Emergent Structured Representations Support In-Context Inference in LLMs

Ningyu Xu[1,2], **Qi Zhang**[1,3], **Xipeng Qiu**[1,3,*], **Xuanjing Huang**[1,3,*]

[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]Institute of Modern Languages and Linguistics, Fudan University
[3]Institute of Trustworthy Embodied Artificial Intelligence, Fudan University
nyxu22@m.fudan.edu.cn, {xpqiu,xjhuang}@fudan.edu.cn

Key Takeaways

LLMs organize task-relevant information into a latent subspace whose relational structure persists across contexts and causally supports in-context inference.

- This subspace provides a promising interface for interpretability and steering beyond individual model components.
- Reliance on the subspace is emergent, helping distinguish abstract inference from shallow heuristics.
- Our methodological framework can be generalized to study internal computation across models and tasks.

The Question

- Whether large language models (LLMs) perform structured inference or merely exploit surface-level statistical associations remains a central debate.
- Here, rather than dissecting individual attention heads or neurons, we isolate **a shared latent subspace** in the residual stream and test its functional role in inference.

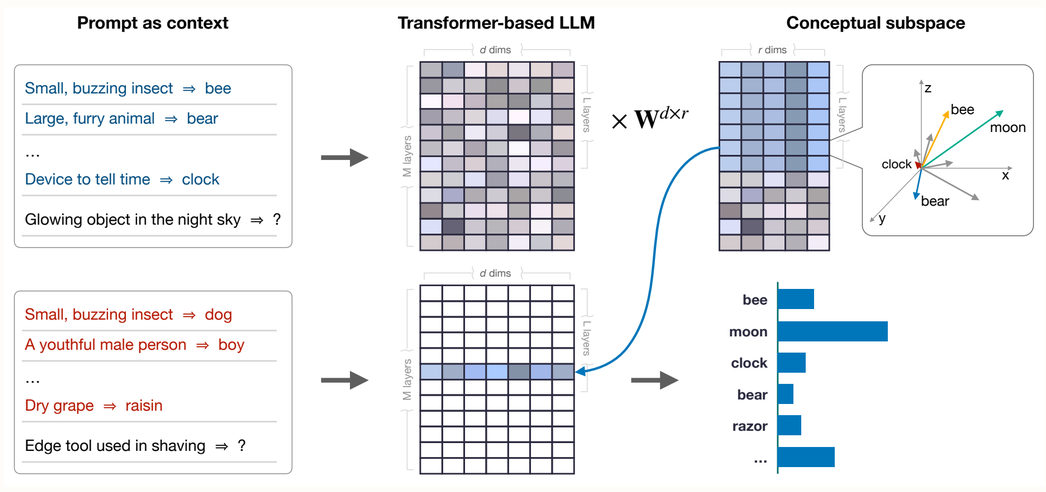

Figure 1: The concept inference task is shown as an illustrative example. In middle-to-late layers, LLMs construct a latent subspace whose relational structure persists across contexts and causally supports inference.

A Shared Latent Subspace Emerges

- Layer-wise SVD reveals high inter-layer overlap in middle-to-late layers, suggesting the emergence of a shared subspace.
- GCCA identifies a common subspace across these layers, with near-unity alignment and smoothly varying projections across depth.
- With more demonstrations, the subspaces become increasingly aligned across contexts, both in their relational geometry and in their overlap, indicating a progressively more coherent and context-invariant structure.

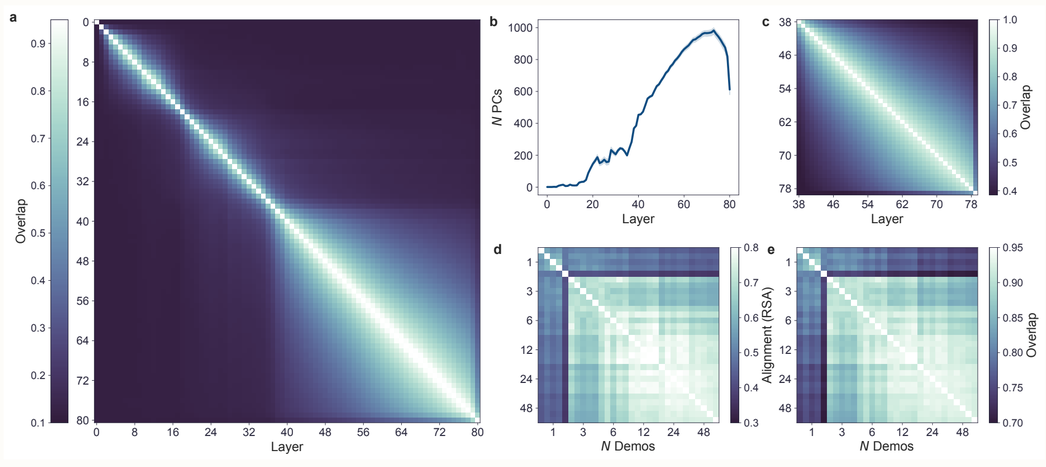

Figure 2: A shared latent subspace emerges in the middle to late layers of Llama-3.1 70B. **a**, Layer-wise similarity of hidden states. **b**, Number of PCs needed to explain 95% variance at each layer. **c**, Overlap between GCCA-derived projection matrices across selected layers. **d–e**, The subspace becomes increasingly stable with more in-context demonstrations.

