Novelty Assessment Report

Paper: Latent Concept Disentanglement in Transformer-based Language Models
PDF URL: https://openreview.net/pdf?id=k3SEVOW2Dg
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

When large language models (LLMs) use in-context learning (ICL) to solve a new task, they must infer latent concepts from demonstration examples. This raises the question of whether and how transformers represent latent structures as part of their computation. Our work experiments with several controlled tasks, studying this question using mechanistic interpretability. First, we show that in transitive reasoning tasks with a latent, discrete concept, the model successfully identifies the latent concept and does step-by-step concept composition. This builds upon prior work that analyzes single-step reasoning. Then, we consider tasks parameterized by a latent numerical concept. We discover low-dimensional subspaces in the model's representation space, where the geometry cleanly reflects the underlying parameterization. Overall, we show that small and large models can indeed disentangle and utilize latent concepts that they learn in-context from a handful of abbreviated demonstrations.

Core Task Landscape

This paper addresses: **Latent Concept Disentanglement in In-Context Learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Latent Variable Inference and Bayesian Perspectives in ICL**
- **Latent Concept Representation and Disentanglement Mechanisms**
- **Prompt Design and Optimization for ICL**
- **Few-Shot Learning with Vision-Language and Multimodal Models**
- **Domain-Specific Few-Shot Applications**
- **Representation Learning for Robustness and Generalization**
- **Specialized ICL Applications and Extensions**

Complete Taxonomy Tree

- Latent Concept Disentanglement in In-Context Learning Survey Taxonomy
- Latent Variable Inference and Bayesian Perspectives in ICL
  - Bayesian and Generative Latent Variable Models for ICL (2 papers)
  - [1] Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning (Wang Xinyi, 2023) View paper
  - [2] Does learning the right latent variables necessarily improve in-context learning? (Mittal, 2024) View paper
  - Variational and Probabilistic Prompt Learning (3 papers)
  - [4] Prompting language-informed distribution for compositional zero-shot learning (Wentao Bao, 2024) View paper
  - [14] VaMP: Variational Multi-Modal Prompt Learning for Vision-Language Models (Silin Cheng, 2025) View paper
  - [19] Prompt Based CVAE Data Augmentation for Few-Shot Intention Detection (Junhao Xue, 2024) View paper
- Latent Concept Representation and Disentanglement Mechanisms
  - Mechanistic Interpretability of Latent Concept Encoding ★ (3 papers)
  - [0] Latent Concept Disentanglement in Transformer-based Language Models (Anon et al., 2026) View paper
  - [3] Provably transformers harness multi-concept word semantics for efficient in-context learning (Dake Bu, 2024) View paper
  - [47] From Context to Concept: Concept Encoding in In-Context Learning (J Song, n.d.) View paper
  - Latent Space Geometry and Semantic Clustering (3 papers)
  - [6] Vocabulary-Defined Semantics: Latent Space Clustering for Improving In-Context Learning (Gu Jian, 2024) View paper
  - [13] The representation landscape of few-shot learning and fine-tuning in large language models (Alessio Ansuini, 2024) View paper
  - [17] NN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference (B Xu, 2023) View paper
  - Disentanglement via Self-Supervision and Weak Supervision (3 papers)
  - [29] Disentangling Latent Shifts of In-Context Learning Through Self-Training (JukiÄ□, 2024) View paper
  - [45] Enhancing the Stability of LLM-based Speech Generation Systems through Self-Supervised Representations (Saez-Trigueros, 2024) View paper
  - Task Recognition versus Task Learning Decomposition (2 papers)
  - [18] Can in-context learners learn a reasoning concept from demonstrations? (Å tefÃ¡nik, 2023) View paper
  - [36] What In-Context Learning â□□Learnsâ□□ In-Context: Disentangling Task Recognition and Task Learning (Chen, 2023) View paper
- Prompt Design and Optimization for ICL
  - Continuous and Soft Prompt Tuning (3 papers)
  - [20] Meta-learning the difference: preparing large language models for efficient adaptation (Hou, 2022) View paper

