Novelty Assessment Report

Paper: Semantic Regexes: Auto-Interpreting LLM Features with a Structured Language

PDF URL: https://openreview.net/pdf?id=6GFznCBscH

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-07

Abstract

Automated interpretability aims to translate large language model (LLM) features into human understandable descriptions. However, natural language feature descriptions are often vague, inconsistent, and require manual relabeling. In response, we introduce semantic regexes, structured language descriptions of LLM features. By combining primitives that capture linguistic and semantic patterns with modifiers for contextualization, composition, and quantification, semantic regexes produce precise and expressive feature descriptions. Across quantitative benchmarks and qualitative analyses, semantic regexes match the accuracy of natural language while yielding more concise and consistent feature descriptions. Their inherent structure affords new types of analyses, including quantifying feature complexity across layers, scaling automated interpretability from insights into individual features to model-wide patterns. Finally, in user studies, we find that semantic regexes help people build accurate mental models of LLM features.

Core Task Landscape

This paper addresses: **Automated Interpretability of Large Language Model Features**

A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

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

- **Feature Extraction and Decomposition Methods**
- **Feature Evaluation and Validation**
- **Feature Application and Analysis**
- **Model Component Analysis**
- **Comprehensive Interpretability Frameworks**
- **Theoretical Foundations and Conceptual Frameworks**
- **Interpretability for Model Control and Steering**
- **Explainability and Downstream Applications**

Complete Taxonomy Tree

- Automated Interpretability of Large Language Model Features Survey Taxonomy
- Feature Extraction and Decomposition Methods
  - Sparse Autoencoder Architectures and Training (5 papers)
  - [6] Sparse Autoencoders Find Highly Interpretable Features in Language Models (Hoagy Cunningham, 2023) View paper
  - [18] A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models (Shu Dong, 2025) View paper
  - [20] The Origins of Representation Manifolds in Large Language Models (Modell, 2025) View paper
  - [25] Improving neuron-level interpretability with white-box language models (Bai Hao, 2024) View paper
  - [27] Route Sparse Autoencoder to Interpret Large Language Models (Wei Shi, 2025) View paper
  - Cross-Layer Feature Alignment and Evolution (1 papers)
  - [21] Mechanistic permutability: Match features across layers (Balagansky, 2024) View paper
  - Domain-Specific Feature Extraction (3 papers)
  - [19] Sparse autoencoders uncover biologically interpretable features in protein language model representations (Onkar Gujral, 2025) View paper
  - [22] From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models (Etowah Adams, 2025) View paper
  - [24] Mechanistic interpretability for steering vision-language-action models (Tomlin, 2025) View paper
- Feature Evaluation and Validation
  - Automated Feature Description Generation (2 papers)
  - [3] Automatically interpreting millions of features in large language models (Paulo Goncalo, 2024) View paper
  - [10] Enhancing Automated Interpretability with Output-Centric Feature Descriptions (Geiger, 2025) View paper
  - Evaluation Without Explanations (1 papers)
  - [30] Evaluating SAE interpretability without explanations (Paulo Goncalo, 2025) View paper
  - Structured Feature Representation Frameworks ★ (1 papers)
  - [0] Semantic Regexes: Auto-Interpreting LLM Features with a Structured Language (Anon et al., 2026) View paper
- Feature Application and Analysis
  - Causal Feature Circuits and Interventions (2 papers)
  - [13] Sparse feature circuits: Discovering and editing interpretable causal graphs in language models (Marks, 2024) View paper
  - [39] Towards unifying interpretability and control: Evaluation via intervention (Bhalla, 2024) View paper
  - Task-Specific Mechanistic Analysis (5 papers)

