Novelty Assessment Report

Paper: Temporal Sparse Autoencoders: Leveraging the Sequential Nature of Language for Interpretability
PDF URL: https://openreview.net/pdf?id=bojVI4l9Kn
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Translating the internal representations and computations of models into concepts that humans can understand is a key goal of interpretability. While recent dictionary learning methods such as Sparse Autoencoders (SAEs) provide a promising route to discover human-interpretable features, they often only recover token-specific, noisy, or highly local concepts. We argue that this limitation stems from neglecting the temporal structure of language, where semantic content typically evolves smoothly over sequences. Building on this insight, we introduce Temporal Sparse Autoencoders (T-SAEs), which incorporate a novel contrastive loss encouraging consistent activations of high-level features over adjacent tokens. This simple yet powerful modification enables SAEs to disentangle semantic from syntactic features in a self-supervised manner. Across multiple datasets and models, T-SAEs recover smoother, more coherent semantic concepts without sacrificing reconstruction quality. Strikingly, they exhibit clear semantic structure despite being trained without explicit semantic signal, offering a new pathway for unsupervised interpretability in language models.

Core Task Landscape

This paper addresses: **Discovering Interpretable Semantic Features in Language Model Representations**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Sparse Dictionary Learning Methods**
- **Embedding Space Analysis and Geometry**
- **Layer-Wise Representation Dynamics**
- **Interpretability Evaluation and Benchmarking**
- **Semantic Representation Learning Foundations**
- **Implicit Semantic Representations in Neural Models**
- **Self-Interpretation and Explanation Generation**
- **Representation Manipulation and Control**
- **Semantic Feature Discovery in Specialized Contexts**
- **Semantic Clustering and Hierarchical Structure**
- ... and 2 more categories

Complete Taxonomy Tree

- Discovering Interpretable Semantic Features in Language Model Representations Survey Taxonomy
- Sparse Dictionary Learning Methods
  - Standard Sparse Autoencoder Approaches (2 papers)
  - [2] Sparse autoencoders find highly interpretable features in language models (Hoagy Cunningham, 2023) View paper
  - [49] Automatically Interpreting Millions of Features in Large Language Models (Paulo Goncalo, 2024) View paper
  - Temporal and Sequential Extensions ★ (1 papers)
  - [0] Temporal Sparse Autoencoders: Leveraging the Sequential Nature of Language for Interpretability (Anon et al., 2026) View paper
  - Multi-Dimensional Feature Discovery (1 papers)
  - [16] Not all language model features are one-dimensionally linear (Engels, 2024) View paper
  - Domain-Specific Sparse Autoencoders (2 papers)
  - [5] Sparse autoencoders uncover biologically interpretable features in protein language model representations (Onkar Gujral, 2025) View paper
  - [29] SAE-RNA: A Sparse Autoencoder Model for Interpreting RNA Language Model Representations (Kim, 2025) View paper
- Embedding Space Analysis and Geometry
  - Semantic Structure and Dimensionality (3 papers)
  - [8] On the sentence embeddings from pre-trained language models (Li, 2020) View paper
  - [20] Semantic Structure in Large Language Model Embeddings (Austin C. Kozlowski, 2025) View paper
  - [34] Shared Global and Local Geometry of Language Model Embeddings (Lee Andrew, 2025) View paper
  - Stylistic and Functional Feature Representation (2 papers)
  - [6] Word embeddings are steers for language models (Chi Han, 2024) View paper
  - [48] Representation of Lexical Stylistic Features in Language Modelsâ Embedding Space (Qing Lyu, 2023) View paper
  - Cross-Modal and Grounded Embeddings (2 papers)
  - [25] Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations (Hao Wu, 2019) View paper

- [43] Explainable semantic space by grounding language to vision with cross-modal contrastive learning (Yizhen Zhang, 2021) View paper
- Specialized Embedding Applications (4 papers)
- [7] Improving uncertainty quantification in large language models via semantic embeddings (Bonilla, 2024) View paper
- [26] Neural semantic tagging for natural language-based search in building information models: Implications for practice (Mehrzad Shahinmoghadam, 2024) View paper
- [31] A full-document analysis of the semantic relation between European Public Assessment Reports and EMA guidelines using a BERT language model (Erik Bergman, 2023) View paper
- [39] Sentiment embeddings with applications to sentiment analysis (Duyu Tang, 2015) View paper

