EMERGENT STRUCTURED REPRESENTATIONS SUPPORT FLEXIBLE IN-CONTEXT INFERENCE IN LARGE LANGUAGE MODELS

Ningyu Xu, Qi Zhang, Xipeng Qiu,* & Xuanjing Huang*
College of Computer Science and Artificial Intelligence
Fudan University
Shanghai, China
nyxu22@m.fudan.edu.cn, {xpqiu,xjhuang}@fudan.edu.cn

ABSTRACT

Large language models (LLMs) exhibit emergent behaviors suggestive of human-like reasoning. While recent work has identified structured conceptual representations within these models, it remains unclear whether they functionally rely on such representations for reasoning. Here we investigate the internal processing of LLMs during in-context inference across diverse tasks. Our results reveal a conceptual subspace emerging in middle to late layers, whose representational structure persists across contexts. Using causal mediation analyses, we demonstrate that this subspace is not merely an epiphenomenon but is functionally central to model predictions, establishing its causal role in inference. We further identify a layer-wise progression where attention heads in early-to-middle layers integrate contextual cues to construct and refine the subspace, which is subsequently leveraged by later layers to generate predictions. Together, these findings provide evidence that LLMs dynamically construct and use structured latent representations in context for inference, offering insights into the computational processes underlying flexible adaptation.

1 INTRODUCTION

Large language models (LLMs) have developed an emergent capacity for in-context learning, which allows them to adapt to novel tasks specified in context through demonstrations or instructions (Brown et al., 2020; Lampinen et al., 2024b). This process, often framed as a form of inference, requires no parameter updates (Xie et al., 2022; Hahn & Goyal, 2023; Lake & Baroni, 2023) and arises naturally from the objective of next-token prediction. The in-context learning abilities have sparked immense interest as they showcase traits of human-like learning and generalization (Griffiths et al., 2025; Russin et al., 2025), enabling impressive, and occasionally human-level, performance across language understanding and reasoning tasks (Wei et al., 2022; Bubeck et al., 2023; Binz & Schulz, 2023; Webb et al., 2023; 2025; Lampinen et al., 2024a; Hagendorff et al., 2023). However, these successes are tempered by brittleness to contextual variations and non-human-like errors (Lewis & Mitchell, 2024; Dentella et al., 2024; Wu et al., 2024). The divergent evidence has fueled ongoing debates over whether LLMs are genuinely approaching structured, human-like inference or merely relying on statistical correlations in their training data (Mitchell, 2025; 2021; Mitchell & Krakauer, 2023; McCoy et al., 2024; Mahowald et al., 2024). Demystifying the internal processes is essential not only for understanding the strengths and limitations of LLM generalization but also for uncovering the computational mechanisms underlying flexible adaptation to changing contexts—a hallmark of natural intelligence (Lake et al., 2017).

Human inference relies on structured knowledge abstracted from experience and adapted to novel situations (Tenenbaum et al., 2011; McClelland et al., 2010; Lake et al., 2017; Johnson-Laird, 2010), a capacity requiring conceptual representations that support symbolic processes while accommodating

*Corresponding authors. Code for reproducing all experiments is available at https://github.com/ningyuxu/llm_structured_icl.

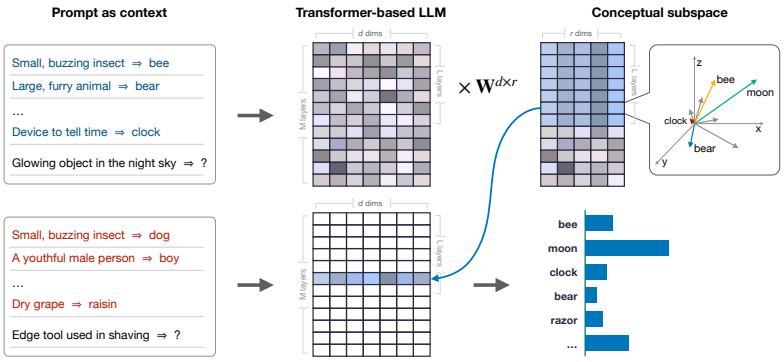

Figure 1: Illustration of how an emergent conceptual subspace supports in-context inference in LLMs. A Transformer-based LLM is presented with a small set of description–word demonstrations followed by a query description. As the model processes the prompt across layers, it integrates contextual cues to construct a shared conceptual subspace that emerges in the middle-to-late layers. Hidden states can be projected into this subspace, where the relational structure among representations persists across layers and across different demonstration contexts. Causal interventions show that the subspace and its internal relational structure are functionally involved in inference and can constrain subsequent computations leading to the final prediction.

noisy input and graded information (Piantadosi et al., 2024). Recent efforts to probe the internal representations of LLMs have revealed emergent structure that capture core properties of human concepts (Xu et al., 2025; Pavlick, 2023), but their functional role during inference remains poorly understood. In parallel, while research in mechanistic interpretability has found evidence for symbolic processes supporting in-context learning (Todd et al., 2024; Yang et al., 2025; Feucht et al., 2025; Yin & Steinhardt, 2025; Olsson et al., 2022), these findings are primarily based on relatively simple or synthesized tasks and may not generalize across broader contexts (Cho et al., 2025; Bakalova et al., 2025; Opiełka et al., 2025). Consequently, it remains unclear whether LLMs rely on abstract, structured representations for inference and how their internal computations adapt to varying contexts.

To address these questions, we applied causal mediation analysis to investigate the internal representations that LLMs construct and use for inference across a diverse set of tasks. We focused on an in-context concept inference task as an informative case study, in which models are guided to derive concepts from descriptions through a few demonstrations (Figure 1). This task has been shown to evoke human-like conceptual representations in LLMs, with the resulting representational structure predicting model performance across a range of understanding and reasoning tasks (Xu et al., 2025). We therefore used it as a simple but meaningful testbed for probing how structured representations are formed and deployed during reasoning, while also evaluating a broader set of inference tasks drawn from Todd et al. (2024) to assess the generality of our findings.

We first traced when task-relevant information emerges across layers, identifying a conceptual subspace that remains approximately isomorphic from middle to late layers. We then employed targeted causal interventions to explore whether this subspace is causally implicated in the model's inference process. Finally, we examined how this subspace is constructed in context by analyzing how model components—primarily attention heads—write into it. Our experiments reveal an emergent conceptual subspace that encodes structured, task-relevant information and is dynamically recruited during inference. Its internal structure becomes more coherent with additional demonstrations and ultimately persists across contexts. These findings suggest that LLMs do not merely rely on surface-level patterns but construct and deploy abstract, structured representations for in-context inference, offering a mechanistic window into the computational basis of their adaptive behavior.

2 A Shared Conceptual Subspace Emerges during Inference

Building on Xu et al. (2025), we adopted the reverse dictionary task as a controlled probe to investigate the mechanisms underlying LLMs' in-context inference, while extending our analysis to a broader

set of inference tasks from Todd et al. (2024) to assess the generality of our findings. The reverse dictionary task simulates the human capacity to identify a specific concept from a description that may be underspecified or vague (Figure 1). For example, given an expression such as "a glowing object in the night sky" or "Earth's natural satellite," a system must infer the intended concept from context and produce the corresponding term (e.g., "moon"). Previous work has shown that LLMs can flexibly derive concepts from descriptions in relation to in-context demonstrations, and their penultimate-layer representations exhibit a stable structure that not only predicts downstream model performance but also aligns well with human behavioral and brain data (Xu et al., 2025). These properties make this task a well-controlled, informative testbed for investigating whether and how LLMs rely on structured conceptual representations to perform inference in context. However, focusing on the penultimate-layer representations leaves open when and where such structure emerges during internal processing. To address this question, we examined the layer-wise evolution of LLM representations. This analysis revealed a shared conceptual subspace whose internal structure persists from middle to late layers. We then looked into this subspace to understand how representations are organized within it.

2.1 METHODS

Let $\mathcal{M}$ be an autoregressive Transformer-based LLM with L layers and a hidden dimension of d. In the reverse dictionary task, the model is presented with a prompt s consisting of N demonstrations in the format of "[Description] ⇒ [Word]", followed by a query description and the delimiter "⇒". We focused on the computations at the last token position, which is responsible for predicting the target word for the query concept. Let $\mathbf{h}_\ell$ denote the hidden state at layer $\ell \in \{0, \ldots, L\}$, where $\mathbf{h}_0$ corresponds to the initial token embedding. The hidden states evolve through the layers as $\mathbf{h}_\ell = \mathbf{h}_{\ell-1} + \sum_{k=1}^{K} \mathbf{a}_{\ell,k} + \mathbf{m}_\ell$, where $\mathbf{a}_{\ell,k} \in \mathbb{R}^d$ is the output of kth attention head (among K heads) at layer ℓ, projected to the hidden state (Elhage et al., 2021; Todd et al., 2024), and $\mathbf{m}_\ell \in \mathbb{R}^d$ is the output of the multilayer perceptron block at layer ℓ. The final hidden state $\mathbf{h}_L$ is mapped by the decoder head to a probability distribution over the model's vocabulary $\mathcal{V}$: $f_{\mathcal{M}}(s) = \text{decode}(\mathbf{h}_L)$, from which the next token is generated.

To characterize the emergence of the representational structure of concepts, we first identified the primary subspace within each layer and compared their relative geometry. For a dataset of n query concepts, let $\mathbf{X}_\ell \in \mathbb{R}^{n \times d}$ denote the centered matrix of hidden states at layer ℓ. We applied singular value decomposition (SVD) to $\mathbf{X}_\ell$ and retained the top k principal components that accounted for 95% of the total variance. We then measured the overlap between layers ℓ and ℓ' by the mean squared cosine of the principal angles $\{\theta_i\}_{i=1}^{k}$ between their respective subspaces: $\frac{1}{k}\sum_{i=1}^{k} \cos^2\theta_i$. This metric captures the extent to which variance-dominant directions are shared across layers, while remaining invariant to basis rotations within each subspace.

Based on the layer-wise alignment patterns, we then employed generalized canonical correlation analysis (GCCA) (Horst, 1961; Kettenring, 1971) to isolate a shared subspace that persists across layers. Given hidden-state matrices $\{\mathbf{X}_\ell\}_{\ell \in \mathcal{L}}$ from a selected set of layers $\mathcal{L}$, GCCA seeks a common latent representation $\mathbf{G} \in \mathbb{R}^{n \times r}$ that maximizes the alignment across layers. Formally, we solve:

$$\min_{\mathbf{G}, \{\mathbf{W}_\ell\}} \sum_{\ell \in \mathcal{L}} \|\mathbf{G} - \mathbf{X}_\ell \mathbf{W}_\ell\|_F^2, \quad \text{s.t. } \mathbf{G}^\top \mathbf{G} = \mathbf{I}, \tag{1}$$

where $\mathbf{W}_\ell \in \mathbb{R}^{d \times r}$ are layer-specific projection matrices, and r is the dimensionality of the shared subspace. We determined r using a non-parametric permutation test (Appendix C.3) to identify a subspace capturing structural alignment beyond noise. Alternative dimensionalities were also evaluated to confirm that the permutation-based choice was appropriate (Appendix D.1).

Finally, we looked into the internal structure of the identified shared subspace. Beyond GCCA alignment—measured as the average correlation between layer-wise projections onto the shared subspace—we employed two complementary measures to compare subspaces across contexts: (1) subspace overlap, quantified via principal-angle similarity between the projection matrices $\mathbf{W}_\ell$, and (2) representational similarity analysis (RSA), which compares the relational structure within each subspace by correlating their (dis)similarity matrices. Detailed definitions and implementation are provided in Appendix C.4.

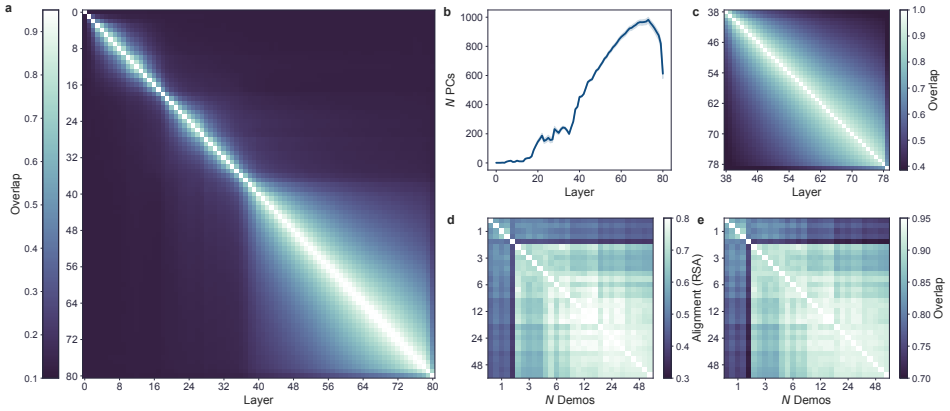

Figure 2: A shared conceptual subspace emerges in the middle to late layers of Llama-3.1 70B. **a**, Layer-wise similarity of hidden states, measured as subspace overlap (mean squared cosine of principal angles) between SVD subspaces explaining 95% variance; averaged over five runs with 24 demonstrations. Axes index layers. **b**, Number of principal components (PCs) needed to explain 95% variance across layers, increasing sharply in the middle layers. Shaded areas represent 95% CIs calculated from 10,000 bootstrap resamples across five runs. **c**, Overlap between GCCA-derived projection matrices across selected layers. **d**–**e**, The conceptual subspace becomes increasingly stable as the number of in-context demonstrations grows. **d**, Cross-context alignment of representational geometry within the GCCA subspace across demonstration sets, measured by RSA. **e**, Overlap between GCCA-derived projection matrices across demonstration sets. Axes in **d**–**e** indicate the number of demonstrations, with each cell representing a single run.

We conducted experiments on a range of open-source decoder-only Transformer LLMs spanning multiple model families and scales, including Llama 3.1, Llama 3 (Grattafiori et al., 2024), and Qwen2.5 (Qwen et al., 2025) (see Appendix C.1 and Table 1 for details). To probe in-context concept inference, we used materials from the THINGS database (Hebart et al., 2019), which comprises 1,854 concrete, nameable object concepts paired with WordNet synset IDs and definitional descriptions. We also included four tasks from Todd et al. (2024): Antonym, Country–Capital, Landmark–Country, and National Parks (Table 2). We focus on concept inference in the main text, with results for the remaining tasks reported in Appendix D.2. For each experimental condition, we randomly sampled N description–word pairs as demonstrations, where $N \in \{1, \ldots, 48\}$. Results were averaged across five independent runs per condition to ensure robustness (Appendix C.2).

2.2 Results

Across all tested models, we observed strong performance on the reverse dictionary task. For Llama-3.1 70B, exact match accuracy increased with the number of demonstrations, rising from 74.56% ($\pm$4.69%) with a single demonstration to 87.95% ($\pm$ 0.32%) with 24 demonstrations, with marginal gains beyond this point. While smaller models achieved lower accuracy uniformly, this monotonic trend remained consistent across models (Appendix Figure 7), replicating prior findings (Xu et al., 2025) and confirming that these models can reliably derive target concepts from linguistic descriptions.

Layer-wise SVD analysis reveals a progressive evolution of models' internal representations across depth (Figure 2a). Early layers (up to approximately layers 36–40) exhibited low cross-layer subspace overlap, indicating rapid representational transformation. In contrast, middle to late layers displayed consistently high inter-layer overlap, suggesting the emergence of a shared subspace that is preserved throughout the remainder of the inference process. This transition coincided with a sharp increase in the number of principal components required to explain 95% of the variance (Figure 2b), suggesting that contextual information is integrated into a higher-dimensional subspace for subsequent computation. Together, these results indicate a qualitative shift in models' internal processing, where most representational structure is established by mid-depth, after which subsequent layers primarily operate within this subspace rather than introducing new dominant directions.

Based on this observed stability, we applied GCCA to isolate a shared subspace spanning the middle to late layers. Both GCCA alignment scores and RSA values approached unity across these layers, indicating strong linear correspondence. At the same time, the layer-specific projection matrices exhibited gradual variation, with higher similarity between adjacent layers (Figure 2c). This pattern suggests that later layers preserve a common representational structure while iteratively refining the subspace to generate predictions.

Finally, we examined how this shared subspace varies across different contexts. As the number of demonstrations increases from 1 to 24, alignment between the internal structure of the subspaces increased steadily, with diminishing improvements beyond this threshold (Figure 2d). A similar trend was observed in the overlap between subspaces extracted from the same layer across contexts (Figure 2e). While this pattern held across models, larger models such as Llama-3.1 70B consistently exhibited higher cross-context alignment (Appendix D.3). These findings suggest that, as LLMs receives more contextual cues, they construct an increasingly coherent and context-invariant relational structure during inference. Such invariance indicates abstraction beyond surface-level details and provides a representational basis for generalization across contexts.

