Novelty Assessment Report

Paper: Segment-Level Attribution for Selective Learning of Long Reasoning Traces

PDF URL: https://openreview.net/pdf?id=C1UD4FLFPL

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Large Reasoning Models (LRMs) achieve strong reasoning performance by generating long chains of thought (CoTs), yet only a small fraction of these traces meaningfully contributes to answer prediction, while the majority contains repetitive or truncated content. Such output redundancy is further propagated after supervised finetuning (SFT), as models learn to imitate verbose but uninformative patterns, which can degrade performance. To this end, we incorporate integrated gradient attribution to quantify each token's influence on final answers and aggregate them into two segment-level metrics: (1) \textit{attribution strength} measures the overall attribution magnitude; and (2) \textit{direction consistency} captures whether tokens' attributions within a segment are uniformly positive or negative (high consistency), or a mixture of both (moderate consistency). Based on these two metrics, we propose a segment-level selective learning framework to identify important segments with high attribution strength but moderate consistency that indicate reflective rather than shallow reasoning. The framework then applies selective SFT on these important segments while masking loss for unimportant ones. Experiments across multiple models and datasets show that our approach improves accuracy and output efficiency, enabling more effective learning from long reasoning traces.

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: **Selective Learning from Long Reasoning Traces Using Segment-Level Attribution**

A total of **8 papers** were analyzed and organized into a taxonomy with **9 categories**.

Taxonomy Overview

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

- **Segment-Level Credit Assignment and Optimization**
- **Attribution-Enhanced Reasoning and Explainability**
- **Structured Reasoning with Hierarchical Segmentation**
- **Safety Assessment in Reasoning Processes**

Complete Taxonomy Tree

- Selective Learning from Long Reasoning Traces Using Segment-Level Attribution Survey Taxonomy
- Segment-Level Credit Assignment and Optimization
  - Segment-Level Attribution for Selective Learning ★ (1 papers)
  - [0] Segment-Level Attribution for Selective Learning of Long Reasoning Traces (Anon et al., 2026) View paper
  - Reinforcement Learning with Segment-Level Advantage Estimation (1 papers)
  - [4] Segment policy optimization: Effective segment-level credit assignment in rl for large language models (Guo, 2025) View paper
  - Direct Reasoning Optimization for Open-Ended Tasks (1 papers)
  - [3] Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks (Xu Yifei, 2025) View paper
- Attribution-Enhanced Reasoning and Explainability
  - Causal Attribution for Faithful Multi-Step Reasoning (1 papers)
  - [2] Towards Faithful Multi-step Reasoning through Fine-Grained Causal-aware Attribution Reasoning Distillation (Z Chu, 2025) View paper
  - Knowledge Graph-Constrained Attribution and Trajectory Reasoning (1 papers)
  - [7] KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision (Wu Rong, 2025) View paper
  - Post-Hoc Attribution via Answer Decomposition (1 papers)
  - [8] Decomposition-Enhanced Training for Post-Hoc Attributions In Language Models (Basu, 2025) View paper
  - Explainable Attribution in Domain-Specific Reasoning (1 papers)
  - [6] TRACING DECEPTION: A TRANSFORMER-BASED MODEL FOR EXPLAINABLE FINANCIAL FRAUD ANALYSIS IN CORPORATE REPORTING (K Hu, 2025) View paper
- Structured Reasoning with Hierarchical Segmentation (1 papers)
  - [1] CoTHSSum: Structured long-document summarization via chain-of-thought reasoning and hierarchical segmentation (Xiaoyong Chen, 2025) View paper
- Safety Assessment in Reasoning Processes (1 papers)
  - [5] SafeRBench: A Comprehensive Benchmark for Safety Assessment in Large Reasoning Models (Xin Gao, 2025) View paper

Narrative

Core task: selective learning from long reasoning traces using segment-level attribution. This field addresses the challenge of training models on extended reasoning sequences by identifying which segments contribute most to correct outcomes. The taxonomy organizes work into four main branches. Segment-Level Credit Assignment and Optimization focuses on methods that attribute credit to individual reasoning steps and optimize learning accordingly, often using reinforcement learning or gradient-based techniques to selectively reinforce valuable segments. Attribution-Enhanced Reasoning and Explainability emphasizes interpretability, developing frameworks that trace causal relationships between reasoning steps and final answers. Structured Reasoning with Hierarchical Segmentation explores how to decompose complex reasoning into meaningful units, often leveraging hierarchical structures or knowledge graphs. Safety Assessment in Reasoning Processes examines how to evaluate and ensure the reliability of multi-step reasoning, particularly in high-stakes domains.

