Novelty Assessment Report

Paper: LLM Pretraining with Continuous Concepts

PDF URL: https://openreview.net/pdf?id=wTGcb3DxOn

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Next token prediction has been the standard training objective used in large language model pretraining. Representations are learned as a result of optimizing for token-level perplexity. We propose Continuous Concept Mixing (CoCoMix), a novel pretraining framework that combines discrete next token prediction with continuous concepts. Specifically, CoCoMix predicts ``continuous concepts'' learned from a pretrained sparse autoencoder and mixes them into the model's hidden state by interleaving with token hidden representations. Through experiments on multiple benchmarks, including language modeling and downstream reasoning tasks, we show that CoCoMix is more sample efficient and consistently outperforms standard next token prediction and knowledge distillation. We find that combining both concept learning and interleaving in an end-to-end framework is critical to performance gains. Furthermore, CoCoMix enhances interpretability and steerability by allowing direct inspection and modification of the predicted concept, offering a transparent way to guide the model’s internal reasoning process.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Language Model Pretraining with Continuous Concept Prediction and Mixing**

A total of **3 papers** were analyzed and organized into a taxonomy with **4 categories**.

Taxonomy Overview

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

- **Continuous Concept Integration in Neural Language Models**
- **Symbolic Concept Extraction and Rule-Based Reasoning**
- **Generative Conceptual Design with Language Models**
- **Cognitive and Linguistic Theories of Conceptual Combination**

Complete Taxonomy Tree

- Language Model Pretraining with Continuous Concept Prediction and Mixing Survey Taxonomy
- Continuous Concept Integration in Neural Language Models
    - Pretraining with Continuous Concept Prediction and Mixing ★ (1 papers)
    - [0] LLM Pretraining with Continuous Concepts (Anon et al., 2026) View paper
- Symbolic Concept Extraction and Rule-Based Reasoning
    - Two-Stage Semantic-to-Symbolic Deconstruction (1 papers)
    - [3] From Semantics to Symbols: A Two-Stage Framework for Deconstructing LLM Reasoning into Concepts and Rules (Yin, 2025) View paper
- Generative Conceptual Design with Language Models
    - Link Prediction-Augmented Conceptual Solution Generation (1 papers)
    - [1] A generation framework of conceptual solutions integrating link prediction and large language model (Zhi-Xing Chang, 2025) View paper
- Cognitive and Linguistic Theories of Conceptual Combination
    - Non-Local Syntactic-Semantic Conceptual Combination (1 papers)
    - [2] Non-Local Conceptual Combination (Alicia Parrish, 2022) View paper

Narrative

Core task: language model pretraining with continuous concept prediction and mixing. The field structure suggested by this taxonomy reflects a diverse landscape where neural, symbolic, generative, and cognitive perspectives converge on how concepts are represented and combined. The main branches include Continuous Concept Integration in Neural Language Models, which focuses on embedding-based and differentiable approaches to concept learning; Symbolic Concept Extraction and Rule-Based Reasoning, which emphasizes discrete structures and logical inference; Generative Conceptual Design with Language Models, which explores creative synthesis and design tasks; and Cognitive and Linguistic Theories of Conceptual Combination, which grounds computational work in human cognition and linguistic theory. These branches relate by offering complementary views on concept representation—some prioritize end-to-end learning in continuous spaces, while others seek interpretability through symbolic abstraction or draw inspiration from psychological models of how humans merge concepts.

A particularly active line of work within the continuous integration branch explores how pretraining objectives can be enriched by predicting and mixing latent concept representations, as exemplified by Continuous Concepts Pretraining[0], which directly targets this goal. This contrasts with approaches like Semantics to Symbols[3], which bridges neural embeddings and symbolic reasoning by extracting discrete concept structures, and Link Prediction LLM Framework[1], which applies language models to structured prediction tasks over knowledge graphs. Meanwhile, Non-Local Conceptual Combination[2] investigates cognitive theories of how concepts interact beyond simple composition, highlighting open questions about whether neural models capture the flexibility of human conceptual blending. Continuous Concepts Pretraining[0] sits squarely within the neural continuous branch, emphasizing differentiable concept mixing during pretraining, and its emphasis on continuous latent spaces distinguishes it from the more symbolic or cognitively grounded directions represented by nearby works.