3 The Conceptual Subspace Causally Mediates Inference

We next examined the functional role of the emergent conceptual subspace in model inference through a series of causal interventions. First, we investigated the contextual information mediated by the subspace through causal mediation analysis under targeted input corruptions (Pearl, 2001; Meng et al., 2022; Wang et al., 2023; Geva et al., 2023; Todd et al., 2024). We then assessed the necessity and sufficiency of the subspace for inference through selective ablation and isolation. Finally, we assessed whether the relational structure encoded in the subspace generalizes across contexts, enabling subspace-level information from one context to modulate inference in another.

3.1 Methods

We define an orthogonal projection operator for each layer ℓ as $\mathbf{P}_\ell = \mathbf{W}_\ell \mathbf{W}_\ell^\top$, where $\mathbf{W}_\ell \in \mathbb{R}^{d\times r}$ is the orthonormal basis spanning the GCCA-derived subspace. This operator allows us to decompose any hidden state $\mathbf{h}_\ell \in \mathbb{R}^d$ into the subspace component $\mathbf{h}_{\ell,\parallel} = \mathbf{P}_\ell \mathbf{h}_\ell$ and its orthogonal complement, $\mathbf{h}_{\ell,\perp} = (\mathbf{I} - \mathbf{P})\mathbf{h}_\ell$. We employed this decomposition throughout the intervention analyses. To contextualize the causal effect of the subspace, we compared it with randomly generated orthonormal bases of the same dimensionality.

We performed activation patching to characterize the contextual information mediated by the conceptual subspace. This intervention estimates the causal contribution of subspace-level representations by transplanting activations from a clean context into a corrupted one. Let $s = [(x_1, y_1), \ldots, (x_N, y_N), x_q]$ denote a prompt with N demonstrations, where x_i and y_i ($i = 1, \ldots, N$) are the description and word in the demonstrations, and x_q is the query description, with y_q denoting its expected correct term. We considered three corruption conditions $\tilde{s}$ to probe different information streams: (i) description corruption ($\tilde{x}_i$), (ii) label (word) corruption ($\tilde{y}_i$), and (iii) query corruption ($\tilde{x}_q$) (Figure 6).

During a forward pass on $\tilde{s}$, we replaced only the subspace-aligned component of the corrupted state $\mathbf{h}_\ell^{\text{corr}}$ with that of the clean state $\mathbf{h}_\ell^{\text{clean}}$: $\mathbf{h}_\ell^{\text{patch}} = \mathbf{h}_\ell^{\text{corr}} + \mathbf{P}_\ell(\mathbf{h}_\ell^{\text{clean}} - \mathbf{h}_\ell^{\text{corr}})$. We evaluated the effect of this intervention using the causal indirect effect (CIE; Todd et al., 2024), calculated as

$$\text{CIE}\big(\mathbf{h}_{\ell,\parallel}\big) = \log f_{\mathcal{M}}\Big(\tilde{s} \mid \mathbf{h}_\ell^{\text{patch}}\Big)[y_q] - \log f_{\mathcal{M}}(\tilde{s})[y_q]\,, \tag{2}$$

where $f_{\mathcal{M}}(\cdot)[y_q]$ denotes the output probability for token y_q from model $\mathcal{M}$. A positive CIE indicates that restoring the subspace component recovers task-relevant information disrupted by the corruption. We also computed the normalized CIE to quantify the fraction of performance recovered:

$$\text{NormCIE}\big(\mathbf{h}_{\ell,\parallel}\big) = \frac{\text{CIE}\big(\mathbf{h}_{\ell,\parallel}\big)}{\log f_{\mathcal{M}}(s)[y_q] - \log f_{\mathcal{M}}(\tilde{s})[y_q]}. \tag{3}$$

This allows us to dissociate the types of contextual information routed through the subspace and to localize their effects across layers.

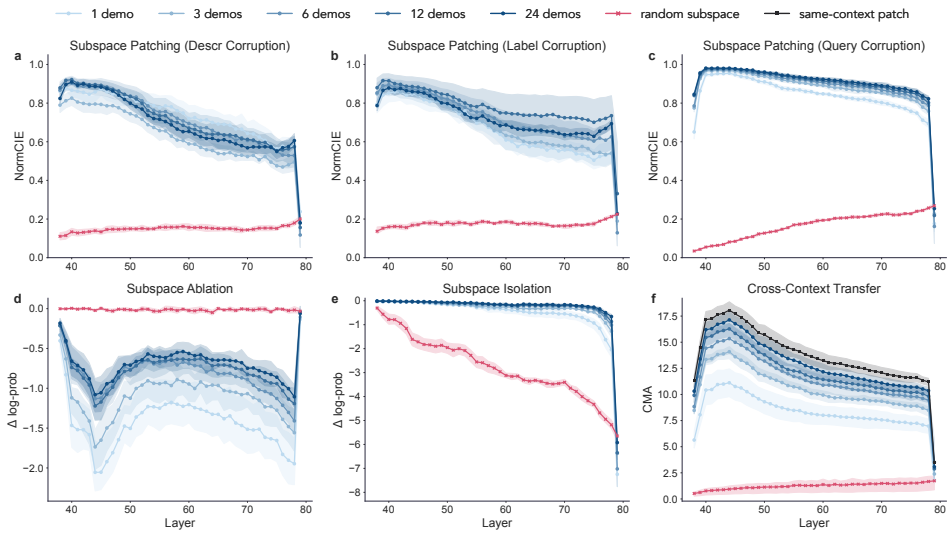

Figure 3: The conceptual subspace causally mediates model inference. **a–c**, Activation patching with N demonstrations under three corruption conditions: description (**a**), label (**b**) and query (**c**). Patching the conceptual subspace (blue) is compared against a random-subspace baseline (red). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d–e**, Subspace necessity and sufficiency tested by ablating the conceptual subspace (**d**) or isolating it (**e**). The y-axis denotes the change in log-probability of the correct token. **f**, Causal effects of cross-context transfer, where the relational structure from a source context is adapted to a target context via an orthogonal transformation. The y-axis reports the causal mediation (CMA) score. Performance is compared against a random-subspace baseline (red) and a same-context patch reference (black) using the conceptual subspace derived from the target context itself. Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

To assess whether the conceptual subspace is necessary for inference, we performed subspace ablation by removing the component of the hidden state lying in the subspace, while leaving the orthogonal complement unchanged: $\mathbf{h}_\ell \leftarrow (\mathbf{I} - \mathbf{P}_\ell)\,\mathbf{h}_\ell$. Similarly, to test for sufficiency, we performed subspace isolation by retaining only the subspace component: $\mathbf{h}_\ell \leftarrow \mathbf{P}_\ell \mathbf{h}_\ell$. The effect of these interventions was measured by the change in the log-probability of the correct term: $\log f_{\mathcal{M}}(s \mid \mathbf{h}'_\ell)[y_q] - \log f_{\mathcal{M}}(s)[y_q]$, where $\mathbf{h}'_\ell$ denotes the modified state. A substantial performance drop after ablation indicates that the subspace is causally necessary for inference, while high performance preservation after isolation suggests that the subspace alone contains sufficient information to support inference.

Finally, to test whether inference relies on a context-invariant relational structure encoded in the conceptual subspace, we performed cross-context subspace transfer. For a given pair of source and target contexts c_i and c_j, each defined by a set of N demonstrations, we learned an orthogonal transformation $\mathbf{Q}_\ell^{(c_i,c_j)} \in \mathbb{R}^{r\times r}$ between their subspaces using a training set of query concepts. We then evaluated cross-context transfer on held-out query concepts. For a shift from concept q_b to q_a, we computed the vector offset in the source context c_i: $\Delta^{c_i}_{\ell,(q_a,q_b)} = \mathbf{W}_\ell^{c_i\top}\left(\mathbf{h}^{c_i}_{\ell,q_a} - \mathbf{h}^{c_i}_{\ell,q_b}\right)$. This offset was then mapped to the target context c_j for query concept q_b: $\mathbf{h}^{c_j}_{\ell,q_b} \leftarrow \mathbf{h}^{c_j}_{\ell,q_b} + \mathbf{W}_\ell^{c_j}\mathbf{Q}_\ell^{(c_i,c_j)}\Delta^{c_i}_{\ell,(q_a,q_b)}$. We quantified the effect of this intervention using the causal mediation score (Wang et al., 2023), which measures the change in the log probability difference between y_{q_a} and y_{q_b} after patching (Appendix C.6). Successful modulation under this transfer indicates that the conceptual subspace supports inference through a functionally aligned relational structure that generalizes across contexts. Besides random orthonormal bases, we additionally compared cross-context transfer against patching using the conceptual subspace derived from the same context, which serves as a within-context reference condition.

3.2 RESULTS

Figure 3 summarizes the effects of targeted causal interventions in Llama-3.1 70B. Activation patching under corrupted demonstrations (Figure 3a–b) shows that model performance can be largely restored by patching the conceptual subspace into middle layers. With 24 in-context demonstrations, the normalized CIE peaked around layer 40 for both description corruption (90.92% $\pm$ 1.28%) and label corruption (87.81% $\pm$ 2.43%), substantially surpassing the random-subspace baseline. Recovery then decreased gradually in later layers, with a sharp drop at layer 79. These results indicate that demonstration-derived contextual information is primarily mediated by the conceptual subspace at mid-depth, where it is integrated for subsequent computation. Patching later layers instead engages computations driven by the corrupted context, which do not fully rely on the conceptual subspace and yield reduced recovery.

Under query corruption (Figure 3c), recovery was even more pronounced and sustained, with normalized CIE peaking at 98.15% ($\pm$0.46%) around layer 42 and remaining robust through layer 78 (82.24% $\pm$ 1.72%). Compared to demonstration corruption, the effect spanned a broader and later window, suggesting that query processing relies on the same conceptual subspace but persists longer in the residual stream. The final-layer decline across all conditions likely reflects a transition from abstract processing to surface-form generation, where the conceptual subspace is less relevant.

Subspace necessity and sufficiency were assessed via ablation and isolation (Figure 3d–e). Ablating the conceptual subspace produced a marked performance drop relative to the random baseline, with the largest effects around layers 44–45 and 77–78 (changes in log-probability exceeding 1), showing that removing this information disrupts inference. However, performance did not collapse entirely, indicating partial redundancy or compensatory pathways. Conversely, isolating the subspace was sufficient to maintain near-original performance across most layers, with degradation confined to the final layers. As the number of demonstrations increases, ablation effects became less pronounced, while isolation continued to preserve performance deeper in the model, suggesting a more robust reliance on the subspace as contextual information accumulates.

Finally, for cross-context transfer, the causal mediation (CMA) score peaked around layer 44 (17.13 $\pm$ 0.90) and remained significant through layer 78 (10.36 $\pm$ 0.38), mirroring the layer-wise profile observed for query corruption. The effect increased with the number of demonstrations, closely approaching the same-context patch reference while substantially exceeding the random-subspace baseline. Together, these results indicate that the model progressively constructs and exploits a more stable, abstract conceptual subspace as contextual cues accumulate, which supports the generalization of relational structure across contexts. Comparable effects were observed across a broader set of inference tasks (Appendix D.2) and were noticeably more pronounced in larger models such as Llama-3.1 70B (Appendix D.4), suggesting that increased scale and compute facilitate the emergence and utilization of such abstract representations for approximating in-context inference.

4 CONTEXTUAL INFORMATION SHAPES THE CONCEPTUAL SUBSPACE

How is the emergent conceptual subspace constructed based on contextual information? To address this question, we focused on attention heads, which are the primary mechanism by which Transformer models integrate contextual information (Vaswani et al., 2017; Elhage et al., 2021). We employed causal mediation analysis to isolate attention heads critical in inference, and then characterized their role by analyzing their attention patterns and output representations, probing how they contribute to constructing and refining the conceptual subspace.

4.1 METHODS

We used activation patching to identify the attention heads that are functionally relevant for inference (Todd et al., 2024; Yang et al., 2025). Following the procedure in Section 3.1, we constructed three corruption conditions $\tilde{s}$ including description, label, and query corruption. Let $\mathbf{a}_{\ell,k} \in \mathbb{R}^d$ represent the output of the kth attention head at layer ℓ, projected into the residual stream. For each corrupted input, we intervened by replacing the corrupted head output $\mathbf{a}_{\ell,k}^{\text{corr}}$ with its counterpart from the clean run $\mathbf{a}_{\ell,k}^{\text{clean}}$. The causal effect of this intervention was then measured via the causal indirect

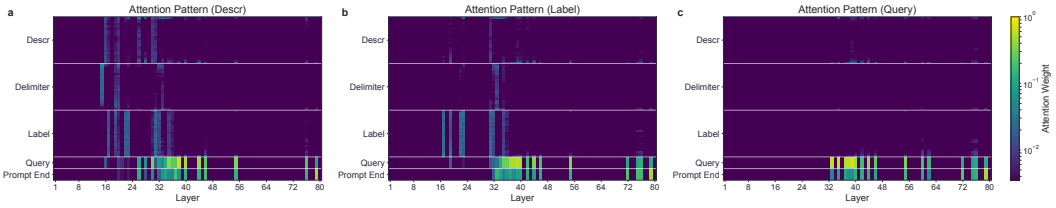

Figure 4: Attention patterns of heads with statistically significant causal indirect effects (CIEs) identified under description (**a**), label (**b**), and query corruption (**c**). Within each layer, attention patterns were averaged over heads with statistically significant CIEs. Layers with no significant heads were assigned zero values. The x-axis indicates layer index, and the y-axis shows attended token spans grouped by source segment. The description, delimiter, and label spans for the 24 in-context demonstrations are ordered top-to-bottom based on their sequence in the prompt. See Figure 21a–c for attention patterns of individual heads.

effect (CIE). We performed a permutation test to control the family-wise error rate and retained only heads with statistically significant effects ($p < 0.05$; Appendix C.6).

We then probed the attention patterns of the identified heads to characterize what contextual information they integrate. For each head, we aggregated attention mass across five token spans: (i) demonstration descriptions, (ii) mapping delimiters (e.g., "⇒"), (iii) demonstration labels, (iv) query tokens, and (v) the final prompt delimiter preceding the prediction. This allows us to localize the specific contextual information the head preferentially attends to during inference.

Finally, to assess the contribution of the identified heads to the shared conceptual subspace, we analyzed both the magnitude and directional consistency of their updates. Let $\mathbf{W}_\ell \in \mathbb{R}^{d\times r}$ denote the GCCA-derived projection matrix for layer ℓ, and $\mathbf{Y}_\ell = \mathbf{X}_\ell \mathbf{W}_\ell \in \mathbb{R}^{n\times r}$ be the projected hidden states of n concepts. For a head k at layer ℓ with stacked outputs $\mathbf{A}_{\ell,k} \in \mathbb{R}^{n\times d}$, we define the head-induced subspace update as: $\Delta\mathbf{Y}_{\ell,k} := \mathbf{A}_{\ell,k}\mathbf{W}_\ell \in \mathbb{R}^{n\times r}$. To characterize the nature of these updates, we computed two complementary metrics: (i) contribution strength ($\alpha_{\ell,k}$) and (ii) directional alignment ($\text{align}_{\ell,k}$). Contribution strength measures the magnitude of the head's update relative to the total subspace energy: $\alpha_{\ell,k} = \frac{\|\Delta\mathbf{Y}_{\ell,k}\|_F^2}{\|\mathbf{Y}_{\ell,k}\|_F^2}$, and directional alignment quantifies the geometric coherence between the head's update and the target subspace representation by their cosine similarity: $\text{align}_{\ell,k} = \frac{\langle \Delta\mathbf{Y}_{\ell,k}, \mathbf{Y}_{\ell,k}\rangle}{\|\Delta\mathbf{Y}_{\ell,k}\|_F \|\mathbf{Y}_{\ell,k}\|_F}$, where $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product. For heads in layers ℓ preceding the first available subspace basis at layer ℓ^*, we calculated $\alpha_{\ell,k}$ and $\text{align}_{\ell,k}$ relative to $\mathbf{W}_{\ell^*}$. This approach was motivated by the hypothesis that information written to the residual stream can persist and remain linearly accessible to subsequent layers.