- [38] SharPT: Shared Latent Space Prompt Tuning (Bo Pang, 2023) View paper
- [48] Latent Prompt Tuning for Text Summarization (Zhang Yubo, 2022) View paper
- Template and Demonstration Engineering (3 papers)
- [26] Decoupling Template Bias in CLIP: Harnessing Empty Prompts for Enhanced Few-Shot Learning (Zhenyu Zhang, 2025) View paper
- [41] StablePT : Towards Stable Prompting for Few-shot Learning via Input Separation (Xiaoming Liu, 2024) View paper
- [46] Embroid: Unsupervised Prediction Smoothing Can Improve Few-Shot Classification (Guha, 2023) View paper
- Prompt Learning for Continual and Transfer Learning (3 papers)
- [25] Generating Prompts in Latent Space for Rehearsal-free Continual Learning (Chengyi Yang, 2024) View paper
- [40] Prompt Transfer for Dual-Aspect Cross-Domain Cognitive Diagnosis (Fei Liu, 2025) View paper
- [42] Latent Domain Prompt Learning for Vision-Language Models (Li Zhixing, 2025) View paper

- Few-Shot Learning with Vision-Language and Multimodal Models
  - Vision-Language Prompt Learning and Disentanglement (3 papers)
  - [7] Locoop: Few-shot out-of-distribution detection via prompt learning (Miyai, 2023) View paper
  - [31] Learning from Orthogonal Space with Multimodal Large Models for Generalized Few-shot Segmentation (Xiaojie Zhou, 2025) View paper
  - [39] Disentangled Prompt Learning for Transferable, Multimodal, Few-Shot Image Classification (John Yang, 2024) View paper
  - Few-Shot Segmentation with Semantic and Structural Guidance (2 papers)
  - [21] SANSA: Unleashing the Hidden Semantics in SAM2 for Few-Shot Segmentation (Trivigno, 2025) View paper
  - [24] Towards Robust Few-shot Point Cloud Semantic Segmentation (Xu Yating, 2023) View paper
  - Few-Shot Learning with Intra-Class Variation and Cluster Prompts (2 papers)
  - [9] MICS: Midpoint Interpolation to Learn Compact and Separated Representations for Few-Shot Class-Incremental Learning (Yuho Jeong, 2024) View paper
  - [11] Exploring intra-class variation factors with learnable cluster prompts for semi-supervised image synthesis (Yunfei Zhang, 2023) View paper
- Domain-Specific Few-Shot Applications
  - Few-Shot Action and Video Recognition (2 papers)
  - [8] Beyond Label Semantics: Language-Guided Action Anatomy for Few-shot Action Recognition (Yao, 2025) View paper
  - [34] TFRS: A task-level feature rectification and separation method for few-shot video action recognition. (Yanfei Qin, 2024) View paper
  - Audio and Speech Few-Shot Learning (2 papers)
  - [5] Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning (Zhao Jun-chuan, 2025) View paper
  - [49] Few-Shot Musical Source Separation (Wang Yu, 2022) View paper
  - Few-Shot Learning in Specialized Domains (3 papers)
  - [27] PBLF: Prompt based learning framework for cross-modal recipe retrieval (Jialiang Sun, 2022) View paper
  - [32] Prompt-SID: Learning Structural Representation Prompt via Latent Diffusion for Single Image Denoising (Huaqiu Li, 2025) View paper
  - [33] Context-aware recommendation using hierarchical attention and positional encoding based on multiple item attributes (CARA) (Hadise Vaghari, 2025) View paper
- Representation Learning for Robustness and Generalization
  - Representation Compactness and Separability (1 papers)
  - [35] JLCSR: Joint Learning of Compactness and Separability Representations for Few-Shot Classification (Sai Yang, 2023) View paper
  - Robustness and Stability in ICL Representations (2 papers)
  - [12] Evaluating Generalization and Representation Stability in Small LMs via Prompting (R Raja, 2025) View paper
  - [37] Enhancing robustness and interpretability in natural language processing through representation learning (Yan, 2024)
  - Fairness and Bias Mitigation in ICL (1 papers)
  - [44] Fair In-Context Learning via Latent Concept Variables (Van, 2024) View paper
- Specialized ICL Applications and Extensions
  - Hierarchical and Structured Prediction in ICL (2 papers)
  - [15] Think-to-talk or talk-to-think? when llms come up with an answer in multi-step arithmetic reasoning (K Kudo, 2024) View paper
  - [23] A circuit for predicting hierarchical structure in-context in Large Language Models (Saanum, 2025) View paper
  - Latent Concept Discovery and Instruction-Following (1 papers)
  - [28] Latent Factor Models Meets Instructions: Goal-conditioned Latent Factor Discovery without Task Supervision (Xie, 2025) View paper
  - Specialized Applications of Latent Representations (5 papers)
  - [10] Cultural Alignment in Large Language Models Using Soft Prompt Tuning (Ferianc, 2025) View paper
  - [16] Rethinking reinforcement learning for recommendation: A prompt perspective (X Xin, 2022) View paper
  - [22] Learn the global prompt in the low-rank tensor space for heterogeneous federated learning. (Lele Fu, 2025) View paper
  - [43] Exploring Challenges in Applying Foundation and Generative Models in AI (Arslan, 2023) View paper
  - [50] Measuring Effects of Steered Representation in Large Language Models (B Park, n.d.) View paper

