Novelty Assessment Report

Paper: StreamingThinker: Large Language Models Can Think While Reading
PDF URL: https://openreview.net/pdf?id=10Iiew095e
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in chain of thought (CoT) reasoning. However, the current LLM reasoning paradigm initiates thinking only after the entire input is available, which introduces unnecessary latency and weakens attention to earlier information in dynamic scenarios. Inspired by human cognition of thinking while reading, we first design a \textit{\textbf{streaming thinking}} paradigm for LLMs, where reasoning unfolds in the order of input and further adjusts its depth once reading is complete. We instantiate this paradigm with \textit{StreamingThinker}, a framework that enables LLMs to think while reading through the integration of streaming CoT generation, streaming-constraint training, and streaming parallel inference. Specifically, StreamingThinker employs streaming reasoning units with quality control for CoT generation, enforces order-preserving reasoning through streaming attention masks and position encoding, and leverages parallel KV caches that decouple input encoding from reasoning generation, thereby ensuring alignment and enabling true concurrency. We evaluate StreamingThinker on the Qwen3 model family across math reasoning, logical reasoning, and context-based QA reasoning tasks. Experimental results show that the StreamingThinker preserves performance comparable to batch thinking, while yielding an 80\% reduction in token waiting before the onset of reasoning and a more than 60\% reduction in time-level latency for producing the final answer, demonstrating the effectiveness of the streaming paradigm for LLM reasoning.

Core Task Landscape

This paper addresses: **Streaming Reasoning with Concurrent Input Processing**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Streaming Inference Architectures and Parallelism**
- **Streaming Reasoning and Semantic Processing**
- **Streaming Language Model Reasoning**
- **Online Learning and Adaptive Models for Streams**
- **Data Stream Analytics and Pattern Mining**
- **Application Domains and Real-Time Systems**
- **Optimization and Control with Streaming Inference**

Complete Taxonomy Tree

- Streaming Reasoning with Concurrent Input Processing Survey Taxonomy
- Streaming Inference Architectures and Parallelism
  - Multi-Device and Distributed Inference Parallelism (5 papers)
  - [13] InSS: An Intelligent Scheduling Orchestrator for Multi-GPU Inference With Spatio-Temporal Sharing (Ziyi Han, 2024) View paper
  - [25] LLMEasyQuant: Scalable Quantization for Parallel and Distributed LLM Inference (Liu, 2024) View paper
  - [30] Top: task-based operator parallelism for asynchronous deep learning inference on gpu (Changyao Lin, 2024) View paper
  - [31] Liger: Interleaving Intra- and Inter-Operator Parallelism for Distributed Large Model Inference (Jiangsu Du, 2024) View paper
  - [40] Kraken: Inherently Parallel Transformers For Efficient Multi-Device Inference (Zhang Hengrui, 2024) View paper
  - Single-Device Concurrent Inference Optimization (3 papers)
  - [5] Optimizing CNN inference speed over big social data through efficient model parallelism for sustainable web of things (Yuhao Hu, 2024) View paper
  - [14] Securely and Efficiently Outsourcing Neural Network Inference via Parallel MSB Extraction (Xin Liu, 2024) View paper
  - [50] Parallel CPU-GPU Execution for LLM Inference on Constrained GPUs (Fan, 2025) View paper
  - Edge and Resource-Constrained Inference (4 papers)
  - [8] A Survey of AI Inference Technologies for On-Device Systems (Wen-Zhu Wang, 2025) View paper
  - [10] A Computing-in-Memory Engine Supporting One-Shot Floating-Point NN Inference and On-Device Fine-Tuning for Edge AI (Haikang Diao, 2025) View paper
  - [23] Real Time Complex Event Processing And Stream Reasoning for Low-Cost IoT Systems (Mouhamet Latyr Ndiaye, 2024) View paper
  - [43] ParaTra: A parallel transformer inference framework for concurrent service provision in edge computing (Fenglong Cai, 2023) View paper
- Streaming Reasoning and Semantic Processing
  - Semantic Stream Integration and Knowledge Reasoning (4 papers)
  - [12] Real-Time Semantic Data Integration and Reasoning in Life- and Time-Critical Decision Support Systems (Andreas Soularidis, 2024) View paper