4.2 Results

Activation patching across corruption conditions reveals a progression of information integration (Figures 4–5; Appendix Figure 18a–c). Under demonstration corruptions (descriptions and labels), statistically significant heads were concentrated in the early-to-middle layers (17–40), coinciding with the depth at which the shared conceptual subspace emerged and began to causally mediate contextual information (around layers 36–40; Section 3.2). The substantial overlap between heads identified under description and label corruptions suggests a joint mechanism that integrates these components to construct an abstract conceptual subspace, with description corruption yielding stronger CIEs (Appendix Figure 18a–c), likely due to its role in contextualizing the query (Bakalova et al., 2025). In contrast, query corruption implicated a later set of heads spanning the middle to late layers, mirroring our observation that query information is mediated by the conceptual subspace over an extended stage of processing. Mid-layer heads (approximately 33–42) likely support query-related computations within the subspace, whereas the final-layer heads (79–80) may implement a readout stage that converts subspace representations into surface-form token predictions.

Attention mass attribution reveals a functional shift from demonstration-centric to query-centric processing (Figure 4). Heads identified under demonstration corruptions and located before layer 38

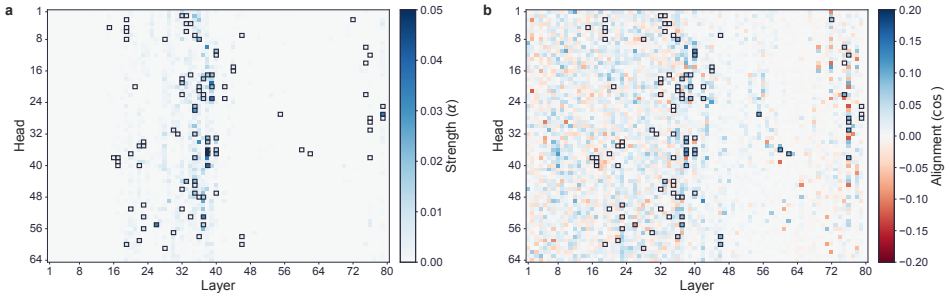

Figure 5: Contribution of attention heads to the conceptual subspace in Llama-3.1 70B. **a**, Contribution strength (α) of attention heads to the conceptual subspace. **b**, Directional alignment (cosine similarity) between attention head outputs and the conceptual subspace. Bordered cells highlight attention heads with statistically significant causal indirect effects (CIEs).

allocated most attention to contextual descriptions, delimiters, and labels, consistent with a role in extracting and integrating demonstration-related cues to construct the subspace. Beyond roughly layer 40, these heads shifted attention toward the query and the final prompt delimiter, signaling a transition from subspace construction to query-conditioned inference. Heads identified under query corruption exhibited heavy attention to query tokens, while the final-layer heads attended predominantly to the last delimiter to facilitate next-token prediction.

The identified heads were also geometrically aligned with the conceptual subspace (Figure 5). Significant heads in early-to-middle layers exhibited both higher contribution strength (α) and greater directional alignment (align) compared to non-significant heads. While contribution strength decreased in later layers, directional alignment remained high, suggesting that late-layer computations stay coordinated with the already-established subspace. Together, these results suggest a two-stage process: (i) Construction, where early-to-middle layers integrate demonstration-derived cues to build and stabilize the conceptual subspace; and (ii) Utilization, where subsequent layers operate within this structured subspace to support inference and produce the final prediction. Results from other LLMs support the generality of our findings (Appendix D.5).

5 Related Work

The in-context learning capacity of LLMs (Brown et al., 2020; Lampinen et al., 2024b) has been studied extensively from both theoretical and empirical perspectives. Theoretically, in-context learning has been framed as approximate Bayesian inference (Xie et al., 2022; Hahn & Goyal, 2023; Wies et al., 2023) or as a form of meta-learning emerging from next-token prediction (Akyürek et al., 2023; Chan et al., 2022; Mahankali et al., 2024). Empirically, research has investigated how various contextual factors can shape model behavior and generalization (Min et al., 2022). Hybrid in-context and in-weight learning in neural networks has been shown to capture aspects of human learning and generalization (Russin et al., 2025; Lake & Baroni, 2023). While these lines of work characterize what problems in-context learning can solve, the internal machinery that implement these computations remains a subject of active debate.

Mechanistic interpretability seeks to decompose LLM computation into interpretable components and circuits (Elhage et al., 2021; Meng et al., 2022; Bakalova et al., 2025; Cho et al., 2025). A growing body of work has identified mechanisms supporting abstract or symbolic-like operations during in-context learning (Olsson et al., 2022; Hendel et al., 2023; Todd et al., 2024; Feng & Steinhardt, 2024; Feucht et al., 2025; Yang et al., 2025). However, much of this evidence comes from relatively simple, synthetic, or tightly controlled settings. Our work complements this literature by moving beyond individual attention heads to characterize the representational substrate of inference. Specifically, we identify an abstract conceptual subspace formed in the residual stream that captures structured, task-relevant information, generalizes across contexts, and is causally implicated in inference over real-world knowledge.

Our work also builds on previous work probing the internal representations learned by LLMs. Evidence suggests that these models encode substantial conceptual knowledge (Forbes et al., 2019; Xu et al., 2025) and can dynamically form representations of world state (Gurnee & Tegmark, 2024; Li et al., 2023; Park et al., 2025). Despite this progress, it remains unclear whether and how such structured representations are formed and used during inference. We bridge this gap by linking representational structure to causal mechanisms, demonstrating how LLMs construct and leverage a conceptual subspace to perform in-context inference.

6 Discussion

In this work, we identify an emergent representational structure that bridges raw contextual input and abstract inference. Our results demonstrate that LLMs construct a dedicated conceptual subspace that abstracts away surface-level details while organizing concepts into a coherent relational structure. This structure persists across contexts and is functionally engaged during inference. These findings inform ongoing debates about the capacity of LLMs for genuine inference: despite being trained solely for next-token prediction, they can learn internal representational substrates that approximate abstract, structured inference rather than relying exclusively on superficial heuristics.

Prior work suggests that middle layers contain substantial structured knowledge and encode meaningful aspects of world state (Skean et al., 2025; Park et al., 2025). Our results extend this view by showing how contextual information and stored knowledge are integrated across middle to late layers to form a coherent conceptual structure that later computations can exploit for task-specific inference. A natural next step is to trace how information encoded in the conceptual subspace is transformed into the target output, and to determine whether analogous substrates support broader forms of reasoning beyond the inference tasks studied here. More generally, connecting such mechanisms to theories of concepts can help clarify how structured inference and reasoning can emerge in both humans and machines (Piantadosi et al., 2024; Johnson-Laird, 2010).

While the conceptual subspace is sufficient and causally central to inference, ablation results also reveal redundancy and compensatory pathways. This raises the possibility that models implement partially overlapping routes to the same behavioral outcome, an important consideration for mechanistic claims and for connecting distributed computation to the symbolic or structured procedures it can approximate. Moreover, our results on smaller models suggest that clean reliance on an abstract conceptual subspace is emergent and scale-dependent. Future work could characterize when models learn to construct and use such an abstract subspace versus alternative pathways, and whether these routes differ in robustness, sample efficiency, or generalization profile.

Our findings also highlight that the mechanisms supporting inference—and the functional roles of model components—shift systematically as contextual evidence accumulates (Lampinen et al., 2026; Lubana et al., 2026). Understanding how internal processing changes across contexts, and how post-training interventions alter models' dependence on contextual cues, is therefore an important direction both for theory-building and for predicting how models will generalize across scenarios, with direct implications for AI safety and reliability.

Finally, our work provides a generalizable framework for studying the internal computation of LLMs at the level of representations, complementing circuit-level analyses that focus on individual components such as heads or neurons. By combining subspace identification with causal intervention, this approach can be naturally extended to other tasks and models. The resulting functionally meaningful subspaces further open a promising interface for steering and monitoring. In particular, our cross-context subspace transfer experiment provides a proof of concept for interventions at the level of task-relevant representations. Exploring whether and how such subspaces can enable flexible, high-level steering of AI systems in more realistic applications and on more complex tasks is a compelling direction for future research (Beaglehole et al., 2026).

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Aleksandra Bakalova, Yana Veitsman, Xinting Huang, and Michael Hahn. Contextualize-then-aggregate: Circuits for in-context learning in gemma-2 2b. In *Second Conference on Language Modeling*, 2025.

Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Mikhail Belkin. Toward universal steering and monitoring of ai models. *Science*, 391(6787):787–792, 2026.

Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. Deep generalized canonical correlation analysis. In Isabelle Augenstein, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei (eds.), *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 1–6, Florence, Italy, August 2019. Association for Computational Linguistics.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891. Curran Associates, Inc., 2022.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. Revisiting in-context learning inference circuit in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Vittoria Dentella, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14 (1):28083, November 2024. ISSN 2045-2322.

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

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024.

Sheridan Feucht, Eric Todd, Byron C Wallace, and David Bau. The dual-route model of induction. In *Second Conference on Language Modeling*, 2025.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*, 2019.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics.

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

Thomas L Griffiths, Brenden M Lake, R Thomas McCoy, Ellie Pavlick, and Taylor W Webb. Whither symbols in the era of advanced neural networks? *arXiv preprint arXiv:2508.05776*, 2025.

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

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838, October 2023. ISSN 2662-8457.

Michael Hahn and Navin Goyal. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*, 2023.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24, 10 2019.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics.

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

Paul Horst. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347, 1961.

Philip N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. ISSN 00063444, 14643510.

Brenden M. Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, November 2023. ISSN 1476-4687.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, July 2024a. ISSN 2752-6542.

Andrew Kyle Lampinen, Stephanie CY Chan, Aaditya K Singh, and Murray Shanahan. The broader spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*, 2024b.

Andrew Kyle Lampinen, Yuxuan Li, Eghbal Hosseini, Sangnie Bhardwaj, and Murray Shanahan. Linear representations in language models can change dramatically over a conversation. *arXiv preprint arXiv:2601.20834*, 2026.

Martha Lewis and Melanie Mitchell. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46, 2024.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.

Ekdeep Singh Lubana, Can Rager, Sai Sumedh R. Hindupur, Valérie Costa, Oam Patel, Sonia Krishna Murthy, Thomas Fel, Greta Tuckute, Daniel Wurgaft, Eric Bigelow, Demba E. Ba, Melanie Weber, and Aaron Mueller. Priors in time: Missing inductive biases for language model interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026.

Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024. ISSN 1364-6613.

James L. McClelland, Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8):348–356, 2010. ISSN 1364-6613.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

Melanie Mitchell. Artificial intelligence learns to reason. *Science*, 387(6740):eadw5211, 2025.

Melanie Mitchell and David C. Krakauer. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Distinguishing antonyms and synonyms in a pattern-based neural network. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 76–85, Valencia, Spain, April 2017. Association for Computational Linguistics.

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

Gustaw Opiełka, Hannes Rosenbusch, and Claire E Stevenson. Analogical reasoning inside large language models: Concept vectors and the limits of abstraction. *arXiv preprint arXiv:2503.03666*, 2025.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041, 2023.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Steven T. Piantadosi, Dyana C.Y. Muller, Joshua S. Rule, Karthikeya Kaushik, Mark Gorenstein, Elena R. Leib, and Emily Sanford. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9):844–856, September 2024. ISSN 1364-6613.

Qwen, :, 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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Jacob Russin, Ellie Pavlick, and Michael J. Frank. Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning. *Proceedings of the National Academy of Sciences*, 122(35):e2510270122, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, September 2023. ISSN 2397-3374.

Taylor W Webb, Keith J Holyoak, and Hongjing Lu. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *PNAS Nexus*, 4(5):pgaf135, 05 2025. ISSN 2752-6542.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 1, 2022. ISSN 2835-8856.

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36637–36651. Curran Associates, Inc., 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1819–1862, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In Chelsea Finn, Yejin Choi, and Marc Deisenroth (eds.), *International Conference on Learning Representations*, 2022.

Ningyu Xu, Qi Zhang, Menghan Zhang, Peng Qian, and Xuanjing Huang. On the tip of the tongue: Analyzing conceptual representation in large language models with reverse-dictionary probe. *arXiv preprint arXiv:2402.14404*, 2024.

Ningyu Xu, Qi Zhang, Chao Du, Qiang Luo, Xipeng Qiu, Xuanjing Huang, and Menghan Zhang. Revealing emergent human-like conceptual representations from language prediction. *Proceedings of the National Academy of Sciences*, 122(44):e2512514122, 2025.

Yukang Yang, Declan Iain Campbell, Kaixuan Huang, Mengdi Wang, Jonathan D. Cohen, and Taylor Whittington Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning*, 2025.

A Limitations

Our work uses concept inference as a simple but informative testbed for probing mechanisms underlying flexible inference in LLMs, which has been shown relevant to broader language understanding and reasoning abilities (Xu et al., 2025; 2024). Although our results suggest that the identified mechanisms generalize beyond this setting to a broader family of inference tasks, we do not yet establish their scope in more complex, real-world tasks, which will be essential for building a fuller picture of model competence and limitations.

We used SVD followed by GCCA as a conservative operational procedure for identifying shared cross-layer representational structure, whose cross-context stability and functional relevance were then validated through representational and causal analyses. While this choice provides an interpretable estimate of shared structure, it inherently favors linear, high-variance, and cross-layer-consistent representations, potentially overlooking task-relevant information that is nonlinear, low-variance, layer-specific, or more distributed. Although sufficient for our claims, future work could explore complementary methods for isolating relevant subspaces, such as nonlinear approaches, task-conditioned subspace discovery, and techniques for capturing multiple interacting latent factors. Our results further suggest that, with sufficient in-context demonstrations, relevant subspaces become increasingly aligned, motivating future work on extracting cross-context shared subspaces for analysis and steering.

B Data, Code, and Materials Availability

This work uses publicly available datasets and materials (Hebart et al., 2019; Todd et al., 2024; Nguyen et al., 2017; Hernandez et al., 2024). Code for reproducing all experiments is available at `https://github.com/ningyuxu/llm_structured_icl`.

C Supplementary Methods

C.1 Models

Our experiments focus on base models—LLMs pretrained solely with next-token prediction, without any post-training or instruction tuning. Table 1 summarizes the open-source LLMs included in our study. In the main paper, we report results for Llama-3.1 70B, which is the largest model we evaluate and the one that achieves the strongest behavioral performance on the reverse dictionary task, making it a natural representative for illustrating the emergence of the conceptual subspace. Results for the remaining models are provided in Appendix D.3–D.5 and Figures 13–23.

C.2 Data

For concept inference, we used data from the THINGS database (Hebart et al., 2019) (available at `https://osf.io/jum2f/`), which contains 1,854 concrete, nameable object concepts selected

to be representative of everyday objects commonly used in American English. Each concept is paired with its WordNet synset ID, definitional descriptions, and multiple associated images, providing a suitable testbed for analyzing conceptual representations derived from natural-language descriptions.

We further extended our analyses to four additional inference tasks from Todd et al. (2024): Antonym, Country–Capital, Landmark–Country, and National Parks (Table 2).

To probe in-context concept inference in LLMs, we randomly designated 20% of the concepts as a training set, from which N description–word pairs were sampled to form the demonstrations. The remaining concepts were used as the test set. For the other inference tasks, we sampled 24 demonstrations at random from each dataset and evaluated model performance on the remaining examples. For each experimental condition, we repeated evaluation over five independent runs, each with a distinct random sample of N demonstrations.

C.3 Generalized Canonical Correlation Analysis (GCCA)

GCCA Eigenvalue Formulation. We employed GCCA to identify a low-dimensional subspace that is shared across model layers. Let $\mathbf{X}_\ell \in \mathbb{R}^{n\times d}$ denote the matrix of hidden states at layer ℓ, where each row $\mathbf{X}_\ell[i,:] = \mathbf{h}_{\ell,i}$ corresponds to the representation of the ith query concept, n is the number of query concepts, and d is the hidden dimensionality of the model. Following standard formulations of GCCA (Horst, 1961; Kettenring, 1971; Benton et al., 2019), the optimization objective in Equation 1 is equivalent to computing the leading eigenvectors of the aggregate projection operator

$$\mathbf{S} = \sum_\ell \mathbf{P}_\ell, \tag{4}$$

where $\mathbf{P}_\ell = \mathbf{X}_\ell \left(\mathbf{X}_\ell^\top \mathbf{X}_\ell + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_\ell^\top$ denotes the regularized projection matrix associated with layer ℓ. Unless otherwise stated, we set the ridge regularization parameter to $\lambda = 0.01$ for numerical stability. The eigenvectors of $\mathbf{S}$ corresponding to its largest eigenvalues define directions that exhibit strong alignment across layers, and thus constitute the shared subspace.