- Layer-Wise Representation Dynamics (2 papers)
- [3] Layer by layer: Uncovering hidden representations in language models (Skean, 2025) View paper
- [4] Context-preserving latent field interpolation for large language model internal state manipulation (Watson, 2025) View paper

- Interpretability Evaluation and Benchmarking (1 papers)
- [1] RAVEL: Evaluating interpretability methods on disentangling language model representations (Geiger, 2024) View paper

- Semantic Representation Learning Foundations
- Causal and Conceptual Representation Learning (2 papers)
- [9] Learning interpretable concepts: Unifying causal representation learning and foundation models (Rajendran, 2024) View paper
- [38] Large Language Models are Interpretable Learners (Wang Ruochen, 2024) View paper
- Distributional and Contextual Semantics (4 papers)
- [22] Word embedding for understanding natural language: a survey (Yang Li, 2017) View paper
- [32] Multi-Relational Hyperbolic Word Embeddings from Natural Language Definitions (Marco Valentino, 2023) View paper
- [33] Exploiting latent semantic information in statistical language modeling (J.R. Bellegarda, 2002) View paper
- [37] Improving lexical embeddings with semantic knowledge (Mo Yu, 2014) View paper
- Semantic Enhancement and Refinement (2 papers)
- [12] Structural Embedding Projection for Contextual Large Language Model Inference (Vincent Enoasmo, 2025) View paper
- [35] Improving the Factuality of Abstractive Text Summarization with Syntactic Structure-Aware Latent Semantic Space (Jianbin Shen, 2024) View paper

- Implicit Semantic Representations in Neural Models (4 papers)
- [10] Implicit representations of meaning in neural language models (Belinda Z. Li, 2021) View paper
- [13] The reasoning-memorization interplay in language models is mediated by a single direction (Yihuai Hong, 2025) View paper
- [15] Emergent representations of program semantics in language models trained on programs (Jin Charles, 2024) View paper
- [44] Mechanistic Interpretability of Socio-Political Frames in Language Models (Asghari Hadi, 2025) View paper

- Self-Interpretation and Explanation Generation (3 papers)
- [41] Definition modeling: Learning to define word embeddings in natural language (Thanapon Noraset, 2017) View paper
- [45] SelfIE: Self-Interpretation of Large Language Model Embeddings (Chen Haozhe, 2024) View paper
- [46] Connecting Concept Layers and Rationales to Enhance Language Model Interpretability (Thomas Bailleux, 2025) View paper

- Representation Manipulation and Control (3 papers)
- [14] Simulated echo shaping in large language models via semantic phase perturbation without intermediate token realignment (Allan, 2025) View paper
- [21] Structural recomposition in large language models through lexico-semantic vector fusion: A computational study (James, 2025) View paper
- [30] Refusal-Aware Red Teaming: Exposing Inconsistency in Safety Evaluations (Yongkang Chen, 2025) View paper

- Semantic Feature Discovery in Specialized Contexts (5 papers)
- [19] Semantic Analysis of test items through Large Language Model embeddings predicts a-priori factorial structure of personality tests (Nicola Milano, 2025) View paper
- [23] Neural models for semantic analysis of handwritten document images (Oliver TÃ¼selmann, 2024) View paper
- [27] Decoding Neural Emotion Patterns through Large Language Model Embeddings (Vos, 2025) View paper
- [28] Meaning modulations and stability in large language models: An analysis of BERT embeddings for psycholinguistic research (Giovanni Cassani, 2023) View paper
- [47] Exploring Embedding Interpretability by Correspondences Between Topic Models and Text Embeddings (M Yuan, 2025) View paper

- Semantic Clustering and Hierarchical Structure (3 papers)
- [40] Semantic Tree Inference on Text Corpa using a Nested Density Approach together with Large Language Model Embeddings (Thomas Haschka, 2025) View paper
- [42] On the Relationship Between RNN Hidden-State Vectors and Semantic Structures (K Aichernig Bernhard, 2024) View paper
- [50] Interpretability of BERT latent space through knowledge graphs (Vito Walter Anelli, 2022) View paper