- [18] V2X-UniPool: Unifying Multimodal Perception and Knowledge Reasoning for Autonomous Driving (Luo, 2025) View paper
- [47] Context-aware query derivation for IoT data streams with DIVIDE enabling privacy by design (Mathias DeÂ Brouwer, 2023) View paper
- [48] Real-Time Semantic Indexing for High-Volume Data Streams (Yeshwanth Raj, 2025) View paper
- Temporal Logic and Rule-Based Stream Reasoning (3 papers)
- [26] Stream reasoning playground (Patrik Schneider, 2022) View paper
- [44] Grounding stream reasoning research (Bonte, 2024) View paper
- [46] Stream reasoning in temporal datalog (Ronca, 2018) View paper

- Streaming Language Model Reasoning
  - Concurrent Input-Output Streaming for LLMs ★ (3 papers)
  - [0] StreamingThinker: Large Language Models Can Think While Reading (Anon et al., 2026) View paper
  - [27] Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model (Ke Hu, 2025) View paper
  - [36] GhostShell: Streaming LLM Function Calls for Concurrent Embodied Programming (Gong Jian, 2025) View paper
  - Streaming Multimodal and Video Understanding (2 papers)
  - [3] Streaming video understanding and multi-round interaction with memory-enhanced knowledge (Xiong Haomiao, 2025) View paper
  - [39] EndoMamba: An Efficient Foundation Model for Endoscopic Videos (Qingyao Tian, 2025) View paper
  - Simultaneous Translation and Speech Processing (1 papers)
  - [7] SimulS2S-LLM: Unlocking Simultaneous Inference of Speech LLMs for Speech-to-Speech Translation (Deng, 2025) View paper
  - Inference-Time Scaling and Reward Modeling (1 papers)
  - [1] Inference-Time Scaling for Generalist Reward Modeling (Liu, 2025) View paper
- Online Learning and Adaptive Models for Streams
  - Online Learning Frameworks and Operationalization (3 papers)
  - [15] Streammlops: Online learning in practice from big data streams & real-time applications (M Barry, 2023) View paper
  - [21] StreamMLOps: Operationalizing online learning for big data streaming & real-time applications (Mariam Barry, 2023) View paper
  - [28] A review on big data real-time stream processing and its scheduling techniques (Nicoleta Tantalaki, 2020) View paper
  - Adaptive Algorithms for Concept Drift (3 papers)
  - [19] Evidential Ensemble Preference-Guided Learning Approach for Real-Time Multimode Fault Diagnosis (Zeyi Liu, 2024) View paper
  - [33] ParaFIS:A new online fuzzy inference system based on parallel drift anticipation (Cl'ement Leroy, 2019) View paper
  - [49] A novel online real-time classifier for multi-label data streams (Venkatesan Rajasekar, 2016) View paper
  - Sparse and Active Online Learning (2 papers)
  - [35] Fast online inference for nonlinear contextual bandit based on Generative Adversarial Network (Da Tsai, 2022) View paper
  - [42] â□□1,2-Norm and CUR Decomposition based Sparse Online Active Learning for Data Streams with Streaming Features (Zhong Chen, 2024) View paper
- Data Stream Analytics and Pattern Mining
  - Trend and Pattern Detection in Streams (2 papers)
  - [4] Scout Sketch+: Finding Both Promising and Damping Items Simultaneously in Data Streams (Guoju Gao, 2024) View paper
  - [20] Real-time suspicious detection framework for financial data streams (Elshan Gadimov, 2025) View paper
  - Data Compression and Efficient Storage for Streams (2 papers)
  - [22] Real-Time Decompression and Rasterization of Massive Point Clouds (Rahul Goel, 2024) View paper
  - [29] Homomorphic data compression for real time photon correlation analysis. (Sebastian Strempfer, 2025) View paper
  - Real-Time Forecasting and Prediction (2 papers)
  - [37] Real-Time Go-Around Prediction: A case study of JFK airport (Liu Ke, 2024) View paper
  - [45] Real-time forecasting of data revisions in epidemic surveillance streams. (Jingjing Tang, 2025) View paper
- Application Domains and Real-Time Systems
  - Autonomous Systems and Robotics (2 papers)
  - [11] Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation (Yi-feng Jiang, 2022) View paper
  - [38] EdgeLoc: A Communication-Adaptive Parallel System for Real-Time Localization in Infrastructure-Assisted Autonomous Driving (Liu, 2024) View paper
  - IoT and Publish-Subscribe Systems (2 papers)
  - [9] Privacy Preserving Publish/Subscribe for Geo-Textual Data Streams (Ya Gao, 2024) View paper
  - [32] Query-Driven Multimodal GraphRAG: Dynamic Local Knowledge Graph Construction for Online Reasoning (Chenyang Bu, 2025) View paper
  - Graph-Based and Dynamic Network Inference (2 papers)
  - [2] Online Scheduling of Edge Multiple-Model Inference with DAG Structure and Retraining (Yifan Zeng, 2025) View paper
  - [17] Helios: Efficient Distributed Dynamic Graph Sampling for Online GNN Inference (Jie Sun, 2025) View paper
  - Communication and Signal Processing (1 papers)
  - [16] Real-time photonic blind interference cancellation (Joshua C. Lederman, 2023) View paper
  - Real-Time Monitoring and Tracking (2 papers)
  - [34] SMART-BARN: Scalable multimodal arena for real-time tracking behavior of animals in large numbers (MÃ¡tÃ© Nagy, 2023) View paper
  - [41] Real-time processing of streaming big data (Ali A. Safaei, 2017) View paper
- Optimization and Control with Streaming Inference (2 papers)
  - [6] High Confidence Level Inference is Almost Free using Parallel Stochastic Optimization (Zhu, 2024) View paper
  - [24] Parallel Control With Event-Based Adaptive Critic Implementation for Robust Optimal Tracking of Uncertain Nonlinear Systems (Shanshan Jiao, 2025) View paper

