Novelty Assessment Report

Paper: Efficient Reasoning with Balanced Thinking
PDF URL: https://openreview.net/pdf?id=cJseWJJ5IM
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Large Reasoning Models (LRMs) have shown remarkable reasoning capabilities, yet they often suffer from overthinking, expending redundant computational steps on simple problems, or underthinking, failing to explore sufficient reasoning paths despite inherent capabilities. These issues lead to inefficiencies and potential inaccuracies, limiting practical deployment in resource-constrained settings. Existing methods to mitigate overthinking, such as suppressing reflective keywords or adjusting reasoning length, may inadvertently induce underthinking, compromising accuracy. Therefore, we propose \textsc{ReBalance}, a training-free framework that achieves efficient reasoning with balanced thinking. \textsc{ReBalance} leverages confidence as a continuous indicator of reasoning dynamics, identifying overthinking through high confidence variance and underthinking via consistent overconfidence. By aggregating hidden states from a small-scale dataset into reasoning mode prototypes, we compute a steering vector to guide LRMs' reasoning trajectories. A dynamic control function modulates this vector's strength and direction based on real-time confidence, pruning redundancy during overthinking, and promoting exploration during underthinking. Extensive experiments conducted on four models ranging from 0.5B to 32B, and across nine benchmarks in math reasoning, general question answering, and coding tasks demonstrate that \textsc{ReBalance} effectively reduces output redundancy while improving accuracy, offering a general, training-free, and plug-and-play strategy for efficient and robust LRM deployment. Code and models will be made publicly available.

Core Task Landscape

This paper addresses: **balancing overthinking and underthinking in large reasoning models**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

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

- **Characterization and Analysis of Reasoning Inefficiencies**
- **Adaptive Reasoning Control and Difficulty-Aware Methods**
- **Reinforcement Learning and Training-Based Optimization**
- **Latent and Compressed Reasoning Representations**
- **Post-Hoc Reasoning Refinement and Optimization**
- **System-Level Efficiency and Serving Optimization**
- **Evaluation Frameworks and Benchmarking**
- **Security and Adversarial Aspects of Reasoning**
- **Survey and Review Papers**
- **Specialized Applications and Domain-Specific Studies**
- ... and 1 more categories

Complete Taxonomy Tree

- balancing overthinking and underthinking in large reasoning models Survey Taxonomy
- Characterization and Analysis of Reasoning Inefficiencies
  - Empirical Studies of Overthinking and Underthinking Patterns (3 papers)
  - [8] Between Underthinking and Overthinking: An Empirical Study of Reasoning Length and correctness in LLMs (Su, 2025) View paper
  - [34] When more is less: Understanding chain-of-thought length in llms (Wu Yuyang, 2025) View paper
  - [43] Think or Not? Exploring Thinking Efficiency in Large Reasoning Models via an Information-Theoretic Lens (Zhou Xiao, 2025) View paper
  - Mechanistic Analysis of Reasoning Dynamics (3 papers)
  - [5] On Reasoning Strength Planning in Large Reasoning Models (Zhang An, 2025) View paper
  - [21] The First Impression Problem: Internal Bias Triggers Overthinking in Reasoning Models (Li, 2025) View paper
  - [31] Lost at the Beginning of Reasoning (Liao, 2025) View paper
  - Reliability and Correctness of Reasoning Trajectories (2 papers)
  - [40] Large Reasoning Models are not thinking straight: on the unreliability of thinking trajectories (Cuesta-Ramirez, 2025) View paper
  - [41] Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification (Zhang Anqi, 2025) View paper
- Adaptive Reasoning Control and Difficulty-Aware Methods
  - Difficulty-Adaptive Reasoning Frameworks (3 papers)
  - [3] DAST: Difficulty-Adaptive Slow-Thinking for Large Reasoning Models (Shen Yi, 2025) View paper
  - [12] Think How to Think: Mitigating Overthinking with Autonomous Difficulty Cognition in Large Reasoning Models (Liu Yongjiang, 2025) View paper
  - [36] Plan and Budget: Effective and Efficient Test-Time Scaling on Large Language Model Reasoning (Lin, 2025) View paper

- Confidence-Based Reasoning Control ★ (2 papers)
- [0] Efficient Reasoning with Balanced Thinking (Anon et al., 2026) View paper
- [25] From "Aha Moments" to Controllable Thinking: Toward Meta-Cognitive Reasoning in Large Reasoning Models via Decoupled Reasoning and Control (Ha Rui, 2025) View paper
- Dynamic Early Exit and Truncation Mechanisms (2 papers)
- [9] Dynamic Early Exit in Reasoning Models (Yang Chenxu, 2025) View paper
- [24] Let LLMs Break Free from Overthinking via Self-Braking Tuning (H Zhao, 2025) View paper