- Latent Factor and Hidden Variable Discovery (3 papers)
- [11] Discovery of the hidden world with large language models (Yongqiang Chen, 2024) View paper
- [18] CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models (Denis McInerney, 2023) View paper
- [36] Uncovering Latent Human Wellbeing in Language Model Embeddings (Pedro Freire, 2024) View paper

- Semantic Representation Evaluation and Analysis (2 papers)
- [17] Analyzing the role of semantic representations in the era of large language models (Zhijing Jin, 2024) View paper
- [24] Interpretable Semantic Representations from Neural Language Models and Computer Vision (Derby, 2022) View paper

Narrative

Core task: Discovering interpretable semantic features in language model representations. The field has organized itself around several complementary perspectives. Sparse Dictionary Learning Methods (including foundational work like Sparse Autoencoders Interpretability[2] and domain-specific extensions such as Protein SAE Features[5] and SAE RNA[29]) aim to decompose dense activations into sparse, human-interpretable feature sets. Embedding Space Analysis and Geometry explores the structural properties of learned representations, examining how semantic relationships manifest in vector spaces (e.g., Word Embedding Survey[22], Hyperbolic Word Embeddings[32]). Layer-Wise Representation Dynamics investigates how meaning evolves across network depth (Layer by Layer[3]), while Interpretability Evaluation and Benchmarking develops systematic ways to assess feature quality (Automatically Interpreting Features[49]). Additional branches address semantic representation learning foundations, implicit representations in neural models (Implicit Meaning Representations[10]), self-interpretation mechanisms (SelfIE[45]), representation manipulation and control

(Word Embeddings Steers[6]), and specialized contexts ranging from emotion decoding (Decoding Emotion Patterns[27]) to sociopolitical frames (Sociopolitical Frames Interpretability[44]).

A particularly active tension exists between static feature extraction and dynamic, context-sensitive approaches. Many studies focus on extracting fixed dictionaries of features, but a growing line of work examines how features evolve temporally or across contexts (Context Preserving Interpolation[4], Reasoning Memorization Direction[13]). Temporal Sparse Autoencoders[0] sits squarely within the Sparse Dictionary Learning branch but extends it to capture sequential dependencies, addressing a key limitation of standard sparse coding methods that treat each activation independently. This positions it alongside RAVEL[1] and other temporal extensions, contrasting with purely spatial decomposition approaches like Sparse Autoencoders Interpretability[2]. The work bridges static interpretability methods and dynamic representation analysis, offering a way to understand how semantic features unfold over processing steps rather than treating them as isolated snapshots.

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 incorporating temporal structure and contrastive losses to model semantic evolution across token sequences, distinguishing it from standard SAE approaches. The sibling subtopics represent orthogonal extensions: domain-specific adaptations (proteins/RNA), multi-dimensional irreducible features, and the baseline standard SAE methods. All share the core goal of discovering interpretable features in language model representations but differ in their architectural innovations and application domains.

Similarities: - All subtopics use sparse autoencoders as the foundational architecture for feature extraction - All aim to discover interpretable semantic features from language model representations - All address limitations of polysemantic neurons through different architectural or methodological innovations - All maintain sparsity as a key inductive bias for interpretability

Differences: - Temporal and Sequential Extensions uniquely model dynamics across token sequences, while siblings focus on static representations - Domain-Specific SAEs adapt to specialized sequence types (biological), whereas the original leaf and other siblings target natural language - Multi-Dimensional Feature Discovery identifies irreducible multi-dimensional features, contrasting with the original leaf's focus on temporal smoothness and standard SAEs' one-dimensional features - The original leaf uses contrastive losses for temporal coherence, while Standard SAEs rely on reconstruction objectives alone - Standard SAEs serve as the baseline without specialized extensions, while all other subtopics represent architectural or domain-specific innovations

Suggested Search Directions: - Hybrid approaches combining temporal modeling with multi-dimensional feature discovery - Cross-domain applications of temporal SAEs to biological sequences - Comparative studies on when temporal structure versus multi-dimensional features better capture semantic properties