Narrative

Core task: streaming reasoning with concurrent input processing. This field addresses the challenge of performing inference or reasoning over data that arrives continuously, often requiring systems to produce outputs while new inputs are still being received. The taxonomy reflects a diverse landscape spanning multiple communities. Streaming Inference Architectures and Parallelism focuses on hardware-

aware designs and model partitioning strategies that enable efficient concurrent execution, as seen in works like Edge Multiple-Model Scheduling[2] and CNN Model Parallelism[5]. Streaming Reasoning and Semantic Processing emphasizes knowledge representation and logic-based methods for integrating temporal or event-driven data, while Streaming Language Model Reasoning targets large language models that must handle incremental or overlapping input-output flows. Online Learning and Adaptive Models for Streams and Data Stream Analytics and Pattern Mining address scenarios where models must update continuously or detect evolving patterns in high-velocity data. Application Domains and Real-Time Systems showcase deployments in robotics, IoT, and multimedia, and Optimization and Control with Streaming Inference explores feedback-driven decision-making under streaming constraints.

Within this landscape, a particularly active line of work centers on concurrent input-output streaming for large language models, where the goal is to interleave token generation with ongoing input reception. StreamingThinker[0] sits squarely in this cluster, proposing mechanisms that allow reasoning to proceed even as new context arrives, a capability also explored by Duplex Speech Modeling[27] in conversational settings and GhostShell[36] in privacy-preserving scenarios. These efforts contrast with more traditional pipeline approaches that separate encoding and decoding phases, and they differ from hardware-centric parallelism studies like Parallel CPU-GPU Inference[50] by emphasizing algorithmic strategies for overlapping computation. Meanwhile, works such as Streaming Video Memory[3] tackle related challenges in vision domains, highlighting that concurrent processing spans modalities. StreamingThinker[0] distinguishes itself by focusing on reasoning tasks that require maintaining coherent intermediate states across dynamically arriving inputs, a setting that remains less explored than pure generation or classification streams.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model