Several active lines of work reveal key trade-offs in this landscape. One cluster, including Direct Reasoning Optimization[3] and Segment Policy Optimization[4], pursues fine-grained credit assignment through policy gradient methods that reward or penalize individual reasoning segments based on their contribution to task success. Another direction, represented by Causal Attribution Distillation[2] and KG-TRACES[7], emphasizes tracing causal dependencies to understand how intermediate steps influence outcomes. The original paper, Segment Attribution Learning[0], sits squarely within the credit assignment branch, sharing the optimization focus of works like Segment Policy Optimization[4] but distinguishing itself through its particular approach to attributing value across long traces. Meanwhile, SafeRBench[5] highlights an orthogonal concern—ensuring that segment-level learning does not compromise reasoning safety—illustrating how selective learning must balance efficiency gains against reliability requirements.

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

All three subtopics address the challenge of improving reasoning quality through selective optimization of reasoning traces, moving beyond simple outcome-based rewards. They share a common focus on granular (segment-level) analysis of reasoning processes, but differ in their core mechanisms: the original leaf uses attribution metrics to identify important segments for selective learning, while siblings use RL-based approaches—one optimizing reasoning directly for open-ended tasks without verifiable rewards, and another using advantage estimation at segment granularity for better credit assignment.

Similarities: - All three operate at segment or sub-trajectory granularity rather than treating entire reasoning traces uniformly - All aim to improve credit assignment or learning efficiency by identifying which parts of reasoning contribute to quality - All address limitations of outcome-only reward signals in reasoning optimization

Differences: - The original leaf uses attribution metrics (likely gradient-based or influence-based) for selection, while siblings use RL policy optimization frameworks - Direct Reasoning Optimization focuses on open-ended tasks without verifiable outcomes, whereas the original leaf and advantage estimation approaches likely assume some form of evaluable quality signal - Segment-Level Advantage Estimation uses RL advantage functions for credit assignment, while the original leaf uses attribution for selective learning (potentially supervised or imitation-based) - The original leaf emphasizes 'selective learning' (choosing what to learn from), while advantage estimation emphasizes improved credit assignment within policy gradient methods

Suggested Search Directions: - Hybrid approaches combining attribution metrics with RL advantage estimation for segment selection - Comparative studies on attribution-based vs. advantage-based segment identification accuracy - Applications to different reasoning domains (verifiable math/code vs. open-ended generation)

Sibling Subtopics

- **Direct Reasoning Optimization for Open-Ended Tasks** (leaves: 1, papers: 1)
- Scope: RL methods that optimize reasoning processes directly without relying on verifiable outcome-based rewards.
- Exclude: Excludes segment-level attribution or advantage estimation approaches; see other optimization methods.
- **Reinforcement Learning with Segment-Level Advantage Estimation** (leaves: 1, papers: 1)
- Scope: RL frameworks that estimate advantages at segment granularity to improve credit assignment in policy optimization.
- Exclude: Excludes attribution-based selection methods and outcome-only reward approaches; see other categories.

Contributions Analysis

Overall novelty summary. The paper proposes segment-level attribution metrics (attribution strength and direction consistency) to identify important reasoning segments for selective supervised finetuning. It occupies the 'Segment-Level Attribution for Selective Learning' leaf within the 'Segment-Level Credit Assignment and Optimization' branch. Notably, this leaf contains only the original paper itself, with no sibling papers identified in the taxonomy. This suggests the specific combination of integrated gradient attribution with segment-level selective SFT represents a relatively sparse research direction within the broader field of selective learning from long reasoning traces.