Sibling Subtopics

- **Domain-Specific Sparse Autoencoders** (leaves: 1, papers: 2)
- Scope: Sparse autoencoders adapted for specialized domains such as protein or RNA sequence representations.
- Exclude: Excludes general natural language applications; those belong in Standard Sparse Autoencoder Approaches.
- **Multi-Dimensional Feature Discovery** (leaves: 1, papers: 1)
- Scope: Methods identifying irreducible multi-dimensional features that cannot be decomposed into independent lower-dimensional components.
- Exclude: Excludes one-dimensional feature extraction; those belong in Standard Sparse Autoencoder Approaches.
- **Standard Sparse Autoencoder Approaches** (leaves: 1, papers: 2)
- Scope: Sparse autoencoders applied to language model activations to extract monosemantic features from polysemantic neurons.
- Exclude: Excludes temporal or multi-dimensional extensions; those belong in their respective specialized categories.

Contributions Analysis

Overall novelty summary. The paper introduces Temporal Sparse Autoencoders (T-SAEs), which extend standard sparse autoencoders by incorporating a contrastive loss to encourage consistent feature activations across adjacent tokens. Within the taxonomy, it occupies the 'Temporal and Sequential Extensions' leaf under 'Sparse Dictionary Learning Methods', where it is currently the sole paper. This places it in a relatively sparse research direction, as the broader 'Sparse Dictionary Learning Methods' branch contains only seven papers across four leaves, with most work concentrated in 'Standard Sparse Autoencoder Approaches' (two papers) and domain-specific applications.

The taxonomy reveals that most interpretability work focuses on static feature extraction or geometric analysis of embedding spaces. The 'Embedding Space Analysis and Geometry' branch (eleven papers across four leaves) and 'Semantic Representation Learning Foundations' (ten papers across three leaves) represent more crowded areas. The paper's temporal approach connects to 'Layer-Wise Representation Dynamics', which examines how representations evolve across network depth, but diverges by focusing on sequential token-level evolution rather than layer-wise progression. The taxonomy's scope notes clarify that temporal modeling distinguishes this work from standard SAEs, which treat activations independently.

Among thirty candidates examined through semantic search, none were found to clearly refute any of the three core contributions. The data-generating process distinguishing semantic and syntactic variables, the T-SAE architecture with contrastive loss, and the empirical validation each examined ten candidates with zero refutable overlaps. This suggests that within the limited search scope, the specific combination of temporal contrastive learning applied to sparse autoencoders for semantic-syntactic disentanglement appears relatively unexplored. However, the search scale (thirty candidates, not hundreds) means substantial prior work outside this sample remains possible.

The analysis indicates the work occupies a genuinely sparse research direction within the taxonomy, with no direct siblings in its leaf and limited temporal extensions elsewhere in sparse dictionary learning. The absence of refutable candidates across thirty examined papers, combined with the taxonomy structure, suggests the temporal contrastive approach represents a distinct methodological contribution. However, the limited search scope and the paper's position in a young subfield mean this assessment reflects current visibility rather than exhaustive coverage of all potentially related work.

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

Contribution 1: Data-generating process distinguishing semantic and syntactic variables

Description: The authors formalize a framework modeling language production through latent variables that separate high-level semantic features (which remain consistent over sequences) from low-level syntactic features (which vary locally). This framework provides theoretical guidance for designing interpretability methods that can recover these distinct types of linguistic information.

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. Unsupervised Disentanglement Learning Model for Exemplar-Guided Paraphrase Generation

URL: View paper

Brief Assessment

Unsupervised Disentanglement Paraphrase[78] focuses on paraphrase generation by disentangling semantic and syntactic representations for sentence reconstruction, not on formalizing a theoretical data-generating process for language production with temporal consistency assumptions as in the original paper.

---

2. The Compositional Architecture of Regret in Large Language Models

URL: View paper

Brief Assessment

Compositional Architecture Regret[76] focuses on identifying regret-specific neurons in LLMs through a data-generating process for regret expressions, not on separating semantic and syntactic variables in general language generation. The paper's framework models regret behavior specifically, not the broader linguistic distinction between semantics and syntax that the original paper addresses.

---

3. DEPT: Decoupled Embeddings for Pre-training Language Models

URL: View paper

Brief Assessment

DEPT[77] focuses on decoupling token embeddings from transformer bodies during pre-training to handle multilingual/multi-domain data heterogeneity. It does not propose a data-generating process framework that models language production through latent semantic and syntactic variables as the original paper does.

