Interpreting Multi-Attribute Confounding through Numerical Attributes in Large Language Models

Hirohane Takagi[1,2,*] **Gouki Minegishi**[1,2,*]
Shota Kizawa[1,2] **Issey Sukeda**[1] **Hitomi Yanaka**[1,2]
[1]The University of Tokyo [2]RIKEN
{htakagi, hyanaka}@is.s.u-tokyo.ac.jp, minegishi@weblab.t.u-tokyo.ac.jp

Abstract

Although behavioral studies have documented numerical reasoning errors in large language models (LLMs), the underlying representational mechanisms remain unclear. We hypothesize that numerical attributes occupy shared latent subspaces and investigate two questions: (1) How do LLMs internally integrate multiple numerical attributes of a single entity? (2) How does irrelevant numerical context perturb these representations and their downstream outputs? To address these questions, we combine linear probing with partial correlation analysis and prompt-based vulnerability tests across models of varying sizes. Our results show that LLMs encode real-world numerical correlations but tend to systematically amplify them. Moreover, irrelevant context induces consistent shifts in magnitude representations, with downstream effects that vary by model size. These findings reveal a vulnerability in LLM decision-making and lay the groundwork for fairer, representation-aware control under multi-attribute entanglement.

1 Introduction

Despite substantial advancements in large language models (LLMs), their capacity to process numerical information remains fragile. Empirical research has documented fundamental misordering, such as the incorrect assertion that "`9.11`" is greater than "`9.9`" (Xie, 2024), and degraded performance on arithmetic word problems when extraneous numbers are present (Shi et al., 2023). Such errors not only undermine reliability in routine arithmetic (Gambardella et al., 2024) but also pose risks in high-stakes domains like financial question answering (Srivastava et al., 2024) and clinical decision support (Hager et al., 2024). These behaviors have been cataloged, yet the internal mechanisms causing numerical errors remain largely unclear.

Mechanistic interpretability methods, principally probing (Belinkov, 2022), are employed to elucidate the encoding of concepts within LLMs' hidden states (Bereska and Gavves, 2024). Previous work shows that LLMs encode numerical attributes, such as geographical coordinates or temporal data, in linear and monotonic internal subspaces (Gurnee and Tegmark, 2024). In particular, Partial Least Squares (PLS; Wold et al., 2001) identifies internal subspaces most correlated with the numerical labels, and the found space is demonstrated to be used for numerical reasoning of comparisons (El-Shangiti et al., 2025). The discovered subspaces are also guaranteed to be causally meaningful as the intervention can modulate the inference of LLMs (Li et al., 2023; Zou et al., 2025).

However, Heinzerling and Inui (2024) reported that intervention in the specific numerical attribute spaces causes side effects on other attributes, indicating that numerical concepts are entangled in the hidden states. Given that LLMs exhibit shared representations—clustering semantically similar concept vectors (Zhao et al., 2025) and aligning across languages and modalities (Wu et al., 2025)—we hypothesize that *multiple numerical attributes likewise occupy overlapping internal subspaces in terms of the magnitude*.

Such shared subspaces imply two key forms of confounding that undermine both interpretability and downstream reliability. First, probing a target attribute (e.g., geographical entity's `population`) risks misattributing correlated numerical information from related attributes (e.g., `area`). Second, realistic prompts often embed multiple numerical values, including irrelevant distractors, which may perturb internal representations and degrade performance. A comprehensive analysis of these cases is therefore essential: it will clarify how inter-attribute entanglement arises within LLMs and inform the design of probing methods and mitigation strategies that account for multi-attribute

[*]Equal Contribution

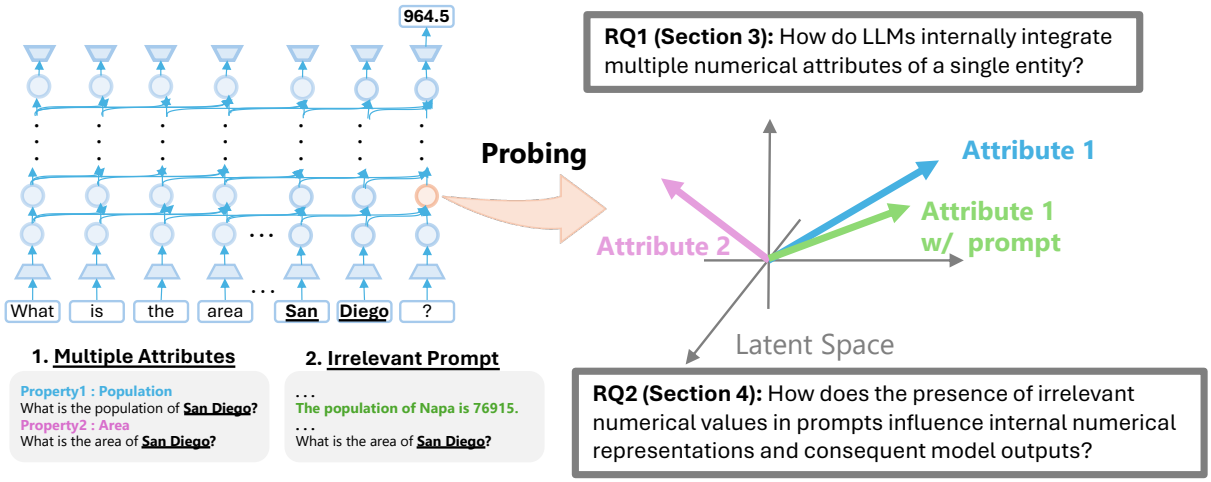

Figure 1: Overview of our approach to analyzing internal representations in LLMs by addressing two research questions (RQs): RQ1 examines how LLMs represent entities with multiple correlated numerical properties (e.g., San Diego's `population` and `area`). RQ2 investigates how irrelevant numerical details in prompts influence these internal representations.

interactions. Given these considerations and the challenges posed by multi-attribute prompts, we formulate the following research questions (RQs), illustrated in Figure 1:

RQ1 How do LLMs internally integrate multiple numerical attributes of a single entity?
RQ2 How does the presence of irrelevant numerical values in prompts influence internal numerical representations and consequent model outputs?

Using extended correlation analyses tailored to the RQs, we quantified the inter-attribute generalization of the PLS probing and obtained the following key insights.For **RQ1**, the internal representation associated with a given numerical attribute often generalizes to predict magnitudes of other related attributes. In some cases, the internal correlations exceed those empirically observed in the data. We also observe asymmetric interference patterns, wherein certain dominant attributes exert disproportionate influence over others. These results imply that numerical attributes that are popular within the same entity are stored as weights in LLMs, but that the handling of numerical magnitude is generalized across attributes. For **RQ2**, LLMs are susceptible to interference from irrelevant numerical inputs, particularly smaller models, where distractors significantly alter internal representations and degrade output reliability. Larger models can mitigate these perturbations through more complex computations. Our findings confirm shared numerical subspaces in LLM internal representations, mechanistically explaining previously observed vulnerabilities such as side effects in intervention and context-induced numerical errors. These analyses motivate refinements to both interpretability methods and deployment practices in numerically-sensitive applications.

We make our code publicly available at `https://github.com/htkg/num_attrs`.

2 Preliminaries

To quantitatively investigate our research questions on the shared subspace hypothesis for numerical attributes, we employ Partial Least Squares (PLS; Wold et al., 2001), which simultaneously performs linear prediction and dimensionality reduction, effectively identifying internal numerical attribute representations (Heinzerling and Inui, 2024; El-Shangiti et al., 2025). To evaluate inter-attribute generalization, we use Spearman's rank correlation, which is robust to scale and offset differences. Additionally, partial correlation analysis accounts for inherent attribute correlations within entities.

