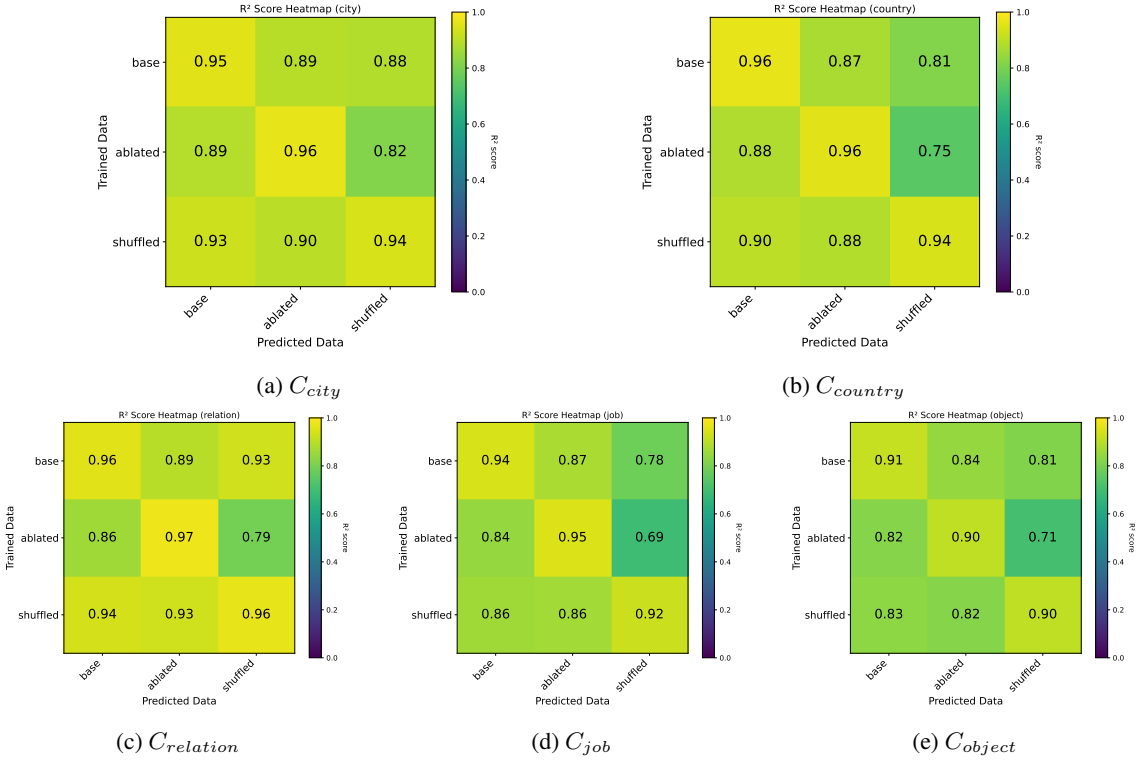
A.17 Consistency Analysis via Index Prediction



(a) C_{city} (b) $C_{country}$

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 46: Cross permutation R^2 scores for index prediction on Llama3-8B-Instruct.



(a) C_{city} (b) $C_{country}$

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 47: Cross permutation R^2 scores for index prediction on Qwen3-8B.

We evaluate whether projection matrices learned under permuted conditions generalize to original setting. Specifically, we use CBR projection matrices learned from the ablated and shuffled datasets to

predict index information in the original datasets, and vice versa, for all discourse contexts. As shown in Figure 46, we observe consistently high predictive performance across contexts, with R^2 scores typically around 0.8.

In addition, this strong performance is not limited to a single LLM family: Qwen3-8B also exhibit similarly high cross permutation prediction accuracy shown in Figure 47. These results demonstrate that the overall geometry of CBR subspace remains stable under both relation ablation and relation shuffling across Llama3-8B-Instruct and Qwen3-8B. This provides further evidence that the CBR subspace primarily reflects relational structure rather than a surface artifact of a particular dataset configuration.

A.18 Effect of Entity Separation on the CBR Subspace

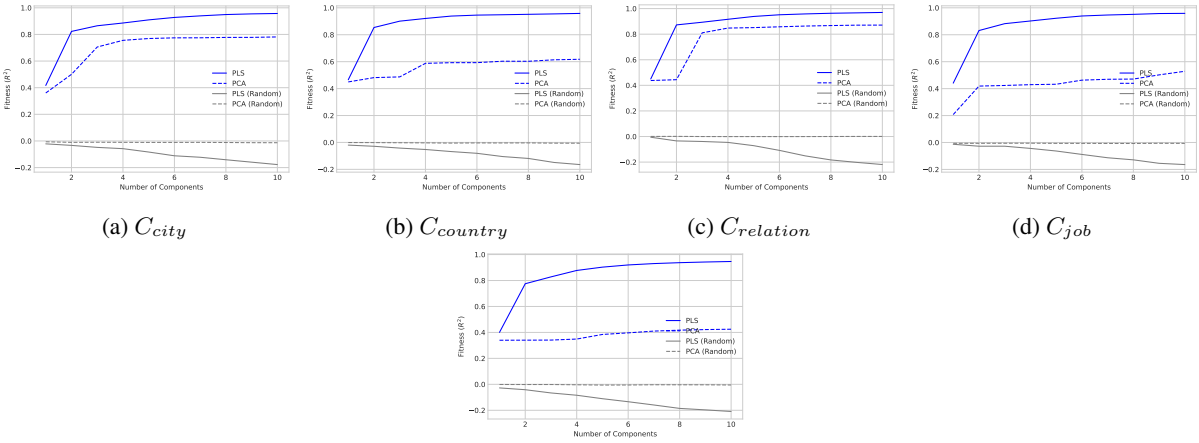
To further analyze the relationship between the CBR subspace and input format variations that reflect diverse discourse patterns observed in real-world text, we also construct additional modified datasets. In the setting (denoted as **separation**), we vary the distance between entity mentions without altering the CBR structure. For example, in Sample 4, the entity "window" is re-mentioned only after the discourse introduces one attribute (denoted as **# separation=1**) for other entities such as "glass" and "keyboard" in contrast to earlier Sample in Figure 1 (a), where "table" is referenced immediately. This modification captures variation in reference distance, a common phenomenon in natural discourse.

(4) **# separation=1**: *The window is crafted in Australia, while the glass is produced in Mexico. The keyboard finds its origins in China. Meanwhile,* ***the window is designed in Germany and exported to Brazil, yet it faces a ban in Spain. The glass is designed in Argentina, exported to Italy, but is also banned in Georgia. The keyboard is designed in France, sent to Russia, but prohibited in Sweden.*** *Together, these products tell a tale of global manufacturing, design, and the complexities of international trade regulations.*

(5) **# separation=2**: *The book, crafted in Israel and designed in Jordan, captivates readers with its unique story. The paper, produced in Argentina and designed in Canada, adds a special touch to the pages. Meanwhile, the ball, made in France and designed in Brazil, brings joy to children everywhere.* ***The book finds its way to Sweden, but it faces a ban in Spain. The paper travels to India, yet it is prohibited in Japan. The ball is exported to Germany, but it is banned in Italy,*** *creating a web of intrigue around these beloved items.*

(6) **# separation=3**: *The basket, crafted in Germany and designed in Iraq, finds its way to Brazil. Meanwhile, the flower, produced in Iran and conceptualized in France, is sent to Turkey. The monitor, made in Japan and styled in Italy, is dispatched to Egypt. However,* ***the journey of the basket comes to an abrupt halt as it is banned in Jordan. The flower faces a similar fate, being banned in Pakistan, while the monitor is restricted and banned in Canada,*** *leaving them all unable to reach certain markets.*

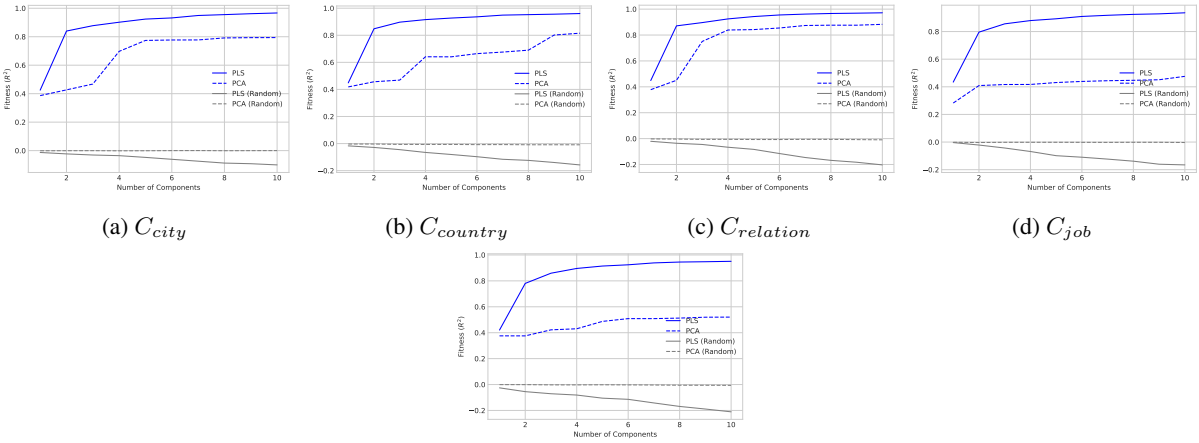
We apply the same PLS-based approach to predict entity and relation indices under each modified setting. The resulting fitness scores are shown in Figure 48, 49, 50, 51, 52 and 53. Across all settings, we observe consistently high R^2 scores, indicating that binding information can still be recovered from a low-dimensional subspace despite these input-level modifications. Furthermore, we evaluate the generality of the learned CBR projection matrices across these settings by using a projection matrix learned in one setting to predict entity and relation indices in another. As shown in Figure 54 and 55, predictive performance remains largely stable under the separation such as for index prediction on the original input. Together, these findings suggest that the overall structure of the CBR subspace is robust to reference distance, supporting the stability of CBR representations across diverse discourse formats.



(a) C_{city} (b) $C_{country}$ (c) $C_{relation}$ (d) C_{job}

(e) C_{object}

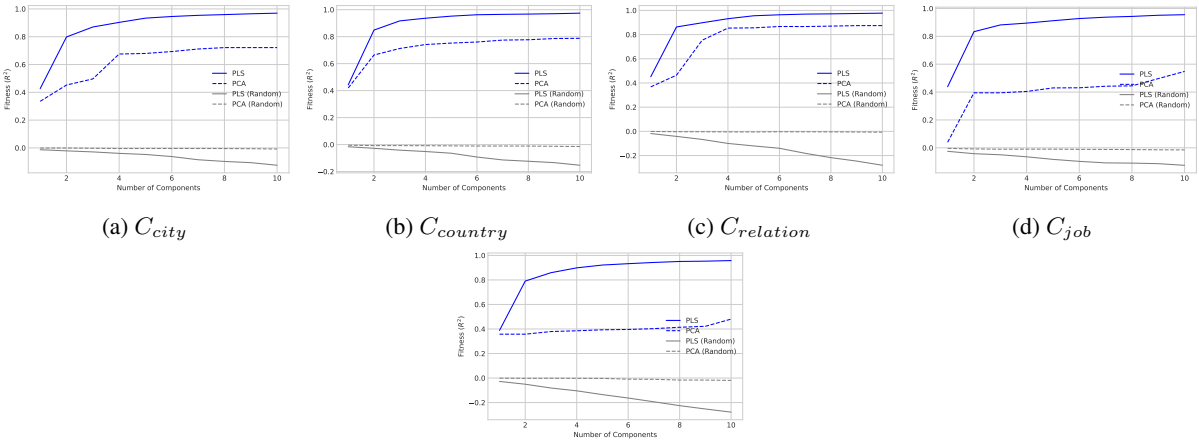
Figure 48: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct on the **# separation=1** setting.



(a) C_{city} (b) $C_{country}$ (c) $C_{relation}$ (d) C_{job}

(e) C_{object}

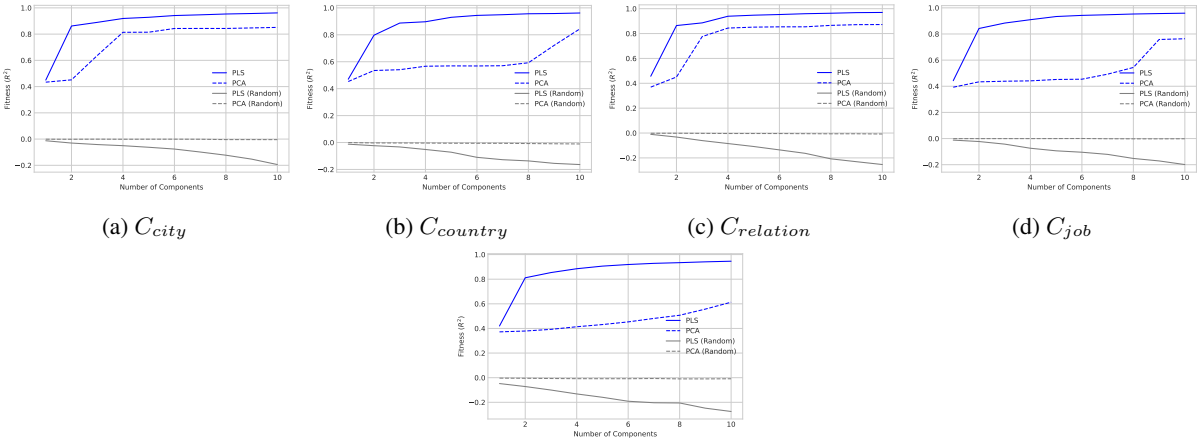
Figure 49: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct on the **# separation=2** setting.



(a) C_{city} (b) $C_{country}$ (c) $C_{relation}$ (d) C_{job}

(e) C_{object}

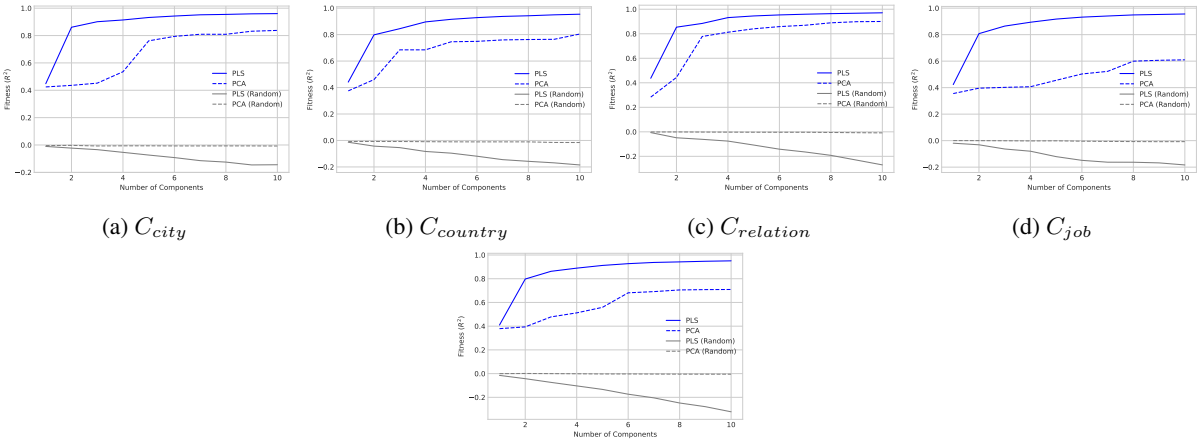
Figure 50: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct on the **# separation=3** setting.