---

4. Decoupled context processing for context augmented language modeling

URL: View paper

Brief Assessment

Decoupled Context Processing[80] focuses on architectural design for context retrieval and encoding in language models, not on modeling language production through latent semantic/syntactic variables or interpretability methods for recovering linguistic information.

---

5. Semantic gradient decoupling for contextual precision in large language models

URL: View paper

Brief Assessment

Semantic Gradient Decoupling[72] focuses on gradient separation during backpropagation for training optimization, not on modeling language production through latent variables or designing interpretability methods for feature recovery.

---

6. Rethinking embedding coupling in pre-trained language models

URL: View paper

Brief Assessment

Rethinking Embedding Coupling[73] focuses on embedding architecture design in pre-trained language models, not on formalizing data-generating processes that distinguish semantic from syntactic variables in language production.

---

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

URL: View paper

Brief Assessment

Latent Cascade Synthesis[75] appears to focus on structure-preserving transformations and contextual precision in LLMs, not on formalizing a data-generating process that separates semantic from syntactic variables for interpretability methods.

---

8. Disentangled representation learning

URL: View paper

Brief Assessment

Disentangled Representation Learning[71] focuses on separating distinct generative factors in visual data (e.g., object color, size, shape) rather than modeling language production through semantic vs. syntactic variables in sequential text.

---

9. Knowledge decoupling via orthogonal projection for lifelong editing of large language models

URL: View paper

Brief Assessment

Knowledge Decoupling Orthogonal[74] focuses on lifelong model editing through orthogonal projection to prevent knowledge interference, not on modeling language generation with separate semantic/syntactic latent variables. The candidate addresses knowledge coupling in sequential edits rather than formalizing a data-generating process for language production.

---

10. Disentangled representation learning for non-parallel text style transfer

URL: View paper

Brief Assessment

Disentangled Style Transfer[79] focuses on separating style (sentiment) and content in text using adversarial and multi-task losses, not on modeling language production through temporal semantic vs. syntactic variables as in the original framework.

Contribution 2: Temporal Sparse Autoencoders with contrastive loss

Description: The authors propose Temporal SAEs (T-SAEs), a modification to standard Sparse Autoencoders that partitions features into high-level and low-level components and incorporates a contrastive loss encouraging high-level features to activate consistently over adjacent tokens. This enables self-supervised disentanglement of semantic from syntactic features without requiring explicit semantic labels.

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. Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral classification

URL: View paper

Brief Assessment

Autoencoder Prototypical Contrastive[69] focuses on hyperspectral image classification using autoencoder combined with prototypical contrastive learning for clustering-based feature extraction. The original paper proposes Temporal SAEs for language model interpretability with temporal consistency constraints on sequential tokens. These are fundamentally different application domains (vision vs. language) and different technical objectives (spatial-spectral feature learning vs. temporal semantic-syntactic disentanglement).

---

2. Causal Differentiating Concepts: Interpreting LM Behavior via Causal Representation Learning

URL: View paper

Brief Assessment

Causal Differentiating Concepts[67] focuses on causal representation learning to identify sparse causal factors mediating LM behavior through constrained contrastive learning, not on modifying SAE architectures with temporal consistency objectives for semantic-syntactic disentanglement.

---

3. Analyzing (in) abilities of saes via formal languages

URL: View paper

Brief Assessment

SAE Formal Languages[62] focuses on analyzing SAE capabilities using formal languages (Dyck-2, expr, English PCFG) and proposes a causal regularization approach based on token interpolation within sequences. This differs fundamentally from the original paper's temporal contrastive loss that partitions features into high/low-level components for semantic-syntactic disentanglement in natural language.

---

4. CMViM: Contrastive Masked Vim Autoencoder for 3D Multi-modal Representation Learning for AD classification

URL: View paper

Brief Assessment

CMViM[64] applies contrastive learning to 3D medical imaging for Alzheimer's disease classification, not to sparse autoencoders for language model interpretability. The technical domains and objectives are fundamentally different.

---

5. Multiobjective models for group recommender systems

URL: View paper

Brief Assessment

Multiobjective Group Recommender[70] applies sparse autoencoders to group recommender systems for aggregating user preferences, not for disentangling semantic from syntactic features in language models. The contrastive loss usage and application domain are fundamentally different.