Probing via PLS Regression Let $h \in \mathbb{Z}_+$ be the hidden dimension of an LLM and $x_i \in \mathbb{R}^h$ be the hidden representation of the i-th sample, i.e., the activation of a specific token at a given layer of the LLM. Collect n samples into $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times h}$ and $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$. When $h > n$, direct regression leads to overfit, necessitating regularization (Gurnee and Tegmark, 2024) or dimension reduction (Heinzerling and Inui, 2024). PLS regression (Wold et al., 2001) addresses this by projecting X onto a low-

dimensional space that maximizes covariance with Y. In a rank-k model, PLS yields $W \in \mathbb{R}^{h\times k}$ to calculate the results of dimension reduction $Z = XW \in \mathbb{R}^{n\times k}$ and $C \in \mathbb{R}^{k}$ to predict $\hat{Y} = ZC \approx Y$, while reconstructing $X \approx ZP^{\top}$ via $P \in \mathbb{R}^{h\times k}$. The model's fitting and its goodness, measured by the R^2 score, are calculated using the algorithm provided in the Scikit-learn (Pedregosa et al., 2011).

Spearman's (Partial) Rank Correlation To evaluate agreement between predictions $\hat{Y}$ and true values Y, we employ Spearman correlation $r(\hat{Y}, Y)$. When Z confounds $\hat{Y}$ and Y, partial correlation (Kim, 2015) is calculated as follows:

$$r(\hat{Y}, Y|Z) = \frac{r(\hat{Y}, Y) - r(\hat{Y}, Z)\, r(Y, Z)}{\sqrt{(1 - r(\hat{Y}, Z)^2)(1 - r(Y, Z)^2)}}.$$

This matches the original correlation $r(\hat{Y}, Y)$ if Y and $\hat{Y}$ are independent of Z, while their correlation reduces the value. All reported correlations are either Spearman's rank or partial rank correlation, computed via scientific libraries (Virtanen et al., 2020; Vallat, 2018).

3 RQ1: Representation of Entities with Multiple Numeric Attributes

This section investigates how LLMs encode the naturally correlated numeric attributes of entities. To address **RQ1**, we particularly focus on:

Preservation: To what extent do LLM representations preserve the natural correlations among multiple numeric attributes of the same entity?

Confounding Effects: When probing on a single attribute, how strongly do other attributes become unintentionally predictable?

3.1 Dataset for Inter-attribute Analysis

Dataset Construction We begin by extending the probing corpus for numeric attributes in Heinzerling and Inui (2024) to construct the *inter-attribute evaluation set*, in which each entity has all of the specified attributes. The original 1,000-sample training split for each attribute to fit the probing model is retained unchanged to ensure comparability. For inter-attribute analysis, we crawled Wikidata (Vrandečić and Krötzsch, 2014).[*]

[*]We retrieved data from https://www.wikidata.org/ on April 15, 2025, under the Creative Commons Attribution-ShareAlike 4.0 International License.

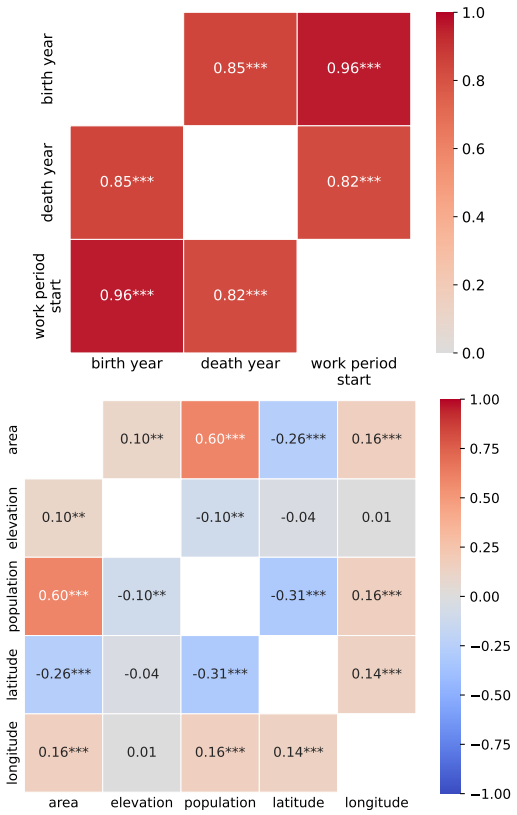

Figure 2: Correlation matrices for human (top) and geographical (bottom) entities. The year attributes of human entities or some attributes of geographical entities are likely to be correlated (with significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

This *inter-attribute evaluation set* includes:

Human entities with `birth year`, `death year`, and `work period start` (402 samples),

Geographical entities with `area`, `elevation`, `population`, `latitude`, and `longitude` (777 samples).

These are not duplicates of the data used for single-attribute fitting or hyperparameter selection. Note that correlations with a larger sample size, including duplicates and missing attributes, are also confirmed in Appendix A to ensure consistency with this limited dataset for computational efficiency.

Observation On *inter-attribute evaluation set*, the pairwise correlations are shown in Figure 2. Attributes of human entities, such as `birth year`, `death year`, and `work period start`, exhibit strong correlations with each other. In addition, `area` and `population` are moderately correlated, and several other pairs, such as `latitude` and other attributes, show non-negligible correlations. These

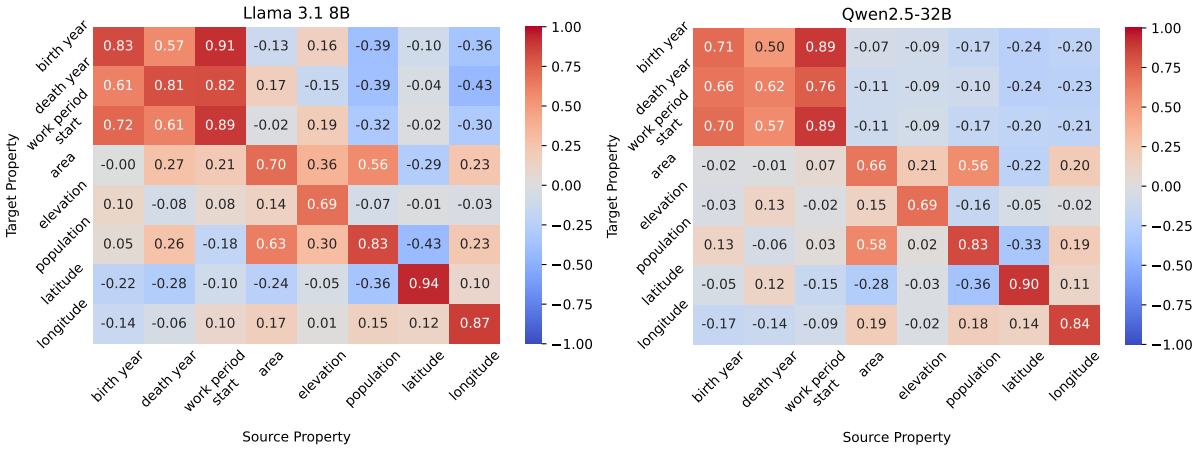

Figure 3: Spearman correlations for Llama 3.1 8B (top) and Qwen2.5-32B (bottom): diagonal is within-attribute, while off-diagonal is inter-attribute. The attributes of human entities can be predicted across attributes. Furthermore, despite the fact that the diagonal components of geographical entities are less than one and reproduction within attributes is incomplete, `area`, `population`, and `latitude` exhibit *unnatural* cross-attribute correlations.

natural correlations can be assumed to be inherent in the training data of LLMs, serving as a baseline for subsequent analyses on inter-attribute effects.

