Novelty Assessment Report

Paper: Transformers Learn Latent Mixture Models In-Context via Mirror Descent
PDF URL: https://openreview.net/pdf?id=SHidElLSVt
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Sequence modelling requires determining which past tokens are causally relevant from the context and their importance: a process inherent to the attention layers in transformers, yet whose underlying learned mechanisms remain poorly understood. In this work, we formalize the task of estimating token importance as an in-context learning problem by introducing a novel framework based on Mixture of Transition Distributions, whereby a latent variable, whose distribution is parameterized by a set of unobserved mixture weights, determines the influence of past tokens on the next. To correctly predict the next token, transformers need to learn the mixture weights in-context. We demonstrate that transformers can implement Mirror Descent to learn the mixture weights from the context. To this end, we give an explicit construction of a three-layer transformer that exactly implements one step of Mirror Descent and prove that the resulting estimator is a first-order approximation of the Bayes-optimal predictor. Corroborating our construction and its learnability via gradient descent, we empirically show that transformers trained from scratch converge to this solution: attention maps match our construction, and deeper models' performance aligns with multi-step Mirror Descent.

Core Task Landscape

This paper addresses: **Learning Latent Mixture Weights In-Context for Sequence Modeling**
A total of **19 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

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

- **Theoretical Foundations of In-Context Learning for Mixture Models**
- **Hidden Markov Models and Sequential Latent Structure Learning**
- **Neural Architectures for Sequential Latent Variable Modeling**
- **Large-Scale Architectures and Applications**

Complete Taxonomy Tree

- Learning Latent Mixture Weights In-Context for Sequence Modeling Survey Taxonomy
- Theoretical Foundations of In-Context Learning for Mixture Models
  - Mirror Descent and Optimization-Based ICL Analysis ★ (1 papers)
  - [0] Transformers Learn Latent Mixture Models In-Context via Mirror Descent (Anon et al., 2026) View paper
  - Bayesian Inference Perspectives on ICL (1 papers)
  - [4] An explanation of in-context learning as implicit bayesian inference (Sang Michael Xie, 2021) View paper
  - Mixture of Linear Regressions ICL Theory (1 papers)
  - [2] In-context Learning for Mixture of Linear Regressions: Existence, Generalization and Training Dynamics (Jin, 2024) View paper
  - Latent Variable Inference and ICL Effectiveness (1 papers)
  - [6] Does learning the right latent variables necessarily improve in-context learning? (Mittal, 2024) View paper
- Hidden Markov Models and Sequential Latent Structure Learning
  - LLM In-Context Learning of HMMs (2 papers)
  - [5] Pre-trained Large Language Models Learn Hidden Markov Models In-context (Dai Yi-jia, 2025) View paper
  - [19] Pre-trained Large Language Models Learn to Predict Hidden Markov Models In-context (Y Dai, n.d.) View paper
  - Contextual HMMs with External Variables (2 papers)
  - [8] Contextual hidden markov models (Mathieu Radenen, 2012) View paper
  - [10] Online learning of contextual hidden markov models for temporal-spatial data analysis (Yuxun Zhou, 2016) View paper
  - Bayesian and Adaptive HMM Parameter Learning (1 papers)
  - [12] Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition (Qiang Huo, 1995) View paper
  - HMM-Based Classification and Mixture of Experts (2 papers)
  - [13] Contextâ□based classification via mixture of hidden Markov model experts with applications in landmine detection (S. E. YÃ¼ksel, 2016) View paper
  - [17] Context-based classification via data-dependent mixtures of logistic and hidden markov model classifiers (Paul Gader, 2011) View paper
- Neural Architectures for Sequential Latent Variable Modeling
  - Generative Models with Dilated Convolutions (1 papers)
  - [3] Stochastic WaveNet: A Generative Latent Variable Model for Sequential Data (Lai, 2022) View paper
  - LSTM-Based Latent Clustering and Dynamics (1 papers)
  - [11] Latent LSTM allocation: Joint clustering and non-linear dynamic modeling of sequence data (Manzil Zaheer, 2017) View paper
  - Amortised Inference in Sequential LVMs (1 papers)
  - [9] Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models (Bayer, 2021) View paper