---

6. One for All, All for One: Learning and Transferring User Embeddings for Cross-Domain Recommendation

URL: View paper

Brief Assessment

Cross Domain Embeddings[66] focuses on cross-domain recommendation systems using contrastive autoencoders for user embeddings across multiple domains (e.g., app installation, video viewing). The original paper addresses interpretability of language models through temporal consistency in sequential token processing, which is a fundamentally different application domain and technical problem.

---

7. Self-supervised user embedding alignment for cross-domain recommendations via multi-LLM co-training

URL: View paper

Brief Assessment

User Embedding Alignment[63] focuses on cross-domain recommendation systems using multi-LLM co-training for user embeddings. The candidate's contrastive loss is applied to align user representations across domains, not to disentangle semantic from syntactic features in language model interpretability.

---

8. Learning Sparse Disentangled Representations for Multimodal Exclusion Retrieval

URL: View paper

Brief Assessment

Sparse Disentangled Multimodal[68] focuses on multimodal retrieval with exclusion queries using sparse autoencoders for word/sentence embeddings, not on temporal consistency or semantic-syntactic disentanglement in language model representations over sequential tokens.

---

9. SparseMVC: Probing Cross-view Sparsity Variations for Multi-view Clustering

URL: View paper

Brief Assessment

SparseMVC[61] addresses multi-view clustering with sparse autoencoders for handling cross-view sparsity variations, not temporal consistency in language models. The candidate uses contrastive learning for cross-view distribution alignment in clustering tasks, fundamentally different from the original's temporal contrastive loss for disentangling semantic from syntactic features in sequential language data.

---

10. A self-supervised contrastive denoising autoencoder-based noise suppression method for micro thrust measurement signals processing

URL: View paper

Brief Assessment

Contrastive Denoising Autoencoder[65] applies contrastive learning to signal denoising in micro thrust measurement, not to disentangling semantic from syntactic features in language models. The domains (signal processing vs. language interpretability) and objectives are fundamentally different.

Contribution 3: Empirical validation of semantic recovery and disentanglement

Description: The authors demonstrate through extensive experiments that T-SAEs recover smoother and more coherent semantic concepts compared to baseline SAEs while maintaining reconstruction quality. They show improved recovery of semantic and contextual information, better disentanglement between feature types, and practical benefits for applications like safety monitoring and model steering.

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. Does higher interpretability imply better utility? A Pairwise Analysis on Sparse Autoencoders

URL: View paper

Brief Assessment

Interpretability Utility Pairwise[56] focuses on the relationship between SAE interpretability scores and steering utility, not on semantic recovery or disentanglement validation. The candidate evaluates whether interpretable features enable effective steering, while the original contribution demonstrates that T-SAEs recover smoother semantic concepts and achieve better disentanglement between feature types.

2. Discriminative reconstruction via simultaneous dense and sparse coding

URL: View paper

Brief Assessment

Discriminative Reconstruction[53] focuses on decomposing signals into dense (smooth) and sparse components for image processing tasks, not on evaluating semantic recovery in sparse autoencoders for language models.

3. Improving Steering Vectors by Targeting Sparse Autoencoder Features

URL: View paper

Brief Assessment

Improving Steering Vectors[52] focuses on using SAEs to measure and improve steering vector effects for model control, not on evaluating semantic recovery and disentanglement properties of SAEs themselves.

4. Disentangling dense embeddings with sparse autoencoders

URL: View paper

Brief Assessment

Disentangling Dense Embeddings[57] focuses on applying SAEs to dense text embeddings from scientific papers, not on temporal consistency in language model activations. The evaluation methods differ fundamentally - the candidate uses probing on embedding spaces while the original evaluates temporal smoothness and consistency over token sequences.

5. Privacy-Aware Traffic Re-Identification with Interpretable Sparse Autoencoders

URL: View paper

Brief Assessment

Privacy Aware Traffic[58] applies sparse autoencoders to computer vision embedding models for traffic re-identification, not to language models for semantic/syntactic feature recovery. The domains and evaluation methods differ fundamentally.

6. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models

URL: View paper

Brief Assessment

Sparse Autoencoders Survey[60] is a survey paper that reviews existing SAE methods and evaluation approaches. It does not present original experimental work demonstrating semantic recovery or disentanglement improvements, and therefore cannot refute the novelty of the original paper's empirical validation contributions.

7. AlignSAE: Concept-Aligned Sparse Autoencoders

URL: View paper

Brief Assessment

AlignSAE[54] focuses on supervised concept alignment with predefined ontologies for factual knowledge, while the original paper addresses unsupervised semantic recovery through temporal consistency in language sequences. These are fundamentally different approaches to interpretability.

8. Towards Interpretable Structure Prediction With Sparse Autoencoders

URL: View paper

Brief Assessment

Interpretable Structure Prediction[51] focuses on protein structure prediction using SAEs on ESM2 models, not on evaluating semantic recovery in language models. The evaluation metrics (contact map prediction, Swiss-Prot concept discovery) are domain-specific to protein biology rather than general semantic/syntactic disentanglement in language.

9. Unmixing Autoencoder for Image Reconstruction from Hyperspectral Data.

URL: View paper

Brief Assessment

Unmixing Autoencoder[59] focuses on spectral unmixing in hyperspectral imaging for chemical component separation, not on semantic feature recovery or disentanglement in language models. The domains and objectives are fundamentally different.

10. On the Theoretical Understanding of Identifiable Sparse Autoencoders and Beyond

URL: View paper

Brief Assessment

Identifiable Sparse Autoencoders[55] focuses on theoretical conditions for SAE identifiability and proposes reweighting strategies to improve monosemanticity. While it validates monosemanticity improvements empirically, it does not address the specific semantic recovery, contextual information extraction, or disentanglement between feature types (semantic vs. syntactic) that T-SAEs demonstrate through temporal consistency objectives.

Appendix: Text Similarity Detection

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

References

- [0] Temporal Sparse Autoencoders: Leveraging the Sequential Nature of Language for Interpretability View paper
- [1] RAVEL: Evaluating interpretability methods on disentangling language model representations View paper