Narrative

Core task: latent concept disentanglement in in-context learning. The field explores how large language models and other foundation models learn to separate and manipulate underlying conceptual factors when presented with few-shot examples. The taxonomy organizes this landscape into several major branches. Latent Variable Inference and Bayesian Perspectives examine how models implicitly perform probabilistic reasoning over hidden variables, with works like LLMs Latent Variables[1] and Right Latent Variables[2] investigating the theoretical underpinnings of this process. Latent Concept Representation and Disentanglement Mechanisms focus on the internal encoding and separation of concepts, including mechanistic interpretability studies that probe how models represent distinct semantic dimensions. Prompt Design and Optimization branches address how to craft inputs that elicit desired disentangled behaviors, while Few-Shot Learning with Vision-Language and Multimodal Models extends these ideas beyond text. Domain-Specific Few-Shot Applications and Representation Learning for Robustness and Generalization branches tackle practical deployment challenges, and Specialized ICL Applications cover niche extensions of the core paradigm.

Particularly active lines of work contrast mechanistic interpretability approaches—which dissect internal representations to understand how concepts are encoded—with methods that optimize prompts or latent spaces to achieve better disentanglement in practice. Some

studies emphasize discovering interpretable latent structures through careful prompt engineering or meta-learning, while others focus on robustness and stability of learned representations across distribution shifts. The original paper, Latent Concept Disentanglement[0], sits within the Mechanistic Interpretability of Latent Concept Encoding cluster, where it likely investigates how transformer architectures internally separate conceptual factors during in-context learning. This positions it closely alongside Multi-Concept Semantics[3], which examines how models handle multiple interacting concepts, and Context to Concept[47], which explores the transformation from contextual examples to abstract concept representations. The emphasis here is on understanding the internal machinery rather than purely optimizing external performance, addressing open questions about what disentangled structures emerge naturally versus what must be explicitly induced.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Provably transformers harness multi-concept word semantics for efficient in-context learning

Authors: Dake Bu, Andi Han, Wei Huang, Atsushi Nitanda, Taiji Suzuki, et al. (7 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ of these models is their in-context learning (ICL) capacity [5], which allows them to learn from a few â¦ concepts, grounded in the concept-specific linear latent geometry observed in LLMs. â¦

Relationship Analysis

Both papers belong to the mechanistic interpretability category, studying how transformers encode and manipulate latent concepts during in-context learning. The original paper focuses on empirical mechanistic analysis of pre-trained models (Gemma-2) using activation patching to reveal discrete latent concept composition in multi-hop reasoning and geometric structure in numerical tasks, while the candidate paper provides theoretical analysis of training dynamics for transformers on concept-specific sparse coding tasks, proving exponential convergence and explaining how multi-concept semantic geometry enables efficient ICL. The key difference is that the original paper emphasizes empirical discovery of existing circuits in trained models, whereas the candidate paper offers provable guarantees about how these representations emerge during training.