Related Works in Same Category

No sibling papers and no sibling subtopics were found under the same parent taxonomy node; the paper appears structurally isolated in the taxonomy.

Contributions Analysis

Overall novelty summary. The paper proposes CoCoMix, a pretraining framework that predicts continuous concepts from sparse autoencoders and interleaves them with token representations during training. According to the taxonomy, this work occupies a singleton leaf node under 'Continuous Concept Integration in Neural Language Models,' with no sibling papers in the same category. This positioning suggests the paper addresses a relatively sparse research direction within the broader landscape of concept-based language modeling, where most related work either focuses on post-hoc symbolic extraction or cognitive theories rather than end-to-end continuous concept integration during pretraining.

The taxonomy reveals that neighboring research directions include symbolic concept extraction (e.g., two-stage semantic-to-symbolic frameworks), generative conceptual design (link prediction-augmented generation), and cognitive theories of conceptual combination. CoCoMix diverges from these by maintaining differentiable concept representations throughout pretraining rather than extracting discrete structures post-training or applying concepts to design tasks. The taxonomy's scope notes explicitly exclude post-hoc extraction and symbolic reasoning from the paper's category, emphasizing that continuous integration during pretraining represents a distinct methodological choice within the field's structure.

Among thirty candidates examined across three contributions, none were identified as clearly refuting the paper's claims. The core CoCoMix framework examined ten candidates with zero refutable matches, as did the concept selection mechanism and interpretability enhancements. This absence of overlapping prior work within the limited search scope suggests that the specific combination of continuous concept prediction from sparse autoencoders with interleaved mixing during pretraining may not have direct precedents among the semantically similar papers retrieved. However, the search scope remains constrained to top-K semantic matches and their citations.

Based on the limited literature search covering thirty candidates, the work appears to occupy a relatively unexplored intersection of sparse autoencoder-based concept extraction and pretraining objectives. The taxonomy structure confirms this is a sparse research direction with no identified siblings, though the broader field includes substantial work on related but methodologically distinct approaches. The analysis cannot rule out relevant work outside the examined candidate set or in adjacent research communities not captured by semantic search.

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

Contribution 1: Continuous Concept Mixing (CoCoMix) pretraining framework

Description: The authors introduce CoCoMix, a new language model pretraining method that augments standard next token prediction by predicting continuous concepts extracted from a pretrained sparse autoencoder and mixing them into the model's hidden state through interleaving with token representations.

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. Series with Pre-trained Language Models

URL: View paper

Brief Assessment

Series Pretrained Language[21] focuses on time series forecasting by transforming temporal signals into token-based representations for language models, not on augmenting language model pretraining with continuous concept prediction from sparse autoencoders.

2. Coevolutionary Continuous Discrete Diffusion: Make Your Diffusion Language Model a Latent Reasoner

URL: View paper

Brief Assessment

Coevolutionary Continuous Discrete Diffusion[18] focuses on diffusion-based language models that jointly denoise in continuous representation and discrete token spaces, rather than augmenting next token prediction with continuous concepts extracted from sparse autoencoders during pretraining.

3. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space

URL: View paper

Brief Assessment

Soft Thinking[15] focuses on inference-time reasoning in continuous concept space without any training, while CoCoMix is a pretraining framework that combines discrete next token prediction with continuous concept prediction during training. These are fundamentally different approaches addressing different stages of the model lifecycle.

4. Controlled Text Generation as Continuous Optimization with Multiple Constraints

URL: View paper

Brief Assessment

Continuous Optimization Constraints[19] focuses on controlled text generation during inference through continuous optimization with multiple constraints, not on pretraining language models with continuous concept prediction combined with discrete token prediction.

