Novelty Assessment Report

Paper: Jet Expansions: Restructuring LLM Computation for Model Inspection
PDF URL: https://openreview.net/pdf?id=u6JLh0BO5h
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Large language models are becoming general knowledge engines for diverse applications. However, their computations are deeply entangled after training, resisting modularization which complicates interpretability, auditing, and long-term maintenance. We introduce Jet Expansions, a framework for expanding computational graphs using jet operators that generalize truncated Taylor series. Our method systematically decomposes language models into explicit input-to-output computational paths and complementary remainders. This functional decomposition provides a principled, knife-like operator for cutting through entanglement in LLMs, enabling scalable model inspection. We demonstrate how Jet Expansions ground and subsume the popular interpretability technique Logit Lens, reveal a (super-)exponential path structure with respect to recursive residual depth, and support several interpretability applications, including sketching a transformer language model with N-gram statistics extracted from its computations and indexing model toxicity levels without curated benchmarks.

Core Task Landscape

This paper addresses: **Decomposing Language Model Computations into Interpretable Paths**

A total of **50 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Activation-Based Decomposition and Feature Extraction**
- **Weight-Based and Circuit-Level Analysis**
- **Reasoning Path Decomposition and Explanation**
- **Specialized Interpretability Applications**
- **Computational Path Formalization and Theory**
- **Architectural and Representational Foundations**
- **Cross-Domain and Emerging Applications**

Complete Taxonomy Tree

- Decomposing Language Model Computations into Interpretable Paths Survey Taxonomy
- Activation-Based Decomposition and Feature Extraction
  - Sparse Autoencoder Approaches (3 papers)
  - [1] Sparse feature circuits: Discovering and editing interpretable causal graphs in language models (Marks, 2024) View paper
  - [4] Improving Sparse Decomposition of Language Model Activations with Gated Sparse Autoencoders (Arthur Conmy, 2024) View paper
  - [5] Not All Language Model Features Are One-Dimensionally Linear (Engels, 2024) View paper
  - Alternative Activation Decomposition Methods (2 papers)
  - [6] Inference-time decomposition of activations (itda): A scalable approach to interpreting large language models (Nanda, 2025) View paper
  - [27] Route sparse autoencoder to interpret large language models (Shi Wei, 2025) View paper
- Weight-Based and Circuit-Level Analysis
  - Circuit Discovery and Causal Analysis (4 papers)
  - [15] Information flow routes: Automatically interpreting language models at scale (Ferrando, 2024) View paper
  - [26] Extracting interpretable task-specific circuits from large language models for faster inference (Mate Alejandro, 2025) View paper
  - [31] Interpretability at scale: Identifying causal mechanisms in alpaca (Wu Zhengxuan, 2023) View paper
  - [33] Efficient automated circuit discovery in transformers using contextual decomposition (Hsu, 2024) View paper
  - Weight Decomposition and Layer-Level Interpretation (4 papers)
  - [12] Bilinear MLPs enable weight-based mechanistic interpretability (Michael T. Pearce, 2024) View paper
  - [14] Weight-based Analysis of Detokenization in Language Models: Understanding the First Stage of Inference Without Inference (Heinzerling, 2025) View paper
  - [22] Neural synthesis through probabilistic layer decomposition in large language models (Phillip Beaumont, 2024) View paper
  - [24] Towards interpretability without sacrifice: Faithful dense layer decomposition with mixture of decoders (Oldfield, 2025) View paper
- Reasoning Path Decomposition and Explanation
  - Chain-of-Thought and Multi-Step Reasoning (6 papers)
  - [2] Selection-inference: Exploiting large language models for interpretable logical reasoning (Creswell, 2022) View paper
  - [10] How interpretable are reasoning explanations from prompting large language models? (Yeo Wei Jie, 2024) View paper