- Large-Scale Architectures and Applications
  - Sparse Mixture-of-Experts Language Models (1 papers)
  - [1] Glam: Efficient scaling of language models with mixture-of-experts (Du, 2022) View paper
  - Dialogue and Knowledge Selection Systems (1 papers)
  - [7] Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue (Kim, 2022) View paper
  - Task Inference and Compositional Reasoning (1 papers)
  - [14] OCEAN: Online Task Inference for Compositional Tasks with Context Adaptation (Ren, 2020) View paper
  - Applied Mixture Models with Textual Context (1 papers)
  - [15] Topic-informed dynamic mixture model for occupational heterogeneity in health risk behaviors (Lorenzo Schiavon, 2025) View paper
  - Full Bayesian ICL and Comprehensive Surveys (2 papers)
  - [16] A Comprehensive Survey on Large Language Models: From Pre-training to Autonomous Agents (J Xu, 2025) View paper
  - [18] In-Context Learning for Full Bayesian Inference (A Reuter, n.d.) View paper

Narrative

Core task: learning latent mixture weights in-context for sequence modeling. The field explores how models—ranging from classical probabilistic frameworks to modern neural architectures—can infer and adapt latent structure on the fly as they process sequential data. The taxonomy organizes this landscape into four main branches. Theoretical Foundations of In-Context Learning for Mixture Models investigates the algorithmic principles underlying in-context adaptation, often connecting transformer mechanics to optimization procedures such as mirror descent or Bayesian inference (e.g., Implicit Bayesian Inference[4]). Hidden Markov Models and Sequential Latent Structure Learning encompasses classical and contextual HMM variants (Contextual Hidden Markov[8], Online Contextual HMM[10]) that explicitly model latent state transitions and mixture dynamics. Neural Architectures for Sequential Latent Variable Modeling includes recurrent and variational approaches (Latent LSTM Allocation[11], Stochastic WaveNet[3]) that embed latent variables within deep networks. Large-Scale Architectures and Applications covers modern transformer-based systems and mixture-of-experts designs (Glam[1]) that scale in-context learning to real-world tasks, bridging theory and practice.

Recent work has intensified around understanding whether and how large language models implicitly perform structured inference over latent variables. A handful of studies (LLMs Learn HMMs[5], Right Latent Variables[6]) demonstrate that transformers can recover hidden Markov structure or mixture components without explicit probabilistic machinery, raising questions about the representational capacity and inductive biases of attention mechanisms. The original paper, Transformers Mirror Descent[0], sits squarely within the optimization-based analysis branch of Theoretical Foundations, arguing that transformer layers implement steps of mirror descent when learning mixture weights in-context. This perspective complements probabilistic views like Implicit Bayesian Inference[4] and contrasts with empirical investigations such as LLMs Learn HMMs[5], which focus on emergent capabilities rather than mechanistic interpretations. Together, these lines of work reveal an active debate over whether in-context learning is best understood through the lens of optimization, Bayesian updating, or emergent neural computation.

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

These theoretical subtopics all analyze how transformers perform in-context learning through different mathematical lenses—optimization algorithms, Bayesian inference, and latent variable learning. The original leaf focuses specifically on proving transformers implement mirror descent and related optimization procedures for learning mixture weights, distinguishing it from Bayesian and regression-specific frameworks. All siblings share a theoretical (rather than empirical) orientation toward understanding ICL mechanisms.

Similarities: - All four categories provide theoretical frameworks for understanding in-context learning mechanisms - All address learning with latent structure (mixture weights, task concepts, latent variables) - All exclude purely empirical studies and focus on mathematical/algorithmic explanations - All are concerned with how transformers implicitly learn from context without parameter updates

Differences: - Mirror Descent leaf focuses on optimization algorithm implementation (mirror descent), while Bayesian Inference uses probabilistic inference frameworks - Mixture of Linear Regressions is task-specific (regression problems), while Mirror Descent and Bayesian perspectives are more general - Latent Variable Inference examines whether correct latent learning is necessary for performance, questioning the assumptions underlying other theoretical frameworks - Mirror Descent explicitly excludes Bayesian and regression-focused analyses, positioning itself as an alternative optimization-based explanation