  - [2] Mechanistic Interpretability of Emotion Inference in Large Language Models (Gratch, 2025) View paper
  - [23] Towards a mechanistic interpretation of multi-step reasoning capabilities of language models (Bosselut, 2023) View paper
  - [26] I Have Covered All the Bases Here: Interpreting Reasoning Features in Large Language Models via Sparse Autoencoders (Druzhinina, 2025) View paper
  - [38] A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis (Alessandro Stolfo, 2023) View paper
  - [44] Unlocking the Future: Exploring Look-Ahead Planning Mechanistic Interpretability in Large Language Models (Cao Pengfei, 2024) View paper
  - Bias and Safety Mechanism Analysis (2 papers)
  - [12] Mechanistic Interpretability with SAEs: Probing Religion, Violence, and Geography in Large Language Models (Simbeck, 2025) View paper
  - [40] On the Role of Attention Heads in Large Language Model Safety (Zhou Zhen-hong, 2024) View paper
  - Failure Mode Investigation (3 papers)
  - [14] Mechanistic understanding and mitigation of language model non-factual hallucinations (Lei Yu, 2024) View paper
  - [35] Understanding Jailbreak Success: A Study of Latent Space Dynamics in Large Language Models (Ball Sarah, 2024) View paper
  - [37] Understanding the Repeat Curse in Large Language Models from a Feature Perspective (Hu Lijie, 2025) View paper
  - Adversarial Interpretability (1 papers)
  - [15] Using Mechanistic Interpretability to Craft Adversarial Attacks against Large Language Models (Addad, 2025) View paper
  - Geospatial and Domain-Specific Concept Analysis (1 papers)
  - [29] Geospatial Mechanistic Interpretability of Large Language Models (Mizzaro, 2025) View paper
- Model Component Analysis
  - Layer-Level Analysis (1 papers)
  - [5] Investigating layer importance in large language models (Yang Zhang, 2024) View paper
  - Attention Mechanism Analysis (1 papers)
  - [32] Mechanistic Fine-tuning for In-context Learning (Cho Hakaze, 2025) View paper
  - Neuron-Level Interpretability (1 papers)
  - [1] Dissociating language and thought in large language models (Kyle Mahowald, 2024) View paper
- Comprehensive Interpretability Frameworks
  - Unified Interpretability Platforms (1 papers)
  - [8] Mechanistic understanding and validation of large AI models with SemanticLens (Maximilian Dreyer, 2025) View paper
  - Interactive Interpretability Tools (1 papers)
  - [42] LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations (Feldhus, 2024) View paper
  - Multimodal Interpretability (1 papers)
  - [48] A survey on mechanistic interpretability for multi-modal foundation models (Lin Zi-hao, 2025) View paper
- Theoretical Foundations and Conceptual Frameworks
  - Cognitive and Philosophical Perspectives (2 papers)
  - [33] Mechanistic Indicators of Understanding in Large Language Models (Beckmann, 2025) View paper
  - [50] The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms (Davies, 2024) View paper
  - Survey and Review Literature (2 papers)
  - [11] Exploring Mechanistic Interpretability in Large Language Models: Challenges, Approaches, and Insights (Sandeep Reddy Gantla, 2025) View paper
  - [16] A practical review of mechanistic interpretability for transformer-based language models (Rai, 2024) View paper
- Interpretability for Model Control and Steering
  - Efficiency and Optimization via Interpretability (1 papers)
  - [36] Saliency-driven Dynamic Token Pruning for Large Language Models (Tao Yao, 2025) View paper
  - Interpretable Model Enhancement (1 papers)
  - [7] Augmenting interpretable models with large language models during training (Chandan Singh, 2023) View paper
- Explainability and Downstream Applications
  - Domain-Specific Explainability Applications (3 papers)
  - [45] Towards Interpretable Mental Health Analysis with Large Language Models (Ananiadou, 2023) View paper
  - [46] CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models (McInerney, 2023) View paper
  - [49] Beyond the black box: Interpretability of llms in finance (Tatsat, 2025) View paper
  - Explainability for Reasoning and Knowledge Tasks (3 papers)
  - [9] Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning (Luo, 2023) View paper
  - [31] Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Graph Representation Learning (HE Xiaoxin, 2023) View paper
  - [43] Large language models for automated data science: Introducing caafe for context-aware automated feature engineering (Hollmann, 2023) View paper
  - Neuroscience and Cognitive Alignment (1 papers)
  - [28] Explanations of Large Language Models Explain Language Representations in the Brain (Rahimi Maryam, 2025) View paper
  - General Explainability Frameworks and Challenges (3 papers)
  - [4] Rethinking Interpretability in the Era of Large Language Models (Singh Chandan, 2024) View paper
  - [17] Exploring Explainability in Large Language Models (Fen Yin, 2025) View paper
  - [34] Explaining black-box behavior in large language models (Rahul Manche, 2022) View paper
  - Scalable Feature Interaction Analysis (2 papers)
  - [41] ProxySPEX: Inference-Efficient Interpretability via Sparse Feature Interactions in LLMs (Butler, 2025) View paper
  - [47] Interpreting learned feedback patterns in large language models (L Marks, 2024) View paper