(a) C_{city} (b) $C_{country}$ (c) $C_{relation}$ (d) C_{job}

(e) C_{object}

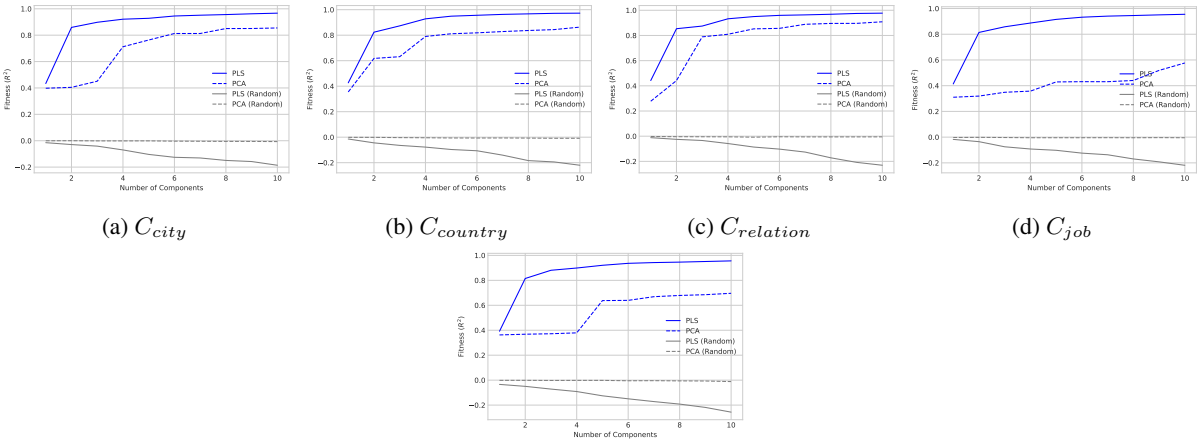
Figure 51: Decoding performance of $[ei, ri]$ from activations of Qwen3-8B on the **# separation=1** setting.



(a) C_{city} (b) $C_{country}$ (c) $C_{relation}$ (d) C_{job}

(e) C_{object}

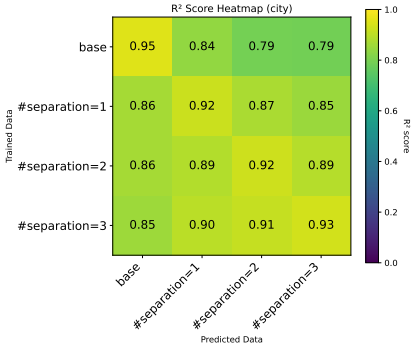
Figure 52: Decoding performance of $[ei, ri]$ from activations of Qwen3-8B on the **# separation=2** setting.



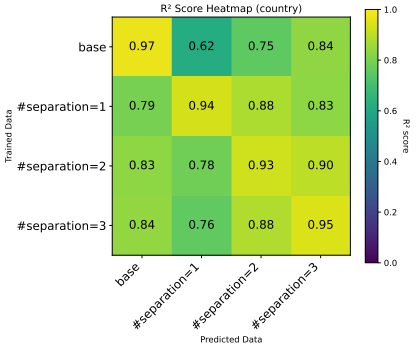
(a) C_{city} (b) $C_{country}$ (c) $C_{relation}$ (d) C_{job}

(e) C_{object}

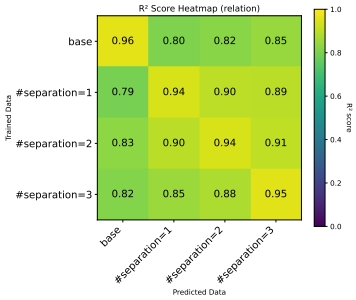
Figure 53: Decoding performance of $[ei, ri]$ from activations of Qwen3-8B on the **# separation=3** setting.



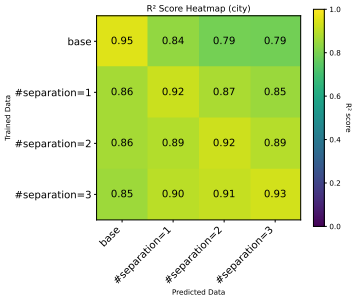
(a) C_{city}



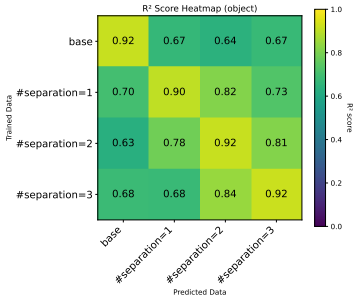
(b) $C_{country}$



(c) $C_{relation}$

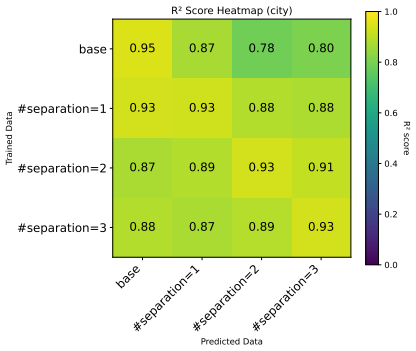


(d) C_{job}

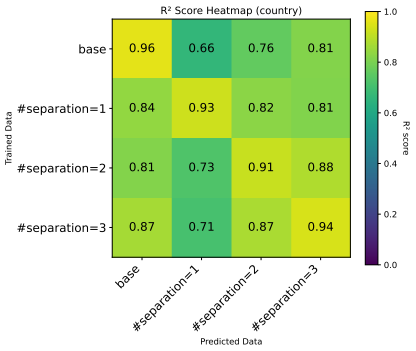


(e) C_{object}

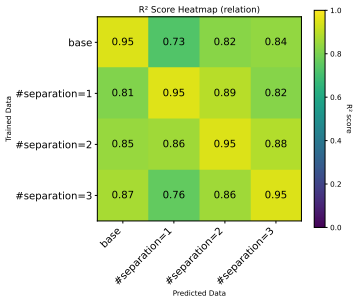
Figure 54: Cross permutation (separation) R^2 scores for index prediction on Llama3-8B-Instruct.



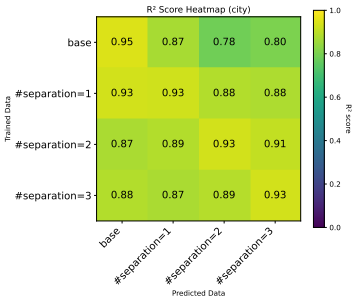
(a) C_{city}



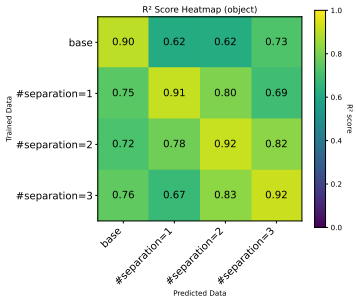
(b) $C_{country}$



(c) $C_{relation}$



(d) C_{job}



(e) C_{object}

Figure 55: Cross permutation (separation) R^2 scores for index prediction on Qwen3-8B.

A.19 Prevalence of CBR Across Diverse Discourse Patterns

To ensure that the observed grid-like CBR representation is not an artifact of a repetitive pattern, we construct 13 discourse patterns with progressively increasing structural complexity (Patt.1–Patt.13, shown below). The patterns vary in attribute count, relational overlap, and the distribution of relations across entities, and each template contains 1,000 naturalistic samples.

- Patt. 1: $(e_1r_1, e_2r_2, e_3r_3, e_3r_4)$
- Patt. 2: $(e_1r_1, e_2r_2, e_2r_3, e_3r_4)$
- Patt. 3: $(e_1r_1, e_1r_2, e_2r_3, e_3r_4)$
- Patt. 4: $(e_1r_1, e_1r_2, e_2r_2, e_2r_3, e_3r_3, e_3r_4)$
- Patt. 5: $(e_1r_1, e_1r_2, e_2r_1, e_2r_3, e_3r_2, e_3r_4)$
- Patt. 6: $(e_1r_1, e_1r_2, e_2r_3, e_2r_4, e_3r_1, e_3r_4)$
- Patt. 7: $(e_1r_1, e_1r_2, e_2r_2, e_2r_3, e_2r_4, e_3r_2, e_3r_3, e_3r_4)$
- Patt. 8: $(e_1r_1, e_1r_2, e_2r_1, e_2r_3, e_2r_4, e_3r_1, e_3r_2, e_3r_4)$
- Patt. 9: $(e_1r_1, e_1r_2, e_2r_1, e_2r_3, e_2r_4, e_3r_1, e_3r_3, e_3r_4)$
- Patt. 10: $(e_1r_1, e_1r_2, e_1r_3, e_2r_1, e_2r_3, e_2r_4, e_3r_2, e_3r_3, e_3r_4)$
- Patt. 11: $(e_1r_1, e_1r_2, e_1r_3, e_2r_1, e_2r_3, e_2r_4, e_3r_1, e_3r_2, e_3r_4)$
- Patt. 12: $(e_1r_1, e_1r_2, e_1r_3, e_2r_1, e_2r_3, e_2r_4, e_3r_1, e_3r_3, e_3r_4)$
- Patt. 13: $(e_1r_1, e_1r_2, e_1r_3, e_2r_1, e_2r_3, e_2r_4, e_3r_1, e_3r_2, e_3r_3)$

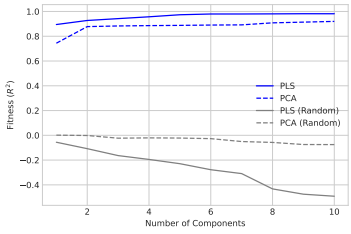
Here, e_ir_j denotes an attribute bound to entity e_i through relation r_j. For example, in the example of Patt. 6 (Sample 8), e_2r_3 corresponds to the attribute "*Perm*", which is connected to entity e_2 (i.e., "*Rick*") via relation r_3 (i.e., *fondness for*). Importantly, these patterns are not simple repetitions of a single structure. Instead, they progressively increase structural density and relational entanglement, covering a wide range of possible entity–relation configurations for a fixed entity set. To illustrate these patterns, we provide examples from the Context: *city*.

(7) **Patt. 1 example.** *Lee, who was born in* $\underline{\textit{Split}}_{e_1r_1}$*, often dreams of traveling to new places. Meanwhile, Rick enjoys his life in* $\underline{\textit{Detroit}}_{e_2r_2}$*, where he explores the city's vibrant culture. Ian, a passionate traveler, expresses his love for* $\underline{\textit{Paris}}_{e_3r_3}$ *but openly shares his dislike for* $\underline{\textit{Austin}}_{e_3r_4}$*, preferring the charm of the French capital instead. The three friends often discuss their different experiences and preferences, bringing their unique perspectives together.*

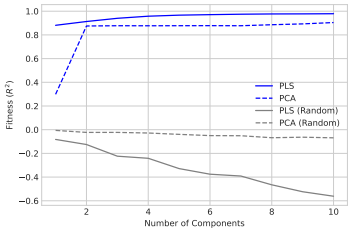
(8) **Patt. 6 example.** *Lee, born in* $\underline{\textit{Split}}_{e_1r_1}$*, enjoys his life in* $\underline{\textit{Hamilton}}_{e_1r_2}$*. Meanwhile, Rick has a fondness for* $\underline{\textit{Perm}}_{e_2r_3}$ *but holds a strong dislike for* $\underline{\textit{Houston}}_{e_2r_4}$*. Ian, originally from* $\underline{\textit{Portland}}_{e_3r_1}$*, also expresses his aversion to* $\underline{\textit{Austin}}_{e_3r_4}$*. Each of them navigates their preferences and experiences, shaping their unique perspectives on the places they call home.*

(9) **Patt. 10 example.** *Lee, born in* $\underline{\textit{Split}}_{e_1r_1}$*, now lives in* $\underline{\textit{Hamilton}}_{e_1r_2}$ *and has a deep affection for* $\underline{\textit{Toronto}}_{e_1r_3}$*. Rick, hailing from* $\underline{\textit{Boston}}_{e_2r_1}$*, holds a fondness for* $\underline{\textit{Perm}}_{e_2r_3}$ *while harboring a dislike for* $\underline{\textit{Houston}}_{e_2r_4}$*. Ian, who currently resides in* $\underline{\textit{Phoenix}}_{e_3r_2}$*, adores* $\underline{\textit{Paris}}_{e_3r_3}$ *but has negative feelings towards* $\underline{\textit{Austin}}_{e_3r_4}$*. Each of them carries their unique experiences and preferences, shaping the places they love and those they do not.*

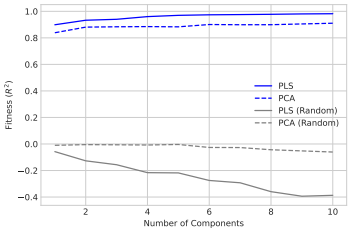
We apply the same PLS-based framework to estimate entity and relation indices across the 13 discourse patterns. The resulting performance scores are shown in Figure 56 and Figure 57. Across all patterns, prediction accuracy remains consistently high even when using a small number of PLS components. This observation suggests that the information required to reconstruct entity–relation bindings is preserved in a low-dimensional subspace, even when the surface structure of the discourse varies substantially.