Suggested Search Directions: - Comparative studies evaluating optimization vs. Bayesian explanations for the same ICL phenomena - Theoretical work unifying optimization and inference perspectives on in-context learning - Extensions of mirror descent analysis to non-mixture or structured prediction tasks

Sibling Subtopics

- **Bayesian Inference Perspectives on ICL** (leaves: 1, papers: 1)
- Scope: Theoretical frameworks explaining in-context learning as implicit Bayesian inference over latent document or task concepts.
- Exclude: Excludes optimization-based explanations and empirical HMM studies; see sibling categories for those approaches.
- **Latent Variable Inference and ICL Effectiveness** (leaves: 1, papers: 1)
- Scope: Investigates whether learning correct task latents necessarily improves in-context learning performance versus exploiting statistical shortcuts.
- Exclude: Excludes studies focused on specific model architectures or HMM applications; see empirical architecture categories.
- **Mixture of Linear Regressions ICL Theory** (leaves: 1, papers: 1)
- Scope: Theoretical analysis of transformer in-context learning for mixture of linear regression tasks with generalization bounds and training dynamics.
- Exclude: Excludes non-regression mixture models and empirical studies; see other theoretical or empirical categories.

Contributions Analysis

Overall novelty summary. The paper formalizes in-context learning of token importance as inference over latent mixture weights in a Mixture of Transition Distributions framework. It resides in the 'Mirror Descent and Optimization-Based ICL Analysis' leaf under 'Theoretical Foundations of In-Context Learning for Mixture Models'. Notably, this leaf contains only the original paper itself—no sibling papers appear in the taxonomy. This isolation suggests the optimization-based mechanistic interpretation of in-context learning for mixture models represents a relatively sparse research direction within the broader field of 19 papers surveyed.

The taxonomy reveals three sibling leaves within Theoretical Foundations: Bayesian Inference Perspectives (1 paper), Mixture of Linear Regressions ICL Theory (1 paper), and Latent Variable Inference and ICL Effectiveness (1 paper). These neighboring directions explore

alternative theoretical lenses—probabilistic inference, regression-specific analysis, and the relationship between latent recovery and performance—rather than explicit optimization algorithms. The paper's focus on mirror descent as the mechanistic substrate distinguishes it from these Bayesian and regression-focused frameworks, while the broader taxonomy shows substantial activity in empirical HMM studies (4 papers) and neural architectures (3 papers), indicating the field balances theory with applied sequential modeling.

Among 26 candidates examined across three contributions, none were flagged as clearly refuting the paper's claims. The MTD framework examined 10 candidates with zero refutations; the three-layer transformer construction examined 10 candidates with zero refutations; the Bayes-optimal connection examined 6 candidates with zero refutations. This absence of overlapping prior work within the limited search scope suggests the specific combination—mixture of transition distributions, explicit transformer construction for mirror descent, and first-order Bayes-optimality proof—has not been directly addressed in the top-30 semantic matches and their citations. However, the search scale is modest and does not cover the full optimization or transformer theory literature.

Given the limited search scope and the paper's unique position as the sole occupant of its taxonomy leaf, the work appears to introduce a novel mechanistic perspective on in-context learning. The optimization-based framing contrasts with existing Bayesian and empirical approaches in neighboring leaves, and the explicit construction offers a concrete algorithmic interpretation. Nonetheless, the analysis reflects top-30 semantic candidates and does not exhaustively survey the broader optimization theory or transformer interpretability communities, leaving open the possibility of related work outside this search boundary.

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

Contribution 1: MTD framework for in-context learning of latent mixture weights

Description: The authors introduce a framework using Mixture of Transition Distributions (MTD) models to formalize token importance estimation as an in-context learning problem. In this framework, latent mixture weights determine the influence of past tokens, and transformers must learn these weights from context to predict the next token correctly.

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. Style-conditional Prompt Token Learning for Generalizable Face Anti-spoofing

URL: View paper

Brief Assessment

Style Conditional Prompt[45] addresses face anti-spoofing through vision-language models and prompt learning for domain generalization, not in-context learning of mixture models or token importance estimation in transformers.