3.2 Method and Models

Representation Extraction In order to rigorously examine confounding across attributes of LLM internal representations for prompts containing numerical attributes, the following two conditions are used. In the *in-question noun* setting, each prompt explicitly asks the model for the value of a specific attribute of a given entity (the question templates are detailed in Appendix A); in the *isolated noun* control, the prompt contains only the entity name without any attribute query. The hidden states of the final tokens at a specified layer are collected and denoted as $X \in \mathbb{R}^{n \times h}$.

Inter-attribute Evaluation For each source attribute s and target attribute t, we define r_s as the Spearman correlation measured in the subspace of s. For example, $r_s(\hat{Y}_t, Y_t)$ denotes the correlation between the ground truth Y_t and the prediction $\hat{Y}_t$ obtained by applying *the PLS model trained on* s to X_t. The overall procedure is as follows:

1. Train PLS on representations X_s from the training split of attribute s, by sweeping the hyperparameters such as layer position l and PLS rank k.
2. Use the single-attribute dataset with 1000 samples to choose the best hyperparameters.
3. On the *inter-attribute evaluation set*, predict Y_t using X_t from the trained PLS model, and compute $r_s(\hat{Y}_t, Y_t)$. When $s \neq t$, we also calculate $r_s(\hat{Y}_s, Y_s|Y_t)$ and $r_s(\hat{Y}_t, Y_t|Y_s)$.

Intuitively, $r_s(\hat{Y}_s, Y_s|Y_t)$ evaluates how well the PLS on s can predict Y_s without interference from Y_t, indicating *attribute fidelity*. On the contrary, $r_s(\hat{Y}_t, Y_t|Y_s)$ evaluates the degree to which the representation on s predicts Y_t after controlling for s, indicating *attribute contamination*. Regarding hyperparameters, we focus on cases where probing is successful by extracting the top five (l, k) pairs with high R^2 in layer-independent analysis. For layer-wise observations, the top three k with high R^2 is extracted. The mean of the metric on the selected PLS models is adopted as the result for robust evaluation.

Models Four transformer-based LLMs: Llama 3.1 8B and Llama 3.1 70B (Grattafiori et al., 2024), and Qwen2.5-3B and Qwen2.5-32B (Yang et al., 2025) are employed to comprehensively examine competitive open models across several sizes.[*]

3.3 Inter-attribute Correlations in Probing

Correlation Heatmaps Figure 3 shows the matrices of $r_s(\hat{Y}_t, Y_t)$ for two representative models, Llama 3.1 8B and Qwen2.5-32B, in the *in-question noun* setting. The heatmaps for the remaining two models are in Appendix B to ensure consistency across the models. Diagonal entries (within-attribute) remain high, confirming that numerical features are recoverable to a certain degree. Off-

[*] All experiments in this paper, including LLM inference, were conducted on an NVIDIA GH200 system (Grace Hopper Superchip), which integrates a 72-core Neoverse V2 CPU and a 120 GB Hopper GPU.

diagonal entries (side effects) are substantial for human entities, reflecting strong year entanglement, and moderate for geographical entities, partially reflecting their natural attribute correlations. Some off-diagonals *unnaturally* match or exceed the natural correlations or the diagonal values, despite incomplete reconstruction of source attributes in the diagonal components.

Discussion on RQ1 Regarding the *preservation* of the correlated structures, the positive/negative correlations in the *inter-attribute evaluation set* are roughly reflected in the correlations through probing. However, some entries exhibit *implausibly* large absolute values, suggesting that the extracted subspaces are overly shared across attributes. For instance, as shown in Figure 3, the PLS model trained on `work period start` predicts `birth year` more accurately than the PLS model trained on `birth year` itself. This implies that the model does not clearly separate these two attributes in the representation space, indicating substantial subspace overlap. Furthermore, certain choices of k yield even larger non-diagonal components, as illustrated in Figure 9 of Appendix C. This suggests that our linear model is not specialized for the source attribute. Although it appears to fit the labels, the extracted subspace retains confounding factors from the knowledge stored in the LLM. In this way, conventional probing methods that focus on a single attribute entangle other attributes and lead to side effects of the intervention (Heinzerling and Inui, 2024). Variations in the correlations between human attributes and geographical attributes across models indicate that training data or architecture influence the degree of subspace overlap.

Impact of Prompt Specificity Table 1 compares *in-question noun* versus *isolated noun* prompts. While the variance of inter-attribute correlations is relatively large, the within-attribute correlations show a clear and consistent increase in the *in-question noun* setting. This indicates that even small contextual information in the prompt can substantially influence the probing results. In contrast, inter-attribute correlations exhibit higher variability, suggesting that the effect is less systematic across unrelated attributes. Overall, these findings confirm that prompt specificity primarily sharpens the probe's focus within each attribute, while minimal contextual cues can still produce measurable side effects.

3.4 Layer-wise Fidelity and Contamination

The partial correlation analysis is performed to investigate the confounding effects in detail at each layer within LLMs, focusing on highly correlated attribute pairs in Figure 3 (`birth year` vs. `work period start` in human entities and `population` vs. `area` in geographical entities).

Observation For the first pair with Llama 3.1 8B (left column of Figure 4), apparent correlation $r_s(\hat{Y}_t, Y_t)$ rapidly rises by layer 10 and plateaus. When predicting `work period start` from `birth year`, attribute contamination remains low (< 0.2) while attribute fidelity peaks at a higher level (> 0.3) than the contamination. Conversely, predicting `birth year` from `work period start` yields high contamination (> 0.4) and lower fidelity (< 0.2). Furthermore, similar patterns can be observed regardless of the model. The middle column of Figure 4 shows Qwen2.5-32B, and the same can be found with less contamination than Llama 3.1 8B. For the second pair with Llama 3.1 8B (right column of Figure 4), similar trends can be observed. The `population` is highly source-faithful (> 0.7), and contamination remains at low levels (< 0.3). On the other hand, there is a certain amount of contamination (> 0.4) in the subspace of `area`, with relatively low fidelity (< 0.6), even though the apparent original correlation is about the same. The results of the remaining models and attribute pairs are shown in Appendix D.

Discussion on RQ1 The asymmetry in human entities indicates that `birth year` information is more strongly encoded, whereas `work period start` is entangled with `birth year`. Also, these results suggest that geographical entities share an overlapping subspace, with `population` dominating `area`. Linked with a prior finding suggesting that LLMs tend to remember popular knowledge in training data with more parametric weights (Mallen et al., 2023), the expression of minor attributes is mixed with the memory of related attributes, leading to *confounding effects*.