5. Unlocking Pretrained LLMs for Motion-Related Multimodal Generation: A Fine-Tuning Approach to Unify Diffusion and Next-Token Prediction

URL: View paper

Brief Assessment

Unify Diffusion Next-Token[17] focuses on motion-related multimodal generation by combining diffusion-based continuous motion generation with autoregressive text prediction, not on language model pretraining with continuous concept prediction from sparse autoencoders.

6. Lightweight Latent Reasoning for Narrative Tasks

URL: View paper

Brief Assessment

Lightweight Latent Reasoning[23] focuses on narrative tasks (plot hole detection, chapter generation) using latent reasoning during RL fine-tuning, not general pretraining with continuous concept prediction as in the original paper.

7. Continuous Entailment Patterns for Lexical Inference in Context

URL: View paper

Brief Assessment

Continuous Entailment Patterns[22] focuses on lexical inference tasks using continuous patterns for fine-tuning pretrained models, not on pretraining frameworks that combine discrete token prediction with continuous concept prediction from sparse autoencoders.

8. Continuous Speech Tokens Makes LLMs Robust Multi-Modality Learners

URL: View paper

Brief Assessment

Continuous Speech Tokens[20] focuses on continuous speech tokens for multi-modal speech-to-speech models using flow matching, not continuous concept prediction from sparse autoencoders for language model pretraining.

9. Large language models are zero-shot time series forecasters

URL: View paper

Brief Assessment

LLMs Time Series Forecasters[14] focuses on zero-shot time series forecasting by encoding numerical values as text for next-token prediction, not on pretraining language models with continuous concept prediction combined with discrete token prediction as in CoCoMix.

10. Latent Reasoning via Sentence Embedding Prediction

URL: View paper

Brief Assessment

Latent Reasoning Embeddings[16] focuses on sentence-level embedding prediction for reasoning tasks, not concept-level prediction mixed with token representations during pretraining. The candidate operates in sentence embedding space for inference efficiency, while the original work augments standard pretraining with sparse autoencoder-derived concepts interleaved into hidden states.

Contribution 2: Concept selection using attribution scores

Description: The authors develop a concept selection mechanism that uses attribution scores to identify which concepts from the sparse autoencoder most influence the model's output, enabling the model to focus on the most relevant semantic features for prediction.

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. Explaining deep convolutional models by measuring the influence of interpretable features in image classification

URL: View paper

Brief Assessment

Interpretable Features Influence[11] focuses on measuring feature influence in image classification using convolutional neural networks, not on selecting concepts from sparse autoencoders for language model pretraining as in the original paper.

2. A survey of feature attribution techniques in explainable AI: taxonomy, analysis and comparison

URL: View paper

Brief Assessment

Feature Attribution Survey[10] focuses on post-hoc explanation methods for interpreting trained model predictions across various architectures, not on pretraining frameworks that use attribution scores to select semantic concepts from sparse autoencoders for continuous concept prediction during training.

3. Evaluating attribution for graph neural networks

URL: View paper

Brief Assessment

Attribution for Graph Networks[7] focuses on evaluating attribution methods for graph neural networks to identify important nodes/ edges in graph-structured data, not on selecting interpretable features from sparse autoencoders for language models as in the original paper.

4. Gradient based feature attribution in explainable ai: A technical review

URL: View paper

Brief Assessment

Gradient Feature Attribution Review[8] focuses on gradient-based feature attribution methods for explaining neural network predictions, not on selecting concepts from sparse autoencoders for language model pretraining. The technical domains and applications are fundamentally different.

5. Do feature attribution methods correctly attribute features?

URL: View paper

Brief Assessment

Feature Attribution Correctness[12] focuses on evaluating whether attribution methods correctly identify ground-truth important features in classification tasks, not on using attribution scores to select semantic concepts for language model pretraining as in the original paper.