---

2. From Context to Concept: Concept Encoding in In-Context Learning

Authors: J Song, S Han, J Gore, P Agrawal | **URL**: View paper

Abstract

â¦ We observe that earlier layers of the model learn to encode the latent concept, whereas â¦ latent concept z that links inputs x to outputs y. For instance, in an ICL task where latent concept â¦

Relationship Analysis

Both papers belong to the Mechanistic Interpretability of Latent Concept Encoding category, using mechanistic interpretability techniques to understand how transformers encode and manipulate latent concepts during in-context learning. They overlap in studying how models disentangle latent concepts through separable representations and use activation patching to establish causal relationships between concept representations and ICL performance. The key difference is that the original paper focuses on multi-hop reasoning with discrete latent concepts (countries, companies) and continuous numerical parameterizations (add-k, circular trajectories), while the candidate paper emphasizes the coupled emergence of concept encoding-decoding mechanisms during training on synthetic tasks and introduces Concept Decodability as a predictive metric for ICL performance across POS tagging and bitwise arithmetic tasks.

Contributions Analysis

Overall novelty summary. The paper investigates how transformers encode and disentangle latent concepts during in-context learning, using mechanistic interpretability to analyze internal representations. It resides in the 'Mechanistic Interpretability of Latent Concept Encoding' leaf, which contains only three papers total (including this one and two siblings: Multi-Concept Semantics and Context to Concept). This is a relatively sparse research direction within the broader taxonomy of 50 papers, suggesting the mechanistic analysis of latent concept encoding remains an emerging area compared to more crowded branches like prompt optimization or vision-language few-shot learning.

The taxonomy reveals several neighboring research directions. The sibling leaf 'Latent Space Geometry and Semantic Clustering' (three papers) explores geometric structures in representations but without the mechanistic focus. The parent branch also includes 'Disentanglement via Self-Supervision' (three papers) and 'Task Recognition versus Task Learning Decomposition' (two papers), which address disentanglement through training objectives rather than interpretability probes. Adjacent branches like 'Bayesian and Generative Latent Variable Models' (two papers) approach latent concepts through probabilistic frameworks, while 'Prompt Design and Optimization' (nine papers across three leaves) focuses on external manipulation rather than internal understanding.

Among 26 candidates examined across three contributions, none were found to clearly refute any claim. The first contribution (two-hop reasoning with latent concepts) examined 10 candidates with zero refutations; the second (low-dimensional geometric structure for numerical tasks) also examined 10 with zero refutations; the third (causal/correlational methodology) examined 6 with zero refutations. This suggests that within the limited search scope, the specific combination of mechanistic interpretability, step-by-step concept composition in transitive reasoning, and geometric analysis of numerical task parameters appears relatively unexplored in prior work.

Based on the top-26 semantic matches examined, the work appears to occupy a distinct position combining mechanistic analysis with controlled task design. The sparse population of its taxonomy leaf and the absence of refuting candidates within the search scope suggest novelty, though this assessment is constrained by the limited literature coverage. A more exhaustive search might reveal additional related work in mechanistic interpretability or geometric representation analysis that was not captured by semantic similarity to this paper's framing.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Mechanistic evidence for latent concept disentanglement in two-hop reasoning tasks

Description: The authors demonstrate that large language models performing two-hop reasoning first resolve an intermediate bridge entity (such as a country) using sparse attention heads, then compose this representation with output concepts to produce the final answer, rather than taking shortcuts directly from source to target.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Relmkg: reasoning with pre-trained language models and knowledge graphs for complex question answering

URL: View paper

Brief Assessment

Relmkg[59] focuses on knowledge graph reasoning for question answering, not on mechanistic interpretability of transformer attention heads or latent concept representations in language models.