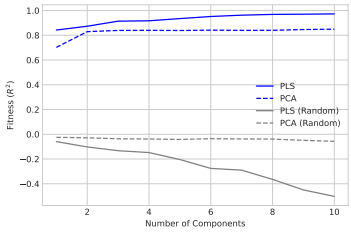
(a) Patt.1



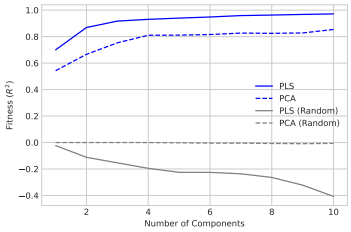
(b) Patt.2



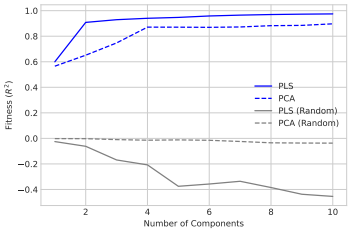
(c) Patt.3



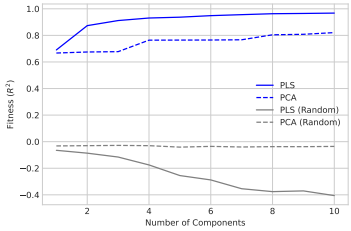
(d) Patt.4



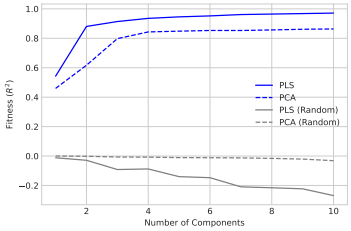
(e) Patt.5



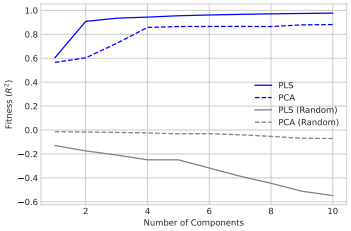
(f) Patt.6



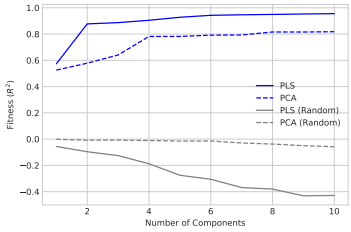
(g) Patt.7



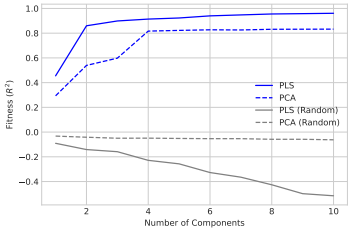
(h) Patt.8



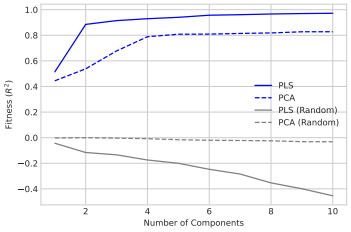
(i) Patt.9



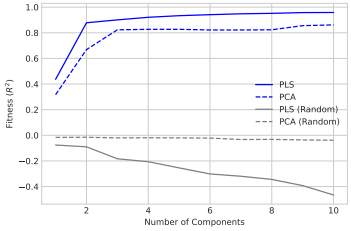
(j) Patt.10



(k) Patt.11

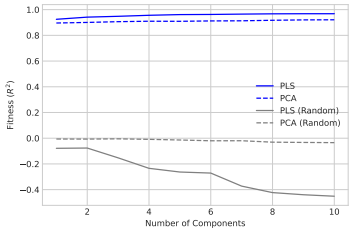


(l) Patt.12

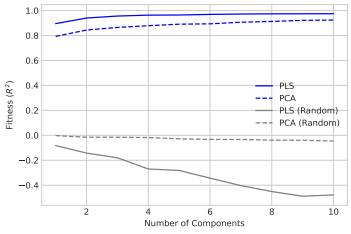


(m) Patt.13

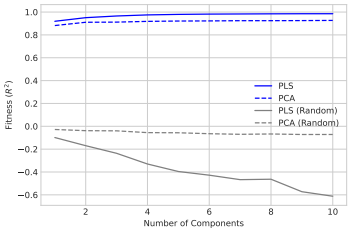
Figure 56: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct on C_{city}.



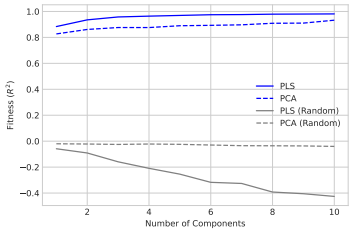
(a) Patt.1



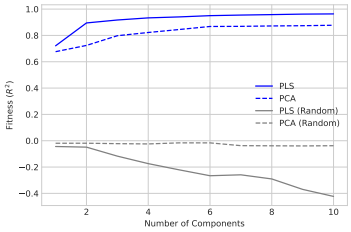
(b) Patt.2



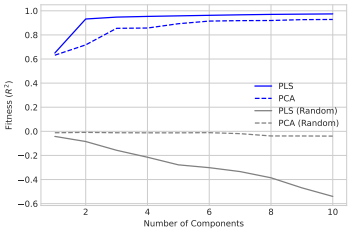
(c) Patt.3



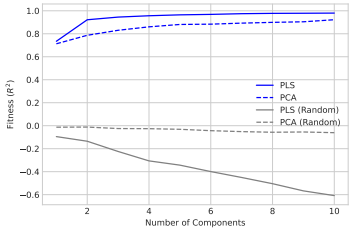
(d) Patt.4



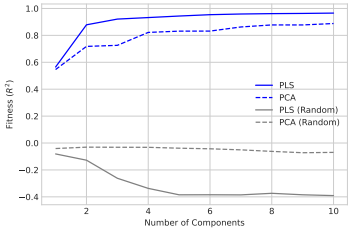
(e) Patt.5



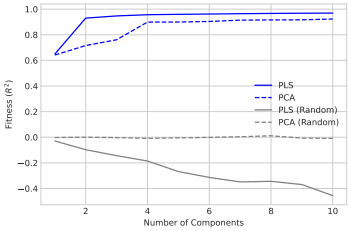
(f) Patt.6



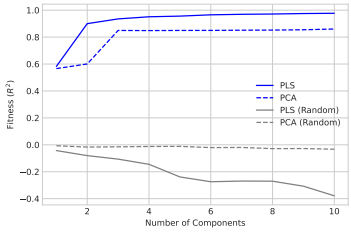
(g) Patt.7



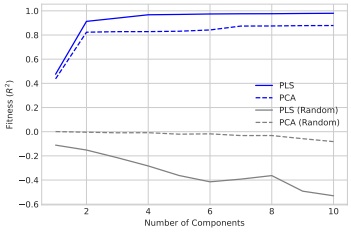
(h) Patt.8



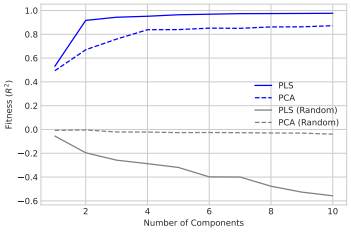
(i) Patt.9



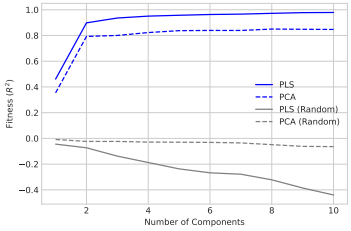
(j) Patt.10



(k) Patt.11

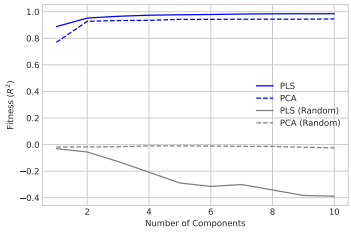


(l) Patt.12

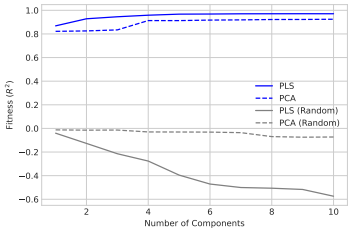


(m) Patt.13

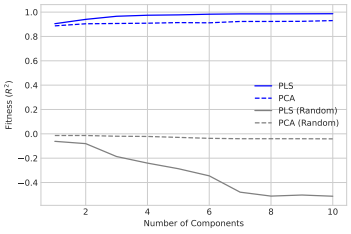
Figure 57: Decoding performance of $[ei, ri]$ from activations of Llama3-8B-Instruct on $C_{country}$.



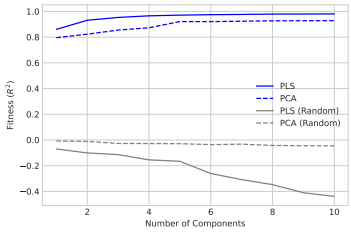
(a) Patt.1



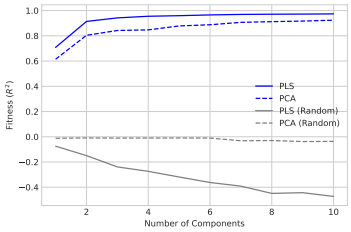
(b) Patt.2



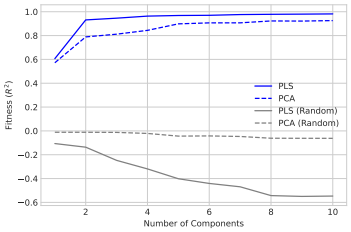
(c) Patt.3



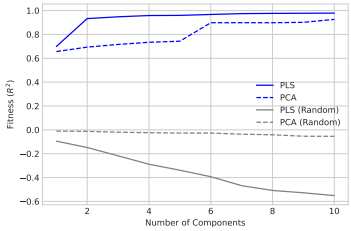
(d) Patt.4



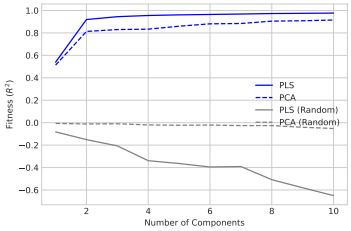
(e) Patt.5



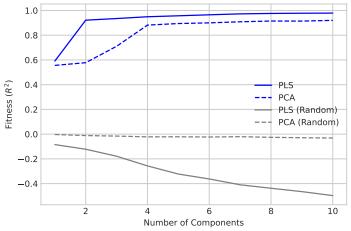
(f) Patt.6



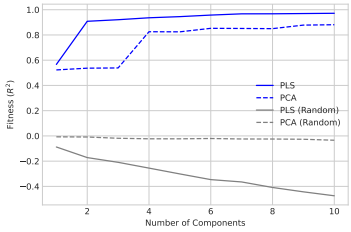
(g) Patt.7



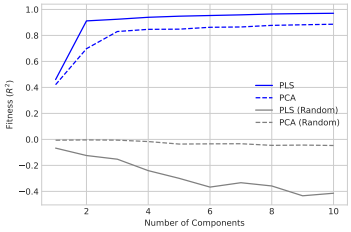
(h) Patt.8



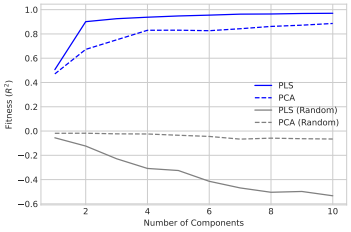
(i) Patt.9



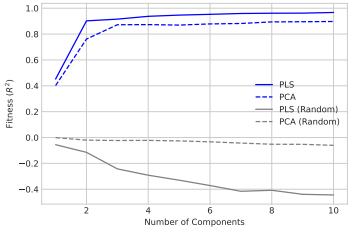
(j) Patt.10



(k) Patt.11

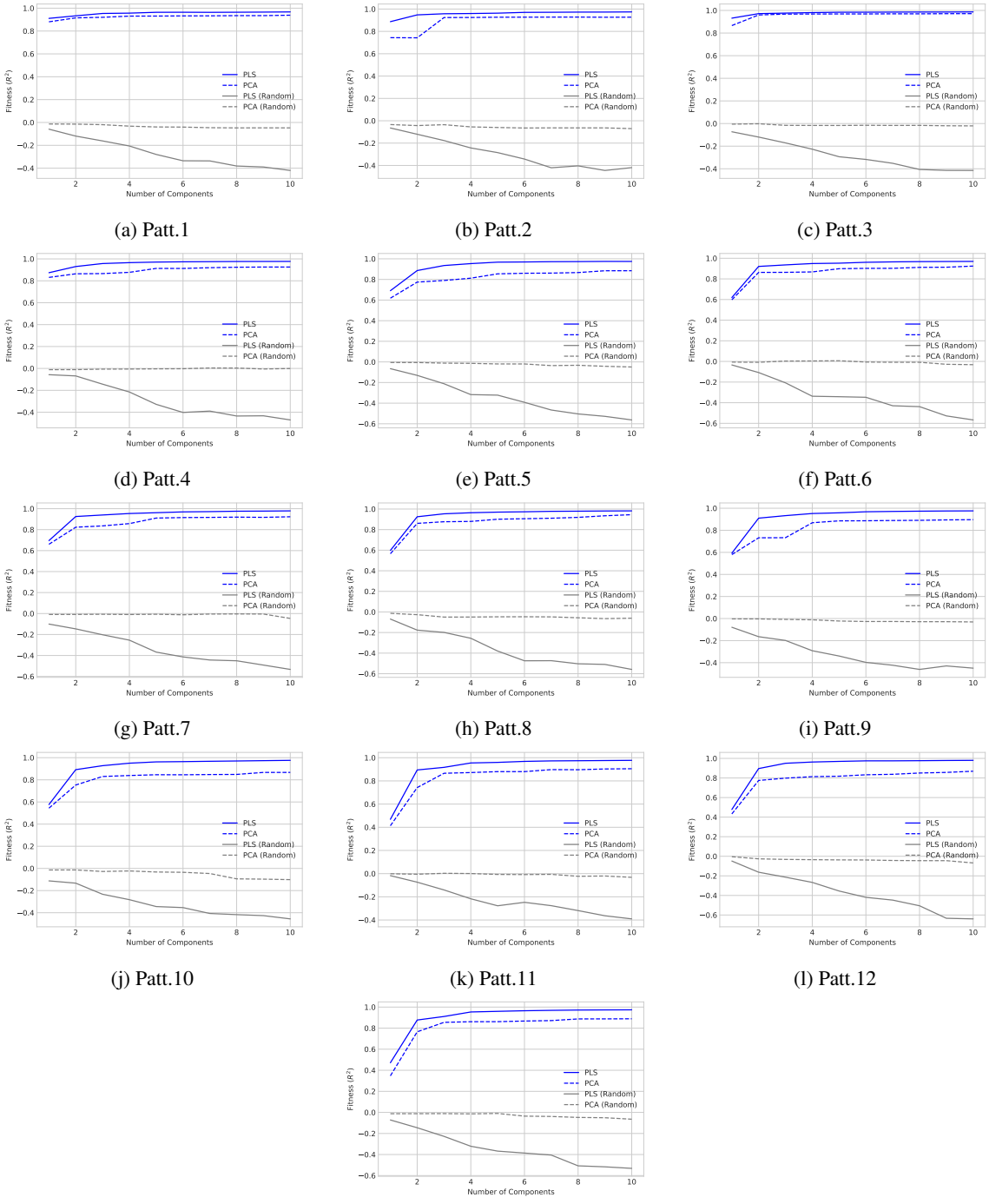


(l) Patt.12



(m) Patt.13

Figure 58: Decoding performance of $[ei, ri]$ from activations of Qwen3-8B on C_{city}.