---

2. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation

URL: View paper

Brief Assessment

Mixture of Recursions[38] focuses on recursive transformers with dynamic token-level computation depths for language modeling, not on in-context learning of latent mixture weights in sequence modeling tasks.

---

3. Mixtures of in-context learners

URL: View paper

Brief Assessment

Mixtures In Context[36] focuses on combining predictions from multiple in-context learning experts using mixture weights, but does not address learning latent mixture weights in Mixture of Transition Distribution models for token importance estimation in sequential data.

---

4. Novel token-level recurrent routing for enhanced mixture-of-experts performance

URL: View paper

Brief Assessment

Token Level Routing[40] focuses on token-level recurrent routing within mixture-of-experts models for dynamic expert selection, not on formalizing token importance estimation as an in-context learning problem using Mixture of Transition Distributions models with latent mixture weights.

---

5. Soft Adaptive Policy Optimization

URL: View paper

Brief Assessment

Soft Adaptive Policy[43] addresses policy optimization in reinforcement learning for LLMs, focusing on token-level importance ratios and gradient clipping strategies. It does not address in-context learning of mixture models or latent variable estimation in sequential prediction tasks.

---

6. On the role of attention in prompt-tuning

URL: View paper

Brief Assessment

Attention Prompt Tuning[42] focuses on prompt-tuning for classification tasks using attention mechanisms to select context-relevant tokens, not on in-context learning of latent mixture weights in sequential models like MTD.

---

7. Merging Multi-Task Models via Weight-Ensembling Mixture of Experts

URL: View paper

Brief Assessment

Weight Ensembling MoE[37] focuses on merging multi-task vision models via mixture of experts for model fusion, not on in-context learning of latent mixture weights in sequence modeling tasks.

---

8. Domain-Specific Pruning of Large Mixture-of-Experts Models with Few-shot Demonstrations

URL: View paper

Brief Assessment

Domain Specific Pruning[44] focuses on pruning mixture-of-experts models for domain specialization using few-shot demonstrations, not on in-context learning of latent mixture weights in sequence modeling tasks.

---

9. MoDES: Accelerating Mixture-of-Experts Multimodal Large Language Models via Dynamic Expert Skipping

URL: View paper

Brief Assessment

MoDES[41] focuses on expert skipping in Mixture-of-Experts multimodal models for computational efficiency, not on in-context learning of mixture weights in sequence modeling tasks.

---

10. Introducing dynamic token embedding sampling of large language models for improved inference accuracy

URL: View paper

Brief Assessment

Dynamic Token Embedding[39] focuses on adjusting token embeddings based on contextual importance in inference, not on formalizing token importance estimation as an in-context learning problem using Mixture of Transition Distributions models with latent mixture weights that transformers must learn from context.

Contribution 2: Explicit three-layer transformer construction implementing one-step Mirror Descent

Description: The authors provide a constructive proof showing that a three-layer disentangled transformer can exactly implement one step of the Mirror Descent algorithm for learning mixture weights. This construction demonstrates how attention mechanisms can compute posterior responsibilities and produce estimates matching the Mirror Descent update rule.

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. Scoop: An Optimizer for Profiling Attacks against Higher-Order Masking

URL: View paper

Brief Assessment

Scoop[22] focuses on second-order optimization for side-channel attacks against masking schemes, not on transformer architectures implementing Mirror Descent for in-context learning of mixture models.

---

2. Time-Dependent Mirror Flows and Where to Find Them

URL: View paper

Brief Assessment

Time Dependent Mirrors[29] focuses on explicit regularization effects on implicit bias through mirror flows in general optimization settings, not on constructive proofs for transformer architectures implementing specific algorithms like Mirror Descent for mixture weight learning.

---

3. MCDDT: Mirror center loss based dual-scale dual-softmax transformer for multi-source subjects transfer learning in motor imagery recognition

URL: View paper

Brief Assessment

MCDDT[20] focuses on motor imagery EEG signal recognition using transformers with mirror center loss for neural activity localization, not on constructing transformers to implement Mirror Descent optimization algorithms for mixture model learning.

---

4. A Unified Approach to Controlling Implicit Regularization Using Mirror Descent

URL: View paper

Brief Assessment

Controlling Implicit Regularization[28] focuses on using Mirror Descent for implicit regularization in regression and classification problems with neural networks, not on constructing transformers that implement Mirror Descent for in-context learning of mixture models.