The Subspace Causally Mediates Inference

- Under corrupted context, patching the subspace into middle layers largely restores performance, showing that the subspace mediates task-relevant information.
- Ablating the subspace causes a marked performance drop, while isolating it preserves near-original performance across most layers.
- The subspace can be transferred across contexts via an orthogonal transformation, suggesting a stable relational structure that generalizes across contexts.

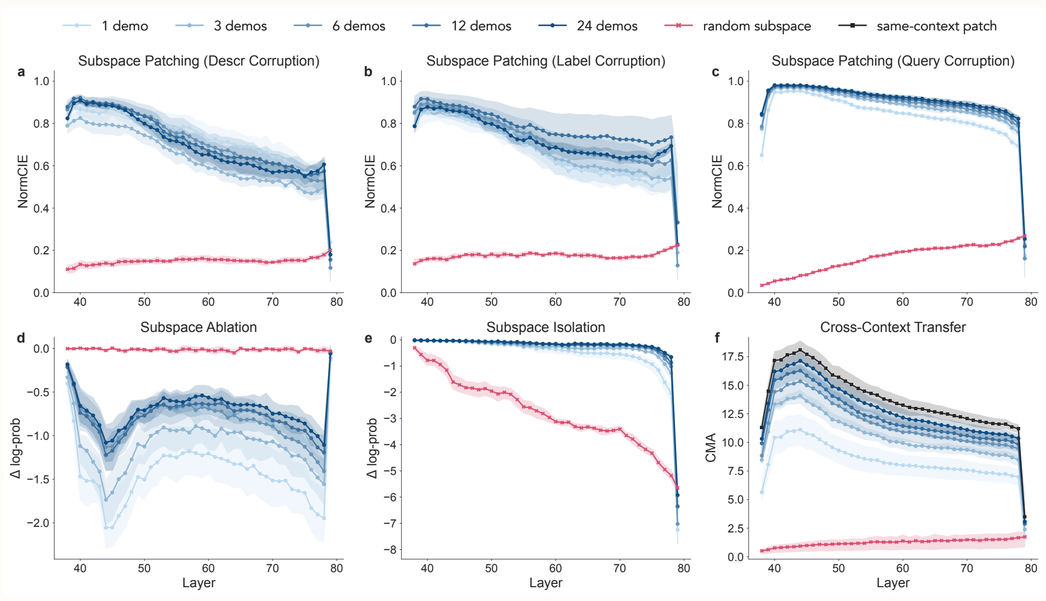

Figure 3: The latent subspace causally mediates inference, as shown by patching (**a–c**), ablation (**d**), isolation (**e**), and cross-context transfer (**f**).

Contextual Information Shapes the Subspace

- Activation patching and attention mass attribution reveal a progression of information integration, shifting from demonstration-centric to query-centric processing.
- Significant heads align more strongly with the latent subspace, particularly in early-to-middle layers.
- Our results suggest a two-stage process: early-to-middle layers construct and stabilize the latent subspace, while later layers operate within it to support inference.

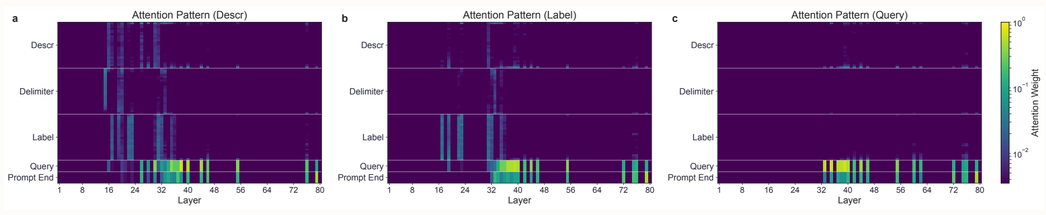

Figure 4: Attention patterns of attention heads with statistically significant CIEs identified under description (**a**), label (**b**), and query corruption (**c**). The description, delimiter, and label spans for the 24 in-context demonstrations are ordered top-to-bottom based on their sequence in the prompt.