The taxonomy reveals that neighboring work pursues related but distinct approaches. The sibling leaf 'Reinforcement Learning with Segment-Level Advantage Estimation' focuses on RL-based policy optimization rather than attribution-guided supervised learning. Another sibling, 'Direct Reasoning Optimization for Open-Ended Tasks', addresses RL without verifiable rewards. The broader 'Attribution-Enhanced Reasoning and Explainability' branch emphasizes interpretability and faithfulness rather than optimization efficiency. The paper thus bridges attribution techniques (typically used for explainability) with selective learning objectives (typically addressed via RL), occupying a distinct methodological niche between these established directions.

Among 30 candidates examined across three contributions, none were found to clearly refute the paper's claims. For the segment-level attribution metrics using integrated gradients, 10 candidates were examined with 0 refutable matches. Similarly, the segment-level selective learning framework and the principled importance definition each examined 10 candidates with no refutations. This suggests that within the limited search scope, the specific combination of integrated gradient attribution at segment granularity for selective SFT appears relatively novel. However, the modest search scale means potentially relevant work outside the top-30 semantic matches may exist.

Based on the limited literature search of 30 candidates, the work appears to introduce a distinctive approach combining attribution-based segment selection with supervised finetuning. The absence of sibling papers in its taxonomy leaf and zero refutations across contributions suggest novelty within the examined scope. However, the analysis does not cover exhaustive citation networks or domain-specific venues, leaving open the possibility of related work in adjacent communities or recent preprints not captured by semantic search.

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

Contribution 1: Segment-level attribution metrics using integrated gradients

Description: The authors introduce two segment-level metrics derived from integrated gradient attribution: attribution strength, which measures the overall magnitude of a segment's influence on model predictions, and direction consistency, which captures whether a segment exhibits uniform or mixed attribution directions. These metrics enable identification of important reasoning segments.

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. Beyond intuition: Rethinking token attributions inside transformers

URL: View paper

Brief Assessment

Rethinking Token Attributions[29] applies integrated gradients to token-level attribution in transformers for model interpretability, not for identifying important reasoning segments in long chain-of-thought traces for selective learning.

2. GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs

URL: View paper

Brief Assessment

GrAInS[32] applies integrated gradients for token-level attribution in LLMs/VLMs to identify influential tokens for inference-time steering, not for segment-level importance analysis in reasoning traces. The original paper aggregates token-level IGs into segment-level metrics (attribution strength and direction consistency) to identify important reasoning segments for selective learning, which is a distinct application.

3. Analyzing Latent Concepts in Code Language Models

URL: View paper

Brief Assessment

Latent Concepts Code[35] applies integrated gradients to code token embeddings for concept discovery and clustering, not for segment-level attribution in reasoning traces. The candidate focuses on interpreting code language models through latent concept analysis, while the original paper uses integrated gradients to measure segment importance in long reasoning chains for selective learning.

4. Explaining pre-trained language models with attribution scores: An analysis in low-resource settings

URL: View paper

Brief Assessment

Attribution Low-Resource[31] focuses on token-level attribution scores for explaining pre-trained language models in classification tasks, not segment-level metrics for reasoning traces. The candidate applies integrated gradients to individual tokens in sentiment/NLI tasks, while the original paper aggregates token attributions into segment-level metrics (attribution strength and direction consistency) specifically for identifying important reasoning segments in long chain-of-thought outputs.

5. An attribution method for Siamese encoders

URL: View paper

Brief Assessment

Siamese Attribution[33] applies integrated gradients to Siamese encoders for comparing two text inputs, producing token-pair attribution matrices. The original paper applies integrated gradients to single reasoning traces for segment-level importance measurement in chain-of-thought reasoning.

6. Discretized Integrated Gradients for Explaining Language Models

URL: View paper

Brief Assessment

Discretized Integrated Gradients[37] focuses on improving integrated gradients for token-level attribution in discrete text spaces (word embeddings), not on segment-level aggregation for reasoning trace analysis. The candidate addresses interpolation path quality in embedding space, while the original paper aggregates token attributions into segment-level metrics for identifying important reasoning segments.