- [2] Sparse autoencoders find highly interpretable features in language models View paper
- [3] Layer by layer: Uncovering hidden representations in language models View paper
- [4] Context-preserving latent field interpolation for large language model internal state manipulation View paper
- [5] Sparse autoencoders uncover biologically interpretable features in protein language model representations View paper
- [6] Word embeddings are steers for language models View paper
- [7] Improving uncertainty quantification in large language models via semantic embeddings View paper
- [8] On the sentence embeddings from pre-trained language models View paper
- [9] Learning interpretable concepts: Unifying causal representation learning and foundation models View paper
- [10] Implicit representations of meaning in neural language models View paper
- [11] Discovery of the hidden world with large language models View paper
- [12] Structural Embedding Projection for Contextual Large Language Model Inference View paper
- [13] The reasoning-memorization interplay in language models is mediated by a single direction View paper
- [14] Simulated echo shaping in large language models via semantic phase perturbation without intermediate token realignment View paper
- [15] Emergent representations of program semantics in language models trained on programs View paper
- [16] Not all language model features are one-dimensionally linear View paper
- [17] Analyzing the role of semantic representations in the era of large language models View paper
- [18] CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models View paper
- [19] Semantic Analysis of test items through Large Language Model embeddings predicts a-priori factorial structure of personality tests View paper
- [20] Semantic Structure in Large Language Model Embeddings View paper
- [21] Structural recomposition in large language models through lexico-semantic vector fusion: A computational study View paper
- [22] Word embedding for understanding natural language: a survey View paper
- [23] Neural models for semantic analysis of handwritten document images View paper
- [24] Interpretable Semantic Representations from Neural Language Models and Computer Vision View paper
- [25] Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations View paper
- [26] Neural semantic tagging for natural language-based search in building information models: Implications for practice View paper
- [27] Decoding Neural Emotion Patterns through Large Language Model Embeddings View paper
- [28] Meaning modulations and stability in large language models: An analysis of BERT embeddings for psycholinguistic research View paper
- [29] SAE-RNA: A Sparse Autoencoder Model for Interpreting RNA Language Model Representations View paper
- [30] Refusal-Aware Red Teaming: Exposing Inconsistency in Safety Evaluations View paper
- [31] A full-document analysis of the semantic relation between European Public Assessment Reports and EMA guidelines using a BERT language model View paper
- [32] Multi-Relational Hyperbolic Word Embeddings from Natural Language Definitions View paper
- [33] Exploiting latent semantic information in statistical language modeling View paper
- [34] Shared Global and Local Geometry of Language Model Embeddings View paper
- [35] Improving the Factuality of Abstractive Text Summarization with Syntactic Structure-Aware Latent Semantic Space View paper
- [36] Uncovering Latent Human Wellbeing in Language Model Embeddings View paper
- [37] Improving lexical embeddings with semantic knowledge View paper
- [38] Large Language Models are Interpretable Learners View paper
- [39] Sentiment embeddings with applications to sentiment analysis View paper
- [40] Semantic Tree Inference on Text Corpa using a Nested Density Approach together with Large Language Model Embeddings View paper
- [41] Definition modeling: Learning to define word embeddings in natural language View paper
- [42] On the Relationship Between RNN Hidden-State Vectors and Semantic Structures View paper
- [43] Explainable semantic space by grounding language to vision with cross-modal contrastive learning View paper
- [44] Mechanistic Interpretability of Socio-Political Frames in Language Models View paper
- [45] SelfIE: Self-Interpretation of Large Language Model Embeddings View paper
- [46] Connecting Concept Layers and Rationales to Enhance Language Model Interpretability View paper
- [47] Exploring Embedding Interpretability by Correspondences Between Topic Models and Text Embeddings View paper
- [48] Representation of Lexical Stylistic Features in Language Modelsâ Embedding Space View paper
- [49] Automatically Interpreting Millions of Features in Large Language Models View paper
- [50] Interpretability of BERT latent space through knowledge graphs View paper
- [51] Towards Interpretable Structure Prediction With Sparse Autoencoders View paper
- [52] Improving Steering Vectors by Targeting Sparse Autoencoder Features View paper
- [53] Discriminative reconstruction via simultaneous dense and sparse coding View paper
- [54] AlignSAE: Concept-Aligned Sparse Autoencoders View paper
- [55] On the Theoretical Understanding of Identifiable Sparse Autoencoders and Beyond View paper
- [56] Does higher interpretability imply better utility? A Pairwise Analysis on Sparse Autoencoders View paper
- [57] Disentangling dense embeddings with sparse autoencoders View paper
- [58] Privacy-Aware Traffic Re-Identification with Interpretable Sparse Autoencoders View paper
- [59] Unmixing Autoencoder for Image Reconstruction from Hyperspectral Data. View paper
- [60] A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models View paper
- [61] SparseMVC: Probing Cross-view Sparsity Variations for Multi-view Clustering View paper
- [62] Analyzing (in) abilities of saes via formal languages View paper
- [63] Self-supervised user embedding alignment for cross-domain recommendations via multi-LLM co-training View paper
- [64] CMViM: Contrastive Masked Vim Autoencoder for 3D Multi-modal Representation Learning for AD classification View paper
- [65] A self-supervised contrastive denoising autoencoder-based noise suppression method for micro thrust measurement signals processing View paper
- [66] One for All, All for One: Learning and Transferring User Embeddings for Cross-Domain Recommendation View paper
- [67] Causal Differentiating Concepts: Interpreting LM Behavior via Causal Representation Learning View paper

- [68] Learning Sparse Disentangled Representations for Multimodal Exclusion Retrieval View paper
- [69] Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral classification View paper
- [70] Multiobjective models for group recommender systems View paper
- [71] Disentangled representation learning View paper
- [72] Semantic gradient decoupling for contextual precision in large language models View paper
- [73] Rethinking embedding coupling in pre-trained language models View paper
- [74] Knowledge decoupling via orthogonal projection for lifelong editing of large language models View paper
- [75] Latent cascade synthesis: Investigating iterative pseudo-contextual scaffold formation in contemporary large language models View paper
- [76] The Compositional Architecture of Regret in Large Language Models View paper
- [77] DEPT: Decoupled Embeddings for Pre-training Language Models View paper
- [78] Unsupervised Disentanglement Learning Model for Exemplar-Guided Paraphrase Generation View paper
- [79] Disentangled representation learning for non-parallel text style transfer View paper
- [80] Decoupled context processing for context augmented language modeling View paper