- [16] Chain-of-Thought for Large Language Model-empowered Wireless Communications (Wang Xudong, 2025) View paper
- [17] Towards interpretable and consistent multi-step mathematical reasoning in large language models (Xinyue Huang, 2025) View paper
- [32] Policy-guided path selection and evaluation in multi-step reasoning with large language models (Pan, 2024) View paper
- [34] Do Cognitively Interpretable Reasoning Traces Improve LLM Performance? (Bhambri, 2025) View paper
- Knowledge Graph-Guided Reasoning (7 papers)
- [8] KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision (Wu Rong, 2025) View paper
- [11] FiDeLiS: Faithful Reasoning in Large Language Model for Knowledge Graph Question Answering (Sui Yuan, 2024) View paper
- [13] Path language modeling over knowledge graphsfor explainable recommendation (Shijie Geng, 2022) View paper
- [19] Paths-over-graph: Knowledge graph empowered large language model reasoning (Xingyu Tan, 2025) View paper
- [20] GraphTrace: A Modular Retrieval Framework Combining Knowledge Graphs and Large Language Models for Multi-Hop Question Answering (Anna Osipjan, 2025) View paper
- [25] DRKG: Faithful and Interpretable Multi-Hop Knowledge Graph Question Answering via LLM-Guided Reasoning Plans (Yan Chen, 2025) View paper
- [43] Towards large-scale interpretable knowledge graph reasoning for dialogue systems (Tuan, 2022) View paper
- Task Decomposition and Modular Reasoning (3 papers)
- [29] Decomposing Complex Questions Makes Multi-Hop QA Easier and More Interpretable (Ruiliu Fu, 2021) View paper
- [37] Text modular networks: Learning to decompose tasks in the language of existing models (Khot, 2021) View paper
- [50] Modular Networks: Learning to Decompose Neural Computation (Kirsch, 2018) View paper

- Specialized Interpretability Applications
  - Task-Specific Mechanistic Analysis (4 papers)
  - [3] How do large language models understand relevance? a mechanistic interpretability perspective (Qi Liu, 2025) View paper
  - [7] Mechanistic interpretability of large language models with applications to the financial services industry (Ashkan Golgoon, 2024) View paper
  - [30] Spectral Journey: How Transformers Predict the Shortest Path (Cohen, 2025) View paper
  - [36] Mechanistic Interpretability for Progress Towards Quantitative AI Safety (Lad, 2024) View paper
  - Evaluation and Generation Analysis (3 papers)
  - [9] DnA-Eval: Enhancing Large Language Model Evaluation through Decomposition and Aggregation (Li Minzhi, 2024) View paper
  - [21] Recap: Transparent inference-time emotion alignment for medical dialogue systems (Srinivasan, 2025) View paper
  - [46] Extractive Fact Decomposition for Interpretable Natural Language Inference in one Forward Pass (Popovic, 2025) View paper
  - Safety and Alignment Applications (3 papers)
  - [45] Transforming Network Intrusion Detection Using Large Language Models (Dongming Wu, 2025) View paper
  - [47] Redefining Experts: Interpretable Decomposition of Language Models for Toxicity Mitigation (Mazhar Abdullah, 2025) View paper
  - [48] Exploring the Generalizability and Explainability of LLMs in Detecting Suicidal Ideation: The Impact of Data Heterogeneity (R Huang, 2025) View paper
- Computational Path Formalization and Theory ★ (2 papers)
  - [0] Jet Expansions: Restructuring LLM Computation for Model Inspection (Anon et al., 2026) View paper
  - [44] Neural-ANOVA: Model Decomposition for Interpretable Machine Learning (Limmer, 2024) View paper
- Architectural and Representational Foundations
  - Compositional Architecture and Modularity (4 papers)
  - [18] Modular Machine Learning: An Indispensable Path towards New-Generation Large Language Models (Wang Xin, 2025) View paper
  - [35] Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference (Yi Tay, 2018) View paper
  - [40] A Compositional Neural Architecture for Language (Andrea E. Martin, 2020) View paper
  - [49] A Hierarchical Language Model For Interpretable Graph Reasoning (Khurana, 2024) View paper
  - Learning Dynamics and Trajectory Analysis (1 papers)
  - [39] The grammar-learning trajectories of neural language models (Leshem Choshen, 2022) View paper
  - Visualization and Exploration Tools (2 papers)
  - [38] Ecco: An open source library for the explainability of transformer language models (J Alammar, 2021) View paper
  - [41] KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts. (Adam Coscia, 2025) View paper
- Cross-Domain and Emerging Applications (3 papers)
  - [23] Towards a translative model of Sperm Whale vocalization (O Paradise, 2025) View paper
  - [28] Situationally-aware path planning exploiting 3d scene graphs (Saad Ejaz, 2025) View paper
  - [42] SMART: A Semantic-Guided Reinforcement Learning for Interpretable Feature Engineering (M Bouadi, 2025) View paper