---

2. Mechanics of Next Token Prediction with Self-Attention

URL: View paper

Brief Assessment

Next Token Mechanics[56] focuses on the fundamental mechanics of single self-attention layers in next-token prediction through hard retrieval and soft composition, not on multi-hop reasoning with latent bridge entities or concept disentanglement in complex reasoning tasks.

---

3. Do Large Language Models Perform Latent Multi-Hop Reasoning without Exploiting Shortcuts?

URL: View paper

Brief Assessment

Latent Multi-Hop Reasoning[55] focuses on evaluating whether LLMs can answer multi-hop queries without shortcuts (co-occurrence exploitation), not on mechanistic analysis of how models internally resolve intermediate concepts through sparse attention heads and step-by-step composition.

---

4. Attention as a Hypernetwork

URL: View paper

Brief Assessment

Attention Hypernetwork[58] focuses on compositional generalization through hypernetwork reformulation of attention mechanisms in abstract reasoning tasks (Raven's matrices), not on mechanistic analysis of two-hop factual reasoning with intermediate bridge entity resolution.

---

5. Latent cascade synthesis: Investigating iterative pseudo-contextual scaffold formation in contemporary large language models

URL: View paper

Brief Assessment

Latent Cascade Synthesis[52] provides only fragmentary text snippets that mention 'rigorous stepwise mapping' and 'compositional capacities' but lacks sufficient detail to assess whether it demonstrates prior work on two-hop reasoning with sparse attention heads resolving intermediate bridge entities before final answer composition.

---

6. Understanding Multi-compositional learning in Vision and Language models via Category Theory

URL: View paper

Brief Assessment

Multi-Compositional Category Theory[51] focuses on compositional learning in vision-language models using category theory, not on mechanistic interpretability of two-hop reasoning in transformers. The candidate addresses compositional zero-shot learning with attributes and primitives, while the original analyzes step-by-step latent concept resolution in language models.

---

7. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision

URL: View paper

Brief Assessment

Neuro-Symbolic Concept Learner[60] focuses on learning visual concepts and semantic parsing from images and questions, not on analyzing internal transformer mechanisms for two-hop reasoning or latent concept identification in language models.

---

8. How does Transformer Learn Implicit Reasoning?

URL: View paper

Brief Assessment

Implicit Reasoning Learning[54] focuses on training transformers from scratch in symbolic environments to study implicit reasoning emergence, while the original paper analyzes pre-trained LLMs using mechanistic interpretability on factual knowledge tasks with explicit bridge entity identification.

---

9. Understanding and patching compositional reasoning in llms

URL: View paper

Brief Assessment

Patching Compositional Reasoning[53] appears to focus on patching techniques for compositional reasoning, but the provided candidate context is too limited (only ellipses and fragments) to determine whether it addresses two-hop reasoning with sparse attention heads resolving intermediate bridge entities. The original paper's specific mechanistic claims about attention head sparsity and step-by-step composition cannot be evaluated against this candidate.

---

10. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization

URL: View paper

Brief Assessment

Grokked Transformers[57] focuses on implicit reasoning through grokking (extended training beyond overfitting) in composition and comparison tasks, not on mechanistic analysis of latent concept identification in two-hop reasoning with world knowledge.

Contribution 2: Discovery of low-dimensional geometric structure in task representations for numerical tasks

Description: For tasks with continuous latent parameters (such as add-k or circular trajectories), the authors find that task vectors lie on smooth low-dimensional manifolds whose geometry mirrors the latent parameter space, enabling interpolation and steering of model behavior.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Manifold-based verbalizer space re-embedding for tuning-free prompt-based classification

URL: View paper

Brief Assessment

Verbalizer Space Re-Embedding[74] focuses on manifold-based re-embedding of verbalizer embeddings for prompt-based classification tasks, not on discovering geometric structures in task representations for numerical in-context learning tasks with continuous parameters.