- Reinforcement Learning and Training-Based Optimization
  - Reinforcement Learning for Reasoning Efficiency (3 papers)
  - [18] S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models (Yang Chenxu, 2025) View paper
  - [22] REA-RL: Reflection-Aware Online Reinforcement Learning for Efficient Large Reasoning Models (Deng, 2025) View paper
  - [49] SAIL-RL: Guiding MLLMs in When and How to Think via Dual-Reward RL Tuning (Shu, 2025) View paper
  - Thinking Pattern and Strategy Optimization (2 papers)
  - [10] Don't Think Longer, Think Wisely: Optimizing Thinking Dynamics for Large Reasoning Models (Anï¼ So-Hyun, 2025) View paper
  - [23] Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement (Zhao WeiXiang, 2025) View paper
- Latent and Compressed Reasoning Representations
  - Latent Space Reasoning (2 papers)
  - [28] A survey on latent reasoning (Zhu, 2025) View paper
  - [29] Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach (Geiping, 2025) View paper
  - Compressed Reasoning Representations (2 papers)
  - [14] Fast thinking for large language models (Zheng Haoyu, 2025) View paper
  - [47] Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching (Baek, 2025) View paper
- Post-Hoc Reasoning Refinement and Optimization
  - External Thought Manipulation and Guidance (2 papers)
  - [19] Thought Manipulation: External Thought Can Be Efficient for Large Reasoning Models (Liu Yu-Le, 2025) View paper
  - [37] ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation (Cao, 2025) View paper
  - Manifold Steering and Representation Intervention (1 papers)
  - [17] Mitigating Overthinking in Large Reasoning Models via Manifold Steering (Huang Yao, 2025) View paper
- System-Level Efficiency and Serving Optimization
  - Multi-Branch Reasoning and Ensembling (2 papers)
  - [7] Instilling parallel reasoning into language models (M Macfarlane, 2025) View paper
  - [45] Thinking Short and Right Over Thinking Long: Serving LLM Reasoning Efficiently and Accurately (Wang, 2025) View paper
- Evaluation Frameworks and Benchmarking
  - Reasoning Efficiency Benchmarks (3 papers)
  - [16] THINK-Bench: Evaluating Thinking Efficiency and Chain-of-Thought Quality of Large Reasoning Models (Li, 2025) View paper
  - [39] OptimalThinkingBench: Evaluating Over and Underthinking in LLMs (Aggarwal, 2025) View paper
  - [48] EffiReason-Bench: A Unified Benchmark for Evaluating and Advancing Efficient Reasoning in Large Language Models (Junquan Huang, 2025) View paper
  - Stress Testing and Robustness Evaluation (1 papers)
  - [44] REST: Stress Testing Large Reasoning Models by Asking Multiple Problems at Once (Pei, 2025) View paper
- Security and Adversarial Aspects of Reasoning
  - Adversarial Attacks on Reasoning Efficiency (2 papers)
  - [30] Excessive Reasoning Attack on Reasoning LLMs (Si, 2025) View paper
  - [38] Hauntattack: When attack follows reasoning as a shadow (Ma Jing-yuan, 2025) View paper
  - Backdoors and Vulnerabilities in Reasoning Models (3 papers)
  - [6] To Think or Not to Think: Exploring the Unthinking Vulnerability in Large Reasoning Models (Zhu, 2025) View paper
  - [13] BadReasoner: Planting Tunable Overthinking Backdoors into Large Reasoning Models for Fun or Profit (Yi Biao, 2025) View paper
  - [42] AI-Driven Privacy Audit Automation and Data Provenance Tracking in Large-Scale Systems (Chang, 2025) View paper
- Survey and Review Papers (6 papers)
  - [1] Stop overthinking: A survey on efficient reasoning for large language models (Sui Yang, 2025) View paper
  - [2] Towards reasoning era: A survey of long chain-of-thought for reasoning large language models (Chen, 2025) View paper
  - [4] Towards concise and adaptive thinking in large reasoning models: A survey (Zhu, 2025) View paper
  - [26] Don't Overthink It: A Survey of Efficient R1-style Large Reasoning Models (Du Yichao, 2025) View paper
  - [35] Efficient inference for large reasoning models: A survey (Liu Yu-e, 2025) View paper
  - [50] Reasoning on a Budget: A Survey of Adaptive and Controllable Test-Time Compute in LLMs (Alomrani, 2025) View paper
- Specialized Applications and Domain-Specific Studies
  - In-Context Learning and Prompting for Reasoning (1 papers)
  - [33] Innate Reasoning is Not Enough: In-Context Learning Enhances Reasoning Large Language Models with Less Overthinking (Ge Yuyao, 2025) View paper
  - Domain-Specific Reasoning Evaluation (1 papers)
  - [27] Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis (Mehta, 2025) View paper
  - Theoretical Foundations and Computational Complexity (1 papers)
  - [46] The Computational Advantage of Depth: Learning High-Dimensional Hierarchical Functions with Gradient Descent (Dandi, 2025) View paper
- Overthinking-Specific Mitigation Methods (4 papers)
  - [11] THOUGHTTERMINATOR: Benchmarking, Calibrating, and Mitigating Overthinking in Reasoning Models (Pu Xiao, 2025) View paper
  - [15] Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models (Xingyu Chen, 2025) View paper
  - [20] Missing Premise exacerbates Overthinking: Are Reasoning Models losing Critical Thinking Skill? (Fan, 2025) View paper

◦ [32] Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs (Chen, 2024) View paper