Narrative

Core task: decomposing language model computations into interpretable paths. The field has organized itself around several complementary perspectives on how to make neural network processing transparent. Activation-Based Decomposition and Feature Extraction focuses on identifying meaningful units within hidden representations, often using sparse autoencoders (Gated Sparse Autoencoders[4]) or feature circuits (Sparse Feature Circuits[1]) to isolate interpretable components. Weight-Based and Circuit-Level Analysis examines the connectivity and parameter structure that determines information flow, while Reasoning Path Decomposition and Explanation targets the step-by-step logic in multi-hop or chain-of-thought settings. Specialized Interpretability Applications adapt these techniques to domains like finance (Financial MI[7]) or knowledge graphs (KG-TRACES[8]), and Computational Path Formalization and Theory provides the mathematical underpinnings—such as decomposition algebras or probabilistic frameworks (Probabilistic Layer Decomposition[22])—that unify diverse methods. Architectural and Representational Foundations study how model design choices shape interpretability, and Cross-Domain and Emerging Applications explore novel settings from wireless networks (CoT Wireless[16]) to biological communication (Sperm Whale Vocalization[23]).

A particularly active line of work centers on formalizing how computations can be rigorously partitioned into additive or multiplicative contributions, with methods like Neural-ANOVA[44] offering variance-based decompositions and Jet Expansions[0] introducing higher-order Taylor-like expansions to capture nonlinear interactions. These theoretical frameworks contrast with more empirical circuit-tracing approaches (Information Flow Routes[15], Task-Specific Circuits[26]) that identify which subnetworks are causally responsible for specific behaviors. Jet Expansions[0] sits squarely within the Computational Path Formalization branch, emphasizing rigorous

mathematical decomposition rather than heuristic feature extraction. Compared to Neural-ANOVA[44], which partitions variance across input dimensions, Jet Expansions[0] extends the toolkit to higher-order terms, enabling finer-grained attribution of model outputs to interactions among features. This formal approach complements activation-based methods (Nonlinear Features[5]) by providing a principled basis for understanding how complex, nonlinear transformations emerge from simpler computational primitives.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Neural-ANOVA: Model Decomposition for Interpretable Machine Learning

Authors: Limmer, Steffen, Udluft, Steffen Limmer, Otte, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The analysis of variance (ANOVA) decomposition offers a systematic method to understand the interaction effects that contribute to a specific decision output. In this paper we introduce Neural-ANOVA, an approach to decompose neural networks into the sum of lower-order models using the functional ANOVA decomposition. Our approach formulates a learning problem, which enables fast analytical evaluation of integrals over subspaces that appear in the calculation of the ANOVA decomposition. Finally, w...

Relationship Analysis

Both papers belong to the Computational Path Formalization and Theory category, focusing on mathematical frameworks for decomposing model computations into interpretable components. While the original paper uses jet operators (generalized Taylor expansions) to decompose residual networks into explicit input-output paths for interpretability, the candidate paper applies functional ANOVA decomposition with automatic integration to neural networks for analyzing interaction effects. The key difference is that the original paper targets path-based decomposition of transformer language models for mechanistic interpretability, whereas the candidate focuses on variance decomposition for sensitivity analysis and feature interaction understanding in general neural networks.

Contributions Analysis

Overall novelty summary. The paper introduces Jet Expansions, a mathematical framework for decomposing language model computations into explicit input-to-output paths using generalized Taylor series operators. It resides in the 'Computational Path Formalization and Theory' leaf, which contains only two papers total. This is one of the sparsest research directions in the taxonomy, indicating a relatively underexplored theoretical niche focused on rigorous mathematical formalizations rather than empirical circuit discovery or application-driven interpretability.