6. Multi-objective feature attribution explanation for explainable machine learning

URL: View paper

Brief Assessment

Multi-Objective Feature Attribution[6] focuses on feature attribution for model explainability in supervised learning, not on selecting semantic concepts from sparse autoencoders for language model pretraining. The attribution scores serve fundamentally different purposes in different domains.

7. A comprehensive survey on self-interpretable neural networks

URL: View paper

Brief Assessment

Self-Interpretable Neural Networks Survey[4] discusses attribution-based methods for feature importance but does not specifically address concept selection from sparse autoencoders using attribution scores for language model pretraining, which is the novel contribution in the original paper's context.

8. A benchmark for interpretability methods in deep neural networks

URL: View paper

Brief Assessment

Interpretability Methods Benchmark[9] focuses on evaluating feature importance estimators for interpreting trained neural networks, not on selecting concepts during model training. The candidate uses attribution scores to evaluate existing interpretability methods, while the original uses them to select semantic concepts during pretraining.

9. Evaluating feature attribution methods in the image domain

URL: View paper

Brief Assessment

Evaluating Feature Attribution Images[5] focuses on evaluating feature attribution methods for explaining image classification models, not on selecting interpretable features from sparse autoencoders for language model pretraining. The attribution scores in [5] are used to evaluate explanation quality, not to select concepts for prediction.

10. How can i explain this to you? an empirical study of deep neural network explanation methods

URL: View paper

Brief Assessment

DNN Explanation Methods Study[13] focuses on comparing user preferences for different explanation methods (LIME, SHAP, Grad-CAM++, etc.) across domains, not on developing attribution-based concept selection mechanisms for sparse autoencoders in language model pretraining.

Contribution 3: Enhanced interpretability and steerability through concept prediction

Description: The framework enables users to directly probe and manipulate predicted concepts during generation, providing transparency into the model's reasoning and allowing controllable text generation by amplifying or modifying specific concept activations.

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. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation

URL: View paper

Brief Assessment

Interpretable Diffusion Directions[27] focuses on discovering interpretable latent directions in diffusion models for text-to-image generation, not language models. The candidate operates on visual semantic spaces (h-space in U-Net), while the original paper works with language model hidden states and sparse autoencoders for text generation.

2. Ctrl: A conditional transformer language model for controllable generation

URL: View paper

Brief Assessment

CTRL Conditional Transformer[31] focuses on controllable generation through domain/style control codes prepended to text, not on probing or manipulating internal concept activations during generation as in the original paper's framework.

3. Taxonomy, opportunities, and challenges of representation engineering for large language models

URL: View paper

Brief Assessment

Representation Engineering Taxonomy[26] focuses on manipulating existing internal representations in LLMs through various steering methods, rather than predicting and mixing continuous concepts during pretraining as the original paper proposes.

4. Unveiling Language Competence Neurons: A Psycholinguistic Approach to Model Interpretability

URL: View paper

Brief Assessment

Language Competence Neurons[32] focuses on neuron-level interpretability through psycholinguistic experiments (sound-shape, sound-gender, implicit causality tasks), not on concept prediction and manipulation during generation as in the original paper's framework.

5. Word embeddings are steers for language models

URL: View paper

Brief Assessment

Word Embeddings Steers[24] focuses on steering language models through linear transformations of output word embeddings for style control (sentiment, toxicity), not on predicting and manipulating internal concepts during generation for transparency into reasoning processes.

6. Editable concept bottleneck models

URL: View paper

Brief Assessment

Editable Concept Bottlenecks[30] focuses on editing trained CBMs by removing/correcting concept labels or data samples using influence functions, not on enabling direct probing and manipulation of predicted concepts during generation for controllable text generation in language models.