(a) Patt.1 (b) Patt.2 (c) Patt.3

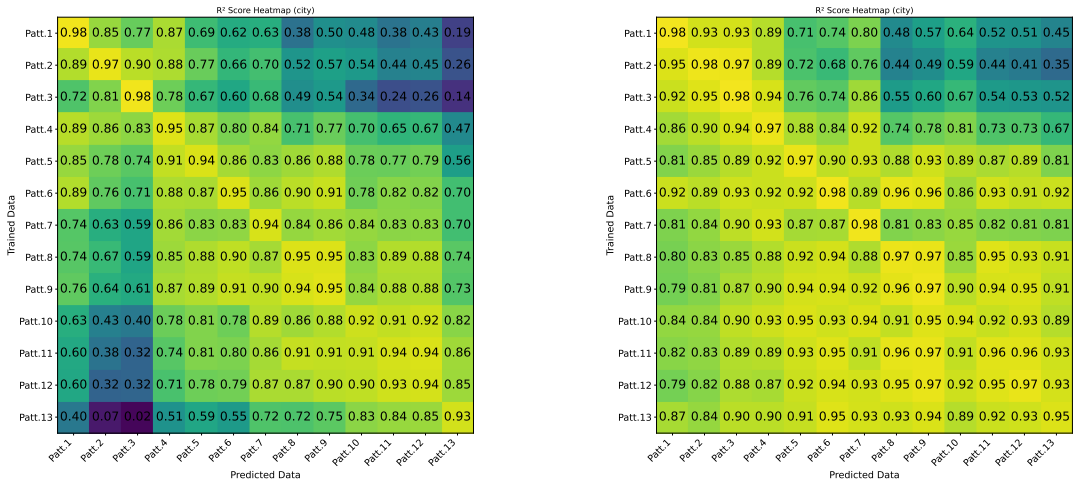
(d) Patt.4 (e) Patt.5 (f) Patt.6

(g) Patt.7 (h) Patt.8 (i) Patt.9

(j) Patt.10 (k) Patt.11 (l) Patt.12

(m) Patt.13

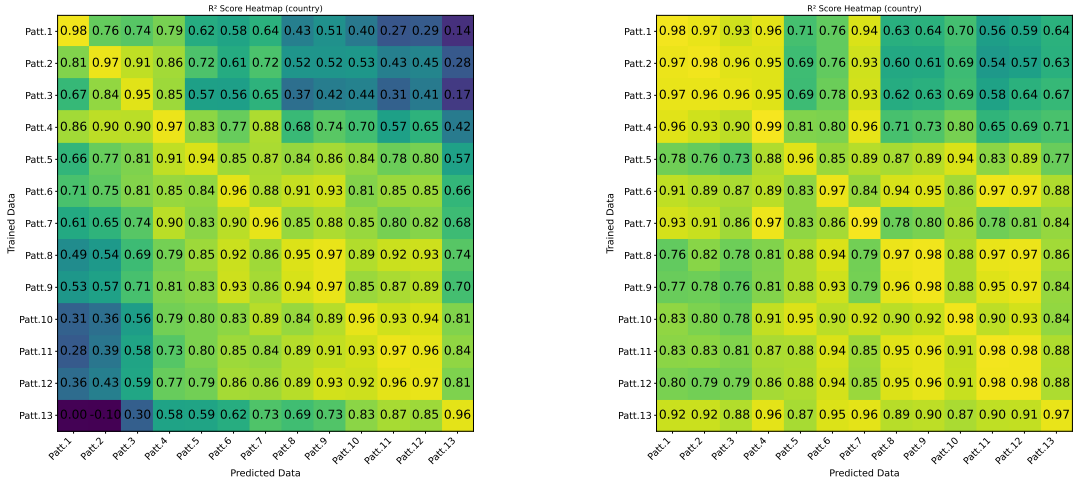
Figure 59: Decoding performance of $[ei, ri]$ from activations of Qwen3-8B on $C_{country}$



(a) Decoding performance of $[ei, ri]$

(b) Decoding performance of $[ri]$

Figure 60: Cross patterns R^2 scores for index prediction from Llama-8B-Instruct on C_{city}.



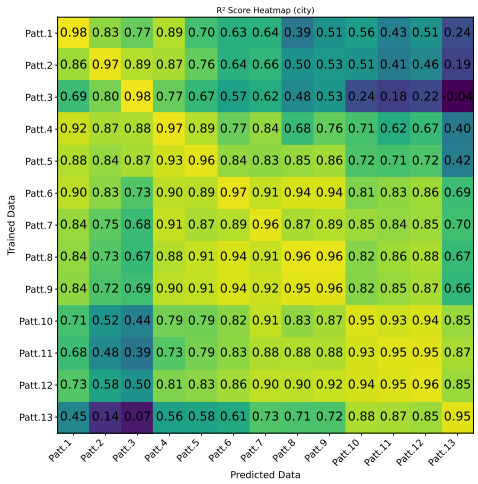
(a) Decoding performance of $[ei, ri]$

(b) Decoding performance of $[ri]$

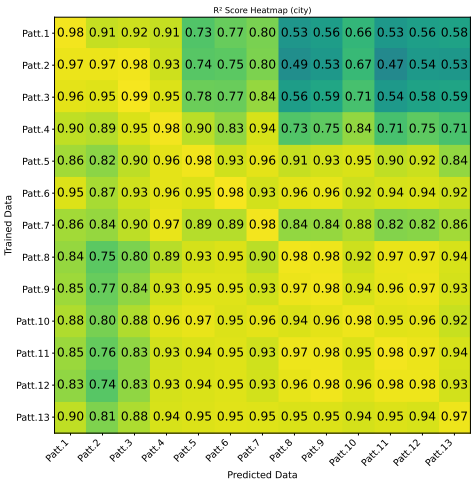
Figure 61: Cross patterns R^2 scores for index prediction from Llama-8B-Instruct on $C_{country}$.

To further examine whether the learned representations generalize across structural variations, we conduct cross-pattern evaluations by applying a projection matrix trained on one pattern to predict indices in another. The results, shown in Figure 60, 61, 62, and 63, indicate that predictive performance transfers well across patterns despite differences in relational selection, repetition, attribute density, and overall structural complexity.

One exception arises for the simplest templates (Patt.1–3), where the entity and relation indices partially overlap. In these cases, the attribute can be uniquely identified using the relation index ri alone, making the entity index ei redundant. This is supported by the PLS results in Figure 56, which show that a single component achieves a high R^2 score for simple patterns such as Patt. 1–4. As a result, a projection matrix trained on more complex patterns (e.g., Patt.13), which learns to jointly decode $[ei, ri]$, does not directly transfer when evaluated on Patt.1–3. However, when decoding is restricted to ri only, the projection matrix again shows strong cross-pattern performance. Therefore, the lower scores observed when transferring from Patt.13 to Patt.1–3 do not contradict our main observation that the CBR representation generalizes across structural variations. Conversely, the lower performance for Patt.1–3 to Patt.8–13 may result from the limited structural complexity of the simpler patterns. Because Patt.1–3 lack the richer relational structures present in patterns such as Patt.10, the learned projection may not generalize to more complex

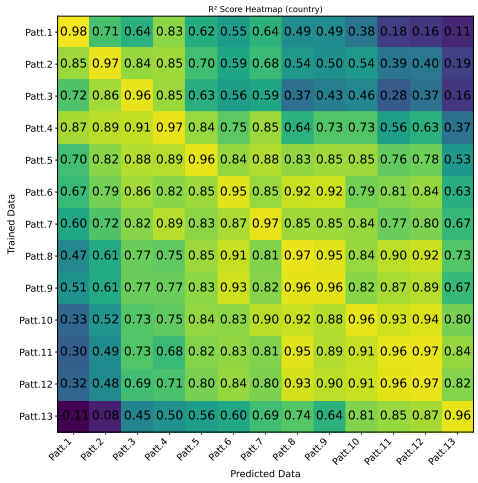


(a) Decoding performance of $[ei, ri]$

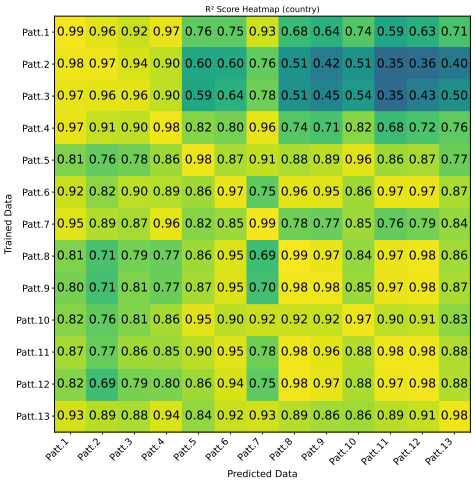


(b) Decoding performance of $[ri]$

Figure 62: Cross patterns R^2 scores for index prediction from Qwen-8B on C_{city}



(a) Decoding performance of $[ei, ri]$



(b) Decoding performance of $[ri]$

Figure 63: Cross patterns R^2 scores for index prediction from Qwen-8B on $C_{country}$.

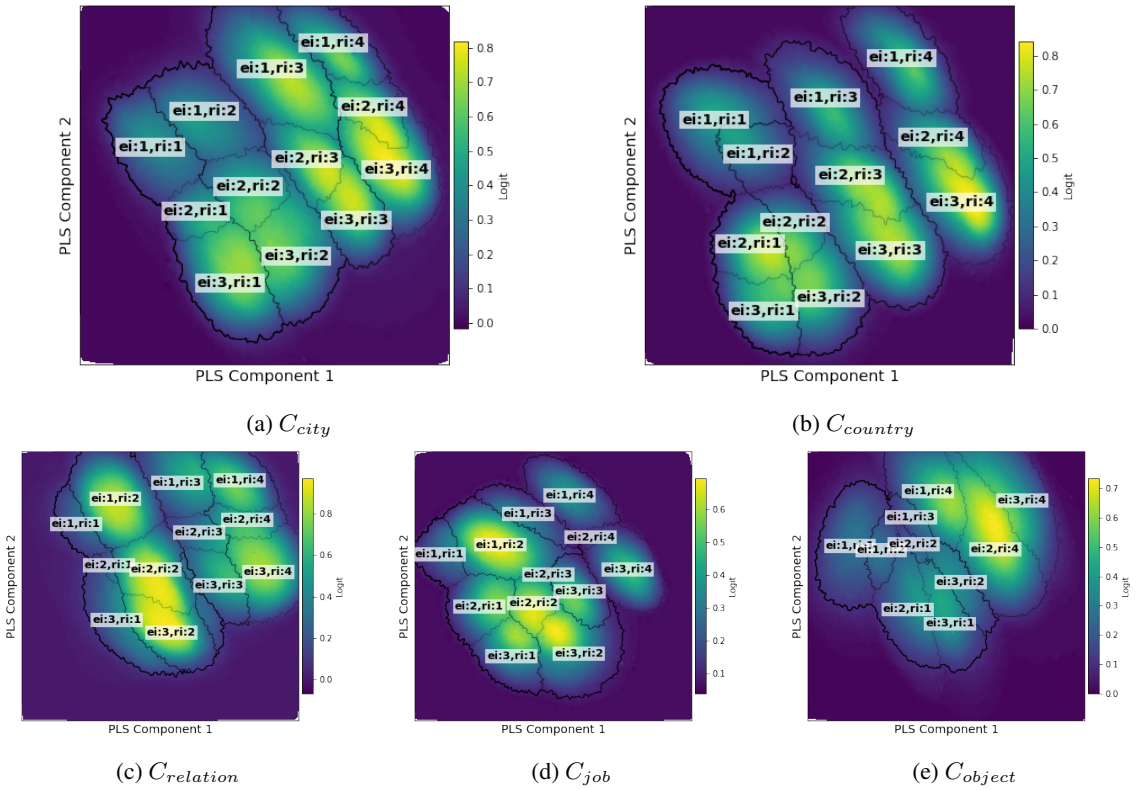
templates.

Taken together, these observations suggest that the overall structure of CBR subspace is stable across a wide range of discourse configurations. The persistence of accurate index prediction under these heterogeneous patterns indicates that the discovered structure does not arise from a particular template design, but instead reflects a broader organizational property of how LLMs encode entity–relation bindings.

A.20 Sampling CBR Subspace across Contexts and LLMs

Section (§3.2) reveals that the CBR subspace is partitioned into regions representing a Voronoi diagram (or cells), where each region corresponds to a specific entity–relation index pair $[ei, ri]$. To evaluate the generality of this Voronoi-like structure, we perform the same CBR-based sampling and visualization procedure across contexts, shown in Figure 64, and on a different model family, Qwen3-8B, shown in Figure 65. In all tested settings, we observe the same characteristic partitioning of the CBR subspace into index-specific regions, indicating that this binding geometry is not context or model specific. This suggests that the Voronoi-like organization of the CBR subspace is a general property of LLMs.

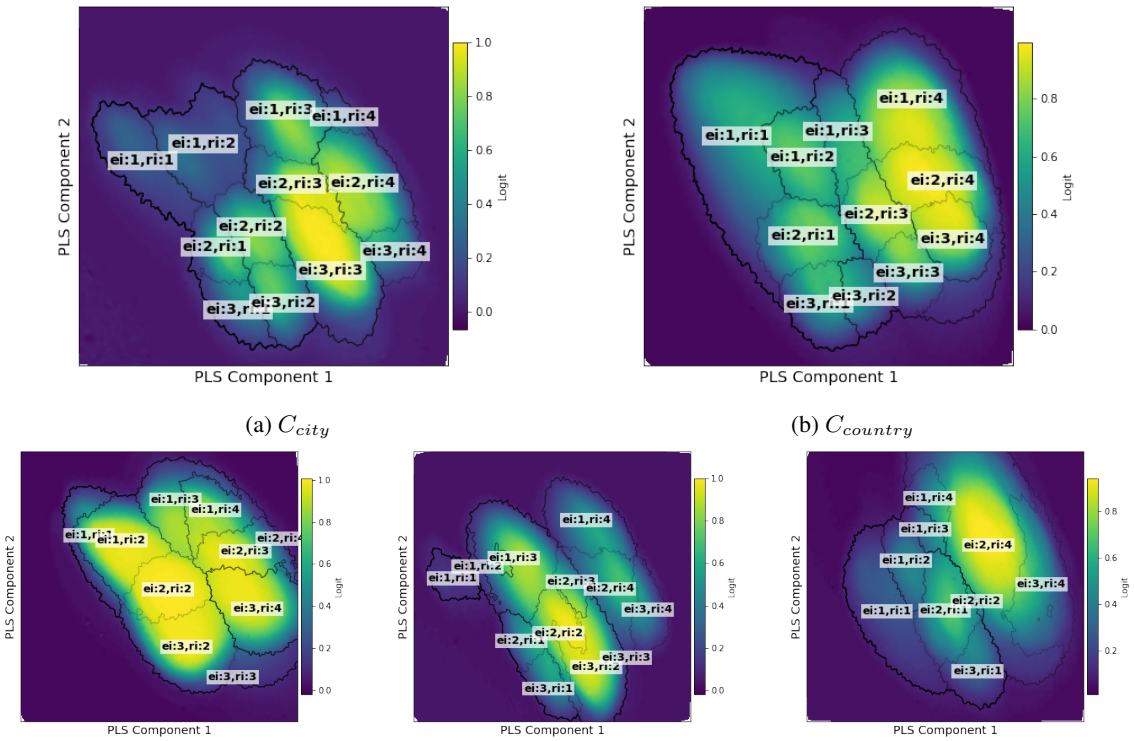
In addition, we also observe that in some contexts, certain entity–relation regions appear to be missing. This likely results from limited sampling density, the low dimensionality of the sampled subspace and improper hyperparameter selection, which may prevent full coverage of all potential regions in those cases. This is further analyzed in the next section.