Rank Selection via Permutation Testing. To determine the dimensionality $\hat{r}$ of the conceptual subspace shared across layers, we employed a non-parametric permutation test. The null hypothesis is that any apparent alignment across layers arises from chance correlations rather than shared structure. Specifically, for each layer ℓ, we independently applied a random permutation π_ℓ to the rows of $\mathbf{X}_\ell$, yielding a permuted matrix $\mathbf{X}_\ell^\pi$. This procedure destroys the correspondence between representations of the same query concept across layers, while preserving the marginal covariance structure within each layer. We repeated this permutation procedure $M = 500$ times. For each permutation, we computed the eigenvalue spectrum $\{\lambda_i^\pi\}$ of the corresponding GCCA operator $\mathbf{S}^\pi$. For each component index i, we defined a significance threshold q_i as the $(1-\alpha)$ quantile of the null distribution of λ_i^π, with $\alpha = 0.05$. The effective rank $\hat{r}$ of the shared subspace was then determined as the number of eigenvalues from the original (unpermuted) data whose magnitude exceeded the corresponding null threshold:

$$\hat{r} = \sum_{i=1}^{r_{\max}} \mathbb{I}(\lambda_i > q_i)\,. \tag{5}$$

To verify that this permutation-based choice of dimensionality was reasonable, we additionally evaluated a range of alternative subspace dimensions, $r \in \{64, 128, 256, 512, 1024, 2048\}$; the corresponding results are reported in Appendix D.1.

C.4 Structural Analysis of Conceptual Subspaces

GCCA Alignment. Given hidden-state matrices $\{\mathbf{X}_\ell\}_{\ell\in\mathcal{L}}$ from a set of layers $\mathcal{L}$, GCCA yields a shared latent representation $\mathbf{G} \in \mathbb{R}^{n\times r}$ and layer-specific projection matrices $\mathbf{W}_\ell \in \mathbb{R}^{d\times r}$ such that $\mathbf{X}_\ell \mathbf{W}_\ell \approx \mathbf{G}$. We denote the layer-wise projected representations as $\mathbf{Y}_\ell = \mathbf{X}_\ell \mathbf{W}_\ell$. To quantify the consistency with which different layers express the shared subspace, we computed the correlation between corresponding dimensions of $\mathbf{Y}_\ell$ across layers. For a pair of layers (ℓ, ℓ'), GCCA alignment is calculated as

$$\text{Align}(\ell, \ell') = \frac{1}{r} \sum_{j=1}^{r} \text{corr}(\mathbf{Y}_{\ell,:,j}, \mathbf{Y}_{\ell',:,j})\,, \tag{6}$$

where $\mathbf{Y}_{\ell,:,j}$ denotes the j-th column of $\mathbf{Y}_\ell$ and $\mathrm{corr}(\cdot,\cdot)$ denotes Pearson correlation.

Subspace Overlap via Principal Angles. To compare subspaces derived under different contexts, we assessed the overlap between the column spaces of the corresponding projection matrices $\mathbf{W}_\ell^c$, where c indexes the set of demonstrations provided in context. For each $\mathbf{W}_\ell^c$, we obtained an orthonormal basis $\mathbf{U}_\ell^c$ via QR decomposition. Given two sets of demonstrations c and c', we computed the principal angles $\{\theta_i\}_{i=1}^r$ between the subspaces spanned by $\mathbf{U}_\ell^c$ and $\mathbf{U}_\ell^{c'}$. Subspace overlap is quantified as the mean squared cosine of these angles: $\frac{1}{r}\sum_{i=1}^r \cos^2\theta_i$. This metric measures whether different contexts induce similar linear subspaces within the model's hidden-state space.

Representational Similarity Analysis (RSA). We employed representational similarity analysis (RSA) to compare the relational structure of concepts encoded within different conceptual subspaces. RSA is nonparametric and has been widely adopted to measure alignment between representation spaces based on their similarity structure. For each context (prompt) s and layer ℓ, we computed a representational dissimilarity matrix (RDM) $\mathbf{D}_\ell^c \in \mathbb{R}^{n\times n}$ from the projected representations $\mathbf{Y}_\ell^c = \mathbf{X}_\ell^c\mathbf{W}_\ell^c$, where each entry $\mathbf{D}_{\ell,ij}^c = 1-\cos\Big(\mathbf{Y}_{\ell,i,:}^c, \mathbf{Y}_{\ell,j,:}^c\Big)$ measures the dissimilarity between concepts i and j. RSA alignment between representations derived under contexts s and s' is then measured by the Spearman's rank correlation between the upper-triangular entries of $\mathbf{D}_\ell^c$ and $\mathbf{D}_\ell^{c'}$.

C.5 Behavioral Evaluation

We evaluated model performance in inference by truncating each generation at the first newline character ("\n") and checking whether the resulting output exactly matched the target word or any of its listed synonyms. We used greedy decoding for all models to ensure a consistent and controlled comparison across model sizes and families.

C.6 Causal Mediation Analysis

Causal Indirect Effect and Causal Mediation Score. We employed log probabilities to measure the causal effect of interventions on the model's predictions. Specifically, the causal indirect effect (CIE) quantifies the change in the model's output after patching a specific component $\mathbf{h}_\ell^{\text{patch}}$ at layer ℓ. We calculated this as the difference in log probabilities for the target (correct) token y_q under the intervention, relative to the original (corrupted) prompt $\tilde{s}$:

$$\mathrm{CIE} = \log f_{\mathcal{M}}\Big(\tilde{s} \mid \mathbf{h}_\ell^{\text{patch}}\Big)[y_q] - \log f_{\mathcal{M}}(\tilde{s})[y_q]\,, \tag{7}$$

where $f_{\mathcal{M}}(\cdot)[y_q]$ denotes the output probability for token y_q from the model $\mathcal{M}$.

For cross-context transfer from a source context c_i to a target context c_j, where y_{q_b} is the expected concept for the query description s_{q_b} in the target context, we measured causal mediation (CMA) score. This score represents the change in the log probability difference between the source-related token y_{q_a} and the target-correct token y_{q_b} after the intervention:

$$\begin{aligned}\mathrm{CMA} = \Big(&\log f_{\mathcal{M}}\Big(c_j, s_{q_b} \mid \mathbf{h}_\ell^{\text{patch}}\Big)[y_{q_a}] - \log f_{\mathcal{M}}\Big(c_j, s_{q_b} \mid \mathbf{h}_\ell^{\text{patch}}\Big)[y_{q_b}]\Big)\\ &- \left(\log f_{\mathcal{M}}(c_j, s_{q_b})[y_{q_a}] - \log f_{\mathcal{M}}(c_j, s_{q_b})[y_{q_b}]\right).\end{aligned} \tag{8}$$

We preferred log probabilities over raw logits because logits are unconstrained outputs that can vary significantly in scale across different conditions. Log probabilities are normalized via the softmax function, ensuring that the metrics reflect the model's relative confidence in its predictions. This makes them a more stable and interpretable objective for quantifying the functional impact of interventions across varying contexts.

Identification of Significant Attention Heads via Permutation Testing. To identify the attention heads that are functionally relevant for inference in LLMs, we performed statistical tests to determine which heads exhibit statistically significant causal effects. We employed a non-parametric permutation test to control the family-wise error rate (FWER) at $\alpha = 0.05$, with the null hypothesis assuming

Table 1: Overview of the large language models (LLMs) used in our experiments. "#Layers" denotes the total number of layers in each model, including the embedding layer (indexed as layer 0). "Sel. Layers (0+)" lists the layers at which GCCA was applied to derive conceptual subspaces, as determined by a layer-wise SVD analysis. "Dim." denotes the dimensionality of the model's hidden states, and "Subspace Dim." reports the dimensionality of the GCCA-derived conceptual subspaces, selected via permutation testing and summarized as median [range] across five runs. All models are publicly available via Hugging Face (Wolf et al., 2019) (`https://huggingface.co`).

Series	Model	#Layers	Sel. Layers (0+)	Dim.	Subspace Dim.
Llama 3.1	`Llama-3.1-70B`	81	38–79	8192	1180 [1176–1181]
	`Llama-3.1-8B`	33	16–31	4096	1250 [1242–1254]
Llama 3	`Meta-Llama-3-8B`	33	16–31	4096	1245 [1244–1249]
Qwen2.5	`Qwen2.5-32B`	65	50–63	5120	1221 [1218–1224]
	`Qwen2.5-7B`	29	22–28	3584	1292 [1290–1299]
	`Qwen2.5-3B`	37	32–36	2048	1537 [1531–1540]

Table 2: Overview of the inference tasks used in our experiments. Data for concept inference were drawn from the THINGS database (Hebart et al., 2019) (`https://osf.io/jum2f/`); data for the remaining four tasks were taken from Todd et al. (2024) (`https://github.com/ericwtodd/function_vectors`).

Task	Examples	Source
Concept Inference	a small very thin pancake ⇒ crepe a small guitar having four strings ⇒ ukulele dried grape ⇒ raisin	Hebart et al. (2019)
Antonym	true ⇒ false difficult ⇒ easy proceed ⇒ halt	Nguyen et al. (2017)
Country–Capital	Austria ⇒ Vienna China ⇒ Beijing Germany ⇒ Berlin	Todd et al. (2024)
Landmark–Country	Achelous River ⇒ Greece Atwell Peak ⇒ Canada Beauvais ⇒ France	Hernandez et al. (2024)
National Parks	Prince William Forest Park ⇒ Virginia Hot Springs National Park ⇒ Arkansas Kings Canyon National Park ⇒ California	Todd et al. (2024)

the absence of any real causal effects. For each permutation, we disrupted the relationship between specific heads and the model's output by applying random sign-flipping to the observed causal indirect effects (CIEs). Specifically, each CIE was multiplied by a value sampled uniformly from $\{-1, 1\}$, and this procedure was repeated 5,000 times. To account for multiple comparisons, we constructed a null distribution based on the maximum CIE observed for each permutation. The significance threshold was then defined as the $(1 - \alpha)$-quantile of this empirical null distribution. Attention heads were classified as significant if their observed CIE exceeded this threshold. We restricted our analysis to these significant heads to focus on those contributing constructively to the in-context concept inference task.

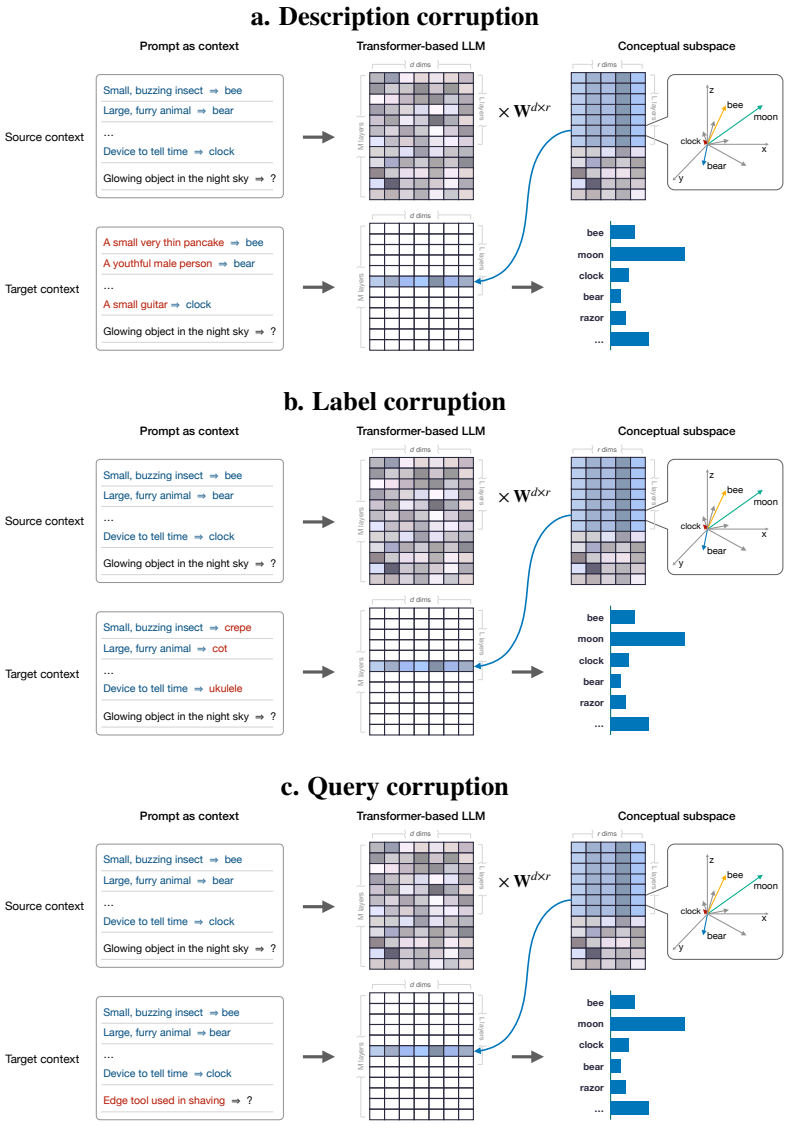

Figure 6: Illustration of causal mediation analysis under three corruption conditions: description (**a**), label (**b**), and query (**c**) corruption. In the source context, the prompt contains correct description–word pairs as demonstrations, together with a query description. In the target context, the corrupted field is replaced with mismatched content of the same token length. We then patch the conceptual subspace from the source context into the target context to test whether the model's prediction shifts systematically toward the expected source-context term. The figure illustrates this procedure using the concept inference task as an example.

D SUPPLEMENTARY RESULTS

D.1 EVALUATING SUBSPACE DIMENSIONALITY

To evaluate whether the permutation-based choice of dimensionality ($\hat{r} = 1180$ for Llama-3.1 70B) was well chosen, we additionally tested alternative subspace dimensions, $r \in \{64, 128, 256, 512, 1024, 2048\}$. For dimensions smaller than $\hat{r}$, the representational results were qualitatively similar: cross-layer alignment remained close to unity (Figure 8a–e), and the GCCA-derived projection matrices varied smoothly across layers (Figure 8g–k). With 24 demonstrations, these lower-dimensional subspaces also exhibited relatively strong alignment in their internal structure (Figure 8m–q) and substantial overlap (Figure 8s–w). In contrast, increasing the dimensionality to 2048 led to markedly weaker cross-layer alignment (Figure 8f) and substantially reduced cross-context alignment (Figure 8r), suggesting that the larger subspace captures additional directions beyond the core shared structure, including information that is less consistently shared and more context-dependent.

Causal intervention analyses showed progressively stronger effects as subspace dimensionality increased. Under corrupted context, patching the GCCA-derived subspace restored a larger portion of performance as more dimensions were retained (Figure 9a–c), while consistently outperforming random-subspace baselines of matched dimensionality. Although $r = 2048$ restored more performance than the permutation-selected $\hat{r}$, the representational analyses suggest that subspaces of this size include additional weaker and more context-dependent directions that may nevertheless remain task-relevant. Ablating the GCCA-derived subspaces (Figure 9d) produced significantly larger performance drops than ablating matched random subspaces, whose effects remained small even at $r = 2048$. These ablation effects also became more pronounced with increasing dimensionality. Likewise, isolating the GCCA-derived subspaces preserved substantially more performance than matched random-subspace baselines (Figure 9e). Once the dimensionality approached the permutation-selected $\hat{r}$, isolation preserved near-original performance across most layers, and the effect was close to that obtained with $r = 2048$. Importantly, across all tested dimensions and intervention settings, the conceptual subspace at the permutation-selected dimensionality consistently outperformed random baselines by a wide margin, supporting the validity of the selected rank.

Together, these results indicate that the permutation-selected dimensionality is a reasonable operating point: it captures the relational structure shared across layers and contexts while retaining most of the task-relevant information functionally used by the model. Increasing the dimensionality yields stronger patching and ablation effects, but at the cost of incorporating information that is less consistently shared and more context-dependent. This pattern suggests that dimensions beyond the permutation-selected rank may contain complementary task-relevant information used by LLMs, an interesting direction for future investigation.

D.2 ASSESSING THE GENERALITY OF CONCEPTUAL SUBSPACES ACROSS INFERENCE TASKS

To assess whether the identified conceptual subspaces generalize beyond concept inference, we extended our analyses to additional inference tasks from Todd et al. (2024), including Antonym, Country–Capital, Landmark–Country, and National Parks (Table 2). Across these tasks, layer-wise representational analyses revealed the same qualitative pattern observed in the concept inference setting: shared subspaces emerged in the middle layers (approximately layers 38–40) and persisted into later layers (Figure 10a–d). The emergence of these subspaces was likewise accompanied by an increase in the number of principal components (PCs) required to explain 95% of the variance (Figure 10e–h), suggesting that contextual information is integrated into a higher-dimensional representational space for subsequent computation. Together, these findings mirror those observed for concept inference (Section 2.2), while also indicating task-dependent variation in dimensionality, which may reflect differences in task complexity.