2. Exploring universal intrinsic task subspace for few-shot learning via prompt tuning

URL: View paper

Brief Assessment

Universal Task Subspace[73] focuses on finding low-dimensional subspaces for prompt tuning across diverse NLP tasks, not on discovering geometric structures in task representations for numerical tasks with continuous latent parameters like add-k or circular trajectories.

3. Parameter Efficient Continual Learning with Dynamic Low-Rank Adaptation

URL: View paper

Brief Assessment

Dynamic Low-Rank[76] focuses on continual learning with dynamic rank allocation for LoRA adapters in vision tasks, not on discovering geometric structures in task representations for numerical in-context learning problems in language models.

4. Espace: Dimensionality reduction of activations for model compression

URL: View paper

Brief Assessment

Espace[71] focuses on dimensionality reduction of activations for model compression in LLMs, not on discovering geometric structures in task representations for in-context learning with continuous parameters.

5. BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models

URL: View paper

Brief Assessment

BLoB[70] focuses on Bayesian uncertainty estimation during fine-tuning of LLMs, not on analyzing geometric structures of task representations in in-context learning settings with continuous parameters.

6. LoPT: Low-Rank Prompt Tuning for Parameter Efficient Language Models

URL: View paper

Brief Assessment

LoPT[75] focuses on parameter-efficient prompt tuning through low-rank optimization of prompt embeddings, not on discovering geometric structures in task representations or analyzing how models encode continuous latent parameters in their hidden states.

7. Training Large Language Models to Reason in a Continuous Latent Space

URL: View paper

Brief Assessment

Continuous Latent Reasoning[67] focuses on reasoning in continuous latent space for logical problem-solving, not on analyzing geometric structure of task representations for numerical ICL tasks with continuous parameters.

8. Gradient boundary infiltration in large language models: A projection-based constraint framework for distributional trace locality

URL: View paper

Brief Assessment

Gradient Boundary Infiltration[72] focuses on gradient update directions and boundary constraints in LLM training, not on task representation geometry or in-context learning with continuous parameters.

9. Not all language model features are one-dimensionally linear

URL: View paper

Brief Assessment

Non-Linear Features[68] focuses on multi-dimensional circular representations (e.g., days of the week, months) as irreducible features requiring multiple dimensions. The original paper studies how task vectors for continuous numerical parameters (add-k, circular trajectories) lie on smooth low-dimensional manifolds that mirror the parameter space, enabling interpolation. These are fundamentally different phenomena: one concerns inherently circular/periodic concepts, the other concerns geometric structure emerging from continuous task parameterization.

10. Sparc: Subspace-aware prompt adaptation for robust continual learning in llms

URL: View paper

Brief Assessment

Sparc[69] focuses on continual learning through PCA-based prompt tuning in a lower-dimensional space for task adaptation, not on discovering geometric structures in task representations for numerical reasoning tasks with continuous latent parameters.

Contribution 3: Causal and correlational methodology for analyzing latent concept manipulation in transformers

Description: The authors develop a systematic approach combining causal mediation analysis (activation patching) and correlational techniques to localize and characterize how transformers represent and compose latent concepts during in-context learning across both discrete and continuous parameterizations.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Causal Mediation Analysis for Interpreting Intermediate Representations in Transformers

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

2. Bridging the Black Box: A Survey on Mechanistic Interpretability in AI

URL: View paper

Brief Assessment

Mechanistic Interpretability Survey[62] provides only fragmentary mentions of activation patching and causal methods without demonstrating a systematic methodology combining causal mediation analysis with correlational techniques for localizing latent concept processing in transformers during in-context learning.

---

3. Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers

URL: View paper

Brief Assessment

Language-Agnostic Representations[63] focuses on multilingual translation tasks and language-agnostic concept representations, not on in-context learning with discrete/continuous parameterizations or latent concept composition across diverse task types.