(a) C_{city} (b) $C_{country}$

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 64: Logit landscape of attribute predictions in the CBR subspace across context on Llama3-8B-Instruct.

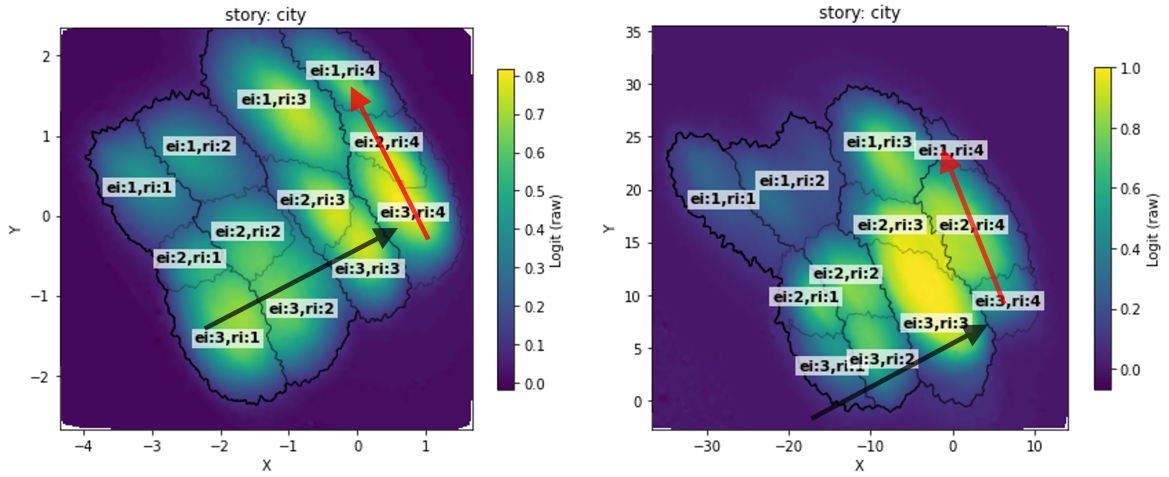


(a) C_{city} (b) $C_{country}$

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

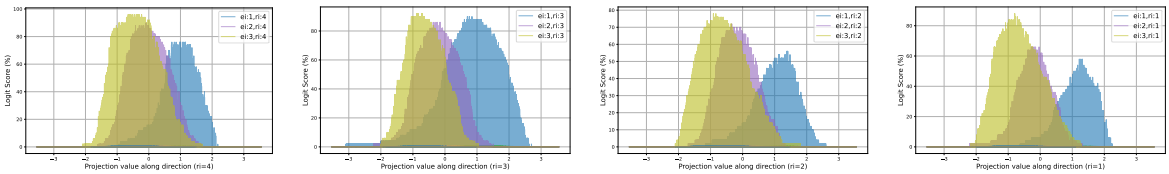
Figure 65: Logit landscape of attribute predictions in the CBR subspace across context on Qwen3-8B.

A.21 Analysis of Logit Score along Entity and Relation Index Directions

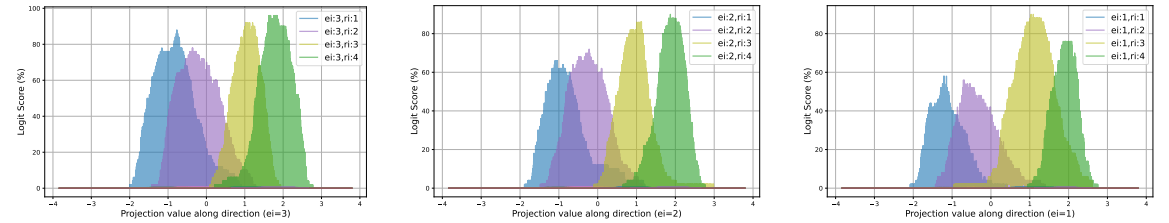


(a) Llama3-8B-Instruct. (b) Qwen3-8B.

Figure 66: ei (red arrow) and ri (black arrow) directions on C_{city}



(a) ei direction with $ri = 4$. (b) ei direction with $ri = 3$. (c) ei direction with $ri = 2$. (d) ei direction with $ri = 1$.

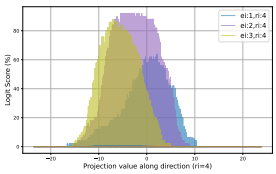


(e) ri direction with $ei = 3$. (f) ri direction with $ei = 2$. (g) ri direction with $ei = 1$.

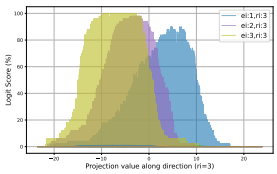
Figure 67: Logit score curves on C_{city} for Llama3-8B-Instruct.

We analyze how the logit score varies as we move along the learned entity-index (ei) and relation-index (ri) directions illustrated in Figure 66a and 66b, and the resulting logit curves are shown in Figure 64 and 68. The results show a characteristic pattern: as we move along one direction, the logit of a particular attribute rises to a peak near the center of its region and then falls off, as the logit of a neighboring attribute begins rising toward its own peak further along the direction. This behavior reinforces our earlier observation that attributes achieve their highest logit near the center of their respective Voronoi-like regions in the CBR subspace.

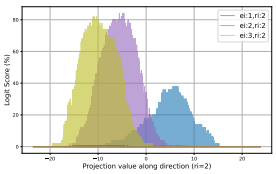
As discussed previously, some regions appear missing in the 2D visualization (e.g., $ei : 3, ri : 1$), as shown in Figure 69. The logit scanning analysis clarifies this phenomenon: in the reduced 2D projection, the logit curves of multiple attributes overlap substantially, indicating that certain regions cannot be fully separated in only two dimensions. This provides further evidence that for a given context (e.g., $C_{relation}$), the 2D visualization could under-sample the true geometry, and that some Voronoi regions become indistinguishable when compressed into a low-dimensional space.



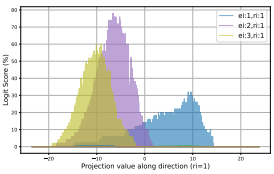
(a) ei direction with $ri = 4$.



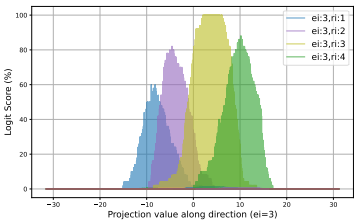
(b) ei direction with $ri = 3$.



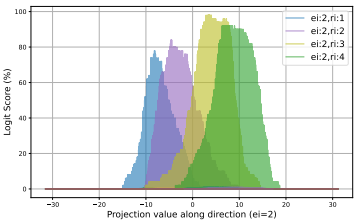
(c) ei direction with $ri = 2$.



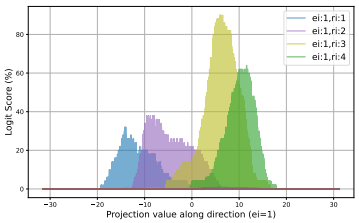
(d) ei direction with $ri = 1$.



(e) ri direction with $ei = 3$.

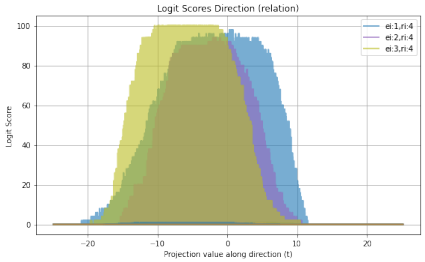


(f) ri direction with $ei = 2$.

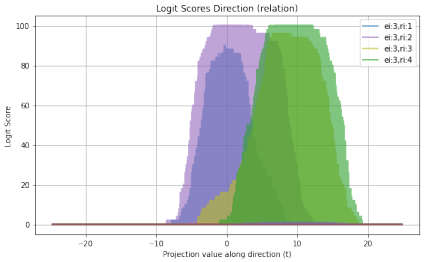


(g) ri direction with $ei = 1$.

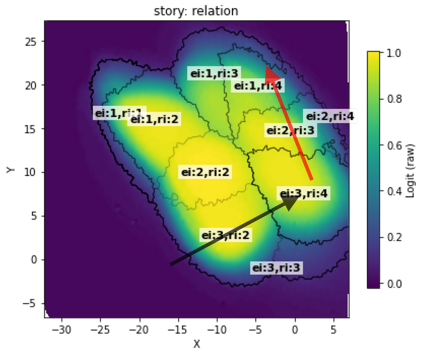
Figure 68: Logit score curves on C_{city} for Qwen3-8B.



(a) Logit score curves along ei direction, where ($ri = 4$).



(b) Logit score curves along ri direction, where ($ei = 3$).

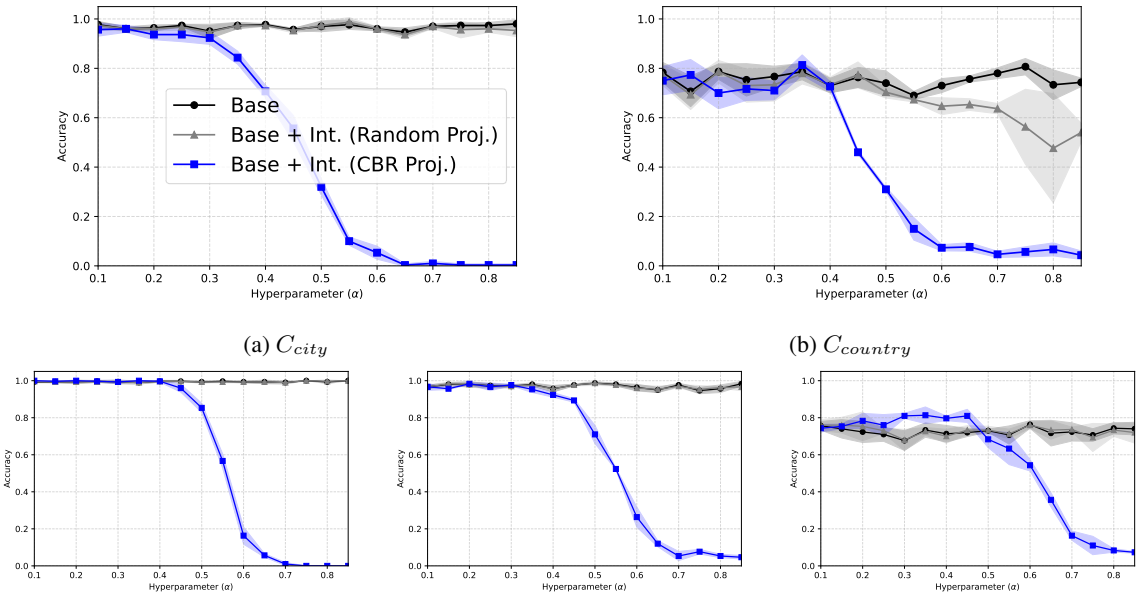


(c) ei (red arrow) and ri (black arrow) directions.

Figure 69: Logit score curves on $C_{relation}$ for Qwen3-8B.

A.22 Perturbing CBR Subspace on Qwen3-8B

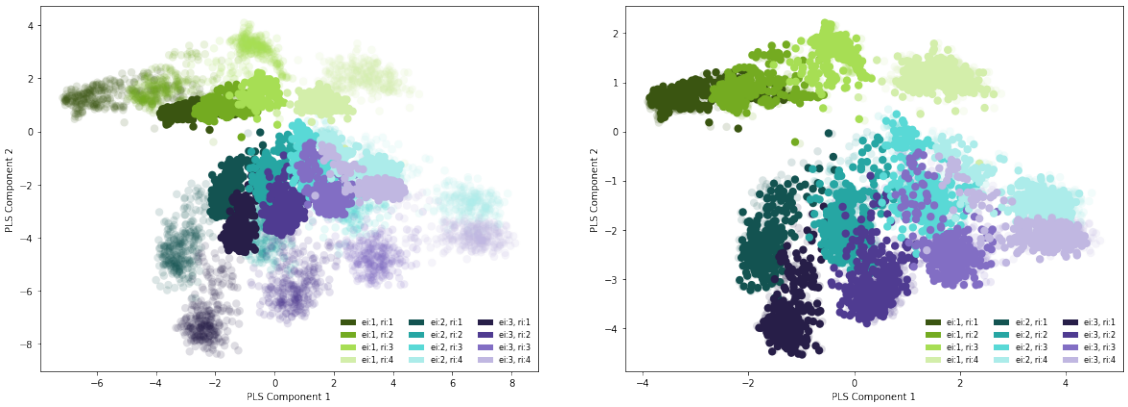
Section (§3.2) mentions that as the perturbation weight increases, the accuracy of attribute predictions decreases significantly. In contrast, perturbations along a random subspace have little or no effect on accuracy. This pattern is also consistently observed in Qwen3-8B, as shown in Figure 70, indicating that this behavior is prevalent across different LLM families rather than being model-specific.



(a) C_{city} (b) $C_{country}$

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 70: Effect of perturbing activations along the CBR subspace versus a random subspace on Qwen3-8B. The X-axis shows the perturbation weight α in Equation 3 and 4, and the Y-axis shows the attribute prediction accuracy. Perturbations along the CBR subspace (i.e., blue line) lead to a significant drop in accuracy, while perturbations along a random subspace (i.e., grey line) have minimal effect. This indicates that LLMs rely on the CBR subspace to make relationally bound predictions.



(a) Perturbing via W_{CBR} (b) Perturbing via W_{rand}

Figure 71: Visualization of the CBR subspace under perturbations along CBR directions using W_{CBR} (Equation 4), and along random directions using W_{rand} (Equation 3) on Llama3-8B-Instruct, where dark colored points denote the distribution before perturbation, while light colored points represent the distribution after perturbation. This demonstrates that the perturbing method using Equation 4 could change the original distribution, while the method using Equation 3 does not.

Figure 72: Causal intervention on the CBR subspace reveals the CBR subspace based mechanism. Steering different components of the CBR subspace produces systematic changes in model behavior. For example, manipulating the relation index in the one-shot attribute activation (e.g., shifting from $ri:1$ to $ri:2$ in the activation of "Australia") redirects the model toward predicting attributes associated with the intervened relation (e.g., changing the output from "China" to "France").

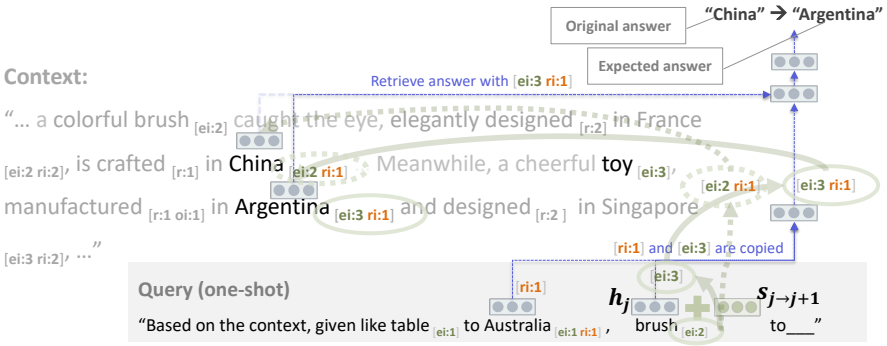


Figure 73: Entity-index Steering.

