Novelty Assessment Report

Paper: Localizing Task Recognition and Task Learning in In-Context Learning via Attention Head Analysis
PDF URL: https://openreview.net/pdf?id=gdvOF1OMa7
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

We investigate the mechanistic underpinnings of in-context learning (ICL) in large language models by reconciling two dominant perspectives: the component-level analysis of attention heads and the holistic decomposition of ICL into Task Recognition (TR) and Task Learning (TL). We propose a novel framework based on Task Subspace Logit Attribution (TSLA) to identify attention heads specialized in TR and TL, and demonstrate their distinct yet complementary roles. Through correlation analysis, ablation studies, and input perturbations, we demonstrate that the identified TR and TL heads independently and effectively capture the TR and TL components of ICL. Via steering experiments with a focus on the geometric analysis of hidden states, we reveal that TR heads promote task recognition through aligning hidden states with the task subspace, while TL heads perform rotations to the hidden states within the subspace towards the correct label to facilitate the correct prediction. We also demonstrate how previous findings in various aspects of ICL's mechanism can be reconciled with our attention-head-level analysis of the TR-TL decomposition of ICL, including induction heads, task vectors, and more. Our framework thus provides a unified and interpretable account of how LLMs execute ICL across diverse tasks and settings.

Core Task Landscape

This paper addresses: **mechanistic analysis of in-context learning through attention head decomposition**

A total of **38 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Attention Head Functional Specialization**
- **Task Recognition and Task Learning Decomposition**
- **Training Dynamics and Emergence**
- **Component-Level Interventions and Interpretability Methods**
- **Theoretical Foundations and Architectural Analysis**
- **Domain-Specific Applications and Extensions**
- **Survey and Methodological Frameworks**

Complete Taxonomy Tree

- mechanistic analysis of in-context learning through attention head decomposition Survey Taxonomy
- Attention Head Functional Specialization
  - Induction and Retrieval Heads (3 papers)
  - [1] Identifying semantic induction heads to understand in-context learning (Guo, 2024) View paper
  - [3] Retrieval head mechanistically explains long-context factuality (Wu Wen-Hao, 2024) View paper
  - [4] In-context learning and induction heads (Olsson, 2022) View paper
  - Task-Specific Head Mechanisms (4 papers)
  - [11] Which Attention Heads Matter for In-Context Learning? (Yin, 2025) View paper
  - [15] Causal Head Gating: A Framework for Interpreting Roles of Attention Heads in Transformers (Nam, 2025) View paper
  - [25] The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation (Kahardipraja, 2025) View paper
  - [32] How do Large Language Models Learn In-Context? Query and Key Matrices of In-Context Heads are Two Towers for Metric Learning (Ananiadou, 2024) View paper
  - Context Management and Attention Allocation (3 papers)
  - [9] Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms (Guo Tianyu, 2024) View paper
  - [14] Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs (Niu, 2025) View paper
  - [19] Focus directions make your language models pay more attention to relevant contexts (Zhu, 2025) View paper
- Task Recognition and Task Learning Decomposition
  - TR-TL Framework and Head-Level Analysis ★ (2 papers)
  - [0] Localizing Task Recognition and Task Learning in In-Context Learning via Attention Head Analysis (Anon et al., 2026) View paper
  - [27] Mechanism of Task-oriented Information Removal in In-context Learning (Cho Hakaze, 2025) View paper
  - Hidden State Geometry and Task Representations (3 papers)
  - [18] Understanding In-context Learning of Addition via Activation Subspaces (Hu, 2025) View paper
  - [23] Learning Task Representations from In-Context Learning (Baturay Saglam, 2025) View paper
  - [24] Unifying Attention Heads and Task Vectors via Hidden State Geometry in In-Context Learning (Yang HaoLin, 2025) View paper