Narrative

Core task: Automated interpretability of large language model features. The field has organized itself around several complementary branches that together address how to extract, validate, and apply interpretable features from LLMs. Feature Extraction and

Decomposition Methods focus on techniques like sparse autoencoders (Sparse Autoencoders Features[6], Sparse Autoencoders Survey[18]) that decompose neural activations into more interpretable components. Feature Evaluation and Validation develops metrics and frameworks to assess whether extracted features are genuinely meaningful, while Feature Application and Analysis explores how these features can be used for tasks ranging from circuit discovery (Sparse Feature Circuits[13]) to domain-specific applications. Model Component Analysis examines specific architectural elements like attention heads or layers, and Comprehensive Interpretability Frameworks integrate multiple methods into unified systems. Theoretical Foundations provide conceptual grounding, while Interpretability for Model Control and Steering and Explainability and Downstream Applications translate insights into practical interventions and real-world uses.

A particularly active tension exists between automated feature discovery methods and structured validation approaches. Many studies pursue scalable extraction techniques that can handle the vast dimensionality of modern LLMs, yet questions remain about how to systematically verify that discovered features correspond to meaningful semantic concepts. Semantic Regexes[0] sits within the Structured Feature Representation Frameworks cluster, emphasizing formal methods for representing and validating feature semantics—a contrast to purely automated approaches like Automatically Interpreting Features[3] that prioritize scalability over structured guarantees. This work shares common ground with efforts like SemanticLens[8] and Output-Centric Features[10] that also seek principled ways to characterize what features represent, but differs in its emphasis on regex-like compositional structures for feature descriptions. The broader challenge across these branches remains balancing automation with interpretability rigor.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on formal, structured representation systems for describing neural network features with precision and compositionality, moving beyond natural language. Its siblings represent alternative approaches to feature interpretability: one uses LLMs to generate natural language explanations at scale with evaluation protocols, while the other bypasses explanations entirely by evaluating interpretability directly from activation patterns. Together, these three approaches span a spectrum from structured formal languages to natural language to explanation-free methods.

Similarities: - All three subtopics address the core challenge of making neural network features interpretable and understandable - Each approach requires some form of evaluation or validation to assess the quality of interpretability - All three are concerned with scalability and systematic methods rather than manual feature analysis

Differences: - Structured Feature Representation Frameworks uses formal languages and structured formats, while Automated Feature Description Generation relies on natural language explanations - Evaluation Without Explanations bypasses the explanation generation step entirely, directly assessing interpretability from activations, whereas the other two subtopics produce explicit descriptions - The original leaf emphasizes precision, consistency, and compositionality through formal representations, while natural language approaches prioritize human readability and accessibility - Automated Feature Description Generation explicitly incorporates LLMs in the pipeline and quantitative evaluation protocols, while the original leaf focuses on representation formalisms themselves

Suggested Search Directions: - Hybrid approaches combining structured representations with natural language generation - Comparative studies evaluating trade-offs between formal precision and natural language accessibility - Methods for translating between structured feature representations and natural language explanations

Sibling Subtopics

- **Automated Feature Description Generation** (leaves: 1, papers: 2)
- Scope: Pipelines using LLMs to generate natural language explanations for features at scale with quantitative evaluation protocols.
- Exclude: Excludes structured or non-natural-language description formats; those belong in Structured Feature Representation Frameworks.
- **Evaluation Without Explanations** (leaves: 1, papers: 1)
- Scope: Methods for assessing feature interpretability directly from activation patterns without requiring intermediate natural language descriptions.
- Exclude: Excludes explanation-based evaluation methods; those belong in Automated Feature Description Generation.

Contributions Analysis

Overall novelty summary. The paper introduces semantic regexes, a structured language for describing LLM features through compositional primitives and modifiers. It resides in the Structured Feature Representation Frameworks leaf, which contains only this single paper within the broader Feature Evaluation and Validation branch. This positioning reflects a relatively sparse research direction focused on formal, structured alternatives to natural language feature descriptions. While the parent branch addresses feature validation more broadly, this specific leaf represents a novel approach to the representation problem itself rather than evaluation metrics or automated explanation generation.