The taxonomy reveals substantial activity in neighboring areas: Activation-Based Decomposition (five papers across two leaves) focuses on extracting features from hidden states, while Weight-Based and Circuit-Level Analysis (eight papers) emphasizes causal subgraph identification. Reasoning Path Decomposition (sixteen papers across three leaves) targets multi-step logic chains. Jet Expansions diverges by providing mathematical foundations for these empirical methods rather than proposing new feature extraction or circuit-tracing techniques. Its scope_note explicitly excludes empirical circuit discovery, positioning it as theoretical infrastructure.

Among twenty-eight candidates examined, the contribution-level analysis shows mixed novelty signals. The core Jet Expansions framework (ten candidates examined, zero refutations) and the function decomposition perspective (ten candidates, zero refutations) appear relatively novel within the limited search scope. However, the claim of grounding existing interpretability tools encountered one refutable candidate among eight examined, suggesting some theoretical overlap with prior formalization efforts. The search scale is modest—top-K semantic matches plus citations—so these findings reflect local rather than exhaustive coverage.

Given the sparse theoretical leaf and limited search scope, the work appears to occupy a distinct formal niche. The framework's mathematical rigor and higher-order expansion machinery differentiate it from variance-based methods like Neural-ANOVA, though the grounding of existing tools shows some precedent. The analysis covers approximately thirty semantically related papers, leaving open whether broader theoretical literature in adjacent fields might reveal additional connections.

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

Contribution 1: Jet Expansions framework for restructuring LLM computations

Description: The authors propose a principled mathematical framework that uses jet operators (functional counterparts of truncated Taylor series) to systematically decompose language models into explicit input-to-output computational paths and complementary remainders. This functional decomposition provides a systematic operator for cutting through entanglement in LLMs, enabling scalable model inspection without requiring additional data or training.

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. Unfolding Videos Dynamics via Taylor Expansion

URL: View paper

Brief Assessment

Unfolding Videos Taylor[59] applies Taylor expansion to video dynamics for self-supervised learning, not to decomposing language model computations for interpretability. The domains (video understanding vs. LLM interpretability) and objectives (learning motion features vs. restructuring computational graphs) are fundamentally different.

---

2. M-Rule: An Enhanced Deep Taylor Decomposition for Multi-model Interpretability

URL: View paper

Brief Assessment

M-Rule[60] focuses on Deep Taylor Decomposition for visual feature attribution in image classification models (DNNs, CNNs, LSTMs), not on restructuring language model computations into explicit paths using jet operators as functional decomposition tools.

---

3. Cat: Interpretable concept-based taylor additive models

URL: View paper

Brief Assessment

Cat[64] focuses on concept-based interpretability for tabular data using Taylor polynomials in a feedforward network (TaylorNet), not on decomposing LLM residual computations using jet operators.

---

4. Explaining nonlinear classification decisions with deep taylor decomposition

URL: View paper

Brief Assessment

Deep Taylor Classification[67] focuses on explaining classification decisions in deep neural networks through Taylor decomposition for interpretability, not on restructuring LLM computations into explicit input-output paths using jet operators as a systematic framework for model inspection.

---

5. Towards explaining anomalies: A deep Taylor decomposition of one-class models

URL: View paper

Brief Assessment

Deep Taylor Anomalies[63] focuses on explaining anomaly detection in one-class models using Taylor decomposition, not on restructuring LLM computations or providing functional decomposition frameworks for language models.

---

6. Hope: High-order polynomial expansion of black-box neural networks

URL: View paper

Brief Assessment

Hope[61] focuses on Taylor expansion of general neural networks for interpretability and function approximation, not specifically on restructuring LLM residual computations using jet operators as functional decomposition tools for model inspection.

---

7. Explaining COVID-19 diagnosis with Taylor decompositions

URL: View paper

Brief Assessment

COVID Taylor Decomposition[65] applies Taylor decomposition to explain COVID-19 diagnosis in medical imaging using deep networks (VGG11, VGG16), not to restructure LLM computations or create input-to-output paths in language models.

---

8. GTEA: Guided Taylor Expansion Approximation Network for Optical Flow Estimation

URL: View paper

Brief Assessment

GTEA[62] applies Taylor expansion to optical flow estimation in computer vision, not to language model interpretability or computational decomposition. The domains and applications are entirely different.