- Training Dynamics and Emergence
  - Multi-Head Attention Training Dynamics (3 papers)
  - [20] In-context learning with representations: Contextual generalization of trained transformers (Yuejie Chi, 2024) View paper
  - [28] In-Context Linear Regression Demystified: Training Dynamics and Mechanistic Interpretability of Multi-Head Softmax Attention (He, 2025) View paper
  - [30] Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality (Chen, 2024) View paper
  - Grokking and Skill Composition (1 papers)
  - [31] Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks (Aritra Das, 2024) View paper
- Component-Level Interventions and Interpretability Methods
  - Ablation and Attribution Methods (3 papers)
  - [22] When Parts Are Greater Than Sums: Individual LLM Components Can Outperform Full Models (Ting Yun Chang, 2024) View paper
  - [29] Attributing Response to Context: A Jensen-Shannon Divergence Driven Mechanistic Study of Context Attribution in Retrieval-Augmented Generation (Li, 2025) View paper
  - [37] Rethinking the Role of Scale for In-Context Learning: An Interpretability-based Case Study at 66 Billion Scale (Hritik Bansal, 2023) View paper
  - Component Reweighting and Steering (2 papers)
  - [34] Soft Injection of Task Embeddings Outperforms Prompt-Based In-Context Learning (Parkï¼Jungwon, 2025) View paper
  - [35] Don't Think of the White Bear: Ironic Negation in Transformer Models Under Cognitive Load (Logan Mann, 2025) View paper
  - Activation and Representation Analysis (1 papers)
  - [12] Decomposing Attention To Find Context-Sensitive Neurons (Gibson, 2025) View paper
- Theoretical Foundations and Architectural Analysis
  - Theoretical Models of ICL (2 papers)
  - [7] How transformers utilize multi-head attention in in-context learning? a case study on sparse linear regression (Xingwu Chen, 2024) View paper
  - [13] In-context language learning: Architectures and algorithms (AkyÃ¼rek, 2024) View paper
  - Scale Effects and Model Behavior (1 papers)
  - [6] Why Larger Language Models Do In-context Learning Differently? (Shi, 2024) View paper
- Domain-Specific Applications and Extensions
  - Multimodal and Vision-Language Models (5 papers)
  - [5] Mechanistic Interpretability of Fine-Tuned Vision Transformers on Distorted Images: Decoding Attention Head Behavior for Transparent and Trustworthy AI (Bahador, 2025) View paper
  - [17] How Do Large Vision-Language Models See Text in Image? Unveiling the Distinctive Role of OCR Heads (Hwan Chang, 2025) View paper
  - [21] Multimodal Task Vectors Enable Many-Shot Multimodal In-Context Learning (Assaf Arbelle, 2024) View paper
  - [33] Mimic In-Context Learning for Multimodal Tasks (Jiang Yuchu, 2025) View paper
  - [36] Mechanistic Finetuning of Vision-Language-Action Models via Few-Shot Demonstrations (Chancharik Mitra, 2025) View paper
  - Specialized Phenomena and Behaviors (2 papers)
  - [8] Repetitions are not all alike: distinct mechanisms sustain repetition in language models (Mahaut, 2025) View paper
  - [38] FACT OR HALLUCINATION? AN ENTROPY-BASED FRAMEWORK FOR ATTENTION-WISE USABLE INFOR (LLMS, n.d.) View paper
- Survey and Methodological Frameworks (4 papers)
  - [2] Attention heads of large language models: A survey (Huang, 2024) View paper
  - [10] Quantitative self-reflection protocols for self-replicating memory chains in large language models: A technical investigation (P McAllister, 2025) View paper
  - [16] Squeezed attention: Accelerating long context length llm inference (Coleman Hooper, 2025) View paper
  - [26] Going Beyond a Basic Attention Head toward an Understanding of Transformer-based Generative AI (Restrepo, 2025) View paper