The taxonomy reveals that most neighboring work pursues different validation strategies. The sibling leaf Automated Feature Description Generation contains papers like Automatically Interpreting Features and Neuron Descriptions that generate natural language explanations at scale, prioritizing automation over structural guarantees. Another sibling, Evaluation Without Explanations, bypasses linguistic descriptions entirely. The broader Feature Extraction branch focuses on discovery methods like sparse autoencoders, while Feature Application explores downstream uses of extracted features. Semantic regexes occupy a distinct niche by providing formal compositional structure for feature descriptions, bridging the gap between automated generation and rigorous representation.

Among twenty-nine candidates examined across three contributions, none clearly refute the core claims. The semantic regex language itself (ten candidates examined, zero refutable) appears novel as a structured formalism combining linguistic and semantic primitives. The primitives and modifiers framework (nine candidates, zero refutable) shows no direct prior work on this specific compositional approach. Model-wide complexity analysis using regex structure (ten candidates, zero refutable) likewise lacks clear precedent. This limited search scope suggests the structured regex formalism represents a genuinely unexplored direction, though the analysis cannot rule out relevant work outside the top-K semantic matches or citation network examined.

Based on this constrained literature search, the work appears to introduce a novel representational framework within a sparsely populated research direction. The absence of sibling papers in its taxonomy leaf and zero refutable candidates across contributions suggest substantive originality, though the twenty-nine-paper scope leaves open the possibility of overlooked related work in formal methods or program synthesis communities outside the core interpretability literature examined here.

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

Contribution 1: Semantic regexes: a structured language for LLM feature descriptions

Description: The authors propose semantic regexes, a structured language that describes LLM features by combining primitives (symbols, lexemes, fields) with modifiers (context, composition, quantification) to produce precise and expressive feature descriptions that are more concise and consistent than natural language.

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. Improving interpretability of deep neural networks with semantic information

URL: View paper

Brief Assessment

Semantic Information Networks[80] focuses on improving interpretability of DNNs for video captioning using topic models extracted from human descriptions, not on creating a structured language for describing LLM features. The candidate addresses a different domain (video analysis) and uses semantic information differently (topic modeling vs. structured regex primitives).

2. Foundations of symbolic languages for model interpretability

URL: View paper

Brief Assessment

Symbolic Languages Foundations[72] focuses on interpretability queries for decision models (decision trees, OBDDs, perceptrons) using first-order logic (FOIL), not on structured languages for describing individual LLM features like semantic regexes.

3. An Interpretable Dynamic Inference System Based on Fuzzy Broad Learning

URL: View paper

Brief Assessment

Fuzzy Broad Learning[77] focuses on fuzzy neural models for classification tasks with interpretable linguistic fuzzy rules, not on structured languages for describing LLM features or neural network interpretability in language models.

4. Neurons to Words: A Novel Method for Automated Neural Network Interpretability and Alignment

URL: View paper

Brief Assessment

Neurons to Words[75] focuses on translating neural network activations into natural language descriptions using LLMs, not on creating a structured language with primitives and modifiers for feature descriptions. The candidate uses character-encoding schemes for activations rather than semantic regexes.

5. Local interpretations for explainable natural language processing: A survey

URL: View paper

Brief Assessment

Local Interpretations Survey[76] focuses on interpretability methods for NLP model predictions and hidden states, not on structured languages for describing individual LLM features extracted via sparse autoencoders.

6. Natural language descriptions of deep visual features

URL: View paper

Brief Assessment

Visual Features Descriptions[71] focuses on generating natural language descriptions for visual features in computer vision models using image captioning techniques, not structured languages for LLM features.

7. Weighted automata extraction and explanation of recurrent neural networks for natural language tasks

URL: View paper

Brief Assessment

Weighted Automata Extraction[79] focuses on extracting weighted automata from RNNs for natural language tasks, not on creating structured languages for describing LLM features. The candidate paper's full text is not available (marked as 'n/a'), preventing detailed comparison.

8. Enhancing Explainability and Accelerating Materials Science Design with Linguistic Summaries

URL: View paper

Brief Assessment

Linguistic Summaries Materials[78] focuses on fuzzy linguistic summaries for materials science optimization, not structured languages for neural network feature interpretability. The domains and technical approaches are entirely different.