Causal interventions revealed a similar pattern across tasks. Under corrupted context, performance was largely restored by patching the conceptual subspace into middle layers, substantially outperforming the random-subspace baseline (Figures 11a–c,g–i; 12a–c,g–i). Ablation and isolation analyses likewise supported the necessity and sufficiency of the subspace for inference (Figures 11d–e,j–k; 12d–e,j–k): ablating the conceptual subspace caused a marked performance drop relative to the random-subspace baseline, whereas isolating it preserved near-original performance across most

layers. In cross-context transfer, patching effects closely approached the same-context reference and remained well above the random-subspace baseline (Figures 11f,i; 12f,i), suggesting that the model constructs and exploits a stable, abstract conceptual subspace whose relational structure is preserved across contexts. Overall, these results indicate that the emergence and functional role of conceptual subspaces are not unique to concept inference, but extend across a broader range of in-context inference tasks, although the precise layers at which the subspace mediates contextual information or supports cross-context abstraction may differ by task.

D.3 Identifying Conceptual Subspaces across Models

Across all LLMs examined, a shared latent subspace consistently emerged in the middle-to-late layers (Figures 13–14). Layer-wise SVD showed that early layers exhibited relatively low cross-layer overlap, whereas middle-to-late layers maintained high inter-layer overlap. This transition coincided with a sharp increase in the number of principal components required to explain 95% of the variance (Figures 13a–b,f–g; 14a–b,f–g,k–l), suggesting the integration of contextual information into a higher-dimensional subspace.

Subspaces identified via GCCA demonstrated near-unity GCCA alignment scores and RSA values across layers. Increasing the number of demonstrations from 1 to 24 led to progressively stronger cross-context alignment of the internal subspace structure, with diminishing gains beyond this threshold (Figures 13d–e,i–j; 14d–e,i–j,n–o). These trends were consistent across model families, though the relative depth at which the shared subspace emerges varied; for example, Qwen2.5 models manifested the shared subspaces at relatively later layers than Llama-3 and Llama-3.1 variants.

Collectively, these results suggest that LLMs construct a conceptual subspace during inference that persists across middle to late layers. While larger models (e.g., Llama-3.1 70B) required higher-dimensional layer-wise subspaces to capture 95% of the variance (Figure 2b), permutation testing identified fewer significant dimensions for GCCA analyses across layers (Table 1). Furthermore, these subspaces exhibited stronger cross-context alignment compared to smaller models. These findings indicate that increased model scale and compute support the formation of richer and more stable relational structures, which are encoded in a more coherent and compressed manner, revealing systematic differences in how LLMs organize internal computations during inference.

D.4 Analyzing the Causal Role of Conceptual Subspaces in Inference across Models

Besides Llama-3.1 70B, we applied the same set of targeted causal interventions to other LLMs; the corresponding results are shown in Figures 15–17. Across models, we observed trends broadly consistent with those reported for Llama-3.1 70B, although the extent to which inference is mediated by the conceptual subspace varied across models.

Specifically, activation patching under corrupted demonstrations and queries (Figures 15a–c,g–i; 16a–c,g–i) showed that demonstration-derived contextual information was primarily mediated by the conceptual subspace at mid-depth, where the subspace began to form, whereas query-related information was mediated at relatively later layers. While patching the conceptual subspace substantially recovered performance compared to random-subspace baselines, the recovery gap was narrower than that observed in Llama-3.1 70B. This suggests that, in these smaller models, task-relevant information may be represented in a more distributed manner, with weaker reliance on a localized, dominant conceptual subspace.

Subspace ablation and isolation experiments (Figures 15d–e,j–k; 16d–e,j–k) revealed patterns similar to those in Llama-3.1 70B. Ablating the conceptual subspace consistently disrupted inference and led to a clear performance drop relative to the random-subspace baselines, although performance did not collapse entirely, indicating partial redundancy or compensatory pathways. Conversely, isolating the subspace preserved near-original performance at middle layers, with degradation in later layers where computation likely shifted toward surface-form generation. As in Llama-3.1 70B, increasing the number of demonstrations attenuated ablation effects, while isolation continued to preserve performance deeper in the model, suggesting a progressively stronger reliance on the subspace as contextual information accumulates.

Finally, cross-context transfer effects increased with the number of demonstrations and consistently exceeded the random-subspace baselines (Figures 15f,l; 16f,l), indicating that these models can construct relational structure that generalizes across contexts. However, the magnitude of this effect was notably reduced in smaller models. This trend was especially pronounced in Qwen2.5 3B (Figure 17), where random-subspace baselines often matched or exceeded the effects of the derived conceptual subspace across interventions. Together, these results suggest a transition in representational strategy: whereas larger models leverage well-defined, abstract conceptual subspaces to approximate in-context inference, smaller models rely more heavily on diffuse or non-subspace-specific representations.

D.5 Analyzing Context Integration into Conceptual Subspaces across Models

Across the LLMs we tested, activation patching across corruption conditions show a staged progression of information integration consistent with the patterns observed in Llama-3.1 70B. Heads with statistically significant effects under demonstration corruptions (descriptions and labels) clustered in early-to-middle layers for Llama-3.1 8B and Llama-3 8B (Figure 18), yet shifted toward middle-to-late layers in the Qwen2.5 models (Figure 19). These depth ranges aligned with the formation of the shared conceptual subspace and its causal mediation of contextual information (Sections D.3–D.4). We also observed substantial overlap between heads identified under description and label corruptions, suggesting a shared mechanism for integrating demonstration-derived cues into the conceptual subspace. In larger models (Llama-3.1 70B and Qwen2.5 32B), description corruption typically implicated more heads and yielded larger causal indirect effects (CIEs) than label corruption. This likely reflects the role of descriptions in contextualizing the query (Bakalova et al., 2025), which appears more pronounced as model capacity increases. Conversely, query corruption consistently implicated a relatively later set of heads, mirroring our observation that query information is mediated by the conceptual subspace over a later stage of processing. These mid-to-late layer heads likely support query-conditioned computations given the established context, whereas final-layer heads may be engaged in a readout stage that maps internal representations to surface-form token predictions.

Attention patterns further reveal a functional shift from demonstration-centric to query-centric processing (Figures 20–21). Heads identified under demonstration corruptions and located in relatively early layers allocated most attention to descriptions, delimiters, and labels, consistent with extracting and integrating demonstration cues to construct or update the subspace. In contrast, later-layer heads shifted attention toward the query and the final prompt delimiter, signaling a transition from subspace construction to query-conditioned inference. Heads identified under query corruption attended heavily to query tokens, while final-layer heads attended more strongly to the last delimiter, consistent with next-token prediction.

The identified heads were also geometrically aligned with the conceptual subspace (Figures 22–23). Significant heads in early-to-middle layers of Llama-3 (or middle-to-late layers of Qwen2.5) exhibited higher contribution strength (α) and stronger directional alignment (align) than non-significant heads. While contribution strength generally decreased in later layers, directional alignment remained relatively high, suggesting that late-layer computations stay coordinated with an already-established subspace (Figure 22). Figure 23 further shows a qualitative difference in how heads interact with this subspace: heads identified under query corruption predominantly wrote coherently into the conceptual subspace (positive directional alignment), suggesting a role in using or stabilizing the subspace for prediction. In contrast, heads identified under demonstration corruption can exhibit substantial write strength (α) but weaker alignment, indicating a role in constructing or reconfiguring the subspace based on contextual evidence rather than simply amplifying existing directions.

In larger models (Llama-3.1 70B and Qwen2.5 32B), we observed notable overlap among the heads identified across corruption conditions; these heads exhibited both high contribution strength and high alignment (Figure 23a,d), indicating a core set of components that integrate relevant contextual information and directly drive model inference. In smaller models, causally relevant heads were more diffuse and can show weaker alignment or contribution strength relative to other heads (Figures 23b–c,e–f; 22g–i). Finally, individual heads in the larger models tended to have lower contribution strengths than those in smaller models, suggesting that computation is distributed across a broader ensemble of moderately weighted heads.

Taken together, these findings support a two-stage account that generalizes across model families: (i) Construction, in which earlier layers integrate demonstration-derived cues to build and stabilize a

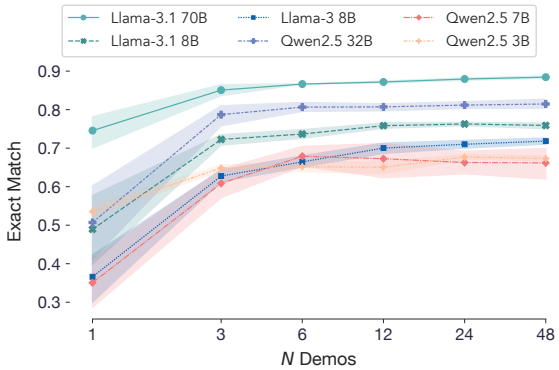

Figure 7: Performance of various LLMs on the reverse dictionary task, evaluated on the THINGS data and measured through exact match accuracy. The models were presented with N demonstrations sampled from the training set and evaluated on an independent test set. Shaded areas denote 95% confidence intervals, calculated from 10,000 resamples across five independent runs.

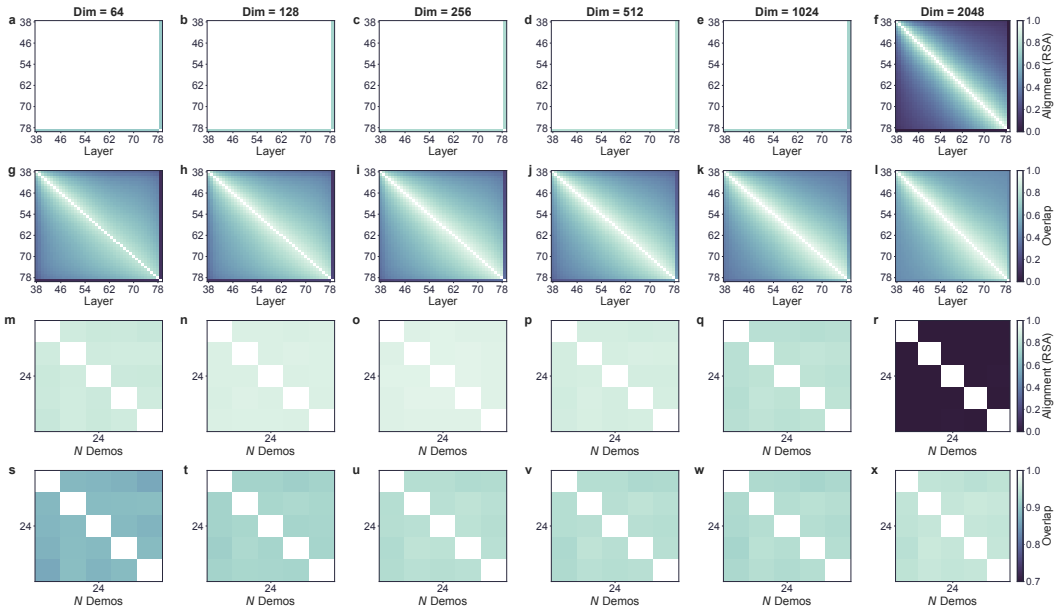

Figure 8: Identifying shared subspaces across layers in Llama-3.1 70B at varying dimensionalities using GCCA. Titles indicate the retained subspace dimensionality. All results are computed with 24 in-context demonstrations and averaged over five runs. **a**–**f**, Alignment of representational geometry between GCCA-derived subspace across selected layers, measured by RSA, with both axes indexing layers. **g**–**l**, Overlap between GCCA-derived projection matrices across selected layers, with both axes indexing layers. **m**–**r**, Cross-context alignment of representational geometry within the GCCA-derived subspace across different demonstration sets, measured by RSA. **s**–**x**, Overlap between GCCA-derived projection matrices across demonstration sets. For **m**–**x**, the axes denote the number of demonstrations, and each matrix entry corresponds to a single run.

conceptual subspace; and (ii) Utilization, in which later layers operate within this structured subspace to perform query-conditioned inference and generate the final prediction. Smaller models appear to rely less consistently on this mechanism, exhibiting more diffuse, non-subspace-specific strategies.

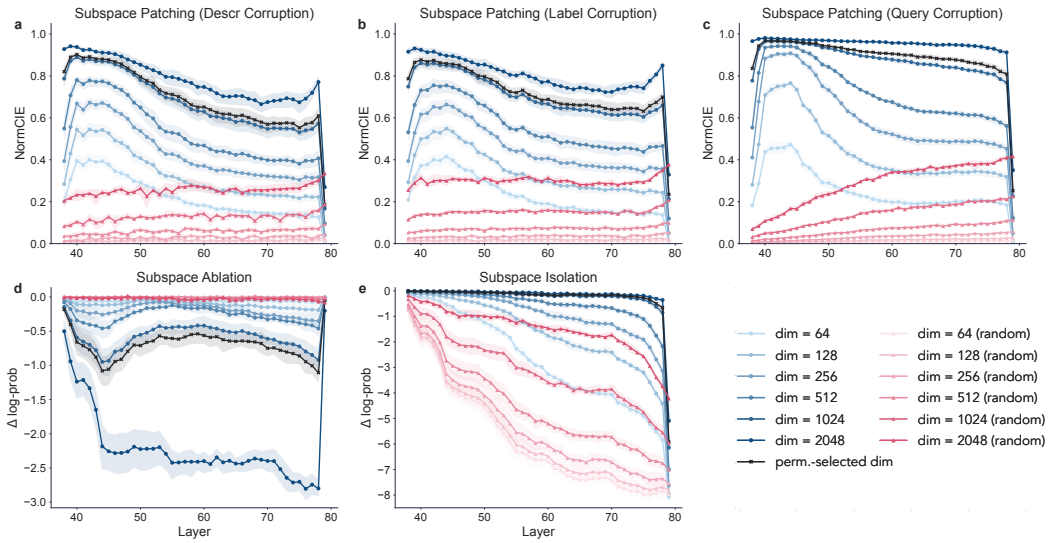

Figure 9: Causal intervention results for Llama-3.1 70B with GCCA-derived subspaces of varying dimensionality. All results are obtained with 24 in-context demonstrations. Results for the GCCA-derived subspace (blue) are compared with dimension-matched random-subspace baselines (red). Black lines denote the permutation-selected dimensionality used in the main text. **a**–**c**, Activation patching under three corruption conditions: description (**a**), label (**b**) and query (**c**). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d**–**e**, Subspace necessity and sufficiency, assessed by ablating the subspace (**d**) or isolating it (**e**). The y-axis denotes the change in log-probability of the correct token. Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

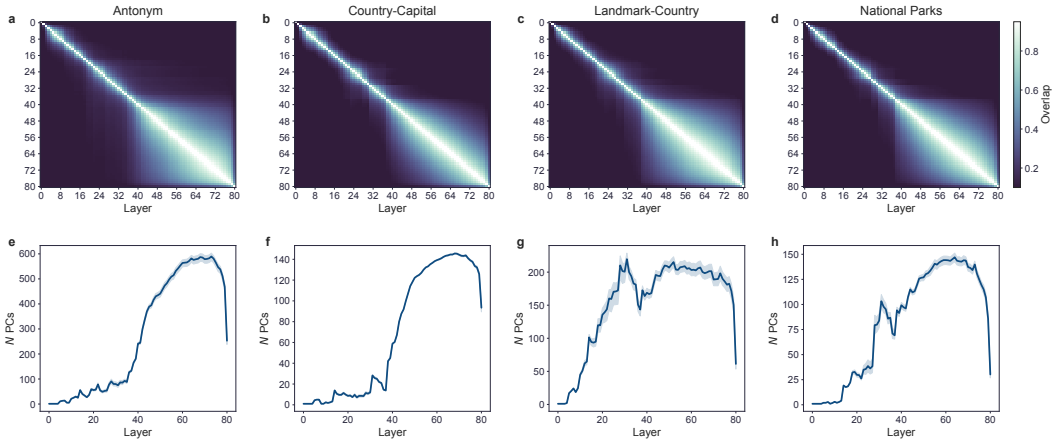

Figure 10: Emergence of conceptual subspaces in Llama-3.1 70B across four in-context inference tasks. **a**–**d**, Layer-wise similarity of hidden states, measured as subspace overlap (mean squared cosine of principal angles) between SVD subspaces explaining 95% variance. Results are averaged over five runs, each with 24 demonstrations. Axes index layers. **e**–**h**, Number of principal components (PCs) needed to explain 95% of the variance at each layer. Shaded areas denote 95% CIs calculated from 10,000 bootstrap resamples over five runs.