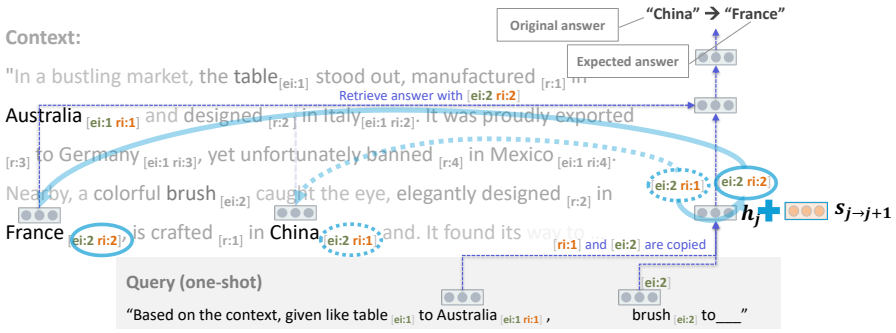


Figure 74: Last-token Steering.

A.23 Activation Steering on other Setting

The results for (b) Entity-index steering illustrated in Figure 73 and (c) Last-token steering illustrated in Figure 74 are shown in Figure 75, 76a and 76b, demonstrating that the logit change pattern is consistent across different steering configurations.

A.24 Activation Steering on Qwen3-8B

Moreover, the same behavior is consistently observed in Qwen3-8B, as show in Figure 77a, 77b, 76a and 78b, indicating that this mechanism is prevalent across different LLM families. Taken together, these results suggest that LLMs rely on an CBR subspace based mechanism to retrieve answers from context.

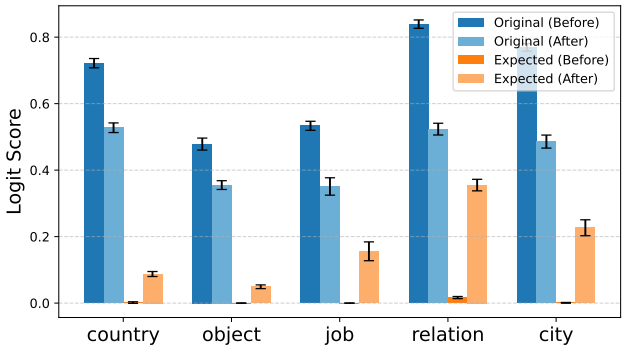
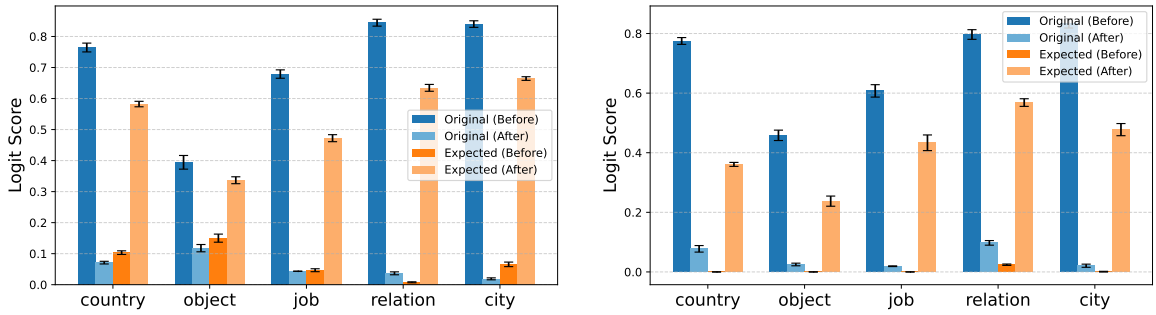


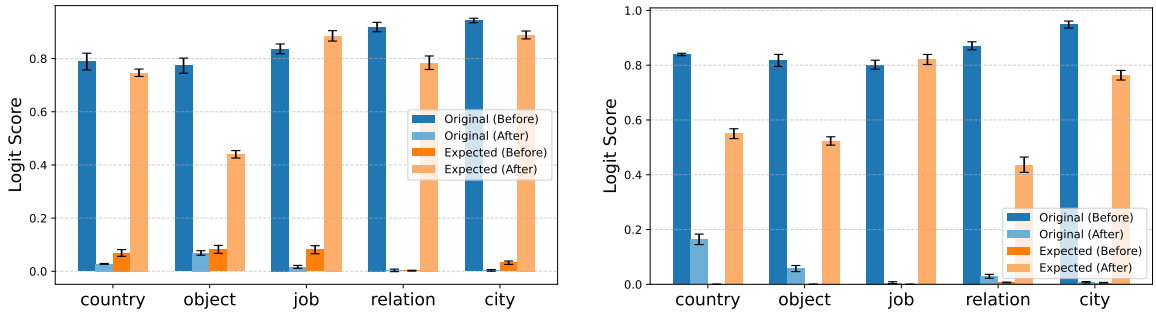
Figure 75: Activation patching via Entity-index (i.e., ei) steering on the activation of the entity token (e.g., "brush") in query part across five contexts on Llama3-8B-Instruct. Each subplot corresponds to one context. We show the change in logit scores for the original answer and the expected answer before and after activation patching, which are denoted as "Original (Before)", "Original (After)", "Expected (Before)" and "Expected (After)" respectively.



(a) Relation-index (i.e., ri) steering.

(b) Entity-index (i.e., ei) steering.

Figure 76: Activation patching on the activation of the last token (i.e., $\mathbf{h}$) across five contexts on Llama3-8B-Instruct.



(a) Relation-index (i.e., ri) steering on the attribute token.

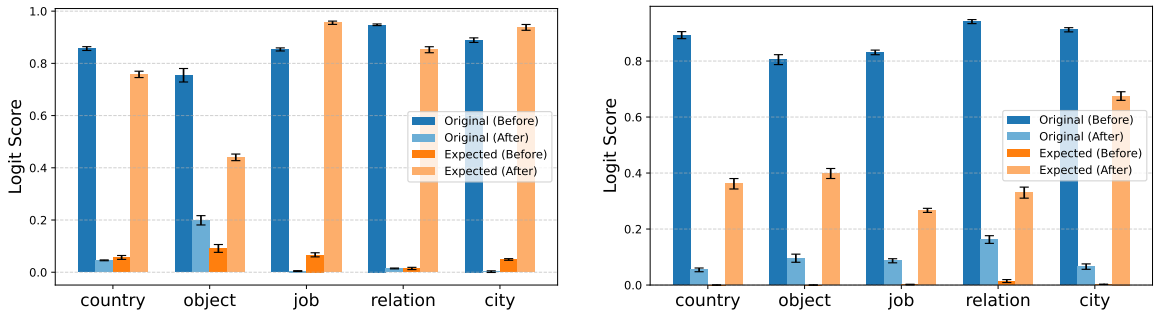
(b) Entity-index (i.e., ei) steering on the entity token.

Figure 77: Activation patching on corresponding token in query part (i.e., $\mathbf{h}$) across five contexts on Llama3-8B-Instruct.

A.25 Activation Steering on Shuffled and Ablated Datasets

Section (§3.2) shows that steering along the CBR subspace direction consistently suppresses the logit of the original answer and increases the logit of the expected answer, in precise alignment with the intended index manipulation. To further test the robustness of this mechanism, we conduct the same AP experiments under both shuffled and ablated settings mentioned in Section (§3.3), where surface-level relations are either reordered or partially removed without altering the underlying indices. As shown in Figure 79a, 79b, 80a and 80b, across these modified settings on both Llama3-8B-Instruct and Qwen3-8B, we observe the same systematic pattern of logit suppression and promotion.

These results indicate that the CBR subspace based mechanism is stable under surface-level perturba-

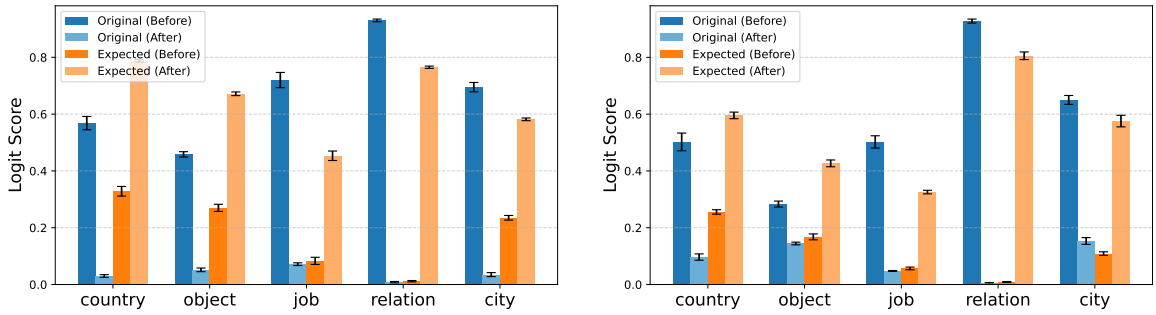


(a) Relation-index (i.e., ri) steering.

(b) Entity-index (i.e., ei) steering.

Figure 78: Activation patching on the activation of the last token (i.e., $\mathbf{h}$) across five contexts on Qwen3-8B.

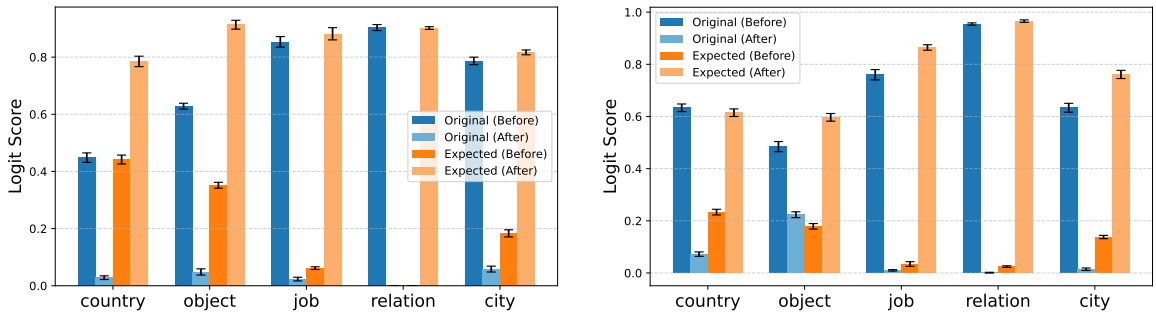
tions and does not depend on superficial modification. Instead, the behavior of the model is governed by the underlying relational structure. This finding strengthens our earlier conclusion that it is the CBR index information, rather than surface form modifications, that drives the organization of the CBR subspace and the relational retrieval behavior of LLMs.



(a) **ablated** dataset

(b) **shuffled** dataset

Figure 79: Activation patching via Relation-index (i.e., ri) steering on the last token (i.e., $\mathbf{h}$) on Llama3-8B-Instruct.



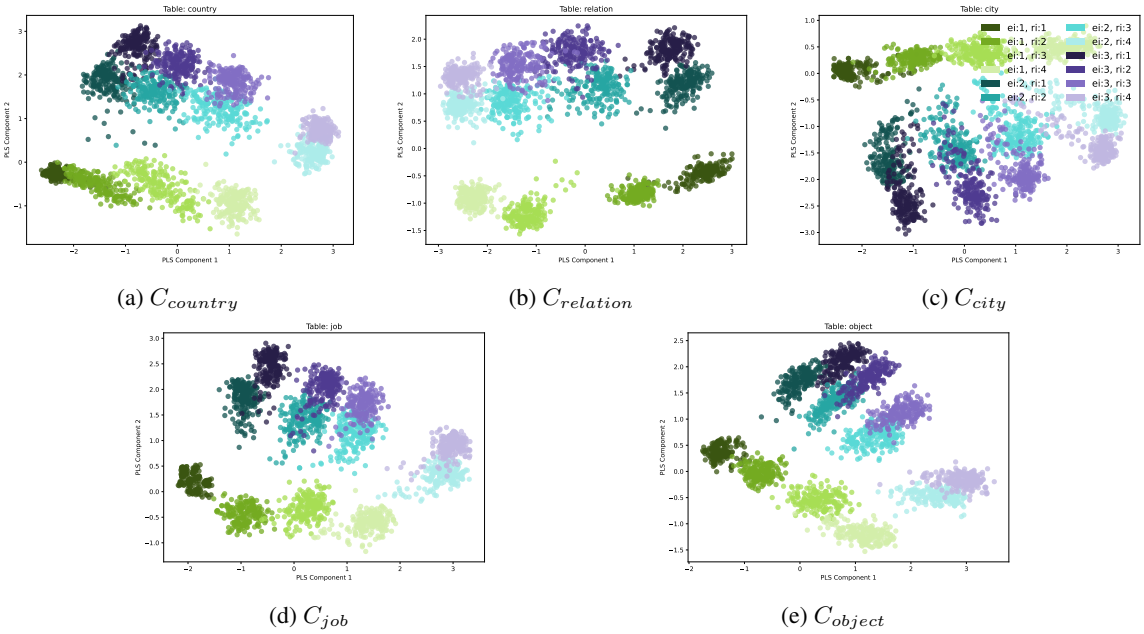
(a) **ablated** dataset

(b) **shuffled** dataset

Figure 80: Activation patching via Relation-index (i.e., ri) steering on the last token (i.e., $\mathbf{h}$) on Qwen3-8B.

A.26 Visualization of CBR Subspace for Template Input

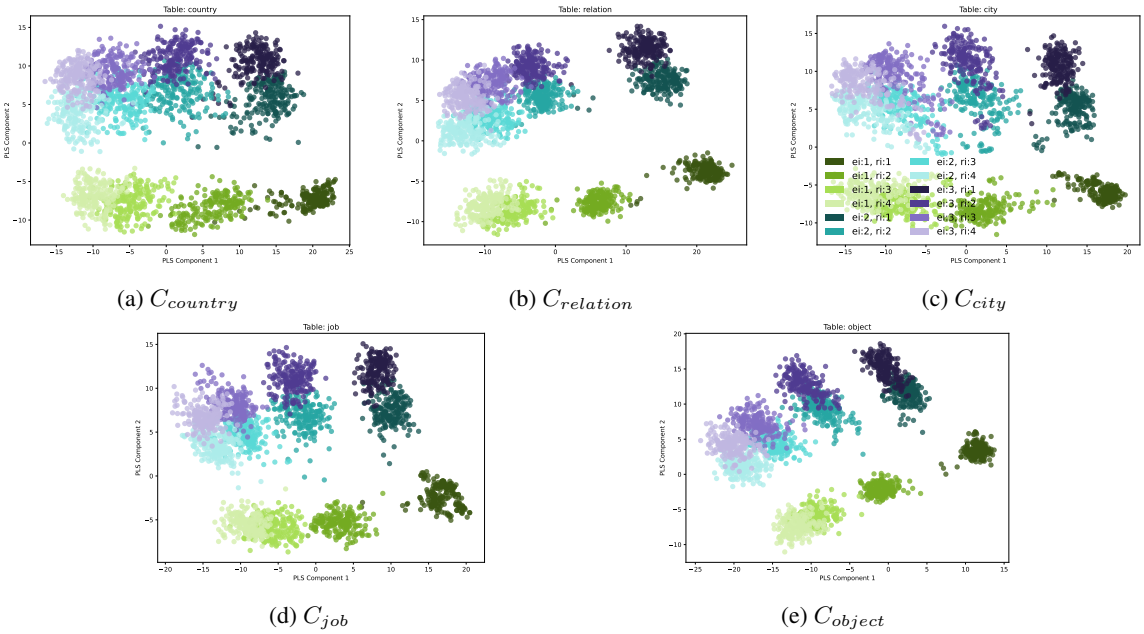
The CBR subspace visualization for Template Input (i.e., Table Template Input as sampled in Table 5, and Discourse Template Input as sampled in Table 6), shown in Figure 81, 82, 83 and 84, are similarly organized along the entity and relation index, indicating that the CBR subspace is a general geometric property common to Table Template Input and Discourse Template Input.