9. Causal abstractions of neural networks

URL: View paper

Brief Assessment

Causal Abstractions[73] focuses on causal abstraction analysis of neural networks using interchange interventions to verify alignment between neural representations and interpretable causal models. This is fundamentally different from the original paper's contribution of semantic regexes, which is a structured language for describing LLM features using primitives like symbols, lexemes, and fields with modifiers for context and composition.

10. Linguistic Interpretability of Transformer-based Language Models: a systematic review

URL: View paper

Brief Assessment

Linguistic Interpretability Review[74] focuses on linguistic knowledge discovery in transformer-based language models using techniques like probing and embedding analysis, not on structured languages for feature descriptions in interpretability pipelines.

Contribution 2: Primitives and modifiers for capturing linguistic and semantic patterns

Description: The authors develop a system of human-interpretable primitives (symbols for exact strings, lexemes for syntactic variants, fields for semantic categories) and modifiers (context, composition, quantification) that enable semantic regexes to express diverse feature activation patterns from simple token detectors to complex linguistic phenomena.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Primitive Generation and Semantic-Related Alignment for Universal Zero-Shot Segmentation

URL: View paper

Brief Assessment

Primitive Generation Segmentation[63] uses primitives for visual feature synthesis in zero-shot segmentation, not for capturing linguistic and semantic patterns in LLM features as in the original paper.

2. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations

URL: View paper

Brief Assessment

Visual-Semantic Embeddings[61] focuses on aligning visual and textual representations through object, attribute, and relation embeddings for cross-modal retrieval, not on developing primitives and modifiers for interpreting LLM feature activation patterns as in the original paper's semantic regex language.

3. Learning Structured Natural Language Representations for Semantic Parsing

URL: View paper

Brief Assessment

Structured Semantic Parsing[64] focuses on learning predicate-argument structures for semantic parsing tasks, not on developing primitives and modifiers for feature interpretability in LLMs. The paper addresses a different problem domain (semantic parsing vs. automated interpretability).

4. Dialectal phraseological units of the Yakut language: structure and semantics

URL: View paper

Brief Assessment

Yakut Phraseological Units[66] focuses on dialectal phraseological units in the Yakut language using traditional linguistic methods (component analysis, semantic-thematic grouping), not on developing computational primitives and modifiers for automated feature interpretation in LLMs.

5. Structural and semantic features of linguistic units of the English-language linguocultural scenario â□□Productsâ□□

URL: View paper

Brief Assessment

Linguocultural Products[65] focuses on identifying structural and semantic features of linguistic units in the thematic area 'Products' using a cognitive-cultural approach. It does not develop a system of primitives and modifiers for feature activation patterns in LLMs.

6. Semantic construction in feature-based TAG

URL: View paper

Brief Assessment

Feature-Based TAG[69] focuses on semantic construction in tree adjoining grammar using primitives like symbols, lexemes, and fields for semantic representation in a formal grammar framework, not on feature activation patterns in neural networks or LLMs as in the original paper.

7. A preferential, pattern-seeking, semantics for natural language inference

URL: View paper

Brief Assessment

Pattern-Seeking Semantics[67] discusses semantic primitives in a theoretical linguistic framework, not a system for automated LLM feature interpretation with structured primitives like symbols, lexemes, and fields combined with modifiers.

8. Using slots and modifiers in logic grammars for natural language

URL: View paper

Brief Assessment

Slots and Modifiers[70] focuses on logic grammars for natural language parsing with Prolog-based semantic interpretation, not on automated interpretability of LLM features or structured descriptions of neural network activation patterns.

9. Structural and semantic features of adjectives across languages and registers

URL: View paper

Brief Assessment

Adjectives Semantic Features[62] analyzes structural and semantic features of adjectives across languages and registers using a linguistic framework. It does not develop a system of primitives and modifiers for automated feature description or semantic regexes.

Contribution 3: Model-wide analysis of feature complexity using semantic regex structure

Description: The authors demonstrate that the inherent structure of semantic regexes enables new types of analyses, such as quantifying feature complexity across model layers by measuring the abstraction level and number of components in semantic regexes, thereby scaling interpretability from individual features to model-wide patterns.

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. Complexity of representations in deep learning

URL: View paper

Brief Assessment

Complexity of Representations[54] analyzes feature complexity across neural network layers using nearest neighbor error rates and data complexity measures, not semantic regex structures or interpretability of individual features.