4 RQ2: Confounding Effects by Prompt

In this section, experiments employing few-shot prompting (Brown et al., 2020) were performed in order to introduce an additional controllable attribute into queries concerning a specific entity attribute. Motivated by previous studies showing that poorly organized few-shot examples can dis-

Prompt Setting	Model	Within-attribute	Inter-attribute
In-question noun	Llama 3.1 8B	0.818 ± 0.088	0.251 ± 0.214
(e.g. "`What is the area of Texas?`")	Qwen2.5-32B	0.766 ± 0.110	0.207 ± 0.205
Isolated noun	Llama 3.1 8B	0.736 ± 0.133	0.206 ± 0.204
(e.g. "`Texas`")	Qwen2.5-32B	0.728 ± 0.108	0.190 ± 0.208

Table 1: Absolute correlation strength (mean $\pm$ standard deviation) for within-attribute and inter-attribute cases, corresponding to diagonal and off-diagonal elements in the correlation matrices of Figure 3. Attribute-specific prompts in the *in-question noun* setting raise the value of the within-attribute correlations by approximately 0.04 to 0.08, and by 0.02 to 0.05 for inter-attribute ones.

tort LLM behavior (Zhao et al., 2021; Ma et al., 2023), we hypothesized that superfluous contextual examples could systematically skew numerical predictions toward the scale of the provided values. The prevalence of this effect was evaluated by varying the number of few-shot examples k. To control for the influence of the original response, partial correlation analysis was applied to quantify how strongly errors co-occur under prompt manipulation. Finally, the characteristics of the observed output deviations were correlated with internal representation metrics, providing a novel interpretation of prompt-induced susceptibility.

4.1 Behavioral Experiments

Method A few-shot prompt is constructed, augmented with k examples of question–answer pairs. The k examples are selected without replacement, unsorted, and have no overlap in entity name with each other or with the target question. Each prompt presents two numeric attributes (the mean of the example answers $\bar{A}_{\text{ref}}$ and the target answers A) within a single prompt context. For each prompt, the model's numeric response is recorded, and the partial Spearman correlation $r(\text{LLM Output}, \bar{A}_{\text{ref}}|A)$ is calculated.

Dataset and Models A subset of the single-attribute dataset with the same question templates provided in Section 3 was used: three attributes in human entities (`birth year`, `death year`, and `work period start`) and three attributes in geographical entities (`area`, `elevation`, and `population`). 1000 questions per attribute were input into the LLM with k examples, and the answers were parsed into numerical values.* The same four LLMs as in Section 3 were employed.

*Responses that failed to parse numerical values from string answers ($< 20\%$) were discarded.

Results and Analysis Figure 5 presents the correlations $r(\text{LLM Output}, \bar{A}_{\text{ref}}|A)$ varying the number of examples and models. A monotonic increase in confounding is observed as the number of examples increases across all models. Furthermore, smaller models (e.g. Llama 3.1 8B, Qwen2.5-3B) exhibit consistently higher correlation values compared to their larger counterparts, indicating greater susceptibility to example bias. These findings suggest that few-shot exposure can systematically skew numerical predictions and that model capacity inversely moderates this effect.

4.2 Linking to Internal Representations

In this experiment, we integrate the behavioral experiment in Subsection 4.1 and the analysis of internal representations using PLS probing in Section 3 to address **RQ2**.

Method For each attribute, the PLS model is fitted by using the training split, which was not included in the previous behavioral experiments. We input the hidden state of the token corresponding to the final question mark (i.e., "?") in each few-shot prompt into this PLS model. The predicted value represents the numerical quantity currently represented by the LLM, as indicated by the numerical-attribute subspace. We collect these and denote them I. Few-shot prompting with $k = 8$ is used to analyze examples where the LLM output is distorted, and $r(\bar{A}_{\text{ref}}, I|A)$ and $r(I, \text{LLM Output}|A)$ are calculated for evaluation.

Results and Analysis Figure 6 shows layer-wise trends for the Llama and Qwen families. The mean of the references distorts the internal state; this distortion is observable in the low-dimensional subspaces extracted by PLS (also see Figure 11 in Appendix E). In Llama, $r(\bar{A}_{\text{ref}}, I)$ increases in early layers, peaks around the middle 20 layers, and then

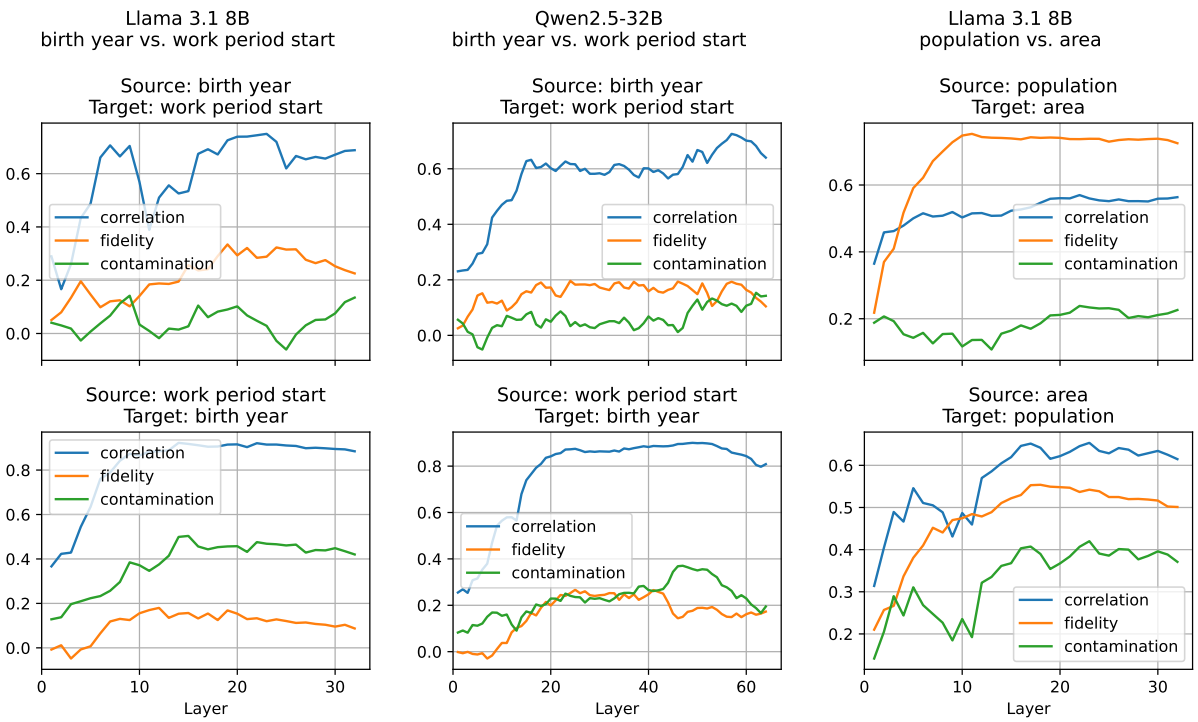

Figure 4: Layer-wise apparent Spearman correlation $r_s(\hat{Y}_t, Y_t)$ (blue), attribute fidelity $r_s(\hat{Y}_s, Y_s \mid Y_t)$ (orange), and attribute contamination $r_s(\hat{Y}_t, Y_t \mid Y_s)$ (green) for Llama 3.1 8B and Qwen2.5-32B, shown for (`birth year`, `work period start`) and (`area`, `population`) pairs. For each column, the upper shows high source-attribute fidelity and low contamination by the target attribute, while the lower side is the opposite with relatively high contamination and low fidelity. Each tick on the horizontal axis corresponds to a Transformer layer from which both the source and target attribute representations were extracted.