---

5. A shallow mirror transformer for subject-independent motor imagery BCI.

URL: View paper

Brief Assessment

Shallow Mirror Transformer[27] focuses on motor imagery BCI applications, not on constructing transformers to implement Mirror Descent algorithms for learning mixture weights in sequence modeling tasks.

---

6. CSFwinformer: Cross-Space-Frequency Window Transformer for Mirror Detection

URL: View paper

Brief Assessment

CSFwinformer[21] focuses on mirror detection in computer vision using spatial-frequency window transformers for texture analysis, not on implementing Mirror Descent optimization algorithms or in-context learning mechanisms.

---

7. Internalizing Tools as Morphisms in Graded Transformers

URL: View paper

Brief Assessment

Internalizing Tools Morphisms[23] focuses on graded transformers with typed block maps and symbolic computation routing, not on constructing transformers that implement Mirror Descent for mixture model learning. The candidate mentions mirror descent only in a geometric context ('mirror descent with bregman divergences') rather than as an algorithmic implementation for learning mixture weights.

---

8. A Universal Banach--Bregman Framework for Stochastic Iterations: Unifying Stochastic Mirror Descent, Learning and LLM Training

URL: View paper

Brief Assessment

Banach Bregman Framework[26] focuses on stochastic optimization in Banach-Bregman geometry for general learning algorithms, not on transformer architectures or their mechanistic implementation of specific algorithms like Mirror Descent for mixture model learning.

---

9. Identifying Equivalent Training Dynamics

URL: View paper

Brief Assessment

Equivalent Training Dynamics[24] focuses on identifying topological conjugacies between training dynamics using Koopman operator theory, not on constructing transformer architectures that implement specific optimization algorithms like Mirror Descent.

---

10. Mat: mixed-strategy game of adversarial training in fine-tuning

URL: View paper

Brief Assessment

Mixed Strategy Adversarial[25] applies entropy mirror descent to adversarial training in fine-tuning NLP models, not to constructing transformers that implement Mirror Descent for learning mixture weights in sequence modeling tasks.

Contribution 3: Theoretical connection between one-step Mirror Descent and Bayes-optimal predictor

Description: The authors establish that the one-step Mirror Descent estimator serves as a first-order approximation to the Bayes-optimal predictor. They prove that the Taylor expansions of both estimators coincide around the no-evidence regime, providing theoretical justification for why this simple non-iterative procedure achieves good performance.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Efficient methods in counterfactual policy learning and sequential decision making

URL: View paper

Brief Assessment

Counterfactual Policy Learning[34] focuses on counterfactual policy learning in continuous action spaces and sequential decision-making, not on Mirror Descent approximating Bayes-optimal predictors in mixture models or in-context learning.

---

2. How rotation invariant algorithms are fooled by noise on sparse targets

URL: View paper

Brief Assessment

Rotation Invariant Noise[33] focuses on rotation invariant algorithms for sparse linear regression problems and does not address the connection between Mirror Descent and Bayes-optimal predictors in the context of in-context learning or mixture models.

---

3. Multi-Expert Distributionally Robust Optimization for Out-of-Distribution Generalization

URL: View paper

Brief Assessment

Multi Expert Robust[35] focuses on distributionally robust optimization for out-of-distribution generalization using multiple expert heads, not on establishing theoretical connections between Mirror Descent and Bayes-optimal predictors in sequential learning contexts.

---

4. Theoretical Foundations of the Deep Copula Classifier: A Generative Approach to Modeling Dependent Features

URL: View paper

Brief Assessment

Deep Copula Classifier[30] focuses on copula-based generative classification for modeling feature dependencies, not on Mirror Descent approximating Bayes-optimal predictors in sequential learning contexts.

---

5. Learning with noisy labels

URL: View paper

Brief Assessment

Noisy Labels[32] addresses learning with random label noise in classification, not sequential modeling or in-context learning. The candidate focuses on unbiased estimators and weighted losses for handling label corruption, which is a fundamentally different problem from approximating Bayes-optimal predictors in mixture transition distributions.