7. A Methodology for Explainable Large Language Models with Integrated Gradients and Linguistic Analysis in Text Classification

URL: View paper

Brief Assessment

Integrated Gradients Linguistic[30] applies integrated gradients at the token level for text classification in clinical contexts, not for segment-level attribution in reasoning traces. The candidate focuses on explaining BERT's classification decisions through token-level linguistic features, whereas the original paper aggregates token attributions into segment-level metrics (attribution strength and direction consistency) specifically for identifying important reasoning segments in long chain-of-thought outputs.

8. Uniform Discretized Integrated Gradients: An effective attribution based method for explaining large language models

URL: View paper

Brief Assessment

Uniform Discretized Gradients[36] focuses on token-level attribution for explaining LLM predictions in NLP tasks (sentiment classification, question answering), not segment-level metrics for identifying important reasoning segments in long chain-of-thought traces.

9. Evaluating attribution methods for explainable nlp with transformers

URL: View paper

Brief Assessment

Evaluating Attribution NLP[34] focuses on evaluating integrated gradients for token-level attribution in NLP tasks like sentiment analysis, not on segment-level aggregation metrics for reasoning chain analysis in LLMs.

10. Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers

URL: View paper

Brief Assessment

Integrated Gradients Adversarial[22] applies integrated gradients to identify influential words in text classification for adversarial attack detection, not for measuring segment-level attribution in reasoning traces or chain-of-thought outputs.

Contribution 2: Segment-level selective learning framework

Description: The authors propose a framework that identifies important segments based on high attribution strength and moderate consistency, then applies selective supervised finetuning only on these segments while masking loss for unimportant ones. This approach acts as implicit regularization by preventing overfitting to uninformative content.

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. Select2Reason: Efficient Instruction-Tuning Data Selection for Long-CoT Reasoning

URL: View paper

Brief Assessment

Select2Reason[11] focuses on selecting high-quality instruction data for fine-tuning based on question difficulty and response length, not on segment-level attribution analysis or selective loss masking during training. The original paper's framework identifies important segments within reasoning traces using integrated gradients and applies selective supervised finetuning by masking loss on unimportant segments, which is a fundamentally different approach from Select2Reason's data selection methodology.

2. Lr-sql: A supervised fine-tuning method for text2sql tasks under low-resource scenarios

URL: View paper

Brief Assessment

LR-SQL[12] focuses on schema linking for text2sql tasks using table slicing and CoT for database reconstruction, not on selective supervised finetuning based on segment importance for reasoning traces.

3. Fine-Grained Preference Optimization Improves Spatial Reasoning in VLMs

URL: View paper

Brief Assessment

Fine-Grained Preference Optimization[13] focuses on vision-language models for spatial reasoning tasks, using segment-level preference optimization to distinguish between descriptive grounding and logical reasoning components. The original paper addresses selective supervised finetuning for long reasoning traces in language models based on attribution strength and consistency metrics, which is a fundamentally different technical approach and application domain.

4. EPiC: Towards Lossless Speedup for Reasoning Training through Edge-Preserving CoT Condensation

URL: View paper

Brief Assessment

EPiC Lossless Speedup[16] focuses on condensing CoT traces by removing middle segments while preserving head and tail portions for efficient training. The original paper's framework identifies important segments based on attribution strength and consistency metrics, then applies selective supervised finetuning with loss masking. These are fundamentally different approaches to reasoning efficiency.

5. Audio Question Answering with GRPO-Based Fine-Tuning and Calibrated Segment-Level Predictions

URL: View paper

Brief Assessment

Audio QA GRPO[17] focuses on audio question answering using calibrated segment-level predictions of acoustic events, not selective supervised finetuning based on segment importance for reasoning traces in language models.

6. Importance weighting can help large language models self-improve

URL: View paper

Brief Assessment

Importance Weighting Self-Improve[10] focuses on filtering self-generated training samples based on distribution shift extent using importance weighting methods, not on segment-level attribution analysis within reasoning traces for selective supervised finetuning.

7. Enhancing Large Language Model Reasoning via Selective Critical Token Fine-Tuning

URL: View paper

Brief Assessment

Critical Token Fine-Tuning[14] operates at the token level using counterfactual perturbations to identify critical tokens, whereas the original paper works at the segment level using integrated gradient attribution. The two approaches differ fundamentally in granularity and identification methodology.