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

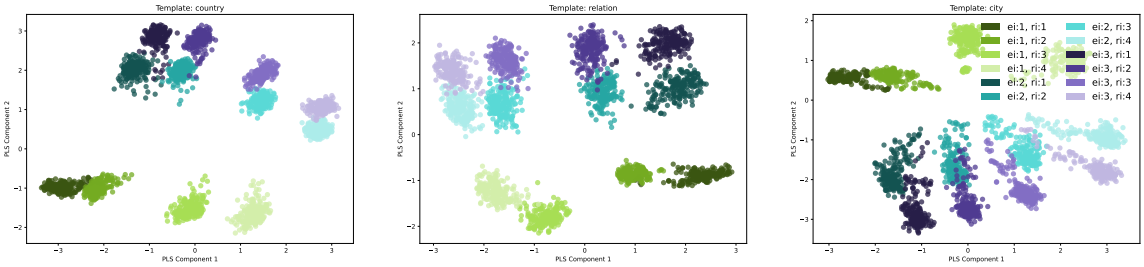
Figure 81: Visualization of the CBR subspace for **Table Template Input** on Llama3-8B-Instruct.



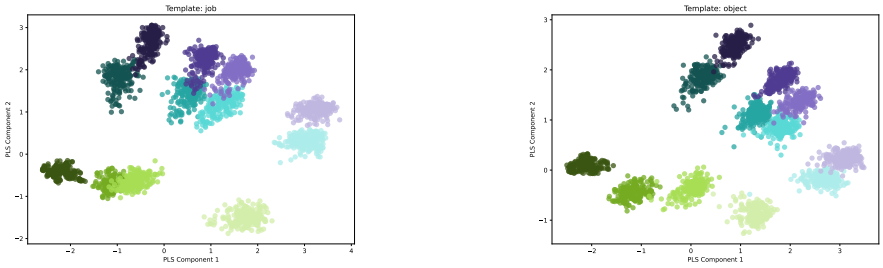
(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

(d) C_{job} (e) C_{object}

Figure 82: Visualization of the CBR subspace for **Table Template Input** on Qwen3-8B.

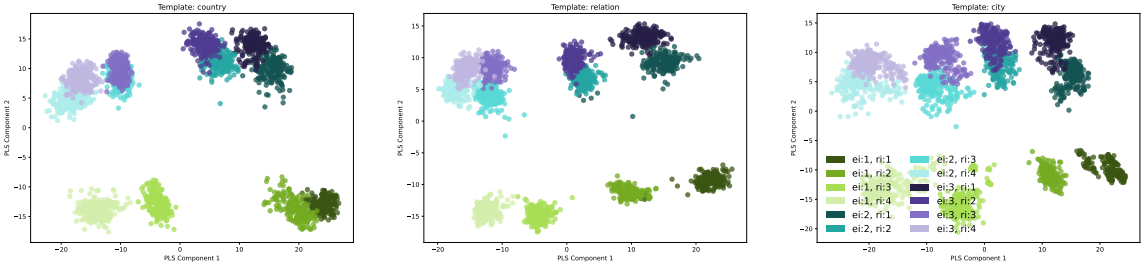


(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

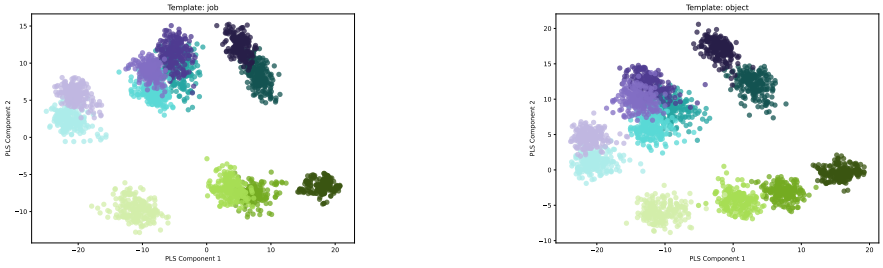


(d) C_{job} (e) C_{object}

Figure 83: Visualization of the CBR subspace for **Discourse Template Input** on Llama3-8B-Instruct.



(a) $C_{country}$ (b) $C_{relation}$ (c) C_{city}

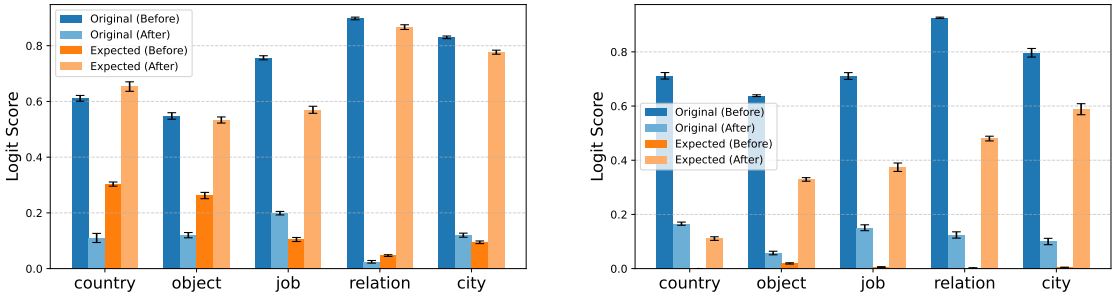


(d) C_{job} (e) C_{object}

Figure 84: Visualization of the CBR subspace for **Discourse Template Input** on Qwen3-8B.

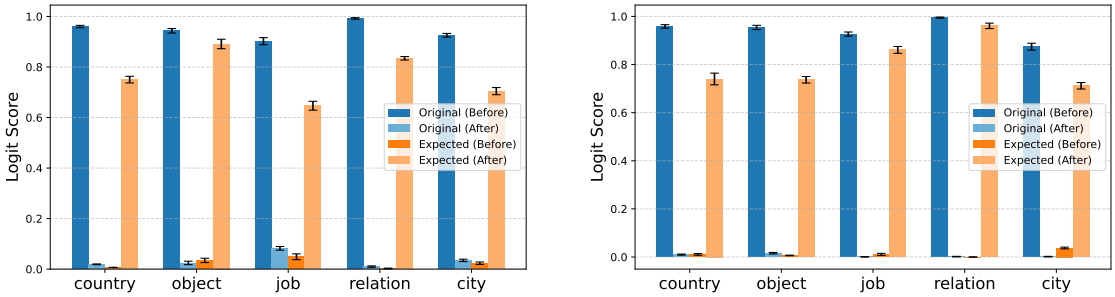
A.27 Activation Steering on Template Input

The results of activation steering on Table Template Input and Discourse Template Input, as described in Section (§A.2), under the one-shot query setting are shown in Figure 85a, 85b, 86a, 86b, 87a, 87b, 88a, and 88b, These results demonstrate that the logit change patterns are consistent across different input formats, further confirming the prevalence of the CBR subspace based mechanism across input formats.



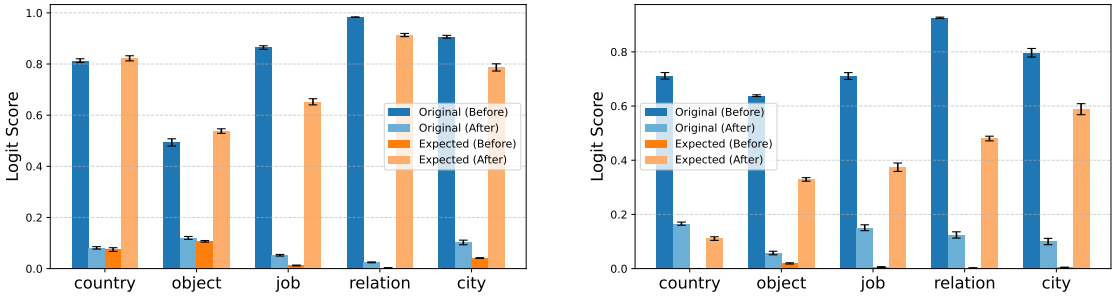
(a) Relation-index (i.e., ri) steering on the attribute token. (b) Entity-index (i.e., ei) steering on the entity token.

Figure 85: Activation patching on **Table Template Input** in query part across five contexts on Llama3-8B-Instruct.



(a) Relation-index (i.e., ri) steering on the attribute token. (b) Entity-index (i.e., ei) steering on the entity token.

Figure 86: Activation patching on **Table Template Input** in query part across five contexts on Qwen3-8B.

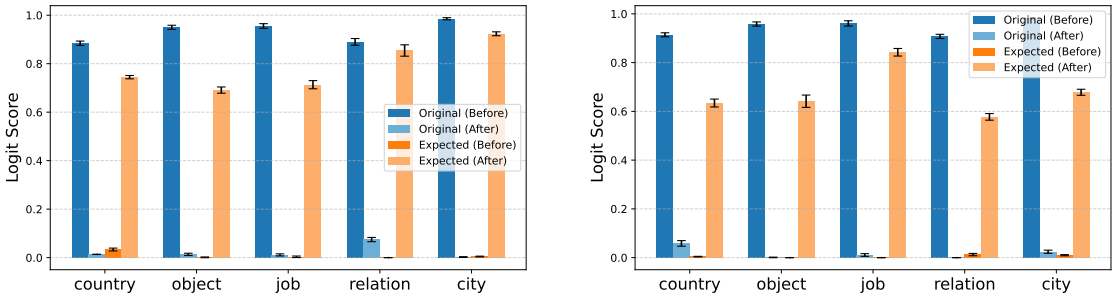


(a) Relation-index (i.e., ri) steering on the attribute token. (b) Entity-index (i.e., ei) steering on the entity token.

Figure 87: Activation patching on **Discourse Template Input** in query part across five contexts on Llama3-8B-Instruct.

A.28 Comparison with Hessian-Based Binding Analysis

Identifying structured subspaces that encode specific functions (e.g., binding) enables more precise monitoring and intervention in model behavior than prompt-level manipulation alone. Feng et al. (2024) propose a Hessian-based algorithm to monitor latent world states and identify binding tokens in LLMs. Their method extracts a binding subspace that encodes entity–attribute associations, enabling the recovery



(a) Relation-index (i.e., ri) steering on the attribute token.

(b) Entity-index (i.e., ei) steering on the entity token.

Figure 88: Activation patching on **Table Template Input** in query part across five contexts on Qwen3-8B.

Condition	Prompt	Hessian	CBR (ours)
pro	1.00†	0.93†	0.95
anti	0.56†	0.83†	**0.94**

Table 8: Comparison with the Hessian-based method. † denotes results estimated from the bar plots in the original paper, as the exact numerical values are not explicitly reported.

of correct bindings from internal representations. For example, in the sample: "*The nurse lives in Singapore. The CEO lives in Canada. The person living in Singapore is male. The person living in Canada is female.*", the method correctly identifies the binding between "*CEO*" and "*female*", rather than the stereotypical association "*CEO–male*". The authors report that this approach outperforms a prompt-based baseline that directly queries the model (e.g., "*The gender of the CEO is*") in gender-bias evaluation settings.

We take the Hessian-based method as a baseline and compare it with our CBR-based framework under the same experimental setting. Specifically, we use the same LLM (Tulu-2-13B), identical templates and entity inventories. As described in Section (§3.1), our approach learns a projection matrix that maps hidden activations into a CBR index space, enabling the matrix to predict bound entity indices (e.g., $ei = 1$ for "*nurse*" and $ei = 2$ for "*CEO*") from the activation of a target token (e.g., "*female*"). To prevent data leakage, the projection matrix is learned using the dataset constructed from different entity inventories, while evaluation is performed on the dataset generated under the same template and settings, which contains 400 samples.

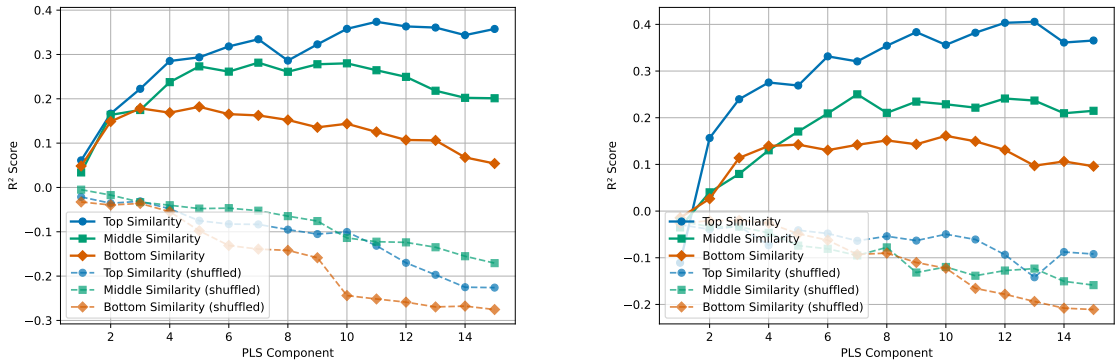
The evaluation is conducted in two conditions: a *pro* condition that follows stereotypical bias (e.g., "nurse–female") and an *anti* condition that counteracts the bias. Table 8 reports the results. While the prompt-based baseline performs well in the stereotypical setting, its accuracy drops substantially in the anti-stereotypical case. The Hessian-based approach improves robustness, particularly under the anti condition. Our CBR-based framework achieves strong performance in both settings, with a notable improvement in the anti condition. These results indicate that the CBR-based framework more reliably captures entity–relation bindings in the model's internal representations. In particular, its strong performance in the anti condition suggests that it can recover correct bindings even when they conflict with stereotypical associations encoded in the model. More broadly, CBR provides a structured way to analyze and monitor internal binding behavior in LLMs, which may support future interpretability and intervention techniques for improving reliability and mitigating bias.