Narrative

Core task: balancing overthinking and underthinking in large reasoning models. The field has organized itself around several complementary perspectives. One major branch focuses on characterizing and analyzing reasoning inefficiencies—understanding when models generate excessive or insufficient reasoning steps and why these patterns emerge. Another set of approaches centers on adaptive reasoning control and difficulty-aware methods, which dynamically adjust computational effort based on problem characteristics; works like DAST[3] and Reasoning Strength Planning[5] exemplify this direction. Reinforcement learning and training-based optimization form a third pillar, aiming to teach models efficient reasoning habits during training. Meanwhile, latent and compressed reasoning representations explore whether reasoning can be internalized or condensed, and post-hoc refinement methods attempt to optimize reasoning traces after generation. System-level efficiency and serving optimization address deployment concerns, while evaluation frameworks and benchmarking provide standardized testbeds for comparing approaches. Security and adversarial aspects examine vulnerabilities introduced by reasoning processes, and specialized applications demonstrate domain-specific tuning.

Within the adaptive control branch, a particularly active line of work uses confidence signals to decide when to stop reasoning. Balanced Thinking[0] sits squarely in this confidence-based reasoning control cluster, alongside Meta-Cognitive Reasoning[25], which similarly leverages model self-assessment to regulate thinking depth. These methods contrast with difficulty-aware approaches like DAST[3], which predict problem hardness upfront rather than monitoring confidence during generation. The central tension across these branches involves trade-offs between accuracy, efficiency, and robustness: some methods prioritize minimizing wasted computation on easy problems, while others focus on ensuring sufficient reasoning for hard cases. Balanced Thinking[0] addresses this by using confidence thresholds to terminate reasoning adaptively, positioning itself as a middle ground that responds to the model's own uncertainty rather than relying solely on external difficulty estimates or fixed budgets.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. From "Aha Moments" to Controllable Thinking: Toward Meta-Cognitive Reasoning in Large Reasoning Models via Decoupled Reasoning and Control

Authors: Ha Rui, Li, Chaozhuo, Pu Rui, Su Sen | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Large Reasoning Models (LRMs) have demonstrated a latent capacity for complex reasoning by spontaneously exhibiting cognitive behaviors such as step-by-step reasoning, reflection, and backtracking, commonly referred to as"Aha Moments". However, such emergent behaviors remain unregulated and uncontrolled, often resulting in overthinking, where the model continues generating redundant reasoning content even after reaching reliable conclusions. This leads to excessive computational costs and increa...

Relationship Analysis

Both papers belong to the Confidence-Based Reasoning Control category, using model confidence signals to guide reasoning trajectories and resource allocation. They overlap in addressing overthinking/underthinking through confidence-based mechanisms: the original paper (ReBalance) uses confidence variance to detect overthinking and consistent overconfidence for underthinking, while the candidate paper (MERA) focuses on meta-cognitive control through explicit reasoning-control decoupling. The key difference is that ReBalance employs training-free steering vectors modulated by dynamic confidence functions, whereas MERA requires supervised fine-tuning and reinforcement learning (CSPO) to learn explicit control strategies at decision points.

Contributions Analysis

Overall novelty summary. The paper proposes ReBalance, a training-free framework that uses confidence signals to detect and mitigate both overthinking and underthinking in large reasoning models. It resides in the 'Confidence-Based Reasoning Control' leaf under 'Adaptive Reasoning Control and Difficulty-Aware Methods', which contains only two papers including this one. This is a relatively sparse research direction within a broader taxonomy of 50 papers across 22 leaf nodes, suggesting the specific approach of using confidence variance and overconfidence patterns for dual-sided reasoning control is not yet heavily explored.

The taxonomy reveals that ReBalance sits within a larger ecosystem of adaptive reasoning methods. Its parent branch includes 'Difficulty-Adaptive Reasoning Frameworks' (3 papers) and 'Dynamic Early Exit and Truncation Mechanisms' (2 papers), which address similar efficiency goals through different signals—explicit difficulty modeling or self-termination criteria. Neighboring branches include 'Reinforcement Learning and Training-Based Optimization' (5 papers) and 'Post-Hoc Reasoning Refinement' (3 papers), which tackle reasoning efficiency through training interventions or post-generation filtering rather than inference-time steering. The taxonomy's scope notes clarify that confidence-based methods like ReBalance are distinguished from difficulty-aware approaches by their reliance on model uncertainty rather than upfront problem characterization.

Among 30 candidates examined, the contribution-level analysis shows mixed novelty signals. 'Confidence as a continuous indicator' examined 10 candidates with 1 refutable match, suggesting some prior work uses confidence for reasoning control but perhaps not in the dual-detection manner proposed here. The 'ReBalance framework' contribution also examined 10 candidates with 1 refutable match, indicating the specific steering vector approach may have precedent. The 'plug-and-play solution' contribution examined 10 candidates with 2 refutable matches, suggesting training-free efficiency improvements are more established. These statistics reflect a limited search scope, not exhaustive coverage of the field.

Given the sparse population of the confidence-based control leaf and the limited search scale, the work appears to occupy a relatively underexplored niche within adaptive reasoning methods. The analysis captures top-30 semantic matches and does not guarantee comprehensive coverage of all relevant prior work, particularly in adjacent areas like difficulty-aware frameworks or latent reasoning representations. The dual focus on both overthinking and underthinking through confidence dynamics may differentiate this work from single-sided approaches, though the limited refutable matches suggest some conceptual overlap exists within the examined candidates.