Narrative

Core task: mechanistic analysis of in-context learning through attention head decomposition. This field seeks to understand how transformer models perform in-context learning by dissecting the roles of individual attention heads and internal components. The taxonomy reflects a multifaceted landscape: one major branch examines Attention Head Functional Specialization, cataloging how heads develop distinct roles such as induction (Induction Heads[4]), retrieval (Retrieval Head Factuality[3]), or semantic pattern matching (Semantic Induction Heads[1]). Another branch focuses on Task Recognition and Task Learning Decomposition, exploring how models separate the recognition of task structure from the execution of task-specific computations. Additional branches address Training Dynamics and Emergence, which trace how these mechanisms arise during learning (Learning to Grok[31]), Component-Level Interventions that test causal importance (Causal Head Gating[15]), Theoretical Foundations linking architectures to algorithmic primitives (ICL Architectures Algorithms[13]), Domain-Specific Applications extending insights to vision or robotics (Vision Transformer Interpretability[5], Mechanistic Finetuning VLA[36]), and Survey and Methodological Frameworks that synthesize interpretability techniques (Attention Heads Survey[2]).

Several active lines of work reveal contrasting emphases and open questions. Some studies pursue fine-grained localization of function, identifying which heads or neurons are causally responsible for specific behaviors (Which Heads Matter[11], Context-Sensitive Neurons[12]), while others investigate higher-level abstractions such as task representations in hidden states (ICL Representations[20], Hidden State Geometry[24]) or the interplay between task recognition and task execution. The original paper, Task Recognition Localization[0], sits squarely within the Task Recognition and Task Learning Decomposition branch, closely aligned with work that disentangles these two phases at the head level. Its emphasis on localizing task recognition mechanisms complements nearby efforts like Task Information Removal[27], which ablates task-related signals, offering a causal counterpart to the localization perspective. This positioning highlights ongoing debates about whether in-context learning emerges from modular, interpretable circuits or from distributed, entangled representations across layers.

Related Works in Same Category

No comparison data available.

Contributions Analysis

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

Contribution 1: Task Subspace Logit Attribution (TSLA) framework for identifying TR and TL heads

Description: The authors introduce TSLA, a theoretically grounded method that identifies attention heads responsible for Task Recognition and Task Learning in in-context learning by measuring head contributions relative to task-label unembeddings in geometric subspace terms, addressing limitations of prior attribution approaches.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Geometric characterization of TR and TL head mechanisms

Description: Through steering experiments and geometric analysis, the authors demonstrate that TR heads align hidden states to the task-label subspace for label-space recognition, while TL heads perform within-subspace rotations toward correct labels to enable accurate prediction.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Unified framework reconciling component-level and holistic ICL perspectives

Description: The authors establish a unified framework that bridges attention-head-level mechanistic analysis with the holistic Task Recognition and Task Learning decomposition, reconciling prior findings on induction heads and task vectors within this integrated perspective.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Induction heads as an essential mechanism for pattern matching in in-context learning

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Localizing Task Recognition and Task Learning in In-Context Learning via Attention Head Analysis View paper
- [1] Identifying semantic induction heads to understand in-context learning View paper
- [2] Attention heads of large language models: A survey View paper
- [3] Retrieval head mechanistically explains long-context factuality View paper
- [4] In-context learning and induction heads View paper
- [5] Mechanistic Interpretability of Fine-Tuned Vision Transformers on Distorted Images: Decoding Attention Head Behavior for Transparent and Trustworthy AI View paper
- [6] Why Larger Language Models Do In-context Learning Differently? View paper
- [7] How transformers utilize multi-head attention in in-context learning? a case study on sparse linear regression View paper
- [8] Repetitions are not all alike: distinct mechanisms sustain repetition in language models View paper
- [9] Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms View paper
- [10] Quantitative self-reflection protocols for self-replicating memory chains in large language models: A technical investigation View paper
- [11] Which Attention Heads Matter for In-Context Learning? View paper
- [12] Decomposing Attention To Find Context-Sensitive Neurons View paper
- [13] In-context language learning: Architectures and algorithms View paper
- [14] Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs View paper
- [15] Causal Head Gating: A Framework for Interpreting Roles of Attention Heads in Transformers View paper
- [16] Squeezed attention: Accelerating long context length llm inference View paper
- [17] How Do Large Vision-Language Models See Text in Image? Unveiling the Distinctive Role of OCR Heads View paper
- [18] Understanding In-context Learning of Addition via Activation Subspaces View paper
- [19] Focus directions make your language models pay more attention to relevant contexts View paper
- [20] In-context learning with representations: Contextual generalization of trained transformers View paper
- [21] Multimodal Task Vectors Enable Many-Shot Multimodal In-Context Learning View paper
- [22] When Parts Are Greater Than Sums: Individual LLM Components Can Outperform Full Models View paper
- [23] Learning Task Representations from In-Context Learning View paper
- [24] Unifying Attention Heads and Task Vectors via Hidden State Geometry in In-Context Learning View paper
- [25] The Atlas of In-Context Learning: How Attention Heads Shape In-Context Retrieval Augmentation View paper
- [26] Going Beyond a Basic Attention Head toward an Understanding of Transformer-based Generative AI View paper
- [27] Mechanism of Task-oriented Information Removal in In-context Learning View paper
- [28] In-Context Linear Regression Demystified: Training Dynamics and Mechanistic Interpretability of Multi-Head Softmax Attention View paper
- [29] Attributing Response to Context: A Jensen-Shannon Divergence Driven Mechanistic Study of Context Attribution in Retrieval-Augmented Generation View paper
- [30] Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality View paper
- [31] Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks View paper