A.29 CBR Analysis on a Real-World Relation Extraction Dataset

To examine whether the CBR representation emerges in real-world data, we analyze the CBR index using the Re-DocRED dataset (Tan et al., 2022), a widely used benchmark for document-level relation extraction. The CBR index is annotated according to the IRS schema described in Section (§3.1). Specifically, the indices are assigned based on the introduction order of relation triples provided in the dataset. An example annotated with the CBR index is shown below. We train the CBR projection matrix on the training set and evaluate it on the development set. If an attribute contains multiple tokens, we average their activations and then decode the corresponding CBR indices.

(10) *William Paul "Bill" Cole III was born May 16, 1956*$_{[ei{:}1,ri{:}1]}$ *and is an American*$_{[ei{:}1,ri{:}2]}$ *businessman, politician and a former Republican*$_{[ei{:}1,ri{:}3]}$ *member of the West Virginia Senate*$_{[ei{:}1,ri{:}4]}$*, representing the 6th district from 2013 to 2017. ...*

To investigate how contextual similarity influences CBR recovery, we measure the embedding similarity between development samples and the training set using a sentence transformer. Based on this similarity score, the development samples are divided into three groups: the Top $k = 50$ most similar samples, the Middle k samples, and the Bottom k least similar samples. We then compute the R^2 score of CBR index prediction for each group. The results are presented in Figure 89. We observe that samples with higher contextual similarity achieve higher R^2 scores. This finding suggests that the CBR signal can be identified in the real-world dataset when the target sample shares strong contextual similarity with examples in the training set.



(a) Decoding performance from Llama3-8B-Instruct

(b) Decoding performance from Qwen3-8B

Figure 89: Relationship between contextual similarity and CBR decoding performance (R^2) on Re-DocRED. "Top/Middle/Bottom Similarity" denote the groups of the Top k, Middle k, and Bottom k samples. "Shuffled" indicates that the index assignments are randomly shuffled. Samples with higher contextual similarity to the training set exhibit higher R^2 scores, indicating that the CBR signal is more reliably recovered when similar contexts are present.

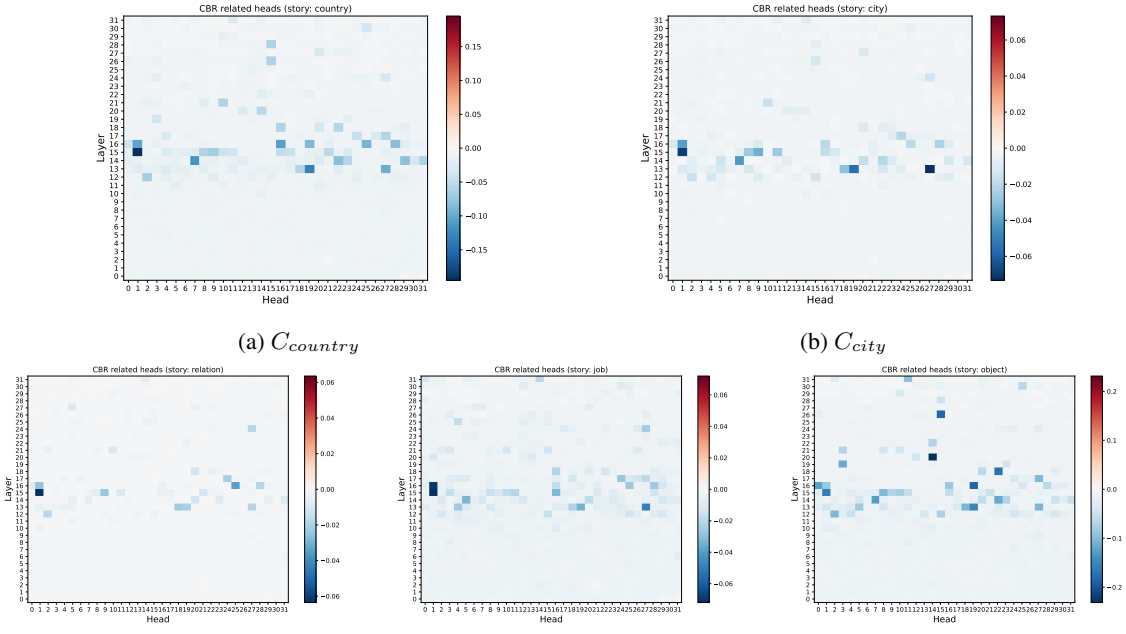
A.30 Detection of CBR related Heads

To identify CBR-related attention heads, we perform head-level activation patching on the final token (i.e., "to" in query part). Specifically, given an input [4] such as Sample 11, we create a corresponding counterfactual input (e.g., Sample 12) that differs in the CBR information associated with the final token. For example, the CBR information of "to" changes from $[ei:2:ri:2]$ to $[ei:2, ri:1]$ in Sample 12. We then patch the activation of each attention head individually from the last token of the counterfactual input into the original input and measure the resulting change in model behavior. Based on this procedure, we define a head-level patching score as $(\text{logit}_{\text{patch}} - \text{logit}_{\text{org}})/\text{logit}_{\text{org}}$, where $\text{logit}_{\text{org}}$ and $\text{logit}_{\text{patch}}$ denote the target logit of answer (e.g., "Berlin") before and after patching, respectively. This score quantifies the contribution of each attention head to encoding CBR information.

(11) **Input**: Sean, who hails from Phoenix, currently resides in Perm. ... Meanwhile, Jose was born in Austin and is now living in Berlin. ... Based on the context, given like Sean to $\text{Perm}_{[ei:1,ri:2]}$, Jose $\text{to}_{[ei:2,ri:2]}$ (Answer: Berlin)

(12) **Counter Input**: Sean, who hails from Perm, currently resides in Phoenix. ... Meanwhile, Jose was born in Berlin and is now living in Austin. ... Based on the context, given like Sean to $\text{Perm}_{[ei:1,ri:1]}$, Jose $\text{to}_{[ei:2,ri:1]}$ (Answer: Berlin)

The patching scores are visualized in Figure 90 and 91, revealing that CBR-related heads are primarily concentrated in the middle layers and are limited to a relatively small subset of heads. To further assess the functional importance of these heads, we ablate them according to their patching score ranking. Using mean ablation (Wang et al., 2022) to selectively knock out the identified heads, we evaluate the resulting performance, with accuracies reported in Figure 92 and 93. The results show that ablating CBR-related heads leads to a substantial degradation in accuracy, providing strong evidence that these heads play a critical role in CBR-based attribute retrieval.

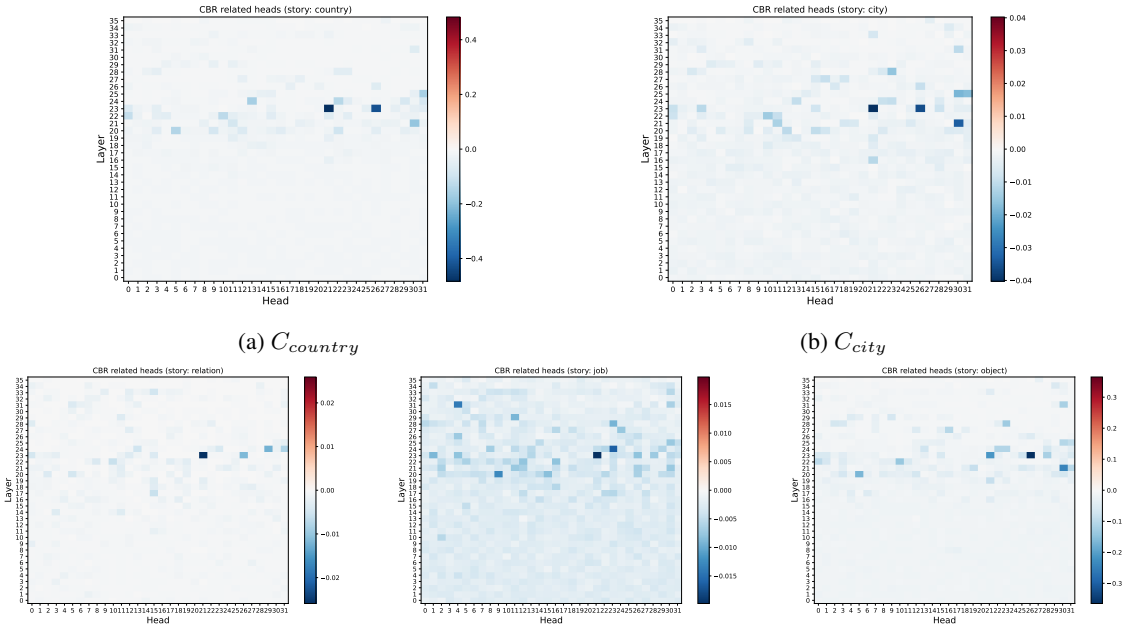


(a) $C_{country}$ (b) C_{city}

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 90: Visualization of the CBR related heads on Llama3-8b-Instruct. Each cell shows the normalized logit change induced by patching a single attention head on the final token. Heads with high scores are primarily concentrated in middle layers, indicating their involvement in CBR-based attribute retrieval.

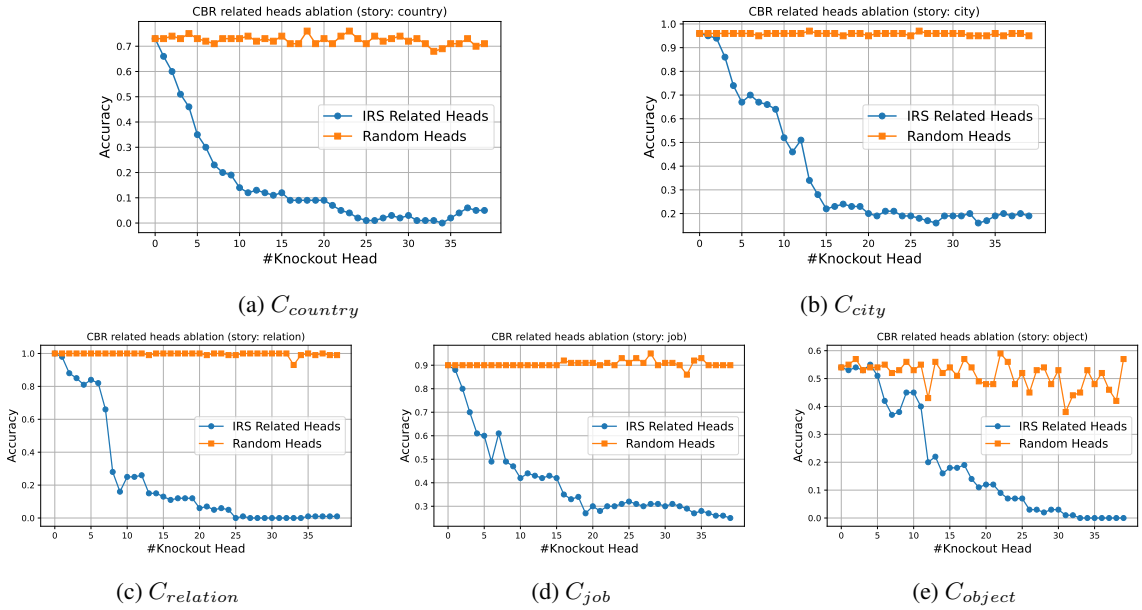
[4] we sampled 300 instances for each context.



(a) $C_{country}$ (b) C_{city}

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

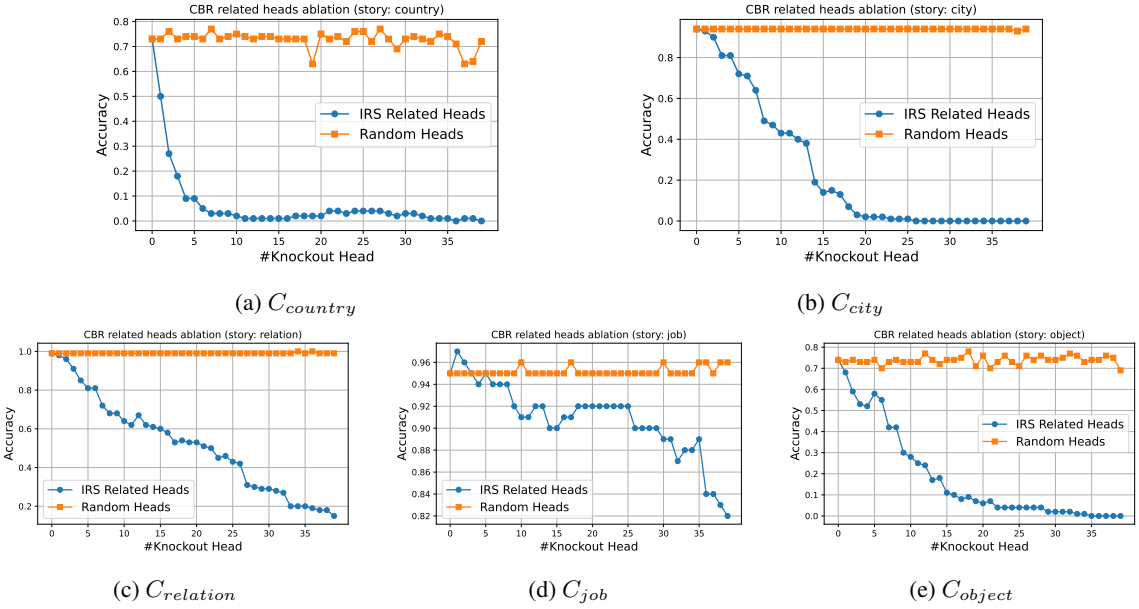
Figure 91: Visualization of the CBR related heads on Qwen3-8b.



(a) $C_{country}$ (b) C_{city}

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 92: CBR related head knockout on Llama3-8b-Instruct, where heads are ablated in descending order of patching score using mean ablation. Comparing with removing Random Heads, removing high-scoring heads causes a sharp drop in accuracy, indicating their critical role in CBR-based retrieval.



(a) $C_{country}$ (b) C_{city}

(c) $C_{relation}$ (d) C_{job} (e) C_{object}

Figure 93: CBR related head knockout on Qwen3-8b.

A.31 Additional Experimental Settings

Machine Causal intervention experiments on CBR subspace are conducted on 48 GB NVIDIA RTX 6000 Ada Generation GPUs.

Other Hyperparameter We set the dimensionality to 15 in Figures 8a and 9, and to 4 in Figure 7. In Equation 2, we set α as -0.4. The value of α in Equation 6 is selected via beam search over the range $[0.4, 1.6]$.

Software The main libraries used in this work include Numpy (Harris et al., 2020), Scikit-learn (Pedregosa et al., 2011), Pytorch (Paszke et al., 2019), Transformers (Wolf et al., 2020) and Matplotlib (Hunter, 2007).

A.32 About AI Assistants for Writing

In this paper, ChatGPT is used to assist with language polishing, grammar checking and minor simplification of visualization code. Its role is only limited to improving wording accuracy and representational clarity. It is important to note that all scientific ideas, methodologies, analyses and discussions are entirely derived from the authors' own research and expertise.