---

4. Understanding Counting Mechanisms in Large Language and Vision-Language Models

URL: View paper

Brief Assessment

Counting Mechanisms[65] focuses on numerical counting tasks using causal mediation and activation patching, while the original paper studies latent concept disentanglement across both discrete (multi-hop reasoning) and continuous (geometric) parameterizations in in-context learning.

---

5. From Context to Concept: Concept Encoding in In-Context Learning

URL: View paper

Brief Assessment

Context to Concept[47] focuses on concept encoding-decoding mechanisms in ICL through synthetic tasks and controlled finetuning experiments, while the original paper emphasizes causal mediation analysis (activation patching) combined with correlational techniques for localizing latent concept representations across discrete and continuous parameterizations. The candidate's approach centers on concept decodability metrics and training dynamics rather than the systematic combination of causal and correlational localization methods described in the original work.

---

6. Causal Intervention Framework for Variational Auto Encoder Mechanistic Interpretability

URL: View paper

Brief Assessment

VAE Causal Intervention[64] focuses on variational autoencoders (VAEs) and generative models, not transformer-based language models performing in-context learning. The technical domains and model architectures are fundamentally different.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Latent Concept Disentanglement in Transformer-based Language Models View paper
- [1] Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning View paper
- [2] Does learning the right latent variables necessarily improve in-context learning? View paper
- [3] Provably transformers harness multi-concept word semantics for efficient in-context learning View paper
- [4] Prompting language-informed distribution for compositional zero-shot learning View paper
- [5] Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning View paper
- [6] Vocabulary-Defined Semantics: Latent Space Clustering for Improving In-Context Learning View paper
- [7] Locoop: Few-shot out-of-distribution detection via prompt learning View paper
- [8] Beyond Label Semantics: Language-Guided Action Anatomy for Few-shot Action Recognition View paper
- [9] MICS: Midpoint Interpolation to Learn Compact and Separated Representations for Few-Shot Class-Incremental Learning View paper
- [10] Cultural Alignment in Large Language Models Using Soft Prompt Tuning View paper
- [11] Exploring intra-class variation factors with learnable cluster prompts for semi-supervised image synthesis View paper
- [12] Evaluating Generalization and Representation Stability in Small LMs via Prompting View paper
- [13] The representation landscape of few-shot learning and fine-tuning in large language models View paper
- [14] VaMP: Variational Multi-Modal Prompt Learning for Vision-Language Models View paper
- [15] Think-to-talk or talk-to-think? when llms come up with an answer in multi-step arithmetic reasoning View paper
- [16] Rethinking reinforcement learning for recommendation: A prompt perspective View paper
- [17] NN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference View paper
- [18] Can in-context learners learn a reasoning concept from demonstrations? View paper
- [19] Prompt Based CVAE Data Augmentation for Few-Shot Intention Detection View paper
- [20] Meta-learning the difference: preparing large language models for efficient adaptation View paper
- [21] SANSA: Unleashing the Hidden Semantics in SAM2 for Few-Shot Segmentation View paper
- [22] Learn the global prompt in the low-rank tensor space for heterogeneous federated learning. View paper
- [23] A circuit for predicting hierarchical structure in-context in Large Language Models View paper
- [24] Towards Robust Few-shot Point Cloud Semantic Segmentation View paper
- [25] Generating Prompts in Latent Space for Rehearsal-free Continual Learning View paper
- [26] Decoupling Template Bias in CLIP: Harnessing Empty Prompts for Enhanced Few-Shot Learning View paper
- [27] PBLF: Prompt based learning framework for cross-modal recipe retrieval View paper
- [28] Latent Factor Models Meets Instructions: Goal-conditioned Latent Factor Discovery without Task Supervision View paper
- [29] Disentangling Latent Shifts of In-Context Learning Through Self-Training View paper
- [30] Disentangling Latent Shifts of In-Context Learning with Weak Supervision View paper
- [31] Learning from Orthogonal Space with Multimodal Large Models for Generalized Few-shot Segmentation View paper
- [32] Prompt-SID: Learning Structural Representation Prompt via Latent Diffusion for Single Image Denoising View paper
- [33] Context-aware recommendation using hierarchical attention and positional encoding based on multiple item attributes (CARA) View paper
- [34] TFRS: A task-level feature rectification and separation method for few-shot video action recognition. View paper