- [32] How do Large Language Models Learn In-Context? Query and Key Matrices of In-Context Heads are Two Towers for Metric Learning View paper
- [33] Mimic In-Context Learning for Multimodal Tasks View paper
- [34] Soft Injection of Task Embeddings Outperforms Prompt-Based In-Context Learning View paper
- [35] Don't Think of the White Bear: Ironic Negation in Transformer Models Under Cognitive Load View paper
- [36] Mechanistic Finetuning of Vision-Language-Action Models via Few-Shot Demonstrations View paper
- [37] Rethinking the Role of Scale for In-Context Learning: An Interpretability-based Case Study at 66 Billion Scale View paper
- [38] FACT OR HALLUCINATION? AN ENTROPY-BASED FRAMEWORK FOR ATTENTION-WISE USABLE INFOR View paper
- [39] Induction heads as an essential mechanism for pattern matching in in-context learning View paper
- [40] DETAIL: Task demonstration attribution for interpretable in-context learning View paper
- [41] The hidden attention of mamba models View paper
- [42] Function vectors in large language models View paper
- [43] Contextcite: Attributing model generation to context View paper
- [44] STAA: Spatio-Temporal Attention Attribution for Real-Time Interpreting Transformer-based AI Video Models View paper
- [45] Otter: A multi-modal model with in-context instruction tuning View paper
- [46] Compositional exemplars for in-context learning View paper
- [47] The learnability of in-context learning View paper
- [48] Are emergent abilities in large language models just in-context learning? View paper
- [49] Rethinking the role of demonstrations: What makes in-context learning work? View paper
- [50] Is mamba capable of in-context learning? View paper
- [51] An explanation of in-context learning as implicit bayesian inference View paper
- [52] Towards more unified in-context visual understanding View paper
- [53] Can mamba learn how to learn? a comparative study on in-context learning tasks View paper
- [54] Towards Multimodal In-context Learning for Vision and Language Models View paper
- [55] The Geometry of Self-Verification in a Task-Specific Reasoning Model View paper
- [56] Gradient boundary infiltration in large language models: A projection-based constraint framework for distributional trace locality View paper
- [57] Attention is all you need View paper
- [58] Latent resonance pathways for large language models through gradient-synchronized semantic fluxion View paper
- [59] Talking Heads: Understanding Inter-layer Communication in Transformer Language Models View paper
- [60] Align attention heads before merging them: An effective way for converting MHA to GQA View paper
- [61] Residual transformer alignment with spectral decomposition View paper
- [62] Analyzing the Mechanism of Attention Collapse in VGGT from a Dynamics Perspective View paper
- [63] Enhancing YOLOv8 with Attention Task Alignment Head for Prohibited Item Detection in Complex X-Ray Images View paper