---

9. Tayloraecnet: A Taylor Style Neural Network For Full-Band Echo Cancellation

URL: View paper

Brief Assessment

TaylorAECNet[68] applies Taylor expansion to acoustic echo cancellation in signal processing, not to decomposing language models for interpretability. The domains and objectives are fundamentally different.

---

10. An integrated model based on feedforward neural network and Taylor expansion for indicator correlation elimination

URL: View paper

Brief Assessment

Indicator Correlation Elimination[66] applies Taylor expansion to eliminate correlations between evaluation indicators in a feedforward neural network context, not to decompose or restructure language model computations for interpretability purposes.

Contribution 2: Treating interpretability as function decomposition

Description: The authors introduce a conceptual shift in interpretability methodology by framing it as a problem of function decomposition rather than traditional data-driven approaches. This perspective enables manipulation of functions directly in function space, requiring no probe datasets or sampling, and allows arbitrary portions of computation to be isolated from the monolithic transformer.

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. Kolmogorov-Arnold Networks for Interpretable and Efficient Function Approximation

URL: View paper

Brief Assessment

KAN[72] focuses on function approximation through univariate transformations based on Kolmogorov-Arnold representation theorems, not on decomposing pre-trained transformer computations for interpretability analysis.

---

2. Neural additive models: Interpretable machine learning with neural nets

URL: View paper

Brief Assessment

Neural Additive Models[75] focuses on decomposing predictions into additive univariate functions for tabular data interpretability, not on decomposing transformer computations in function space without probe datasets as in the original paper.

---

3. Beyond the Black Box: A Review of Quantitative Metrics for Neural Network Interpretability and Their Practical Implications

URL: View paper

Brief Assessment

Black Box Metrics[74] focuses on quantitative evaluation metrics for interpretability (fidelity, complexity, robustness) rather than function decomposition methods. The paper reviews post-hoc interpretability techniques and metrics, not methods for decomposing neural network computations into explicit input-output paths.

---

4. Interpretable basis decomposition for visual explanation

URL: View paper

Brief Assessment

Interpretable Basis Decomposition[76] focuses on decomposing neural network activations into semantically interpretable components using pre-trained concept vectors from annotated datasets. This is fundamentally different from the original paper's approach of manipulating functions directly in function space using jet operators without requiring probe datasets or sampling.

---

5. A comprehensive survey on self-interpretable neural networks

URL: View paper

Brief Assessment

Self-Interpretable Networks Survey[73] focuses on categorizing existing self-interpretable neural network architectures across multiple paradigms (attribution, function, concept, prototype, rule-based). While it discusses function-based methods including functional decomposition, it does not present this as a novel contribution but rather surveys existing approaches. The survey's scope is descriptive taxonomy rather than introducing new methodological frameworks for interpretability.

---

6. A survey on kolmogorov-arnold network

URL: View paper

Brief Assessment

Kolmogorov-Arnold Survey[70] focuses on KAN architectures that use learnable spline-parameterized functions for function approximation, not on decomposing existing transformer computations into interpretable paths as the original paper does.

---

7. Tensor Product Neural Networks for Functional ANOVA Model

URL: View paper

Brief Assessment

Tensor Product Networks[77] focuses on functional ANOVA decomposition for interpretability, which decomposes functions into sums of lower-dimensional components. The ORIGINAL paper's contribution is about manipulating transformer computations directly in function space without probe datasets, which is a different technical approach and application domain.

---

8. Tensorization of neural networks for improved privacy and interpretability

URL: View paper

Brief Assessment

Tensorization Neural Networks[69] focuses on tensor train/MPS decomposition for neural networks with applications to privacy obfuscation and topological phase estimation, not on general function decomposition methods for interpretability without input attribution as described in the original paper.

---

9. Multilevel wavelet decomposition network for interpretable time series analysis

URL: View paper

Brief Assessment

Wavelet Decomposition Network[71] focuses on decomposing time series into frequency components using wavelet transforms for interpretability, not on general function decomposition methods for neural network interpretability without input attribution as described in the original paper.

---

10. Neural basis models for interpretability

URL: View paper

Brief Assessment

Neural Basis Models[78] focuses on decomposing shape functions in GAMs using shared basis functions for tabular/image data interpretability, not on general function decomposition methods for transformer LLMs without input attribution as described in the original paper.