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

Contribution 1: Confidence as a continuous indicator of reasoning dynamics

Description: The authors demonstrate that stepwise confidence and confidence variance can reliably indicate when large reasoning models exhibit overthinking (high variance) or underthinking (persistent overconfidence), providing a foundation for dynamic control of reasoning behavior.

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. Alignment for efficient tool calling of large language models

URL: View paper

Brief Assessment

Tool Calling Alignment[68] focuses on using confidence to determine when to invoke external tools versus answering directly, not on identifying overthinking/underthinking patterns in reasoning trajectories. The candidate addresses tool invocation efficiency rather than reasoning process dynamics.

2. Let LLMs Break Free from Overthinking via Self-Braking Tuning

URL: View paper

Brief Assessment

Self-Braking Tuning[24] focuses on identifying overthinking through step-level analysis and keyword markers rather than using confidence signals. The candidate does not demonstrate prior work using confidence variance to indicate overthinking or underthinking in reasoning models.

3. m1: Unleash the Potential of Test-Time Scaling for Medical Reasoning with Large Language Models

URL: View paper

Brief Assessment

m1[72] focuses on test-time scaling in medical reasoning and identifies overthinking through token budget analysis, but does not propose confidence or confidence variance as continuous indicators for detecting overthinking/underthinking dynamics.

4. Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know?

URL: View paper

Brief Assessment

Reasoning about Uncertainty[70] focuses on calibration of reasoning models' self-verbalized confidence estimates across different benchmarks, not on using confidence variance to detect overthinking/underthinking dynamics during the reasoning process itself.

5. Lexical Hints of Accuracy in LLM Reasoning Chains

URL: View paper

Brief Assessment

Lexical Hints[71] focuses on lexical markers (hedging words, sentiment) in chain-of-thought as post-hoc calibration signals for answer correctness, not on using confidence variance to dynamically control overthinking/underthinking during reasoning generation.

6. Stop overthinking: A survey on efficient reasoning for large language models

URL: View paper

Brief Assessment

Stop Overthinking Survey[1] focuses on categorizing existing efficient reasoning methods rather than proposing confidence as an indicator. The survey discusses various approaches but does not claim to introduce confidence-based reasoning control as a novel contribution.

7. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models

URL: View paper

Brief Assessment

Long Chain Survey[2] focuses on surveying long chain-of-thought reasoning paradigms and does not specifically investigate confidence signals as indicators of overthinking or underthinking in the manner proposed by the original paper.

8. Concise: Confidence-guided compression in step-by-step efficient reasoning

URL: View paper

Prior Art Analysis

Concise[69] demonstrates that confidence can serve as a continuous indicator of reasoning dynamics in large reasoning models, specifically identifying when models exhibit overthinking or underthinking behaviors. The paper explicitly frames 'confidence deficit' (where low confidence causes reflection on correct steps) and 'termination delay' (where reflection continues despite high confidence in the answer) as key patterns. This work predates the ORIGINAL paper and provides a confidence-guided perspective on reasoning behavior, including the observation that 'confidence deficit' involves undertrusting correct steps and 'termination delay' involves persistent reflection after reaching confident answers—concepts that closely parallel the ORIGINAL paper's claims about confidence variance indicating overthinking and persistent overconfidence indicating underthinking.

Evidence

Evidence 1 - **Rationale**: Both papers identify confidence as an indicator of reasoning problems. The ORIGINAL paper claims high variance indicates overthinking, while Concise[69] describes 'confidence deficit' where low confidence causes unnecessary reflection—both characterizing overthinking through confidence metrics. - **Original**: we can observe that the confidence values correlate with lrms' reasoning behaviors. specifically, high confidence variance may reflect frequent indecisive switching between different reasoning paths, causing redundant steps and delayed answer convergence,i.e., overthinking. conversely, consistent ov... - **Candidate**: confidence deficit. one major source of redundancy in lrms stems from their tendency to undertrust their correct intermediate steps. lrms often display unexpected reflection despite exhibiting fine-grained reasoning capabilities and achieving high stepwise accuracy, triggering reflection even on sim...

Evidence 2 - **Rationale**: Both papers use confidence to characterize problematic reasoning patterns. The ORIGINAL paper claims 'consistent overconfidence' indicates underthinking, while Concise[69] describes 'termination delay' where high confidence after answers still leads to continued reflection—both using confidence states to identify reasoning inefficiencies. - **Original**: conversely, consistent overconfidence can lead to premature commitment to incorrect reasoning paths,i.e., underthinking. thus, confidence can be leveraged as an indicator of reasoning dynamics. - **Candidate**: termination delay. lrms exhibit another important redundant reasoning pattern we term termination delay: after producing a confident final answer, the model is expected to conclude with minimal additional reasoning. however, it often continues to generate unnecessary reflection steps even after repe...