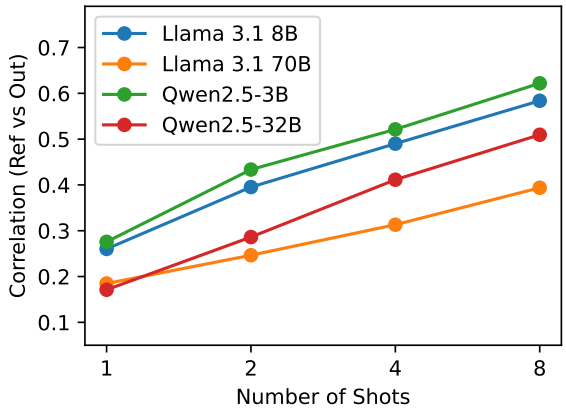

Figure 5: Correlations between the model outputs and the reference means in prompts. All models show higher correlations as the number of few-shot examples increases, with smaller models being more susceptible.

declines. This aligns with Hendel et al. (2023), suggesting that Llama acquires the essential internal representations through in-context learning in early layers, then uses them for inference. In contrast, Qwen shows an initial drop in $r(\bar{A}_{\text{ref}}, I)$, followed by a slower rise and a late-stage secondary peak. This suggests that context processing differs by model architecture or training.

Discussion on RQ2 The correlations between the internal states and the actual outputs are also observed, though weaker than those with the inputs. Notably, this gap widens with model size. These findings, together with Subsection 4.1, indicate that larger models retrieve memorized information robustly against contextual perturbations, leveraging nonlinear or high-dimensional mechanisms beyond linear subspace analysis. In summary, numerical attributes referenced in context are transiently modulated in a shared internal subspace, and model robustness differs with respect to whether such perturbations propagate to the output.

5 Related Work

Probing Methods for LLMs While binary classifiers are employed to probe language models (Belinkov, 2022) to identify specific noun concepts (Burns et al., 2023; Zhao et al., 2025) embedded in internal representations, probing with regression models suggests that numerical concepts such as spatial and temporal information are linearly encoded (Liétard et al., 2021; Gurnee and Tegmark, 2024; Heinzerling and Inui, 2024). Other analyses

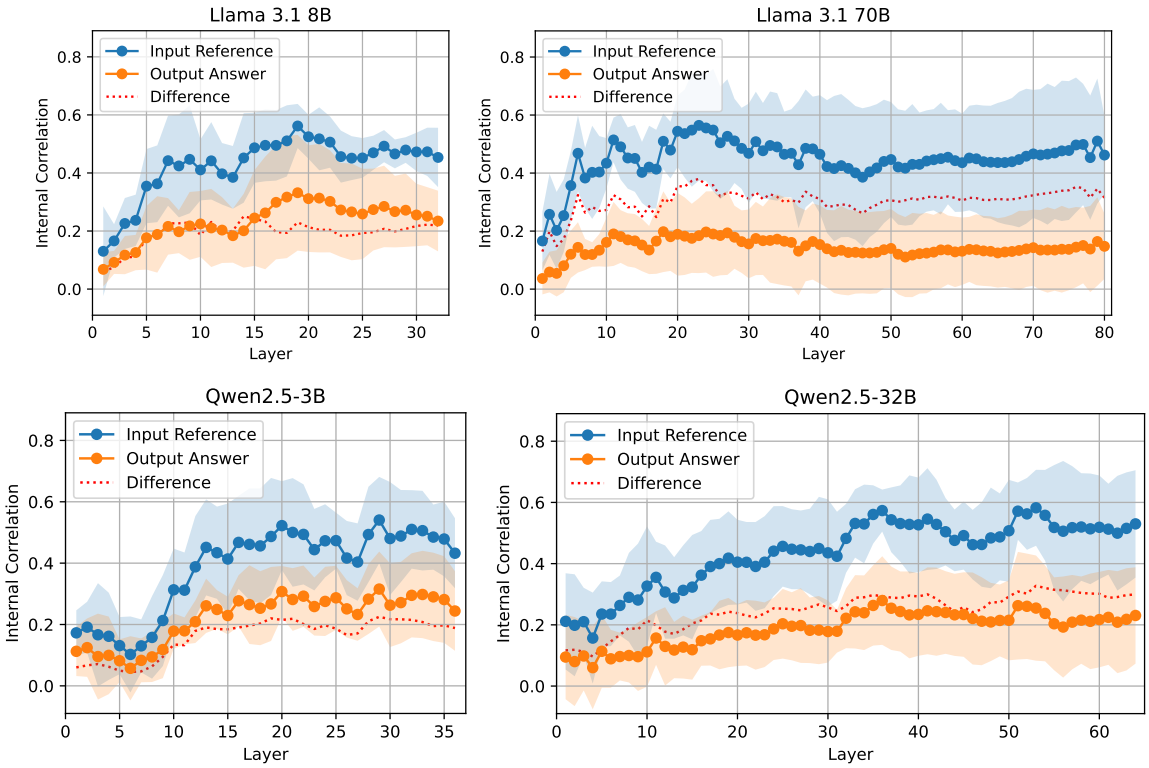

Figure 6: Layer-wise Spearman correlations of PLS predictions with (blue) example's answer mean and (orange) few-shot model output. The shaded area shows the standard deviation for 18 samples, using the top three PLS rank from Subsection 3.4 for six attributes. The (red) line shows the difference between the two. The input references correlate well with the intermediate layers, and smaller models show a smaller the gap between these values and their correlations with the output, indicating easier context propagation.

report circular representations for modular arithmetic (Engels et al., 2025), and helical embeddings for periodic numerical attributes (Kantamneni and Tegmark, 2025; Lei and Cooper, 2025). Through these analyses, we can assume that numerical attributes are linearly represented in terms of magnitude, allowing linear models for probing. While prior work typically probed isolated concepts individually, our study makes a unique contribution by examining relationships across multiple correlated numerical attributes and offering insights through the comprehensive layer-wise analysis.

Confounding in LLM Representations While several works have shown that LLMs' internal representations entangle multiple concepts (Wu et al., 2025; Zhao et al., 2025), the direct impact of this entanglement on downstream interventions remains unexplored. In the related domain of knowledge editing, similar unintended side effects have been documented when modifying stored knowledge (Li et al., 2024b; Duan et al., 2025). To mitigate these effects and improve response accuracy for inserted target knowledge, methods such as conflict-aware edits (Jung et al., 2025) and orthogonal editing directions (Fang et al., 2025) have been proposed. In contrast, our study quantifies the confounding influence of mutually dependent numerical attributes via a novel partial-correlation setup for interpretability. We also highlight the limitations of conventional linear probing when assessing these intertwined subspaces.

Prompt Vulnerabilities Prompt engineering (Sahoo et al., 2025) has become central to controlling LLM behavior, with methods such as few-shot prompting improving accuracy and format consistency (Brown et al., 2020). However, prompts remain susceptible to manipulation and misunderstanding—ranging from conflicts with a model's internal knowledge (Xie et al., 2024) to prompt-injection exploits (Liu et al., 2024) and jailbreak attacks (Shen et al., 2024). Although several studies document that numerical contexts can induce degraded or erratic LLM outputs (Shi et al., 2023; Zhao et al., 2024), the internal mechanisms governing these failures under realistic and complex prompts have not been elucidated. In this work,

we close that gap by investigating how numerically confounding prompts influence internal representations across different model scales.

6 Conclusions and Future Work

In this work, we systematically applied linear probing models to dissect how large language models encode multiple numerical attributes, integrating confounding analysis and prompt-based vulnerability testing within numerical contexts. Our probing results reveal that LLMs not only capture real-world numerical correlations but tend to exaggerate these relationships across internal representations. Moreover, we demonstrate that irrelevant numerical information embedded in prompts can induce significant representation drift, with pronounced variability across model scales.

These findings elucidate concrete risks for LLM-powered data-driven decision support in numerical contexts and suggest avenues for extending the analysis to biases at the intersection of multiple social factors in language models (Lalor et al., 2022; Li et al., 2024a). Our investigation can also link to representation–based hallucination detection and fidelity enhancement (Li et al., 2023; Du et al., 2024; Sriramanan et al., 2024). Integrating these techniques with our approach could enable the targeted mitigation of biases from interacting attributes in prompts before output generation, thereby advancing fairness and reliability in complex contexts.