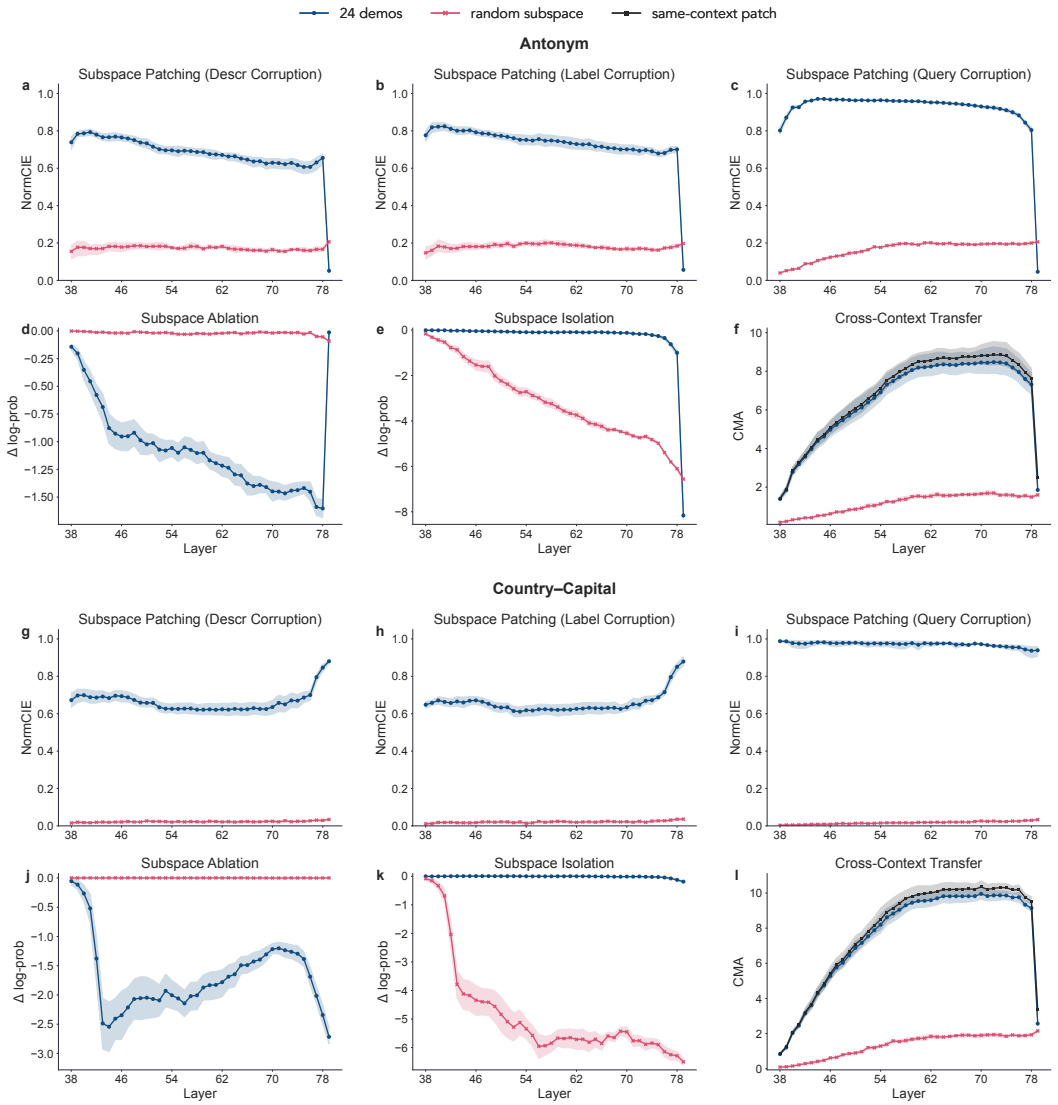

Figure 11: Causal intervention results for Llama-3.1 70B on two in-context inference tasks (Antonym and Country–Capital). **a**–**c**, **g**–**i**, Activation patching with 24 demonstrations under three corruption conditions: description (**a**, **g**), label (**b**, **h**) and query (**c**, **i**). The conceptual subspace (blue) is compared against a random-subspace baseline (red). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d**–**e**, **j**–**k**, Subspace necessity and sufficiency tested by ablating the conceptual subspace (**d**, **j**) or isolating it (**e**, **k**). The y-axis denotes the change in log-probability of the correct token. **f**, **l**, Causal effects of cross-context transfer, where the relational structure from a source context is adapted to a target context via an orthogonal transformation. The y-axis reports the causal mediation (CMA) score. Performance is compared against a random-subspace baseline (red). Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

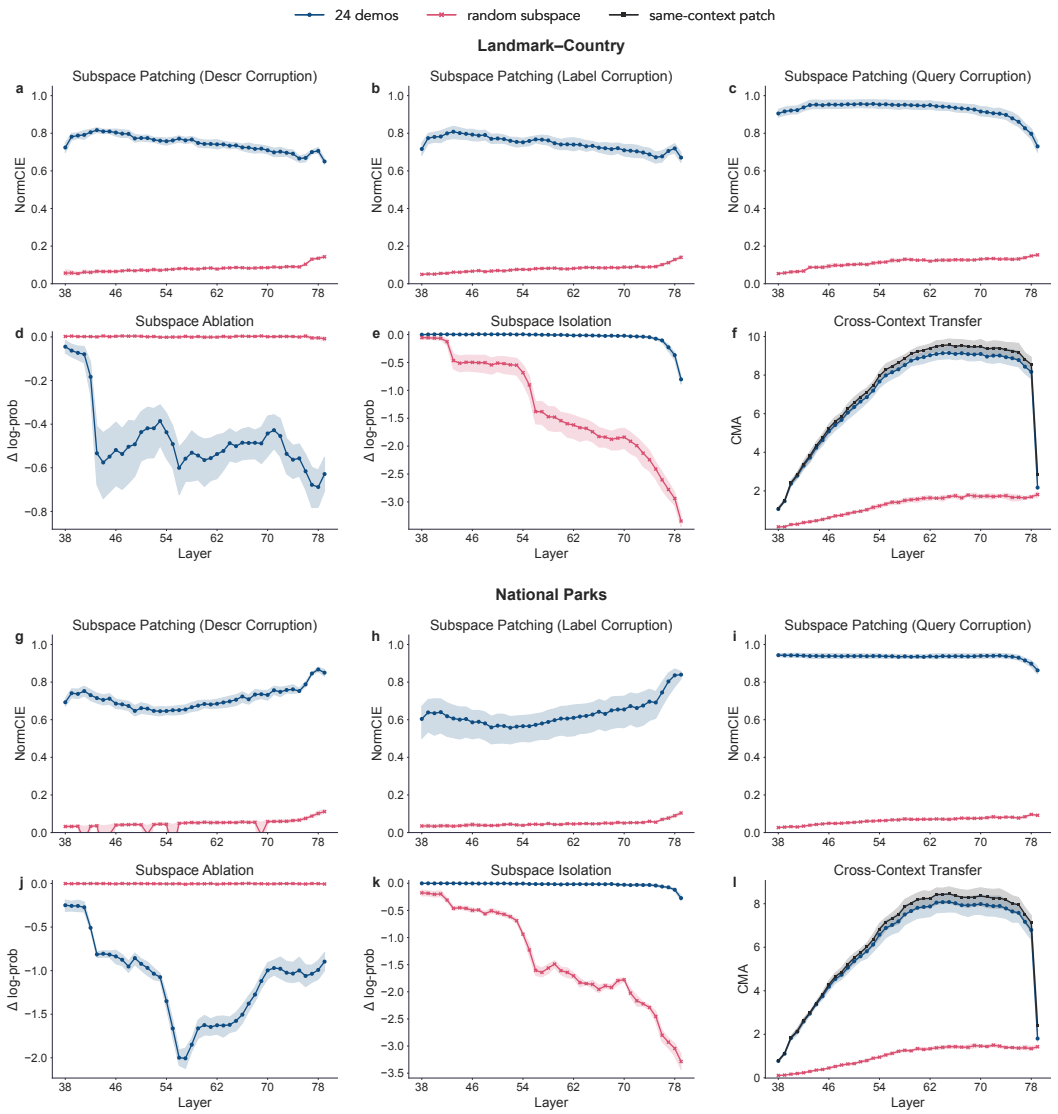

Figure 12: Causal intervention results for Llama-3.1 70B on two in-context inference tasks (Landmark–Country and National Parks). **a–c**, **g–i**, Activation patching with 24 demonstrations under three corruption conditions: description (**a**, **g**), label (**b**, **h**) and query (**c**, **i**). The conceptual subspace (blue) is compared against a random-subspace baseline (red). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d–e**, **j–k**, Subspace necessity and sufficiency tested by ablating the conceptual subspace (**d**, **j**) or isolating it (**e**, **k**). The y-axis denotes the change in log-probability of the correct token. **f**, **l**, Causal effects of cross-context transfer, where the relational structure from a source context is adapted to a target context via an orthogonal transformation. The y-axis reports the causal mediation (CMA) score. Performance is compared against a random-subspace baseline (red). Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

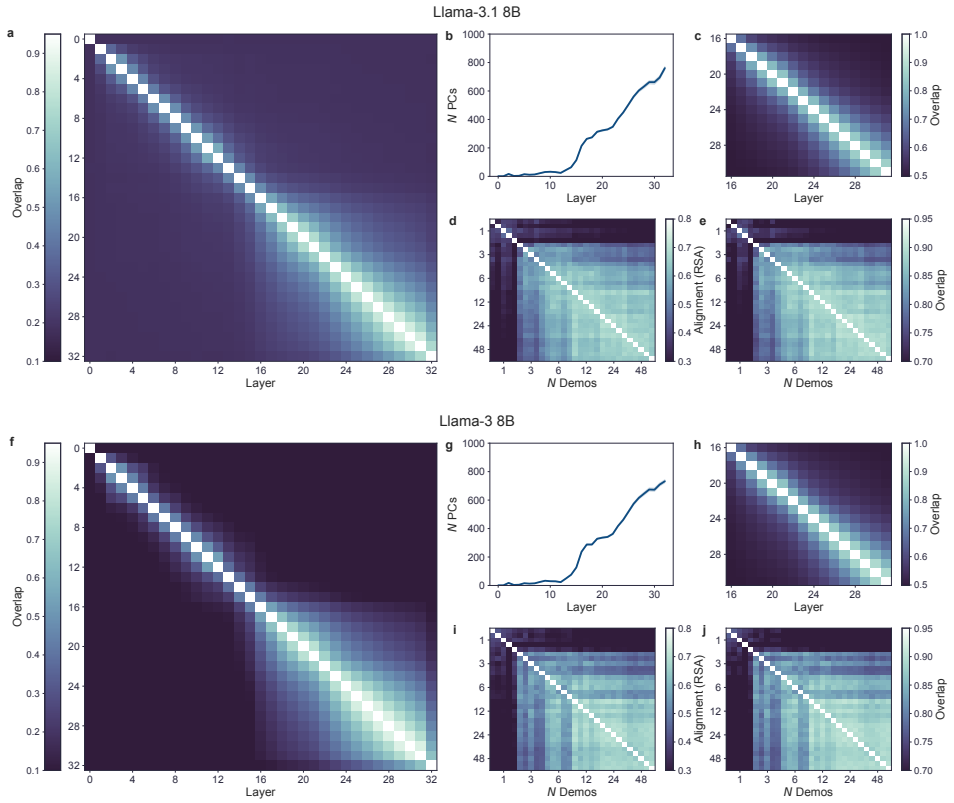

Figure 13: Emergence of conceptual subspaces in Llama-3.1 8B (**a**–**e**) and Llama-3 8B (**f**–**j**). **a**, **f**, Layer-wise similarity of hidden states, measured as subspace overlap (mean squared cosine of principal angles) between SVD subspaces explaining 95% variance; results are averaged over five runs with 24 demonstrations. Axes index layers. **b**, **g**, Number of principal components (PCs) required to explain 95% of the variance across layers. Shaded areas denote 95% CIs calculated from 10,000 bootstrap resamples across five runs. **c**, **h**, Overlap between GCCA-derived projection matrices across selected layers. **d**–**e**, **i**–**j**, Conceptual subspaces becomes increasingly stable as the number of in-context demonstrations grows. **d**, **i**, Cross-context alignment of representational geometry within the GCCA subspace across demonstration sets, measured by RSA. **e**, **j**, Overlap between GCCA-derived projection matrices across demonstration sets. Axes in **d**–**e** and **i**–**j** indicate the number of demonstrations, with each cell corresponding to a single run.

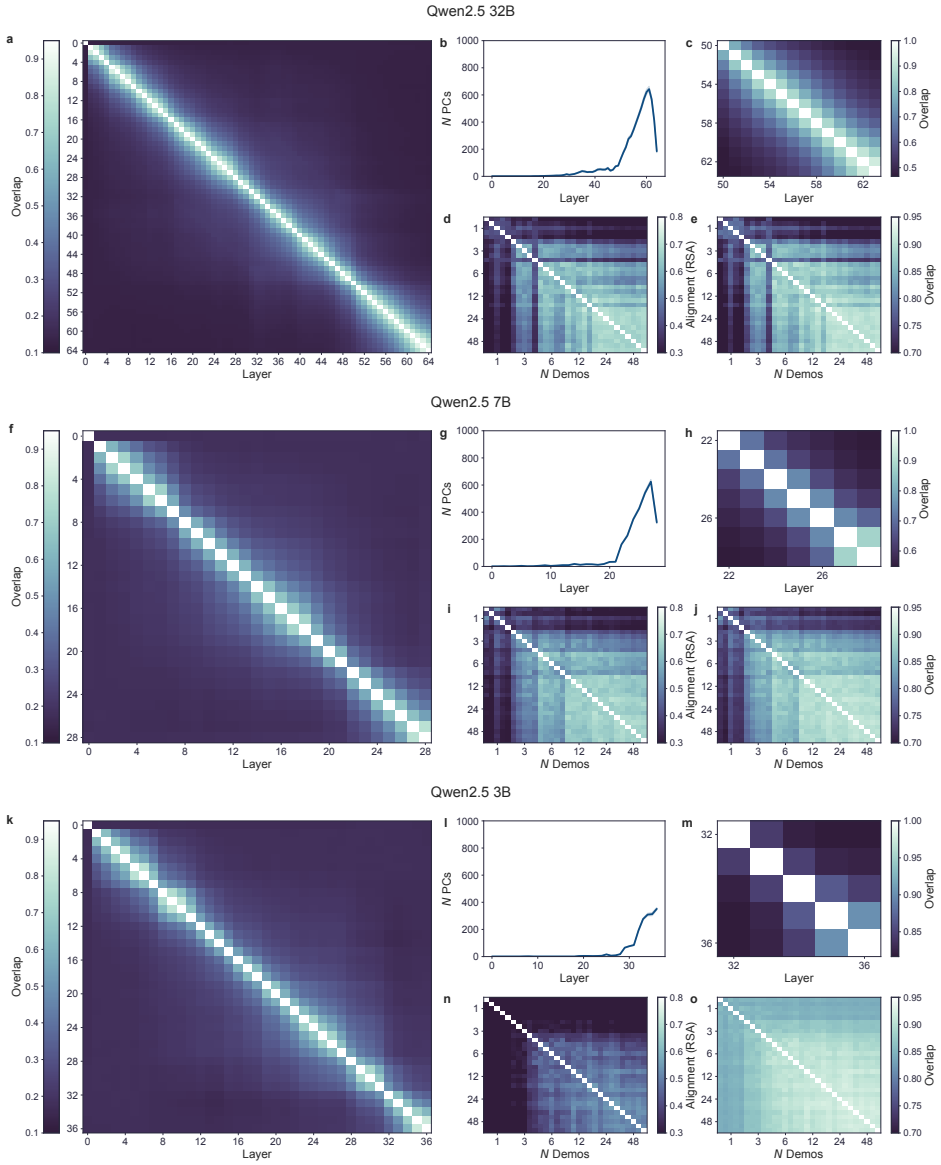

Figure 14: Emergence of conceptual subspaces in Qwen2.5 32B (**a**–**e**), Qwen2.5 7B (**f**–**j**), and Qwen2.5 3B (**k**–**o**). **a**, **f**, **k**, Layer-wise similarity of hidden states, measured as subspace overlap (mean squared cosine of principal angles) between SVD subspaces explaining 95% variance; results are averaged over five runs with 24 demonstrations. Axes index layers. **b**, **g**, **l**, Number of principal components (PCs) required to explain 95% of the variance across layers. Shaded areas denote 95% CIs calculated from 10,000 bootstrap resamples across five runs. **c**, **h**, **m**, Overlap between GCCA-derived projection matrices across selected layers. **d**–**e**, **i**–**j**, **n**–**o**, Conceptual subspaces becomes increasingly stable as the number of in-context demonstrations grows. **d**, **i**, **n**, Cross-context alignment of representational geometry within the GCCA subspace across demonstration sets, measured by RSA. **e**, **j**, **o**, Overlap between GCCA-derived projection matrices across demonstration sets. Axes in **d**–**e**, **i**–**j** and **n**–**o** indicate the number of demonstrations, with each cell corresponding to a single run.