Contribution 3: Theoretical grounding of existing interpretability tools

Description: The authors establish a rigorous mathematical foundation using jet operators that subsumes and generalizes existing interpretability techniques like Logit Lens and path expansion methods. This framework provides formal justification for these tools and extends them to new instantiations such as extracting n-gram probability tables directly from LLMs without requiring corpus data.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Metropolis-Hasting based Expanded Path Size Logit model for cyclists' route choice using GPS data

URL: View paper

Brief Assessment

Metropolis-Hasting Path Size[55] addresses cyclist route choice modeling using GPS data and path sampling algorithms, which is entirely unrelated to LLM interpretability frameworks or jet operators.

---

2. Small Vectors, Big Effects: A Mechanistic Study of RL-Induced Reasoning via Steering Vectors

URL: View paper

Brief Assessment

Steering Vectors[53] focuses on mechanistic analysis of RL-trained steering vectors using logit-lens and path-patching, not on establishing mathematical frameworks that unify or generalize interpretability methods like Logit Lens and path expansion.

---

3. A framework for the interpretation of first-order interaction in logit modeling.

URL: View paper

Brief Assessment

The candidate paper (First-Order Interaction Logit[58]) focuses on logit modeling and first-order interactions in a different context, not on interpretability methods for language models like Logit Lens or path expansion techniques.

---

4. Mechanistic Interpretability in the Presence of Architectural Obfuscation

URL: View paper

Brief Assessment

Architectural Obfuscation[54] focuses on privacy-preserving techniques that scramble model representations, not on mathematical frameworks that unify interpretability methods like Logit Lens and path expansion.

---

5. Towards unifying interpretability and control: Evaluation via intervention

URL: View paper

Brief Assessment

Intervention Evaluation[51] focuses on unifying interpretability methods (SAEs, Logit Lens, Tuned Lens, probing) for intervention evaluation, not on providing mathematical foundations using jet operators or generalizing path expansion methods as claimed in the original paper.

---

6. Jet expansions of residual computation

URL: View paper

Prior Art Analysis

Jet Expansions Residual[52] demonstrates that similar theoretical grounding work exists. Both papers establish mathematical frameworks using jet operators to formalize and generalize existing interpretability techniques like Logit Lens. The candidate paper explicitly states it 'grounds and subsumes logit lens' and provides formal justification for these tools, directly paralleling the original's claim of establishing 'rigorous mathematical foundation using jet operators that subsumes and generalizes existing interpretability techniques like Logit Lens.' Both papers use jet operators to provide theoretical foundations for the same interpretability methods, indicating prior work exists in this area.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly claim to ground and subsume the Logit Lens interpretability technique using their jet expansion frameworks, demonstrating that similar theoretical grounding work exists. - **Original**: we show that jetexpansions encompass existing interpretability tools such as the logit lens (nostalgebraist, 2021b), and extend them to new instantiations such as extractingn-gram probability table from llms. - **Candidate**: we demonstrate how our framework grounds and subsumes logit lens, reveals a (super-)exponential path structure in the recursive residual depth and opens up several applications.

Evidence 2 - **Rationale**: Both papers provide formal mathematical frameworks using jet operators to ground the Logit Lens technique, showing that the theoretical grounding claimed by the original paper was already established in prior work. - **Original**: a principled theoretical framework, based on jet operators, formally grounding existing tools such aslogit lens(nostalgebraist, 2021b;a) andpath expansion(elhage et al., 2021). - **Candidate**: jet lenses and logit lens. the logit lens (nostalgebraist, 2021b; geva et al., 2021; 2022; merullo et al., 2023; belrose et al., 2023) is an interpretability method that consists in applying the decoder to intermediate representations as follows: logitlensl(z) = uγ (hl(z)) = j0υ(hl(z))(hl(z)).

Evidence 3 - **Rationale**: Both papers demonstrate how jet operators formally ground the Logit Lens by showing it is equivalent to a zeroth-order jet expansion, establishing the same theoretical connection between jets and existing interpretability tools. - **Original**: thelogit lens(nostalgebraist, 2021b; geva et al., 2021; 2022; merullo et al., 2023; belrose et al., 2023) is a widely used mechanistic interpretability tool that applies the decoder to intermediate hidden states: logitlensℓ(z) =uν(h ℓ(z)) = dec(hℓ(z)). aimed at highlighting the iterative refinement ... - **Candidate**: it is immediate to verify that logitlensl is equivalent to the expansion yielded by jet expand(q, l, {hl}, 0). this suggests two generalizations, which we dub iterative and joint jet lenses, respectively.