2. Deep grokking: Would deep neural networks generalize better?

URL: View paper

Brief Assessment

Deep Grokking[59] focuses on analyzing feature rank evolution across layers in deep neural networks during training dynamics and grokking phenomena, not on automated interpretability or structured language descriptions of features.

3. Going beyond neural network feature similarity: The network feature complexity and its interpretation using category theory

URL: View paper

Brief Assessment

Network Feature Complexity[55] focuses on quantifying redundancy in neural network features through functionally equivalent features and category theory, not on interpretability methods using structured language descriptions like semantic regexes for LLM features.

---

4. How deep is deep enough?--Quantifying class separability in the hidden layers of deep neural networks

URL: View paper

Brief Assessment

Class Separability Depth[57] focuses on quantifying class separability across layers using the GDV metric in classification networks, not on analyzing feature complexity through structured language descriptions like semantic regexes.

---

5. Formation of Representations in Neural Networks

URL: View paper

Brief Assessment

Formation of Representations[53] focuses on analyzing neural network layers through alignment relations between representations, weights, and gradients, not on quantifying feature complexity or interpretability of individual features across layers.

---

6. Pan-microalgal dark proteome mapping via interpretable deep learning and synthetic chimeras

URL: View paper

Brief Assessment

Dark Proteome Mapping[58] focuses on protein sequence classification in microalgae using transformer models and interpretability tools like HELIX and DeepLift. It does not address feature complexity quantification across neural network layers using structured language descriptions like semantic regexes.

---

7. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream

URL: View paper

Brief Assessment

Ventral Stream Complexity[60] focuses on quantifying feature complexity across layers in deep neural networks modeling the visual ventral stream, not on interpretability of LLM features using structured language descriptions like semantic regexes.

---

8. Intrinsic dimension of data representations in deep neural networks

URL: View paper

Brief Assessment

Intrinsic Dimension[52] focuses on measuring the intrinsic dimensionality of data representations across neural network layers using geometric properties of manifolds, not on analyzing feature complexity through structured language descriptions like semantic regexes.

---

9. How transferable are features in deep neural networks?

URL: View paper

Brief Assessment

Transferable Features[51] analyzes layer-wise feature transferability in neural networks but does not propose semantic regexes or structured language descriptions. The paper focuses on transfer learning experiments rather than interpretability methods for quantifying feature complexity.

---

10. Hybrid parallel fuzzy CNN paradigm: Unmasking intricacies for accurate brain MRI insights

URL: View paper

Brief Assessment

Hybrid Fuzzy CNN[56] focuses on medical image analysis for brain MRI diagnosis using fuzzy logic and CNNs, not on interpretability methods for quantifying feature complexity across neural network layers using structured language descriptions.

Appendix: Text Similarity Detection

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

References

- [0] Semantic Regexes: Auto-Interpreting LLM Features with a Structured Language View paper
- [1] Dissociating language and thought in large language models View paper
- [2] Mechanistic Interpretability of Emotion Inference in Large Language Models View paper
- [3] Automatically interpreting millions of features in large language models View paper
- [4] Rethinking Interpretability in the Era of Large Language Models View paper
- [5] Investigating layer importance in large language models View paper
- [6] Sparse Autoencoders Find Highly Interpretable Features in Language Models View paper
- [7] Augmenting interpretable models with large language models during training View paper
- [8] Mechanistic understanding and validation of large AI models with SemanticLens View paper
- [9] Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning View paper
- [10] Enhancing Automated Interpretability with Output-Centric Feature Descriptions View paper
- [11] Exploring Mechanistic Interpretability in Large Language Models: Challenges, Approaches, and Insights View paper
- [12] Mechanistic Interpretability with SAEs: Probing Religion, Violence, and Geography in Large Language Models View paper
- [13] Sparse feature circuits: Discovering and editing interpretable causal graphs in language models View paper
- [14] Mechanistic understanding and mitigation of language model non-factual hallucinations View paper
- [15] Using Mechanistic Interpretability to Craft Adversarial Attacks against Large Language Models View paper
- [16] A practical review of mechanistic interpretability for transformer-based language models View paper
- [17] Exploring Explainability in Large Language Models View paper
- [18] A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models View paper
- [19] Sparse autoencoders uncover biologically interpretable features in protein language model representations View paper