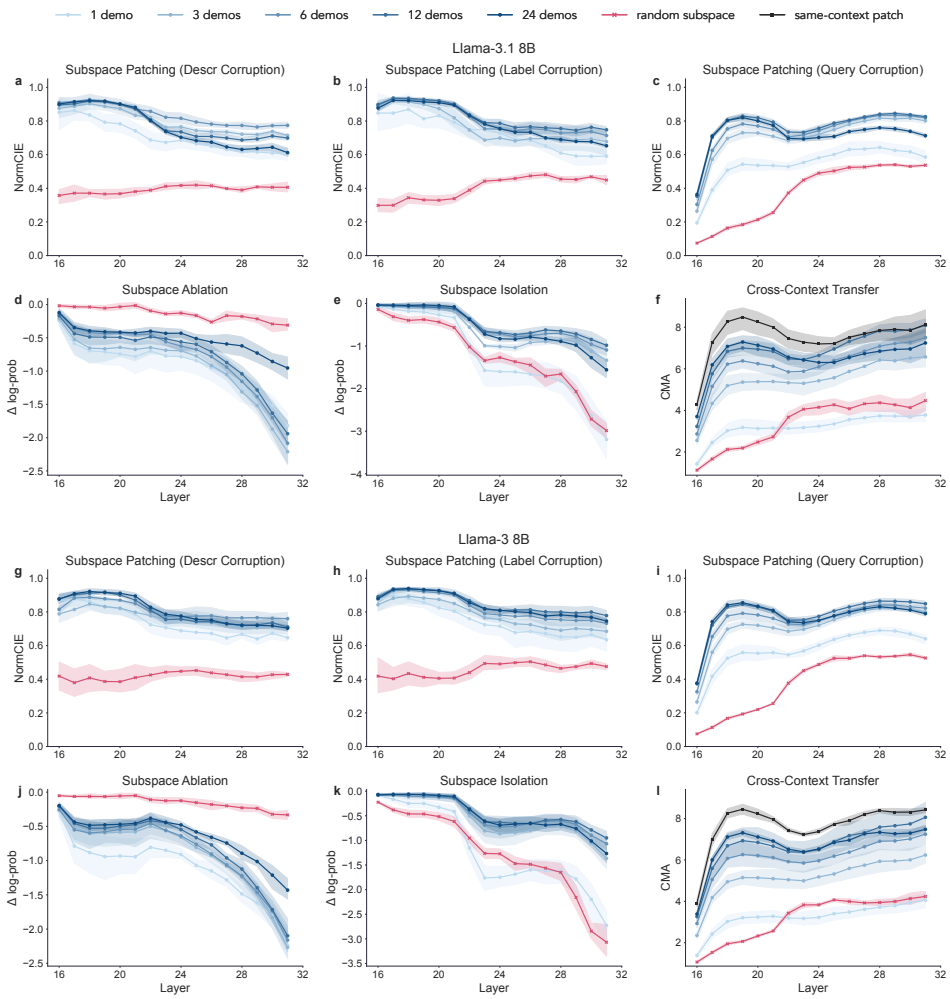

Figure 15: Causal intervention results for Llama-3.1 8B (**a**–**f**) and Llama-3 8B (**g**–**l**). **a**–**c**, **g**–**i**, Activation patching with N demonstrations under three corruption conditions: description (**a**, **g**), label (**b**, **h**) and query (**c**, **i**). Patching the conceptual subspace (blue) is compared against a random-subspace baseline (red). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d**–**e**, **j**–**k**, Subspace necessity and sufficiency tested by ablating the conceptual subspace (**d**, **j**) or isolating it (**e**, **k**). The y-axis denotes the change in log-probability of the correct token. **f**, **l**, Causal effects of cross-context transfer, where the relational structure from a source context is adapted to a target context via an orthogonal transformation. The y-axis reports the causal mediation (CMA) score. Performance is compared against a random-subspace baseline (red) and a same-context patch reference (black) using the conceptual subspace derived from the target context itself. Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

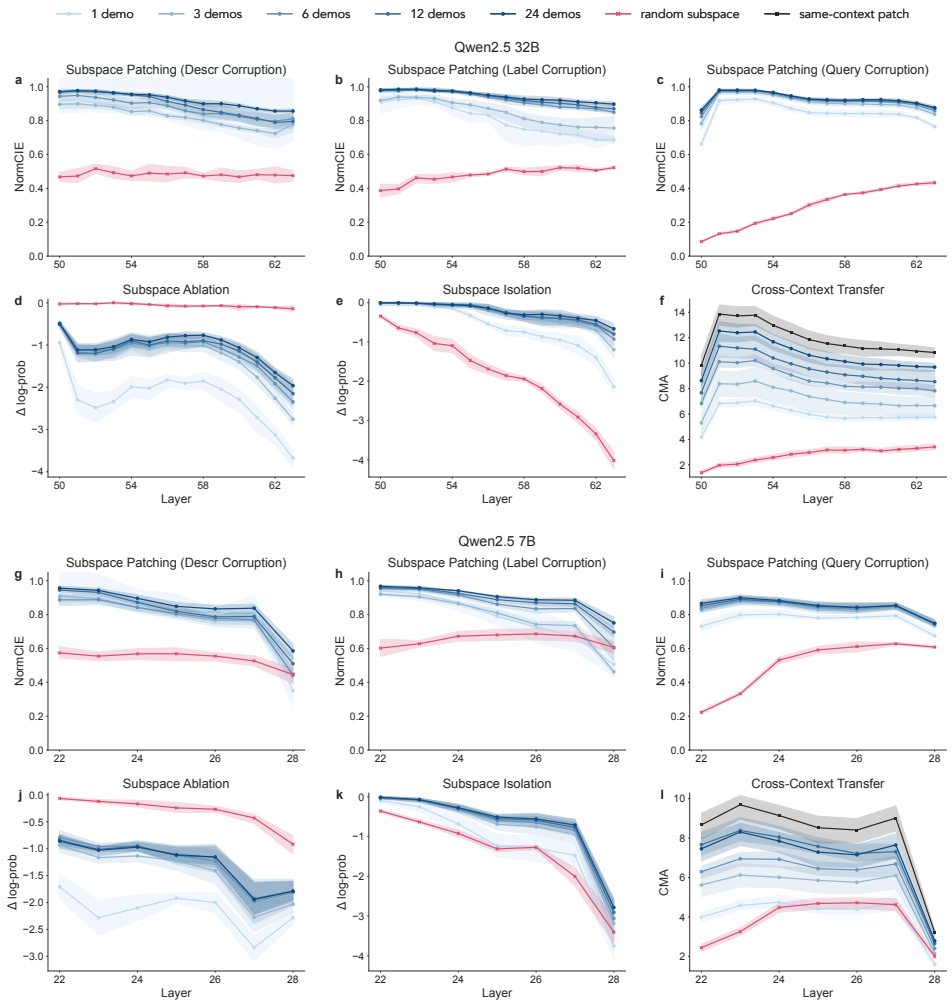

Figure 16: Causal intervention results for Qwen2.5 32B (**a**–**f**) and Qwen2.5 7B (**g**–**l**). **a**–**c**, **g**–**i**, Activation patching with N demonstrations under three corruption conditions: description (**a**, **g**), label (**b**, **h**) and query (**c**, **i**). Patching the conceptual subspace (blue) is compared against a random-subspace baseline (red). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d**–**e**, **j**–**k**, Subspace necessity and sufficiency tested by ablating the conceptual subspace (**d**, **j**) or isolating it (**e**, **k**). The y-axis denotes the change in log-probability of the correct token. **f**, **l**, Causal effects of cross-context transfer, where the relational structure from a source context is adapted to a target context via an orthogonal transformation. The y-axis reports the causal mediation (CMA) score. Performance is compared against a random-subspace baseline (red) and a same-context patch reference (black) using the conceptual subspace derived from the target context itself. Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

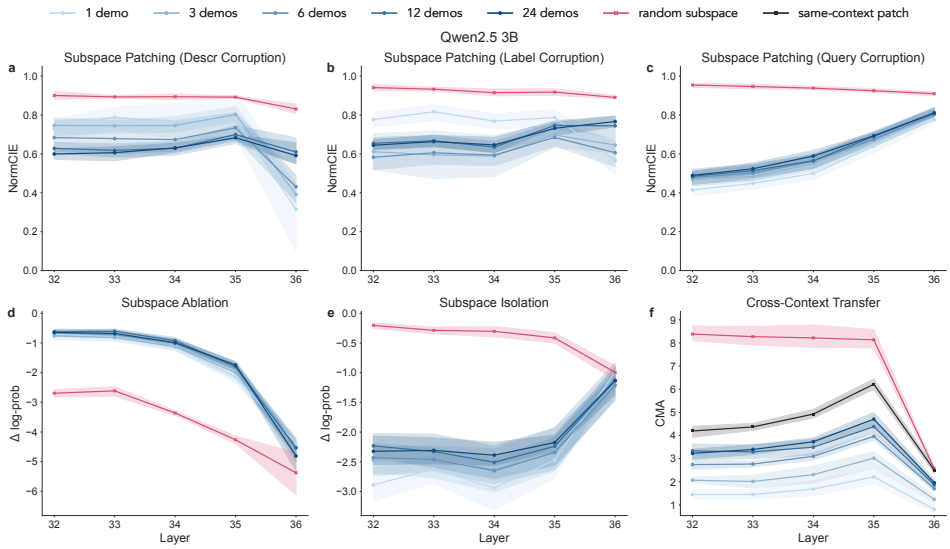

Figure 17: Causal intervention results for Qwen2.5 3B. **a–c**, Activation patching with N demonstrations under three corruption conditions: description (**a**), label (**b**) and query (**c**). Patching the conceptual subspace (blue) is compared against a random-subspace baseline (red). The x-axis indexes layers, and the y-axis shows the normalized causal indirect effect (CIE). **d–e**, Subspace necessity and sufficiency tested by ablating the conceptual subspace (**d**) or isolating it (**e**). The y-axis denotes the change in log-probability of the correct token. **f**, Causal effects of cross-context transfer, where the relational structure from a source context is adapted to a target context via an orthogonal transformation. The y-axis reports the causal mediation (CMA) score. Performance is compared against a random-subspace baseline (red) and a same-context patch reference (black) using the conceptual subspace derived from the target context itself. Shaded regions indicate 95% CIs computed from 10,000 bootstrap resamples across five runs.

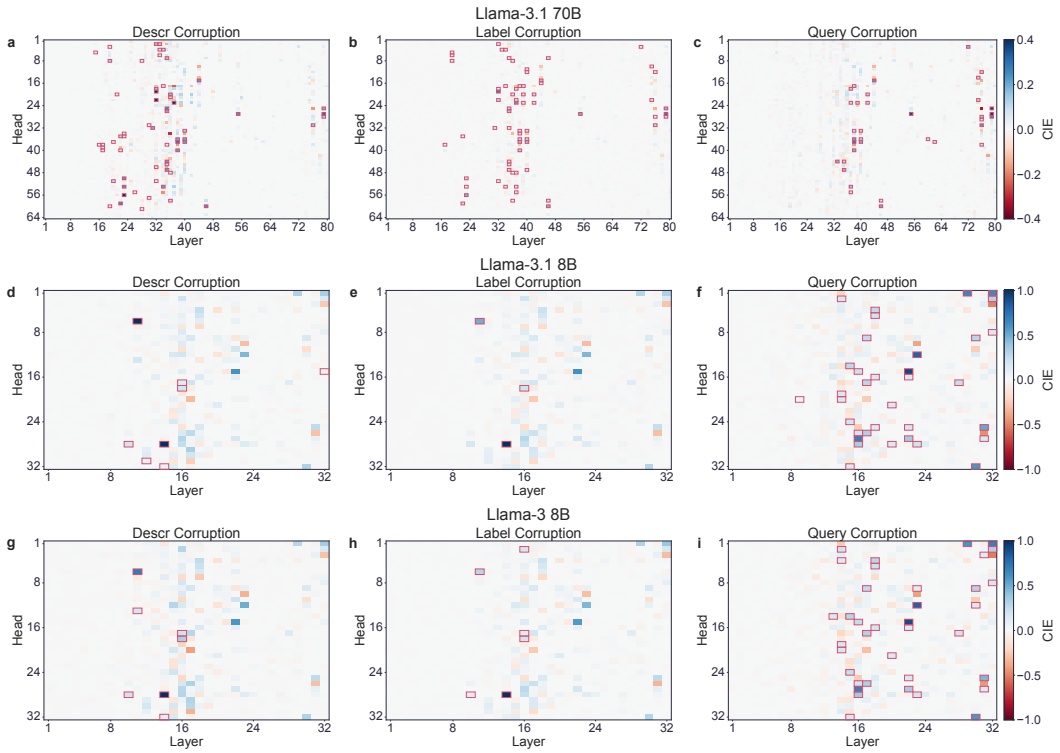

Figure 18: Causal indirect effects (CIEs) of attention heads across Llama-3.1 70B (**a**–**c**), Llama-3.1 8B (**d**–**f**), and Llama-3 8B (**g**–**i**). Within each row, columns represent CIEs under description, label, and query corruption, respectively. The x axis indexes model layers, and the y-axis denotes attention head indices. Bordered cells highlight attention heads with statistically significant effects under the respective corruption conditions.

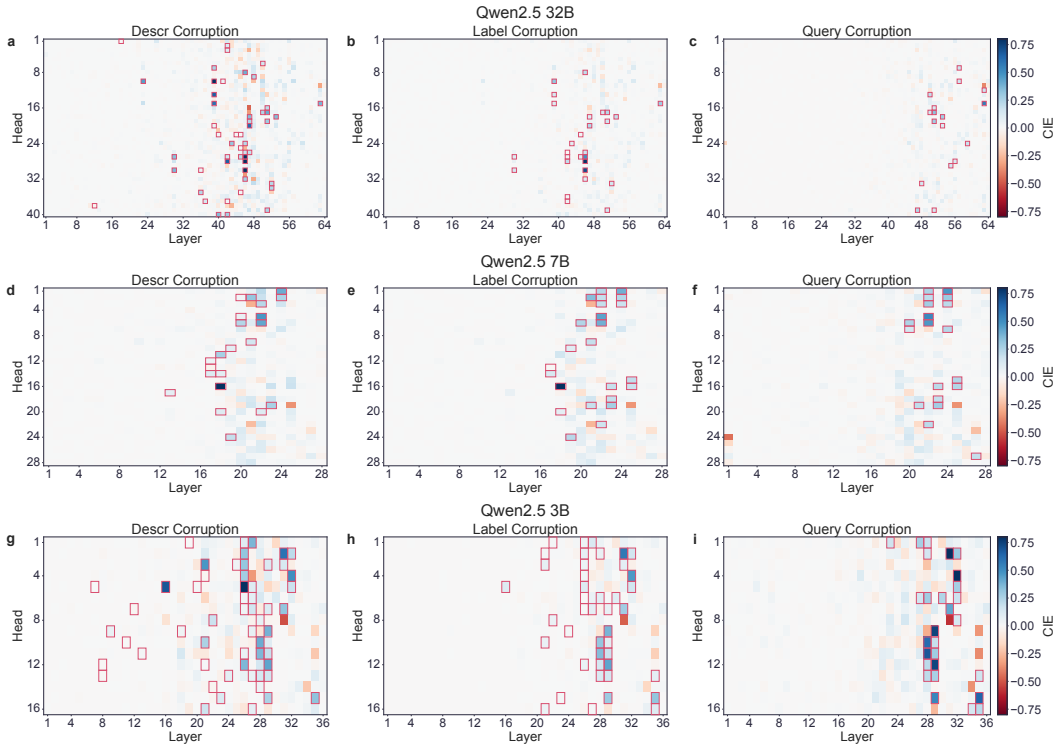

Figure 19: Causal indirect effects (CIEs) of attention heads across Qwen2.5 32B (**a**–**c**), Qwen2.5 7B (**d**–**f**), and Qwen2.5 3B (**g**–**i**). Within each row, columns represent CIEs under description, label, and query corruption, respectively. The x axis indexes model layers, and the y-axis denotes attention head indices. Bordered cells highlight attention heads with statistically significant effects under the respective corruption conditions.