---

6. Bayesian online natural gradient (BONG)

URL: View paper

Brief Assessment

BONG[31] focuses on variational Bayes for sequential inference in neural networks, not on mixture of transition distributions or in-context learning in transformers. The theoretical connection established in BONG[31] concerns conjugate exponential families in online learning, which is a different setting from the original paper's MTD framework for transformers.

Appendix: Text Similarity Detection

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

References

- [0] Transformers Learn Latent Mixture Models In-Context via Mirror Descent View paper
- [1] Glam: Efficient scaling of language models with mixture-of-experts View paper
- [2] In-context Learning for Mixture of Linear Regressions: Existence, Generalization and Training Dynamics View paper
- [3] Stochastic WaveNet: A Generative Latent Variable Model for Sequential Data View paper
- [4] An explanation of in-context learning as implicit bayesian inference View paper
- [5] Pre-trained Large Language Models Learn Hidden Markov Models In-context View paper
- [6] Does learning the right latent variables necessarily improve in-context learning? View paper
- [7] Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue View paper
- [8] Contextual hidden markov models View paper
- [9] Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models View paper
- [10] Online learning of contextual hidden markov models for temporal-spatial data analysis View paper
- [11] Latent LSTM allocation: Joint clustering and non-linear dynamic modeling of sequence data View paper
- [12] Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition View paper
- [13] Contextâ□based classification via mixture of hidden Markov model experts with applications in landmine detection View paper
- [14] OCEAN: Online Task Inference for Compositional Tasks with Context Adaptation View paper
- [15] Topic-informed dynamic mixture model for occupational heterogeneity in health risk behaviors View paper
- [16] A Comprehensive Survey on Large Language Models: From Pre-training to Autonomous Agents View paper
- [17] Context-based classification via data-dependent mixtures of logistic and hidden markov model classifiers View paper
- [18] In-Context Learning for Full Bayesian Inference View paper

- [19] Pre-trained Large Language Models Learn to Predict Hidden Markov Models In-context View paper
- [20] MCDDT: Mirror center loss based dual-scale dual-softmax transformer for multi-source subjects transfer learning in motor imagery recognition View paper
- [21] CSFwinformer: Cross-Space-Frequency Window Transformer for Mirror Detection View paper
- [22] Scoop: An Optimizer for Profiling Attacks against Higher-Order Masking View paper
- [23] Internalizing Tools as Morphisms in Graded Transformers View paper
- [24] Identifying Equivalent Training Dynamics View paper
- [25] Mat: mixed-strategy game of adversarial training in fine-tuning View paper
- [26] A Universal Banach--Bregman Framework for Stochastic Iterations: Unifying Stochastic Mirror Descent, Learning and LLM Training View paper
- [27] A shallow mirror transformer for subject-independent motor imagery BCI. View paper
- [28] A Unified Approach to Controlling Implicit Regularization Using Mirror Descent View paper
- [29] Time-Dependent Mirror Flows and Where to Find Them View paper
- [30] Theoretical Foundations of the Deep Copula Classifier: A Generative Approach to Modeling Dependent Features View paper
- [31] Bayesian online natural gradient (BONG) View paper
- [32] Learning with noisy labels View paper
- [33] How rotation invariant algorithms are fooled by noise on sparse targets View paper
- [34] Efficient methods in counterfactual policy learning and sequential decision making View paper
- [35] Multi-Expert Distributionally Robust Optimization for Out-of-Distribution Generalization View paper
- [36] Mixtures of in-context learners View paper
- [37] Merging Multi-Task Models via Weight-Ensembling Mixture of Experts View paper
- [38] Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation View paper
- [39] Introducing dynamic token embedding sampling of large language models for improved inference accuracy View paper
- [40] Novel token-level recurrent routing for enhanced mixture-of-experts performance View paper
- [41] MoDES: Accelerating Mixture-of-Experts Multimodal Large Language Models via Dynamic Expert Skipping View paper
- [42] On the role of attention in prompt-tuning View paper
- [43] Soft Adaptive Policy Optimization View paper
- [44] Domain-Specific Pruning of Large Mixture-of-Experts Models with Few-shot Demonstrations View paper
- [45] Style-conditional Prompt Token Learning for Generalizable Face Anti-spoofing View paper