Evidence 3 - **Rationale**: Both papers explicitly claim that confidence serves as a continuous indicator for characterizing reasoning dynamics. Concise[69] adopts a 'confidence-guided perspective' to understand reflection generation, while the ORIGINAL paper claims confidence as a 'continuous and reliable signal' for characterizing overthinking and underthinking. - **Original**: as the current methods struggle to balance between overthinking and underthinking, we identify that confidence can serve as a continuous and reliable signal for characterizing both overthinking and underthinking in lrms, enabling fine-grained behavioral control. - **Candidate**: to overcome existing limitations, we aim to construct compact, coherent reasoning chains as training datasets by precisely removing redundant reflections, ensuring lrms do not suffer performance degradation after fine-tuning. to this end, based on the understanding that reflections are not solely de...

9. Innate Reasoning is Not Enough: In-Context Learning Enhances Reasoning Large Language Models with Less Overthinking

URL: View paper

Brief Assessment

In-Context Learning[33] focuses on how external prompting (zero-shot and few-shot CoT) regulates reasoning behavior in reasoning LLMs, rather than proposing confidence as an intrinsic indicator of overthinking/underthinking dynamics.

---

10. On Reasoning Strength Planning in Large Reasoning Models

URL: View paper

Brief Assessment

Reasoning Strength Planning[5] focuses on pre-allocated direction vectors in activation space that control reasoning token numbers, not on confidence signals indicating overthinking/underthinking dynamics during generation.

Contribution 2: REBALANCE framework for dynamic reasoning control

Description: The authors introduce REBALANCE, a training-free method that extracts steering vectors from hidden states and applies a dynamic control function to adjust reasoning trajectories in real-time based on confidence levels, balancing between overthinking and underthinking without requiring additional training.

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. Deep think with confidence

URL: View paper

Brief Assessment

Deep Think Confidence[51] focuses on filtering reasoning traces using confidence metrics for ensemble voting scenarios, not on real-time steering vector-based control of individual reasoning trajectories as in REBALANCE.

---

2. Fractional Reasoning via Latent Steering Vectors Improves Inference Time Compute

URL: View paper

Brief Assessment

Fractional Reasoning[55] focuses on continuous control of reasoning intensity via latent steering vectors to improve test-time compute strategies (best-of-n, majority voting). REBALANCE addresses a different problem: balancing overthinking and underthinking during reasoning by dynamically adjusting hidden states based on confidence levels, without requiring test-time aggregation methods.

---

3. Efficient inference for large reasoning models: A survey

URL: View paper

Brief Assessment

Efficient Inference Survey[35] is a survey paper that categorizes existing efficient reasoning methods but does not propose a specific training-free dynamic control framework using steering vectors extracted from hidden states based on confidence levels.

---

4. SEAL: Steerable Reasoning Calibration of Large Language Models for Free

URL: View paper

Prior Art Analysis

SEAL[53] demonstrates prior work on training-free dynamic reasoning control using steering vectors extracted from hidden states. Both papers extract steering vectors from latent representations to guide reasoning trajectories without additional training. SEAL[53] categorizes thoughts into execution, reflection, and transition types, extracting steering vectors from these categories in the latent space for on-the-fly calibration. This approach predates and overlaps substantially with REBALANCE's claim of being first to use steering vectors from hidden states for training-free dynamic reasoning control.

Evidence

Evidence 1 - **Rationale**: Both papers describe training-free methods that extract steering vectors from latent/hidden states for dynamic reasoning control. SEAL[53] explicitly describes extracting steering vectors in latent space for on-the-fly calibration, which directly overlaps with REBALANCE's claimed novelty of using steering vectors from hidden states. - **Original**: we propose rebalance, a training-free method that achieves efficient reasoning with balanced thinking. to achieve dynamic control between overthinking and underthinking, we first identify reasoning steps indicating overthinking and underthinking from a smallscale seen dataset, aggregate their corres... - **Candidate**: we introduce seal (steerable reasoning calibration), a training-free approach that seamlessly calibrates the cot process, improving accuracy while demonstrating significant efficiency gains. seal consists of an offline stage for extracting the reasoning steering vector in the latent space, followed ...

Evidence 2 - **Rationale**: Both papers aggregate hidden/latent states from different reasoning categories to compute steering vectors. SEAL[53] identifies thought categories with clear separation in latent space, similar to REBALANCE's aggregation of hidden states into reasoning mode prototypes. - **Original**: by aggregating hidden states from a small-scale dataset into reasoning mode prototypes, we compute a steering vector to guide lrms' reasoning trajectories. a dynamic control function modulates this vector's strength and direction based on real-time confidence - **Candidate**: our analysis reveals that excessive reflection and transition thoughts are strongly correlated with failure cases and these thought categories exhibit clear separation in the latent space. based on these, we introduce seal (steerable reasoning calibration), a training-free approach that seamlessly c...

Evidence 3 - **Rationale**: Both methods use an offline stage to extract steering vectors from hidden/latent states, then apply them during inference. This demonstrates that SEAL[53] already employed the core mechanism of offline steering vector extraction that REBALANCE claims as novel. - **Original**: confidence-based steering vector extraction. in this section, based on the modeling of overthinking and underthinking introduced in sec 3.2, we extract prototypical representations of both reasoning modes from the hidden states of lrms via an offline, single forward pass. - **Candidate**: seal consists of an offline stage for extracting the reasoning steering vector in the latent space, followed by an on-the-fly calibration of the reasoning trace through representation intervention using the steering vector.