- [35] JLCSR: Joint Learning of Compactness and Separability Representations for Few-Shot Classification View paper
- [36] What In-Context Learning â□□Learnsâ□□ In-Context: Disentangling Task Recognition and Task Learning View paper
- [37] Enhancing robustness and interpretability in natural language processing through representation learning
- [38] SharPT: Shared Latent Space Prompt Tuning View paper
- [39] Disentangled Prompt Learning for Transferable, Multimodal, Few-Shot Image Classification View paper
- [40] Prompt Transfer for Dual-Aspect Cross-Domain Cognitive Diagnosis View paper
- [41] StablePT : Towards Stable Prompting for Few-shot Learning via Input Separation View paper
- [42] Latent Domain Prompt Learning for Vision-Language Models View paper
- [43] Exploring Challenges in Applying Foundation and Generative Models in AI View paper
- [44] Fair In-Context Learning via Latent Concept Variables View paper
- [45] Enhancing the Stability of LLM-based Speech Generation Systems through Self-Supervised Representations View paper
- [46] Embroid: Unsupervised Prediction Smoothing Can Improve Few-Shot Classification View paper
- [47] From Context to Concept: Concept Encoding in In-Context Learning View paper
- [48] Latent Prompt Tuning for Text Summarization View paper
- [49] Few-Shot Musical Source Separation View paper
- [50] Measuring Effects of Steered Representation in Large Language Models View paper
- [51] Understanding Multi-compositional learning in Vision and Language models via Category Theory View paper
- [52] Latent cascade synthesis: Investigating iterative pseudo-contextual scaffold formation in contemporary large language models View paper
- [53] Understanding and patching compositional reasoning in llms View paper
- [54] How does Transformer Learn Implicit Reasoning? View paper
- [55] Do Large Language Models Perform Latent Multi-Hop Reasoning without Exploiting Shortcuts? View paper
- [56] Mechanics of Next Token Prediction with Self-Attention View paper
- [57] Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization View paper
- [58] Attention as a Hypernetwork View paper
- [59] Relmkg: reasoning with pre-trained language models and knowledge graphs for complex question answering View paper
- [60] The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision View paper
- [61] Causal Mediation Analysis for Interpreting Intermediate Representations in Transformers View paper
- [62] Bridging the Black Box: A Survey on Mechanistic Interpretability in AI View paper
- [63] Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers View paper
- [64] Causal Intervention Framework for Variational Auto Encoder Mechanistic Interpretability View paper
- [65] Understanding Counting Mechanisms in Large Language and Vision-Language Models View paper
- [66] Interpret and Improve In-Context Learning via the Lens of Input-Label Mappings View paper
- [67] Training Large Language Models to Reason in a Continuous Latent Space View paper
- [68] Not all language model features are one-dimensionally linear View paper
- [69] Sparc: Subspace-aware prompt adaptation for robust continual learning in llms View paper
- [70] BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models View paper
- [71] Espace: Dimensionality reduction of activations for model compression View paper
- [72] Gradient boundary infiltration in large language models: A projection-based constraint framework for distributional trace locality View paper
- [73] Exploring universal intrinsic task subspace for few-shot learning via prompt tuning View paper
- [74] Manifold-based verbalizer space re-embedding for tuning-free prompt-based classification View paper
- [75] LoPT: Low-Rank Prompt Tuning for Parameter Efficient Language Models View paper
- [76] Parameter Efficient Continual Learning with Dynamic Low-Rank Adaptation View paper