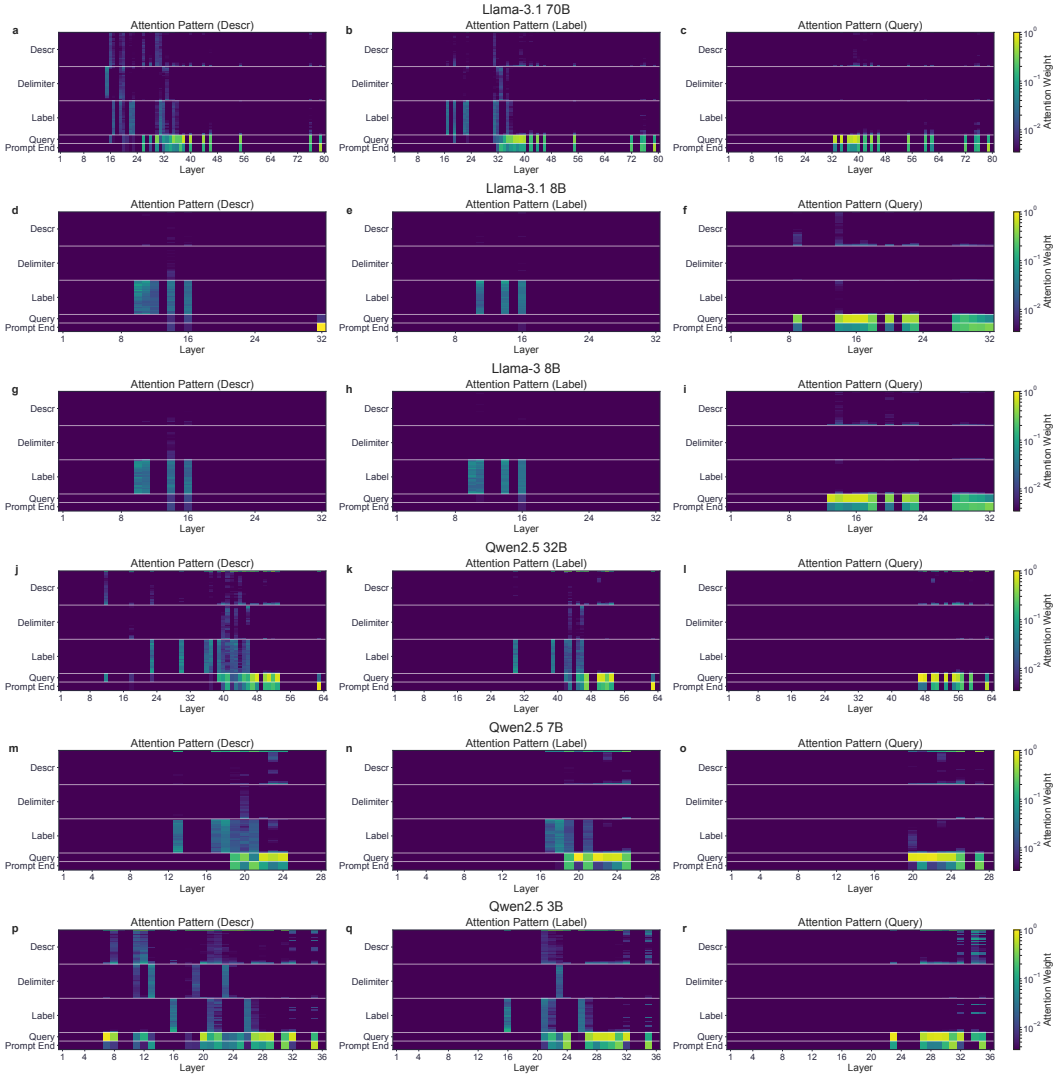

Figure 20: Attention patterns of heads with statistically significant causal indirect effects (CIEs) in Llama-3.1 70B (**a**–**c**), Llama-3.1 8B (**d**–**f**), Llama-3 8B (**g**–**i**), Qwen2.5 32B (**j**–**l**), Qwen2.5 7B (**m**–**o**), and Qwen2.5 3B (**p**–**r**). Within each model (row), columns show the attention patterns of heads identified under description, label, and query corruption, respectively. Within each layer, attention patterns were averaged over heads with statistically significant CIEs. Layers with no significant heads were set to zero. The x-axis indicates layer index, and the y-axis shows attended token spans grouped by source segment. The description, delimiter, and label spans for the 24 in-context demonstrations are ordered top-to-bottom based on their sequence in the prompt. See Figure 21 for attention patterns of individual heads.

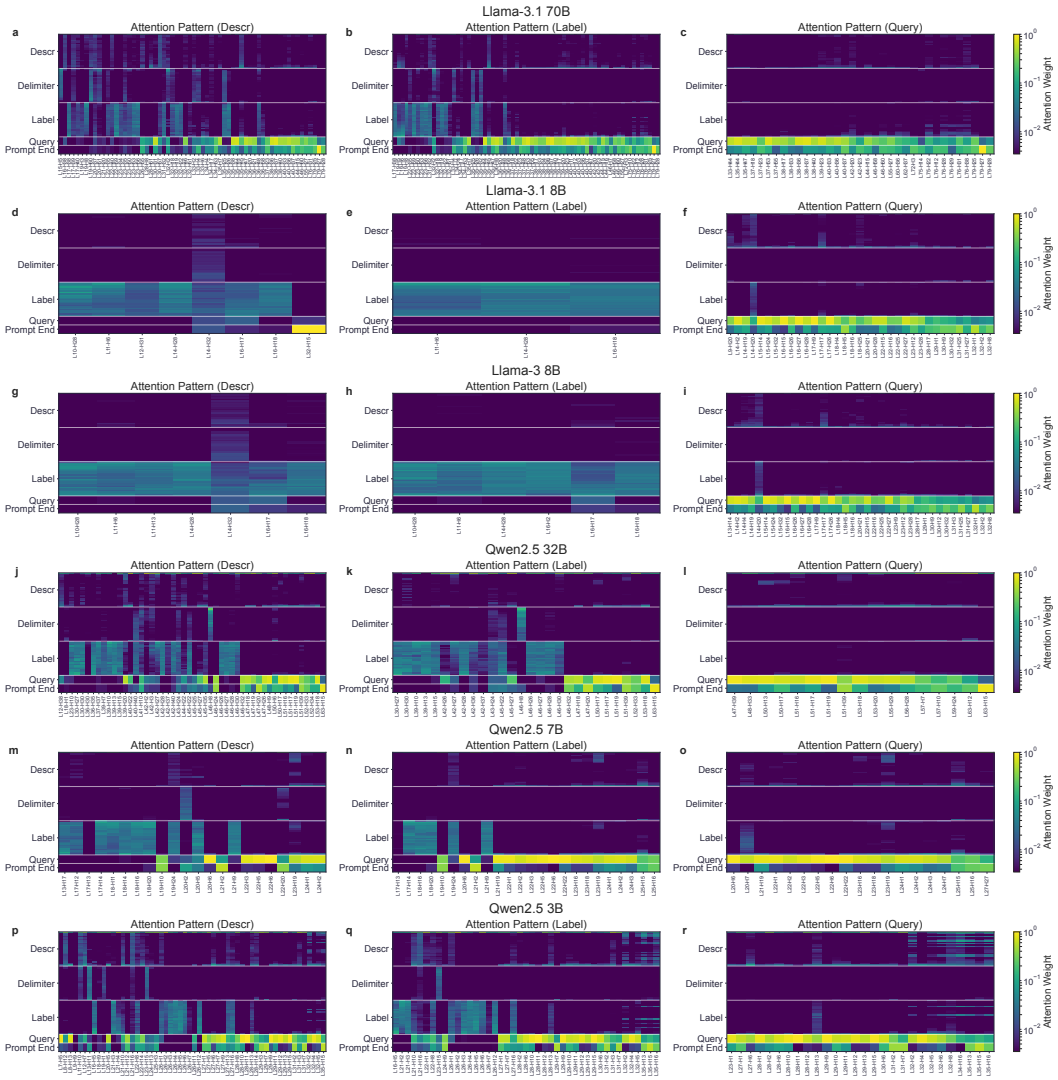

Figure 21: Attention patterns of heads with statistically significant causal indirect effects (CIEs) in Llama-3.1 70B (**a–c**), Llama-3.1 8B (**d–f**), Llama-3 8B (**g–i**), Qwen2.5 32B (**j–l**), Qwen2.5 7B (**m–o**), and Qwen2.5 3B (**p–r**). Within each model (row), columns show the attention patterns of heads identified under description, label, and query corruption, respectively. The x-axis denotes attention heads by their (layer, head) index, and the y-axis shows attended token spans grouped by source segment. The description, delimiter, and label spans for the 24 in-context demonstrations are ordered top-to-bottom based on their sequence in the prompt.

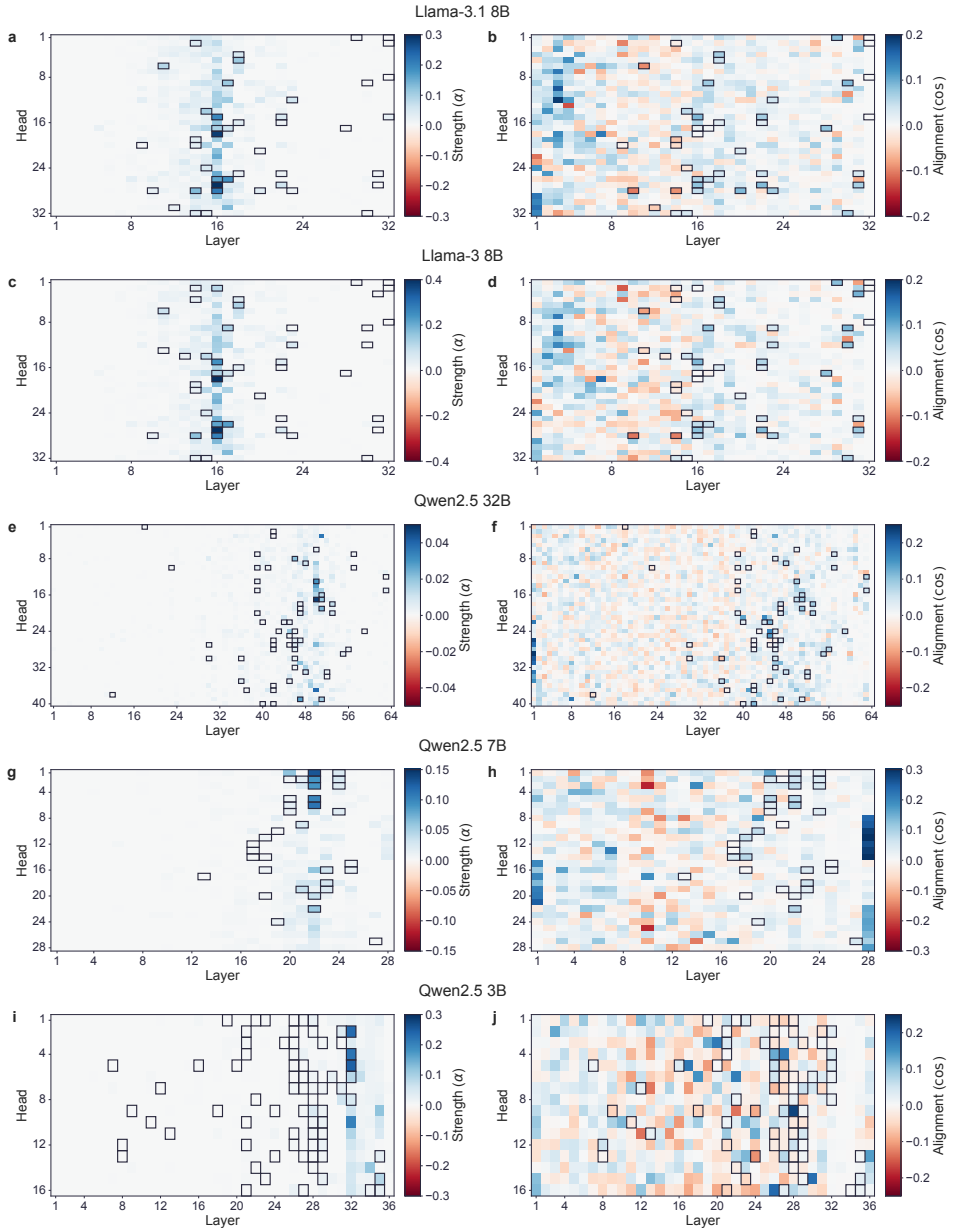

Figure 22: Contribution of attention heads to the conceptual subspace across Llama-3.1 8B (**a**–**b**), Llama-3 8B (**c**–**d**), Qwen2.5 32B (**e**–**f**), Qwen2.5 7B (**g**–**h**), and Qwen2.5 3B (**i**–**j**). For each model, the left column reports the contribution strength (α) of each head to the conceptual subspace, and the right column shows the directional alignment (cosine similarity) between head outputs and the subspace. Bordered cells highlight attention heads with statistically significant causal indirect effects (CIEs) under any corruption condition.

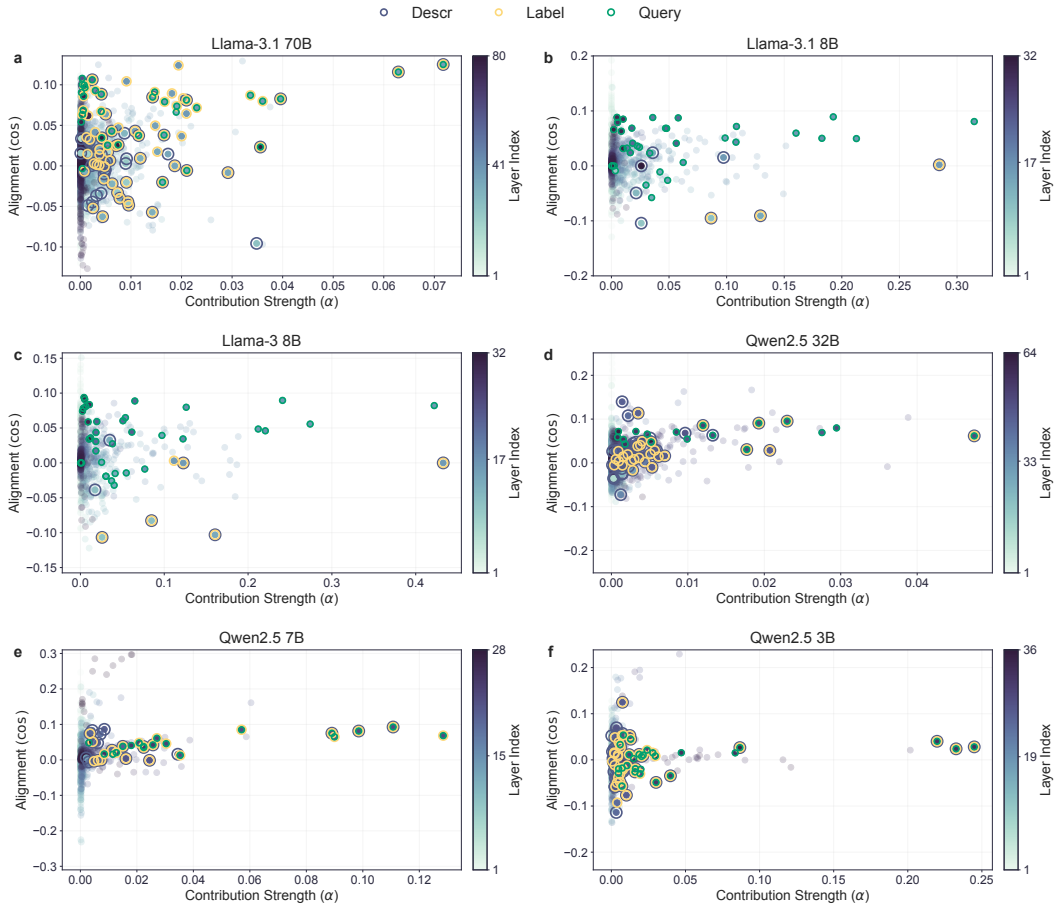

Figure 23: Contribution–alignment relationship for attention heads across Llama-3.1 70B (**a**), Llama-3.1 8B (**b**), Llama-3 8B (**c**), Qwen2.5 32B (**d**), Qwen2.5 7B (**e**), and Qwen2.5 3B (**f**). Each point corresponds to an attention head, with the x-axis denoting its contribution strength to the conceptual subspace (α) and the y-axis showing its directional alignment (cosine similarity) with that subspace. Points are colored by layer index. Outlines mark heads with statistically significant causal indirect effects (CIEs); the outline color indicates the corruption condition(s) (description, label, query) under which significance was detected.