---

5. Calibrating reasoning in language models with internal consistency

URL: View paper

Brief Assessment

Internal Consistency[57] focuses on measuring agreement between intermediate layer predictions to calibrate reasoning paths, whereas REBALANCE uses steering vectors from hidden states with confidence-based dynamic control to balance overthinking and underthinking.

---

6. Controlling Thinking Speed in Reasoning Models

URL: View paper

Brief Assessment

Controlling Thinking Speed[56] focuses on steering vectors for fast/slow thinking transitions based on difficulty estimation, while the original paper addresses overthinking/underthinking balance using confidence-based dynamic control. These represent distinct technical approaches to reasoning efficiency.

---

7. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas

URL: View paper

Brief Assessment

Spatial Reasoning VLMs[58] focuses on vision-language models' spatial reasoning through attention mechanism interventions, not on general reasoning models' overthinking/underthinking control via hidden state steering vectors.

---

8. Confident adaptive language modeling

URL: View paper

Brief Assessment

Confident Adaptive[54] focuses on early exit decoding in language models with calibrated confidence measures for textual consistency, not on steering vectors for reasoning trajectory control in LRMs.

---

9. Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification

URL: View paper

Brief Assessment

Self-Verification[41] focuses on probing hidden states to verify intermediate answer correctness and enable early exit decisions, whereas REBALANCE uses confidence-based steering vectors to dynamically control reasoning trajectories between overthinking and underthinking modes. The technical approaches and objectives differ fundamentally.

---

10. Improving reasoning performance in large language models via representation engineering

URL: View paper

Brief Assessment

Representation Engineering[52] focuses on extracting control vectors from hidden states to modulate model behavior at inference time, but applies this to improve reasoning performance through representation space manipulation rather than confidence-based dynamic control of reasoning trajectories. The technical approaches differ fundamentally in their control mechanisms and indicators.

Contribution 3: Plug-and-play solution improving efficiency and accuracy

Description: The authors validate that REBALANCE simultaneously reduces reasoning length and improves accuracy across multiple models (0.5B to 32B parameters) and nine benchmarks spanning math, science, commonsense, and coding tasks, providing a general and practical deployment strategy.

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. Logit arithmetic elicits long reasoning capabilities without training

URL: View paper

Brief Assessment

Logit Arithmetic[62] focuses on eliciting long reasoning capabilities through logit-level guidance at inference time, not on reducing reasoning length or mitigating overthinking/underthinking as REBALANCE does.

---

2. Two Experts Are All You Need for Steering Thinking: Reinforcing Cognitive Effort in MoE Reasoning Models Without Additional Training

URL: View paper

Brief Assessment

Two Experts[59] focuses on MoE architectures and identifying specialized 'cognitive experts' using NPMI, whereas the original paper addresses general LRMs through confidence-based steering vectors. The technical approaches and architectural requirements differ fundamentally.

---

3. Specreason: Fast and accurate inference-time compute via speculative reasoning

URL: View paper

Brief Assessment

Specreason[67] focuses on speculative reasoning for LRM inference acceleration using a small model to speculate intermediate steps, while REBALANCE addresses balanced thinking through confidence-based steering to mitigate both overthinking and underthinking. These are distinct technical approaches to improving reasoning efficiency.

---

4. Thinking slow, fast: Scaling inference compute with distilled reasoners

URL: View paper

Brief Assessment

Thinking Slow Fast[66] focuses on distilling reasoning capabilities into subquadratic architectures (Mamba models) to improve inference throughput, not on training-free methods for balancing overthinking/underthinking in existing reasoning models.

---

5. Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement

URL: View paper

Prior Art Analysis

Self-Guided Efficiency[23] demonstrates that similar training-free methods for improving reasoning efficiency and accuracy were proposed prior to REBALANCE. Both papers present training-free frameworks that reduce reasoning length while maintaining or improving accuracy across multiple models and benchmarks. Self-Guided Efficiency[23] introduces 'efficiency steering, a training-free activation steering technique' and validates it on 'seven lrm backbones across multiple mathematical reasoning benchmarks' showing that methods 'significantly reduce reasoning length while preserving or improving task performance.' This directly challenges the novelty claim that REBALANCE was the first to provide a general, training-free, plug-and-play strategy for simultaneous efficiency and accuracy improvements.

Evidence

Evidence 1 - **Rationale**: Both papers propose training-free frameworks for improving reasoning efficiency, with Self-Guided Efficiency[23] explicitly introducing a training-free technique prior to REBALANCE's publication. - **Original**: rebalance, a training-free framework that achieves efficient reasoning with balanced thinking - **Candidate**: we introduce efficiency steering, a training-free activation steering technique that modulates reasoning behavior via a single direction in the model's representation space.