Limitations

We hypothesized that numerical magnitude is encoded in shared linear subspaces across attributes. However, our experiments did not identify a single subspace that uniformly represents arbitrary numerical features. While we extended standard probing with partial-correlation analysis under a single-confounder assumption, this method does not scale to multiple concurrent confounders. Future work should employ multivariate causal inference or factor-analytic techniques to disentangle complex attribute relationships.

We used few-shot prompts to introduce irrelevant numerical attributes into the context but focused only on fact-retrieval tasks. We did not explore the dynamics of arithmetic reasoning under in-context learning. A broader study—including generative and multi-step arithmetic reasoning scenarios—is needed to fully understand how LLMs manipulate numerical information in their internal representations for realistic prompts.

Acknowledgement

We thank the three anonymous reviewers for their helpful comments and feedback. This work was partially supported by JST CREST Grant Number JPMJCR2565, Japan. In this paper, LLMs were used to polish writing and coding.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *Advances in Neural Information Processing Systems*, volume 37, pages 102948–102972. Curran Associates, Inc.

Zenghao Duan, Wenbin Duan, Zhiyi Yin, Yinghan Shen, Shaoling Jing, Jie Zhang, Huawei Shen, and Xueqi Cheng. 2025. Related knowledge perturbation matters: Rethinking multiple pieces of knowledge editing in same-subject. 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 363–373, Albuquerque, New Mexico. Association for Computational Linguistics.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. In *ACL (Short Papers)*, pages 85–91.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622.

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, Bangkok, Thailand. Association for Computational Linguistics.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Dahyun Jung, Jaehyung Seo, Jaewook Lee, Chanjun Park, and Heuiseok Lim. 2025. CoME: An unlearning-based approach to conflict-free model editing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6410–6422, Albuquerque, New Mexico. Association for Computational Linguistics.

Subhash Kantamneni and Max Tegmark. 2025. Language models use trigonometry to do addition. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Seongho Kim. 2015. Ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods*, 22(6):665–674.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Ge Lei and Samuel J. Cooper. 2025. The representation and recall of interwoven structured knowledge in llms: A geometric and layered analysis. *Preprint*, arXiv:2502.10871.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Victoria R Li, Yida Chen, and Naomi Saphra. 2024a. ChatGPT doesn't trust chargers fans: Guardrail sensitivity in context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6327–6345, Miami, Florida, USA. Association for Computational Linguistics.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2024b. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations*.

Bastien Liétard, Mostafa Abdou, and Anders Søgaard. 2021. Do language models know the way to Rome? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 510–517, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847, Philadelphia, PA. USENIX Association.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 43136–43155. Curran Associates, Inc.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 1671–1685, New York, NY, USA. Association for Computing Machinery.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda,

and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. 2024. Evaluating llms' mathematical reasoning in financial document question answering. In *ACL (Findings)*, pages 3853–3878.

Raphael Vallat. 2018. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Svante Wold, Michael Sjöström, and Lennart Eriksson. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130. PLS Methods.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In *The Twelfth International Conference on Learning Representations*.

Zikai Xie. 2024. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting for large-language-models. *Preprint*, arXiv:2408.05093.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. 2025. Beyond single concept vector: Modeling concept subspace in LLMs with gaussian distribution. In *The Thirteenth International Conference on Learning Representations*.

Siyan Zhao, Tung Nguyen, and Aditya Grover. 2024. Probing the decision boundaries of in-context learning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 130408–130432. Curran Associates, Inc.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Attribute Pair	Correlation	n
(birth year, death year)	0.964***	12578
(birth year, work period start)	0.980***	8094
(death year, work period start)	0.849***	3936
(area, elevation)	0.127***	5643
(area, population)	0.574***	9722
(area, latitude)	-0.263***	10196
(area, longitude)	0.039***	10196
(elevation, population)	-0.093***	5587
(elevation, latitude)	-0.095***	5995
(elevation, longitude)	0.009	5995
(population, latitude)	-0.286***	9997
(population, longitude)	0.230***	9997
(latitude, longitude)	0.053***	14735

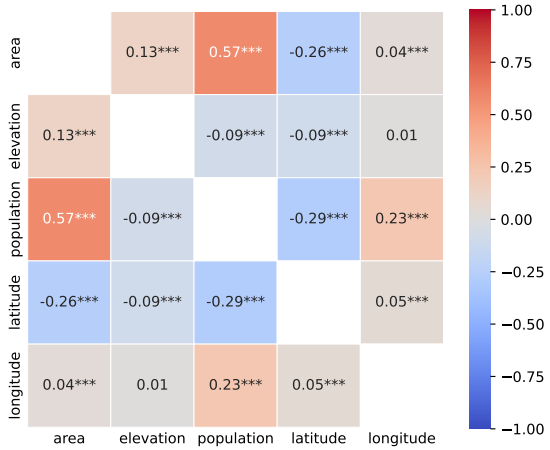

Figure 7: (Left) Spearman correlation and sample size n. Significance: $^{***}p < 0.001$. Note that significance is easily achieved due to large sample size. (Right) Correlation matrix between attributes of geographical entities.

A Detailed Configuration of Experiments

Natural Correlations in Larger Data By collecting entities with each attribute pair in Wikidata (Vrandečić and Krötzsch, 2014), the correlation between two attributes with a larger sample size is evaluated, resulting in Figure 7. For visibility, the heatmap of the correlation matrix is also displayed in Figure 7 for geographical entities with the five attributes. These correlations align with the trend observed in the correlation matrices in Figure 2. Based on these observations, similar correlation patterns may also be present in the large corpora on which LLMs are trained.

Question Templates For the probing in the *In-question noun* setting, the prompt for extracting a specific numerical attribute of an entity is as follows, where the input string to the LLM is a question containing the entity name `{Noun}`.

`birth year` In what year was `{Noun}` born?
`death year` In what year did `{Noun}` die?
`work period start` In what year did `{Noun}` start working?
`area` What is the area of `{Noun}`?
`elevation` How high is `{Noun}`?
`population` What is the population of `{Noun}`?
`latitude` What is the latitude of `{Noun}`?
`longitude` What is the longitude of `{Noun}`?

Prompt Example The question sentence in the original noun setting and the prompt to the LLM in the few-shot prompting setting are as follows. In this instance, the number of shots is set to $k = 4$. As illustrated in Brown et al. (2020), the Q and A are explicitly shown and connected by line breaks.

Original Question:

```
"What is the area of Sapporo?"
```

Question with Four Irrelevant Examples:

```
"Q: What is the area of Anaheim?
A: 131

Q: What is the area of Saanen?
A: 120

Q: What is the area of Yazd?
A: 131

Q: What is the area of Gdynia?
A: 135

Q: What is the area of Sapporo?
A: "
```

This constitutes the few-shot prompting format adopted throughout our experiments. To evaluate the robustness of our setup in Subsection 4.1, we conducted a supplementary evaluation by testing several combinations formed by combining three factors: (i) prompt layout (separated Q–A pairs with line breaks vs. compact sequences without them), (ii) answer value order (randomized vs. ascending) and (iii) answer value diversity (narrow vs. wide range). Across all six numerical attributes, we observed no systematic differences in the trend of LLM outputs for the first two factors. Accordingly, we adopt the line break-separated, randomly ordered format as the default in this work.