7. How do large language models understand relevance? a mechanistic interpretability perspective

URL: View paper

Brief Assessment

LLM Relevance Understanding[25] focuses on mechanistic interpretability of relevance judgment in IR tasks using activation patching, not on concept prediction and manipulation for controllable text generation.

8. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification

URL: View paper

Brief Assessment

Language in Bottle[28] focuses on interpretable image classification using concept bottleneck models with vision-language alignment (CLIP), not on language model pretraining or text generation control as in the original paper.

9. Concept bottleneck large language models

URL: View paper

Brief Assessment

Concept Bottleneck LLMs[29] focuses on building interpretable models through concept bottleneck layers for classification and generation tasks, not on continuous concept mixing during pretraining. The original paper's contribution involves predicting and mixing continuous concepts from sparse autoencoders during pretraining to enhance sample efficiency and reasoning, which is a different technical approach and application context.

10. Towards Uncovering How Large Language Model Works: An Explainability Perspective

URL: View paper

Brief Assessment

Uncovering LLM Works[33] is a survey paper reviewing existing explainability techniques for LLMs, including representation engineering and model editing. It does not present a novel framework for concept prediction and manipulation during generation like the original paper's CoCoMix.

Appendix: Text Similarity Detection

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

References

- [0] LLM Pretraining with Continuous Concepts View paper
- [1] A generation framework of conceptual solutions integrating link prediction and large language model View paper
- [2] Non-Local Conceptual Combination View paper
- [3] From Semantics to Symbols: A Two-Stage Framework for Deconstructing LLM Reasoning into Concepts and Rules View paper
- [4] A comprehensive survey on self-interpretable neural networks View paper
- [5] Evaluating feature attribution methods in the image domain View paper
- [6] Multi-objective feature attribution explanation for explainable machine learning View paper
- [7] Evaluating attribution for graph neural networks View paper
- [8] Gradient based feature attribution in explainable ai: A technical review View paper
- [9] A benchmark for interpretability methods in deep neural networks View paper
- [10] A survey of feature attribution techniques in explainable AI: taxonomy, analysis and comparison View paper
- [11] Explaining deep convolutional models by measuring the influence of interpretable features in image classification View paper
- [12] Do feature attribution methods correctly attribute features? View paper
- [13] How can i explain this to you? an empirical study of deep neural network explanation methods View paper
- [14] Large language models are zero-shot time series forecasters View paper
- [15] Soft thinking: Unlocking the reasoning potential of llms in continuous concept space View paper
- [16] Latent Reasoning via Sentence Embedding Prediction View paper
- [17] Unlocking Pretrained LLMs for Motion-Related Multimodal Generation: A Fine-Tuning Approach to Unify Diffusion and Next-Token Prediction View paper
- [18] Coevolutionary Continuous Discrete Diffusion: Make Your Diffusion Language Model a Latent Reasoner View paper
- [19] Controlled Text Generation as Continuous Optimization with Multiple Constraints View paper
- [20] Continuous Speech Tokens Makes LLMs Robust Multi-Modality Learners View paper
- [21] Series with Pre-trained Language Models View paper
- [22] Continuous Entailment Patterns for Lexical Inference in Context View paper
- [23] Lightweight Latent Reasoning for Narrative Tasks View paper
- [24] Word embeddings are steers for language models View paper
- [25] How do large language models understand relevance? a mechanistic interpretability perspective View paper
- [26] Taxonomy, opportunities, and challenges of representation engineering for large language models View paper
- [27] Self-discovering interpretable diffusion latent directions for responsible text-to-image generation View paper
- [28] Language in a bottle: Language model guided concept bottlenecks for interpretable image classification View paper
- [29] Concept bottleneck large language models View paper
- [30] Editable concept bottleneck models View paper
- [31] Ctrl: A conditional transformer language model for controllable generation View paper
- [32] Unveiling Language Competence Neurons: A Psycholinguistic Approach to Model Interpretability View paper
- [33] Towards Uncovering How Large Language Model Works: An Explainability Perspective View paper