Evidence 2 - **Rationale**: Self-Guided Efficiency[23] presents lightweight, training-free methods for enhancing LRM efficiency, which directly overlaps with REBALANCE's claim of offering a training-free, plug-and-play strategy. - **Original**: rebalanceeffectively reduces output redundancy while improving accuracy, offering a general, training-free, and plug-and-play strategy for efficient and robust lrm deployment. - **Candidate**: we propose two lightweight methods to enhance lrm efficiency. first, we introduce efficiency steering, a training-free activation steering technique

---

6. Mindstar: Enhancing math reasoning in pre-trained llms at inference time

URL: View paper

Brief Assessment

Mindstar[63] focuses on inference-time tree search for math reasoning using process-supervised reward models, not on training-free methods that reduce reasoning length while improving accuracy across diverse tasks like REBALANCE.

---

7. Advancing language model reasoning through reinforcement learning and inference scaling

URL: View paper

Brief Assessment

Reinforcement Learning Scaling[64] focuses on scaling RL training and test-time inference through repeated sampling and longer generation, not on training-free deployment strategies that simultaneously reduce reasoning length and improve accuracy across diverse benchmarks.

---

8. Adaptthink: Reasoning models can learn when to think

URL: View paper

Prior Art Analysis

Adaptthink[61] demonstrates a similar plug-and-play approach that simultaneously reduces reasoning length and improves accuracy across multiple models and benchmarks without requiring additional training. The candidate paper shows that their method reduces average response length by 53.0% for the 1.5B model and 40.1% for the 7B model while improving average accuracy by 2.4% and 2.3% respectively across math benchmarks. This directly challenges the novelty claim of REBALANCE being the first to provide a general, training-free, plug-and-play strategy that improves both efficiency and accuracy.

Evidence

Evidence 1 - **Rationale**: Both papers claim to simultaneously reduce reasoning length while improving accuracy across multiple models, demonstrating that similar prior work exists for this contribution. - **Original**: extensive experiments conducted on four models ranging from 0.5b to 32b, and across nine benchmarks in math reasoning, general question answering, and coding tasks demonstrate that rebalanceeffectively reduces output redundancy while improving accuracy, offering a general, training-free, and plug-an... - **Candidate**: compared to the original 1.5b and 7b models, adaptthink reduces the average response length by 53.0% and 40.1%, respectively, while also improves the average accuracy by 2.4% and 2.3%, demonstrating that adaptthink enables significantly more efficient reasoning without compromising and even enhancin...

Evidence 2 - **Rationale**: Both papers demonstrate that their methods reduce output length while improving accuracy across multiple benchmarks, showing that the claimed novelty of simultaneously achieving both goals was already demonstrated in prior work. - **Original**: rebalancenot only reduces output length but also improves the accuracy. to summarize, our contributions are as follows: • as the current methods struggle to balance between overthinking and underthinking, we identify that confidence can serve as a continuous and reliable signal for characterizing bo... - **Candidate**: our experiments demonstrate that adaptthink effectively enables reasoning models to adaptively select the optimal thinking mode based on problem difficulty, leading to substantial reductions in inference cost compared to prior approaches, while consistently enhancing model accuracy. for instance, on...

Evidence 3 - **Rationale**: Both papers validate their methods across multiple models and benchmarks, showing simultaneous improvements in both efficiency and accuracy, which refutes the claim of being the first to provide such a solution. - **Original**: extensive experiments across four models ranging from 0.5b to 32b, and on nine benchmarks covering math reasoning, general question answering, and coding tasks, demonstrate the effectiveness and strong generalization capabilities of rebalance. notably, rebalancenot only reduces output length but als... - **Candidate**: table 1 presents the evaluation results of different methods on gsm8k, math500, and aime 2024. compared to the original 1.5b and 7b models, adaptthink reduces the average response length by 53.0% and 40.1%, respectively, while also improves the average accuracy by 2.4% and 2.3%, demonstrating that a...

---

9. LAPO: Internalizing Reasoning Efficiency via Length-Adaptive Policy Optimization

URL: View paper

Brief Assessment

LAPO[65] uses reinforcement learning to internalize reasoning length control during training, while REBALANCE is a training-free, plug-and-play framework. These represent fundamentally different approaches to efficiency improvement.

---

10. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency

URL: View paper

Brief Assessment

InternVL3.5[60] focuses on multimodal model optimization through cascade RL and visual resolution routing for vision-language tasks, not on general reasoning model inference efficiency improvements without training as described in the original contribution.

Appendix: Text Similarity Detection

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

References

- [0] Efficient Reasoning with Balanced Thinking View paper
- [1] Stop overthinking: A survey on efficient reasoning for large language models View paper
- [2] Towards reasoning era: A survey of long chain-of-thought for reasoning large language models View paper
- [3] DAST: Difficulty-Adaptive Slow-Thinking for Large Reasoning Models View paper