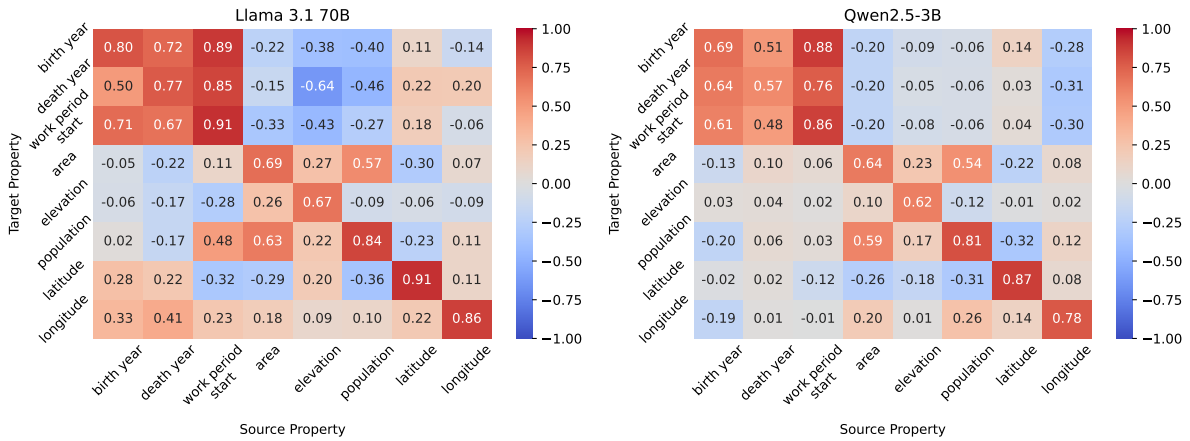

Figure 8: Spearman correlations for Llama 3.1 70B (top) and Qwen2.5-3B (bottom): diagonal is within- attribute, off-diagonal is inter-attribute.

Prompt Setting	Model	Within-attribute	Inter-attribute
In-question noun (e.g. "`What is the area of Texas?`")	Llama 3.1 70B	0.807 ± 0.092	0.291 ± 0.210
	Qwen2.5-3B	0.731 ± 0.115	0.196 ± 0.204
Isolated noun (e.g. "`Texas`")	Llama 3.1 70B	0.736 ± 0.124	0.196 ± 0.205
	Qwen2.5-3B	0.674 ± 0.131	0.178 ± 0.184

Table 2: Absolute correlation strength (mean $\pm$ standard deviation) for within-attribute and inter-attribute cases, corresponding to diagonal and off-diagonal elements in the correlation matrices of Figure 8. While attribute-specific prompts improve the within-attribute correlation, Llama shows larger side effects on the inter-attribute correlation.

For the third factor, increasing the diversity of example answer values within a single context raises within-context variety. However, when comparing behaviors across multiple contexts to assess correlations, it reduces the between-context variance of the mean answer values. When compared at k = 8, the correlation slightly decreased, but the same qualitative tendency was observed. Although the composition quality of few-shot examples can affect model behavior in certain cases (Yao et al., 2024), the overall results reported in this paper remain robust for single-answer numerical attribute tasks.

B Correlation Matrices for Other Models

In addition to the results in Figure 3 and Table 1, the experiments in Subsection 3.3 were conducted for Llama 3.1 70B and Qwen2.5-3B. The correlation matrices and their summaries are displayed in Figure 8 and Table 2. These results are indicative of the natural statistical dependencies present within the dataset, especially among attributes belonging to the same entities. However, correlations between attributes across human and geographical entities reflect model-dependent behavior. This occasionally manifests as moderate inter-attribute values in the Llama models.

C Potential of Misalignment in Linear Probing

Instead of selecting hyperparameters (k, l) to best fit the source attributes, we computed, for each source–target attribute pair, the average of the top five correlations with the highest absolute values. These aggregated scores are visualized as heatmaps in Figure 9. These results demonstrate that linear probing can achieve high correlation scores even with simple models, highlighting its utility in assessing representational structure. However, they also reveal that such scores may arise from a misalignment between the probed attribute and the internal representations. Specifically, we observe cases where a target attribute can be predicted from representations primarily encoding a different source attribute.

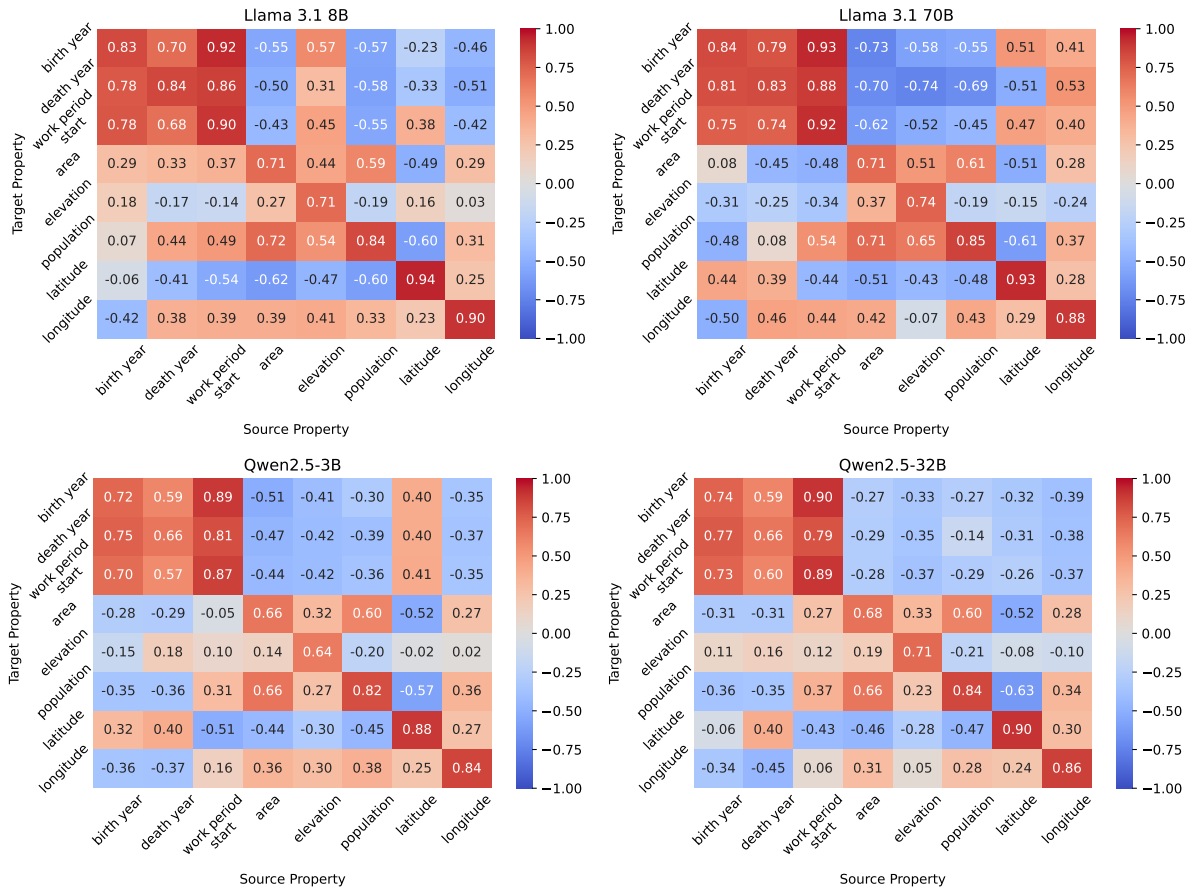

Figure 9: Maximized correlations of probing model predicting target attributes from source attributes. In Llama models, inter-attribute correlations grow with model size, suggesting entangled internal representations, whereas in Qwen, larger models exhibit increased emphasis only on diagonal components, indicating more disentangled attribute representations.