- [20] The Origins of Representation Manifolds in Large Language Models View paper
- [21] Mechanistic permutability: Match features across layers View paper
- [22] From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models View paper
- [23] Towards a mechanistic interpretation of multi-step reasoning capabilities of language models View paper
- [24] Mechanistic interpretability for steering vision-language-action models View paper
- [25] Improving neuron-level interpretability with white-box language models View paper
- [26] I Have Covered All the Bases Here: Interpreting Reasoning Features in Large Language Models via Sparse Autoencoders View paper
- [27] Route Sparse Autoencoder to Interpret Large Language Models View paper
- [28] Explanations of Large Language Models Explain Language Representations in the Brain View paper
- [29] Geospatial Mechanistic Interpretability of Large Language Models View paper
- [30] Evaluating SAE interpretability without explanations View paper
- [31] Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Graph Representation Learning View paper
- [32] Mechanistic Fine-tuning for In-context Learning View paper
- [33] Mechanistic Indicators of Understanding in Large Language Models View paper
- [34] Explaining black-box behavior in large language models View paper
- [35] Understanding Jailbreak Success: A Study of Latent Space Dynamics in Large Language Models View paper
- [36] Saliency-driven Dynamic Token Pruning for Large Language Models View paper
- [37] Understanding the Repeat Curse in Large Language Models from a Feature Perspective View paper
- [38] A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis View paper
- [39] Towards unifying interpretability and control: Evaluation via intervention View paper
- [40] On the Role of Attention Heads in Large Language Model Safety View paper
- [41] ProxySPEX: Inference-Efficient Interpretability via Sparse Feature Interactions in LLMs View paper
- [42] LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations View paper
- [43] Large language models for automated data science: Introducing caafe for context-aware automated feature engineering View paper
- [44] Unlocking the Future: Exploring Look-Ahead Planning Mechanistic Interpretability in Large Language Models View paper
- [45] Towards Interpretable Mental Health Analysis with Large Language Models View paper
- [46] CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models View paper
- [47] Interpreting learned feedback patterns in large language models View paper
- [48] A survey on mechanistic interpretability for multi-modal foundation models View paper
- [49] Beyond the black box: Interpretability of llms in finance View paper
- [50] The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms View paper
- [51] How transferable are features in deep neural networks? View paper
- [52] Intrinsic dimension of data representations in deep neural networks View paper
- [53] Formation of Representations in Neural Networks View paper
- [54] Complexity of representations in deep learning View paper
- [55] Going beyond neural network feature similarity: The network feature complexity and its interpretation using category theory View paper
- [56] Hybrid parallel fuzzy CNN paradigm: Unmasking intricacies for accurate brain MRI insights View paper
- [57] How deep is deep enough?--Quantifying class separability in the hidden layers of deep neural networks View paper
- [58] Pan-microalgal dark proteome mapping via interpretable deep learning and synthetic chimeras View paper
- [59] Deep grokking: Would deep neural networks generalize better? View paper
- [60] Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream View paper
- [61] Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations View paper
- [62] Structural and semantic features of adjectives across languages and registers View paper
- [63] Primitive Generation and Semantic-Related Alignment for Universal Zero-Shot Segmentation View paper
- [64] Learning Structured Natural Language Representations for Semantic Parsing View paper
- [65] Structural and semantic features of linguistic units of the English-language linguocultural scenario â□□Productsâ□□ View paper
- [66] Dialectal phraseological units of the Yakut language: structure and semantics View paper
- [67] A preferential, pattern-seeking, semantics for natural language inference View paper
- [68] Naive semantics for natural language understanding View paper
- [69] Semantic construction in feature-based TAG View paper
- [70] Using slots and modifiers in logic grammars for natural language View paper
- [71] Natural language descriptions of deep visual features View paper
- [72] Foundations of symbolic languages for model interpretability View paper
- [73] Causal abstractions of neural networks View paper
- [74] Linguistic Interpretability of Transformer-based Language Models: a systematic review View paper
- [75] Neurons to Words: A Novel Method for Automated Neural Network Interpretability and Alignment View paper
- [76] Local interpretations for explainable natural language processing: A survey View paper
- [77] An Interpretable Dynamic Inference System Based on Fuzzy Broad Learning View paper
- [78] Enhancing Explainability and Accelerating Materials Science Design with Linguistic Summaries View paper
- [79] Weighted automata extraction and explanation of recurrent neural networks for natural language tasks View paper
- [80] Improving interpretability of deep neural networks with semantic information View paper