8. Not All Thoughts Matter: Selective Attention for Efficient Reasoning

URL: View paper

Brief Assessment

Selective Attention Reasoning[18] focuses on inference-time KV-cache management for efficient reasoning during generation, not on selective supervised finetuning based on segment importance attribution. The candidate addresses computational efficiency during inference by maintaining only first and last windows of attention, while the original contribution concerns training-time selective loss masking based on integrated gradient attribution to identify important reasoning segments.

9. Towards Efficient Medical Reasoning with Minimal Fine-Tuning Data

URL: View paper

Brief Assessment

Efficient Medical Reasoning[15] focuses on sample-level selection for medical reasoning datasets using difficulty and gradient influence, not segment-level selective learning within long reasoning traces. The two approaches operate at different granularities and address distinct problems.

10. Lens: Learning to segment anything with unified reinforced reasoning

URL: View paper

Brief Assessment

Lens Unified Reasoning[9] focuses on reinforcement learning for vision-language segmentation tasks, not on selective supervised finetuning based on segment importance for reasoning traces in language models. The technical domains and objectives are fundamentally different.

Contribution 3: Principled importance definition using integrated gradients

Description: The authors adopt integrated gradients as a principled method to measure segment importance, capturing both direct and indirect influences on final answer prediction. This approach addresses limitations of sequential appending or leave-one-out methods that underestimate indirectly contributive segments.

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. TIMING: Temporality-Aware Integrated Gradients for Time Series Explanation

URL: View paper

Brief Assessment

TIMING[23] applies integrated gradients to time series explanation tasks, focusing on temporal attribution and point-wise significance. The original paper uses integrated gradients for segment-level attribution in reasoning traces for language models, addressing different technical challenges in distinct domains.

2. Guided integrated gradients: An adaptive path method for removing noise

URL: View paper

Brief Assessment

Guided Integrated Gradients[19] focuses on reducing noise in visual attribution maps by adapting the integration path, not on measuring segment-level importance for reasoning traces in language models. The domains and applications are fundamentally different.

3. Contrastive Integrated Gradients: A Feature Attribution-Based Method for Explaining Whole Slide Image Classification

URL: View paper

Brief Assessment

Contrastive Integrated Gradients[25] applies integrated gradients to whole slide image classification in computational pathology, not to measuring segment importance in reasoning traces for language models.

4. Assessing the Reliability of Integrated Gradients-Based Saliency Maps for 3D Point Cloud Semantic Segmentation Models

URL: View paper

Brief Assessment

Point Cloud Saliency[27] applies integrated gradients to 3D point cloud semantic segmentation, not to measuring segment importance in reasoning traces for language models.

5. IG2: Integrated Gradient on Iterative Gradient Path for Feature Attribution

URL: View paper

Brief Assessment

IG2[24] focuses on feature attribution for neural network inputs (images, text features) to explain model predictions at the instance level, not on measuring segment importance in reasoning traces for selective learning.

6. A rigorous study of integrated gradients method and extensions to internal neuron attributions

URL: View paper

Brief Assessment

Rigorous Integrated Gradients[20] focuses on theoretical foundations and axioms of integrated gradients for input feature attribution, not on segment-level attribution for reasoning traces in language models. The candidate addresses mathematical properties and internal neuron attributions in neural networks, while the original applies integrated gradients to identify important segments in chain-of-thought reasoning for selective learning.

7. Manifold Integrated Gradients: Riemannian Geometry for Feature Attribution

URL: View paper

Brief Assessment

Manifold Integrated Gradients[21] focuses on improving integrated gradients for feature attribution in vision models through Riemannian geometry, addressing visualization noise and adversarial robustness. It does not address segment-level attribution for reasoning traces or measuring segment importance in language model outputs.

8. Integrated Gradients for Feature Assessment in Point Cloud-Based Data Sets

URL: View paper

Brief Assessment

Point Cloud Feature Assessment[28] applies integrated gradients to point cloud semantic segmentation for feature attribution, not to measuring segment importance in reasoning traces for language models.

9. Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers

URL: View paper

Brief Assessment

Integrated Gradients Adversarial[22] uses integrated gradients for word-level attribution in text classification tasks, not for measuring segment importance in reasoning model outputs or capturing indirect influences on answer prediction.

10. Xrai: Better attributions through regions

URL: View paper

Brief Assessment

XRAI[26] uses integrated gradients for pixel-level attribution in image saliency tasks, not for measuring segment importance in reasoning traces. The original paper applies IG to identify important reasoning segments in long chain-of-thought sequences, which is a fundamentally different application domain and methodology.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 3 similarity segment(s) across 3 paper(s).

The following **3 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

2. Discretized Integrated Gradients for Explaining Language Models

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

3. A rigorous study of integrated gradients method and extensions to internal neuron attributions

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Segment-Level Attribution for Selective Learning of Long Reasoning Traces View paper
- [1] CoTHSSum: Structured long-document summarization via chain-of-thought reasoning and hierarchical segmentation View paper
- [2] Towards Faithful Multi-step Reasoning through Fine-Grained Causal-aware Attribution Reasoning Distillation View paper
- [3] Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks View paper
- [4] Segment policy optimization: Effective segment-level credit assignment in rl for large language models View paper
- [5] SafeRBench: A Comprehensive Benchmark for Safety Assessment in Large Reasoning Models View paper
- [6] TRACING DECEPTION: A TRANSFORMER-BASED MODEL FOR EXPLAINABLE FINANCIAL FRAUD ANALYSIS IN CORPORATE REPORTING View paper
- [7] KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision View paper
- [8] Decomposition-Enhanced Training for Post-Hoc Attributions In Language Models View paper
- [9] Lens: Learning to segment anything with unified reinforced reasoning View paper
- [10] Importance weighting can help large language models self-improve View paper
- [11] Select2Reason: Efficient Instruction-Tuning Data Selection for Long-CoT Reasoning View paper
- [12] Lr-sql: A supervised fine-tuning method for text2sql tasks under low-resource scenarios View paper
- [13] Fine-Grained Preference Optimization Improves Spatial Reasoning in VLMs View paper
- [14] Enhancing Large Language Model Reasoning via Selective Critical Token Fine-Tuning View paper
- [15] Towards Efficient Medical Reasoning with Minimal Fine-Tuning Data View paper
- [16] EPiC: Towards Lossless Speedup for Reasoning Training through Edge-Preserving CoT Condensation View paper
- [17] Audio Question Answering with GRPO-Based Fine-Tuning and Calibrated Segment-Level Predictions View paper
- [18] Not All Thoughts Matter: Selective Attention for Efficient Reasoning View paper
- [19] Guided integrated gradients: An adaptive path method for removing noise View paper
- [20] A rigorous study of integrated gradients method and extensions to internal neuron attributions View paper
- [21] Manifold Integrated Gradients: Riemannian Geometry for Feature Attribution View paper
- [22] Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers View paper
- [23] TIMING: Temporality-Aware Integrated Gradients for Time Series Explanation View paper
- [24] IG2: Integrated Gradient on Iterative Gradient Path for Feature Attribution View paper
- [25] Contrastive Integrated Gradients: A Feature Attribution-Based Method for Explaining Whole Slide Image Classification View paper
- [26] Xrai: Better attributions through regions View paper
- [27] Assessing the Reliability of Integrated Gradients-Based Saliency Maps for 3D Point Cloud Semantic Segmentation Models View paper
- [28] Integrated Gradients for Feature Assessment in Point Cloud-Based Data Sets View paper
- [29] Beyond intuition: Rethinking token attributions inside transformers View paper
- [30] A Methodology for Explainable Large Language Models with Integrated Gradients and Linguistic Analysis in Text Classification View paper
- [31] Explaining pre-trained language models with attribution scores: An analysis in low-resource settings View paper
- [32] GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs View paper
- [33] An attribution method for Siamese encoders View paper
- [34] Evaluating attribution methods for explainable nlp with transformers View paper
- [35] Analyzing Latent Concepts in Code Language Models View paper
- [36] Uniform Discretized Integrated Gradients: An effective attribution based method for explaining large language models View paper
- [37] Discretized Integrated Gradients for Explaining Language Models View paper