Evidence 4 - **Rationale**: Both papers use the same mathematical foundation (jet operators from Ehresmann, 1951) to expand residual computations, showing that the theoretical approach claimed as novel by the original paper was already employed in prior work. - **Original**: noting that llms are particular types of residual networks (he et al., 2016; vaswani et al., 2017), our key idea is to recursively expand residual computations usingjet operators(ehresmann, 1951), the functional counterpart of truncated taylor series. - **Candidate**: to tackle non-linearities and enable expansions in general residual networks similar to that of equation (5), we turn to jets (ehresmann, 1951), which generalize taylor expansions.

---

7. Closed-Form Bayesian Inferences for the Logit Model via Polynomial Expansions

URL: View paper

Brief Assessment

Bayesian Logit Expansions[57] focuses on closed-form Bayesian inference for logit models using polynomial expansions of the likelihood function, not on interpretability techniques for LLMs like Logit Lens or path expansion methods.

---

8. nnterp: A Standardized Interface for Mechanistic Interpretability of Transformers

URL: View paper

Brief Assessment

nnterp[56] focuses on providing a standardized software interface for accessing transformer internals across architectures, not on mathematical frameworks that unify interpretability methods like Logit Lens and path expansion.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 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. Jet expansions of residual computation

Detected in: Contribution: contribution_3

⚠ **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] Jet Expansions: Restructuring LLM Computation for Model Inspection View paper
- [1] Sparse feature circuits: Discovering and editing interpretable causal graphs in language models View paper
- [2] Selection-inference: Exploiting large language models for interpretable logical reasoning View paper
- [3] How do large language models understand relevance? a mechanistic interpretability perspective View paper
- [4] Improving Sparse Decomposition of Language Model Activations with Gated Sparse Autoencoders View paper
- [5] Not All Language Model Features Are One-Dimensionally Linear View paper
- [6] Inference-time decomposition of activations (itda): A scalable approach to interpreting large language models View paper
- [7] Mechanistic interpretability of large language models with applications to the financial services industry View paper
- [8] KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision View paper
- [9] DnA-Eval: Enhancing Large Language Model Evaluation through Decomposition and Aggregation View paper
- [10] How interpretable are reasoning explanations from prompting large language models? View paper
- [11] FiDeLiS: Faithful Reasoning in Large Language Model for Knowledge Graph Question Answering View paper
- [12] Bilinear MLPs enable weight-based mechanistic interpretability View paper