D Layer-wise Confounding Analysis

The layer-wise correlation trends for models that were not included in Figure 4 of Subsection 3.4 are shown in Figure 10. For the `birth year` attribute, PLS models show consistently higher fidelity and lower contamination when `work start period` is treated as a confounding factor. Interestingly, this relationship often reverses when the source and target attributes are swapped. Predicting `work start period` from `birth year` yields substantially higher fidelity across layers, whereas predicting `birth year` from `work start period` results in weaker fidelity and stronger contamination. In the case of geographical entities, models fitted to `area` are more susceptible to confounding from `population`, resulting in a smaller gap between fidelity and contamination. These asymmetric patterns suggest that the strength and direction of attribute-specific encoding may reflect how prevalent or salient the corresponding knowledge is in the training corpus.

E Detailed Analysis on Prompt-induced Perturbations of Attribute Subspaces

We further investigate the dimensionality required to capture the effects of contextual attribute perturbations within the low-dimensional subspaces extracted by PLS, as introduced in Section 4.2. While Figure 6 reports partial correlations using the top-3 ranks selected based on their performance in single-attribute probing (see Section 3.2), Figure 11 explores how the required dimensionality changes when capturing context-induced effects. Specifically, the odd-numbered rows show results using the top-3 ranks from the single-attribute setup. The even-numbered rows correspond to the top-3 ranks per layer that maximize partial correlation (i.e., $r(\bar{A}_{\text{ref}}, I|A)$ or $r(I, \text{LLM Output}|A)$). These results indicate that magnitude-related contextual effects are captured in compact, low-dimensional subspaces, suggesting that prompt-level numerical interference is encoded along low-rank directions.

Figure 10: Layer-wise apparent Spearman correlation $r_s(\hat{Y}_t, Y_t)$ (blue), attribute fidelity $r_s(\hat{Y}_s, Y_s|Y_t)$ (orange), and attribute contamination $r_s(\hat{Y}_t, Y_t|Y_s)$ (green) for Qwen2.5-3B and Llama 3.1 70B, shown for `birth year`/`work period start` and `area`/`population` pairs.

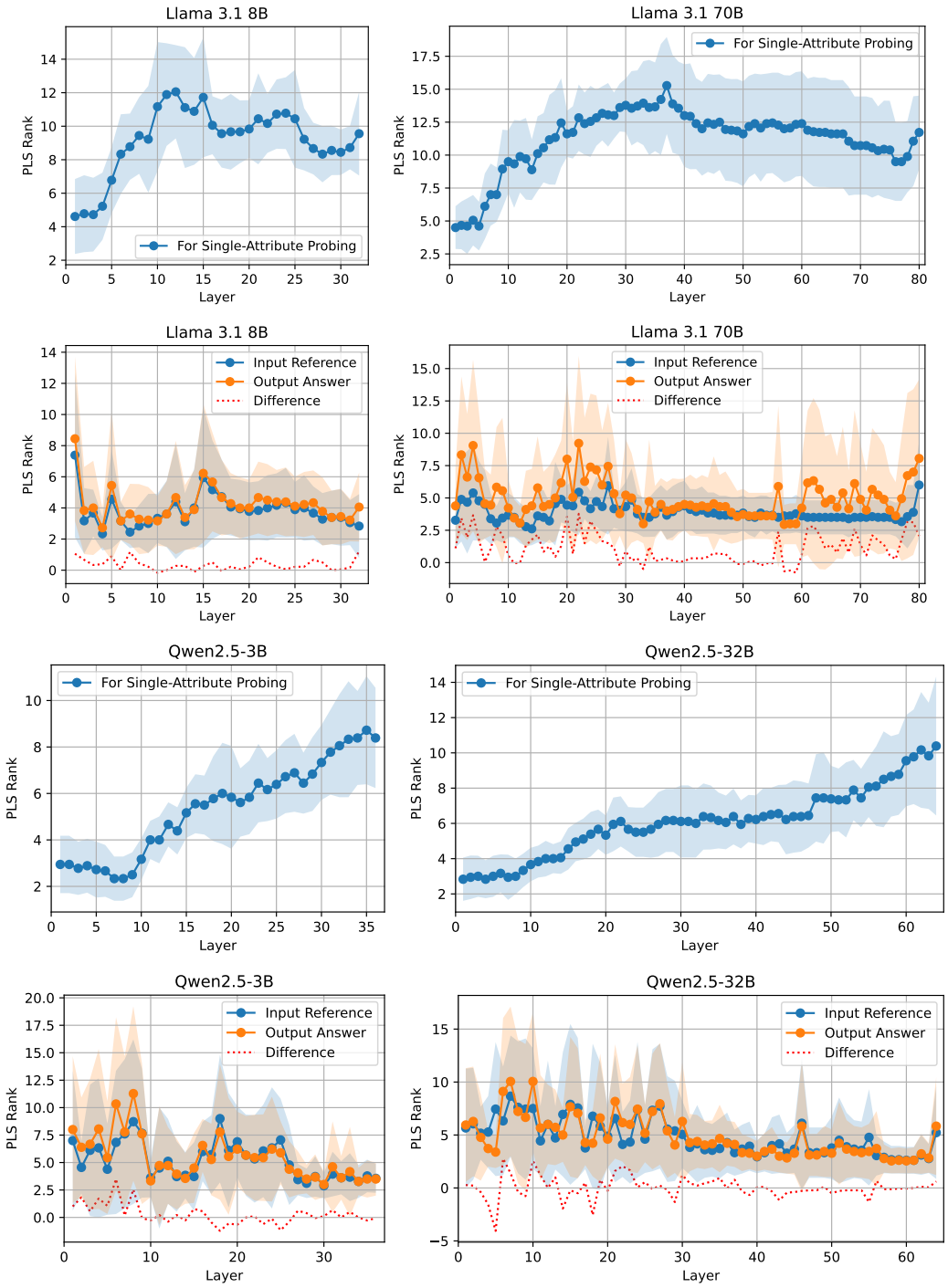

Figure 11: Optimal layer-wise PLS ranks across models. For each of six attributes, the top three ranks (18 samples total) are averaged and their standard deviations plotted. Odd-numbered rows display ranks yielding the highest R^2 in conventional single-attribute PLS probing; even-numbered rows display ranks with the largest context effects measured by $r(\bar{A}_{\text{ref}}, I|A)$ (Input Reference) and $r(I, \text{LLM Output}|A)$ (Output Answer), and the difference between the latter and the former is shown. These results show that the subspace dimensions capturing $r(\bar{A}_{\text{ref}}, I|A)$ and $r(I, \text{LLM Output}|A)$ differ only slightly, but they are much smaller than in single-attribute settings. The effectiveness of PLS in extracting low-dimensional representations is also highlighted.

F Broader Impacts

Our findings reveal significant risks of numerical attribute confounding in LLMs. Amplified numerical correlations and sensitivity to irrelevant numerical cues can lead to erroneous or biased outputs, posing serious concerns in high-stakes domains such as finance, healthcare, and policy. Smaller models, in particular, are more susceptible to such context-driven distortions and should be evaluated with caution before use in sensitive applications. At the same time, understanding these vulnerabilities opens avenues for mitigation through better prompt design, robust training, and improved interpretability. These strategies can enhance the reliability and fairness of LLMs in numerically intensive settings.