- [4] Towards concise and adaptive thinking in large reasoning models: A survey View paper
- [5] On Reasoning Strength Planning in Large Reasoning Models View paper
- [6] To Think or Not to Think: Exploring the Unthinking Vulnerability in Large Reasoning Models View paper
- [7] Instilling parallel reasoning into language models View paper
- [8] Between Underthinking and Overthinking: An Empirical Study of Reasoning Length and correctness in LLMs View paper
- [9] Dynamic Early Exit in Reasoning Models View paper
- [10] Don't Think Longer, Think Wisely: Optimizing Thinking Dynamics for Large Reasoning Models View paper
- [11] THOUGHTTERMINATOR: Benchmarking, Calibrating, and Mitigating Overthinking in Reasoning Models View paper
- [12] Think How to Think: Mitigating Overthinking with Autonomous Difficulty Cognition in Large Reasoning Models View paper
- [13] BadReasoner: Planting Tunable Overthinking Backdoors into Large Reasoning Models for Fun or Profit View paper
- [14] Fast thinking for large language models View paper
- [15] Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models View paper
- [16] THINK-Bench: Evaluating Thinking Efficiency and Chain-of-Thought Quality of Large Reasoning Models View paper
- [17] Mitigating Overthinking in Large Reasoning Models via Manifold Steering View paper
- [18] S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models View paper
- [19] Thought Manipulation: External Thought Can Be Efficient for Large Reasoning Models View paper
- [20] Missing Premise exacerbates Overthinking: Are Reasoning Models losing Critical Thinking Skill? View paper
- [21] The First Impression Problem: Internal Bias Triggers Overthinking in Reasoning Models View paper
- [22] REA-RL: Reflection-Aware Online Reinforcement Learning for Efficient Large Reasoning Models View paper
- [23] Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement View paper
- [24] Let LLMs Break Free from Overthinking via Self-Braking Tuning View paper
- [25] From "Aha Moments" to Controllable Thinking: Toward Meta-Cognitive Reasoning in Large Reasoning Models via Decoupled Reasoning and Control View paper
- [26] Don't Overthink It: A Survey of Efficient R1-style Large Reasoning Models View paper
- [27] Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis View paper
- [28] A survey on latent reasoning View paper
- [29] Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach View paper
- [30] Excessive Reasoning Attack on Reasoning LLMs View paper
- [31] Lost at the Beginning of Reasoning View paper
- [32] Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs View paper
- [33] Innate Reasoning is Not Enough: In-Context Learning Enhances Reasoning Large Language Models with Less Overthinking View paper
- [34] When more is less: Understanding chain-of-thought length in llms View paper
- [35] Efficient inference for large reasoning models: A survey View paper
- [36] Plan and Budget: Effective and Efficient Test-Time Scaling on Large Language Model Reasoning View paper
- [37] ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation View paper
- [38] Hauntattack: When attack follows reasoning as a shadow View paper
- [39] OptimalThinkingBench: Evaluating Over and Underthinking in LLMs View paper
- [40] Large Reasoning Models are not thinking straight: on the unreliability of thinking trajectories View paper
- [41] Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification View paper
- [42] AI-Driven Privacy Audit Automation and Data Provenance Tracking in Large-Scale Systems View paper
- [43] Think or Not? Exploring Thinking Efficiency in Large Reasoning Models via an Information-Theoretic Lens View paper
- [44] REST: Stress Testing Large Reasoning Models by Asking Multiple Problems at Once View paper
- [45] Thinking Short and Right Over Thinking Long: Serving LLM Reasoning Efficiently and Accurately View paper
- [46] The Computational Advantage of Depth: Learning High-Dimensional Hierarchical Functions with Gradient Descent View paper
- [47] Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching View paper
- [48] EffiReason-Bench: A Unified Benchmark for Evaluating and Advancing Efficient Reasoning in Large Language Models View paper
- [49] SAIL-RL: Guiding MLLMs in When and How to Think via Dual-Reward RL Tuning View paper
- [50] Reasoning on a Budget: A Survey of Adaptive and Controllable Test-Time Compute in LLMs View paper
- [51] Deep think with confidence View paper
- [52] Improving reasoning performance in large language models via representation engineering View paper
- [53] SEAL: Steerable Reasoning Calibration of Large Language Models for Free View paper
- [54] Confident adaptive language modeling View paper
- [55] Fractional Reasoning via Latent Steering Vectors Improves Inference Time Compute View paper
- [56] Controlling Thinking Speed in Reasoning Models View paper
- [57] Calibrating reasoning in language models with internal consistency View paper
- [58] Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas View paper
- [59] Two Experts Are All You Need for Steering Thinking: Reinforcing Cognitive Effort in MoE Reasoning Models Without Additional Training View paper
- [60] InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency View paper
- [61] Adaptthink: Reasoning models can learn when to think View paper
- [62] Logit arithmetic elicits long reasoning capabilities without training View paper
- [63] Mindstar: Enhancing math reasoning in pre-trained llms at inference time View paper
- [64] Advancing language model reasoning through reinforcement learning and inference scaling View paper
- [65] LAPO: Internalizing Reasoning Efficiency via Length-Adaptive Policy Optimization View paper
- [66] Thinking slow, fast: Scaling inference compute with distilled reasoners View paper
- [67] Specreason: Fast and accurate inference-time compute via speculative reasoning View paper
- [68] Alignment for efficient tool calling of large language models View paper
- [69] Concise: Confidence-guided compression in step-by-step efficient reasoning View paper
- [70] Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know? View paper

- [71] Lexical Hints of Accuracy in LLM Reasoning Chains View paper
- [72] m1: Unleash the Potential of Test-Time Scaling for Medical Reasoning with Large Language Models View paper