- [13] Path language modeling over knowledge graphsfor explainable recommendation View paper
- [14] Weight-based Analysis of Detokenization in Language Models: Understanding the First Stage of Inference Without Inference View paper
- [15] Information flow routes: Automatically interpreting language models at scale View paper
- [16] Chain-of-Thought for Large Language Model-empowered Wireless Communications View paper
- [17] Towards interpretable and consistent multi-step mathematical reasoning in large language models View paper
- [18] Modular Machine Learning: An Indispensable Path towards New-Generation Large Language Models View paper
- [19] Paths-over-graph: Knowledge graph empowered large language model reasoning View paper
- [20] GraphTrace: A Modular Retrieval Framework Combining Knowledge Graphs and Large Language Models for Multi-Hop Question Answering View paper
- [21] Recap: Transparent inference-time emotion alignment for medical dialogue systems View paper
- [22] Neural synthesis through probabilistic layer decomposition in large language models View paper
- [23] Towards a translative model of Sperm Whale vocalization View paper
- [24] Towards interpretability without sacrifice: Faithful dense layer decomposition with mixture of decoders View paper
- [25] DRKG: Faithful and Interpretable Multi-Hop Knowledge Graph Question Answering via LLM-Guided Reasoning Plans View paper
- [26] Extracting interpretable task-specific circuits from large language models for faster inference View paper
- [27] Route sparse autoencoder to interpret large language models View paper
- [28] Situationally-aware path planning exploiting 3d scene graphs View paper
- [29] Decomposing Complex Questions Makes Multi-Hop QA Easier and More Interpretable View paper
- [30] Spectral Journey: How Transformers Predict the Shortest Path View paper
- [31] Interpretability at scale: Identifying causal mechanisms in alpaca View paper
- [32] Policy-guided path selection and evaluation in multi-step reasoning with large language models View paper
- [33] Efficient automated circuit discovery in transformers using contextual decomposition View paper
- [34] Do Cognitively Interpretable Reasoning Traces Improve LLM Performance? View paper
- [35] Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference View paper
- [36] Mechanistic Interpretability for Progress Towards Quantitative AI Safety View paper
- [37] Text modular networks: Learning to decompose tasks in the language of existing models View paper
- [38] Ecco: An open source library for the explainability of transformer language models View paper
- [39] The grammar-learning trajectories of neural language models View paper
- [40] A Compositional Neural Architecture for Language View paper
- [41] KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts. View paper
- [42] SMART: A Semantic-Guided Reinforcement Learning for Interpretable Feature Engineering View paper
- [43] Towards large-scale interpretable knowledge graph reasoning for dialogue systems View paper
- [44] Neural-ANOVA: Model Decomposition for Interpretable Machine Learning View paper
- [45] Transforming Network Intrusion Detection Using Large Language Models View paper
- [46] Extractive Fact Decomposition for Interpretable Natural Language Inference in one Forward Pass View paper
- [47] Redefining Experts: Interpretable Decomposition of Language Models for Toxicity Mitigation View paper
- [48] Exploring the Generalizability and Explainability of LLMs in Detecting Suicidal Ideation: The Impact of Data Heterogeneity View paper
- [49] A Hierarchical Language Model For Interpretable Graph Reasoning View paper
- [50] Modular Networks: Learning to Decompose Neural Computation View paper
- [51] Towards unifying interpretability and control: Evaluation via intervention View paper
- [52] Jet expansions of residual computation View paper
- [53] Small Vectors, Big Effects: A Mechanistic Study of RL-Induced Reasoning via Steering Vectors View paper
- [54] Mechanistic Interpretability in the Presence of Architectural Obfuscation View paper
- [55] Metropolis-Hasting based Expanded Path Size Logit model for cyclists' route choice using GPS data View paper
- [56] nnterp: A Standardized Interface for Mechanistic Interpretability of Transformers View paper
- [57] Closed-Form Bayesian Inferences for the Logit Model via Polynomial Expansions View paper
- [58] A framework for the interpretation of first-order interaction in logit modeling. View paper
- [59] Unfolding Videos Dynamics via Taylor Expansion View paper
- [60] M-Rule: An Enhanced Deep Taylor Decomposition for Multi-model Interpretability View paper
- [61] Hope: High-order polynomial expansion of black-box neural networks View paper
- [62] GTEA: Guided Taylor Expansion Approximation Network for Optical Flow Estimation View paper
- [63] Towards explaining anomalies: A deep Taylor decomposition of one-class models View paper
- [64] Cat: Interpretable concept-based taylor additive models View paper
- [65] Explaining COVID-19 diagnosis with Taylor decompositions View paper
- [66] An integrated model based on feedforward neural network and Taylor expansion for indicator correlation elimination View paper
- [67] Explaining nonlinear classification decisions with deep taylor decomposition View paper
- [68] Tayloraecnet: A Taylor Style Neural Network For Full-Band Echo Cancellation View paper
- [69] Tensorization of neural networks for improved privacy and interpretability View paper
- [70] A survey on kolmogorov-arnold network View paper
- [71] Multilevel wavelet decomposition network for interpretable time series analysis View paper
- [72] Kolmogorov-Arnold Networks for Interpretable and Efficient Function Approximation View paper
- [73] A comprehensive survey on self-interpretable neural networks View paper
- [74] Beyond the Black Box: A Review of Quantitative Metrics for Neural Network Interpretability and Their Practical Implications View paper
- [75] Neural additive models: Interpretable machine learning with neural nets View paper
- [76] Interpretable basis decomposition for visual explanation View paper
- [77] Tensor Product Neural Networks for Functional ANOVA Model View paper
- [78] Neural basis models for interpretability View paper