Authors: Ke Hu, Ehsan Hosseini Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, et al. (10 authors total) | **Year/Venue**: 2025 • Interspeech | **URL**: View paper

Abstract

Spoken dialogue is an intuitive form of human-computer interaction, yet current speech language models often remain constrained to turn-based exchanges, lacking real-time adaptability such as user barge-in. We propose a novel duplex speech to speech (S2S) architecture featuring continuous user inputs and codec agent outputs with channel fusion that directly models simultaneous user and agent streams. Using a pretrained streaming encoder for user input enables the first duplex S2S model without r...

Relationship Analysis

Both papers belong to the concurrent input-output streaming category, enabling LLMs to process inputs while generating outputs simultaneously. StreamingThinker focuses on streaming reasoning where the model generates chain-of-thought reasoning incrementally as text input arrives sentence-by-sentence, using streaming attention masks and parallel KV caches to enable thinking while reading. In contrast, the candidate paper addresses duplex speech-to-speech modeling with continuous audio streams for both user input and agent output, using channel fusion and separate architectures for codec-based voice generation rather than text-based reasoning.

2. GhostShell: Streaming LLM Function Calls for Concurrent Embodied Programming

Authors: Gong Jian, Huang, Youwei, Yuan Bo, Zhu Ming, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We present GhostShell, a novel approach that leverages Large Language Models (LLMs) to enable streaming and concurrent behavioral programming for embodied systems. In contrast to conventional methods that rely on pre-scheduled action sequences or behavior trees, GhostShell drives embodied systems to act on-the-fly by issuing function calls incrementally as tokens are streamed from the LLM. GhostShell features a streaming XML function token parser, a dynamic function interface mapper, and a multi...

Relationship Analysis

Both papers belong to the concurrent input-output streaming category, enabling LLMs to process inputs and generate outputs simultaneously. They overlap in addressing latency reduction and real-time reasoning during streaming scenarios. However, StreamingThinker focuses on streaming chain-of-thought reasoning with depth-adjustable thinking aligned to sequential input order, while GhostShell emphasizes streaming function call execution for embodied robotic systems with multi-channel concurrent behavioral programming across physical components.

Contributions Analysis

Overall novelty summary. The paper introduces a streaming thinking paradigm enabling LLMs to reason concurrently with input reception, instantiated through the StreamingThinker framework. It resides in the 'Concurrent Input-Output Streaming for LLMs' leaf, which contains only three papers including this one. This leaf sits within the broader 'Streaming Language Model Reasoning' branch, indicating a relatively sparse but emerging research direction. The taxonomy reveals that while streaming inference architectures are well-studied across hardware and distributed systems, concurrent reasoning during input arrival for LLMs remains underexplored compared to adjacent areas like multimodal streaming or simultaneous translation.

The taxonomy positions this work at the intersection of streaming inference and reasoning-specific challenges. Neighboring leaves address streaming multimodal understanding and simultaneous translation, which share the concurrent processing goal but target different modalities or tasks. The parent branch excludes batch-based reasoning and inference-time scaling without streaming, clarifying that StreamingThinker's novelty lies in its order-preserving reasoning during input arrival rather than post-hoc computation scaling. Sibling papers in the same leaf explore duplex communication and privacy-preserving scenarios, suggesting the field is fragmenting into specialized concurrent processing contexts rather than converging on unified frameworks.

Across three contributions, the analysis examined 29 candidate papers with zero refutable pairs identified. The streaming thinking paradigm examined 9 candidates with no refutations, the StreamingThinker framework examined 10 with none refutable, and the streaming CoT generation pipeline examined 10 with none refutable. This suggests that among the top-30 semantically similar works retrieved, none provide directly overlapping prior art for the specific combination of streaming reasoning, order-preserving CoT generation, and parallel KV cache mechanisms. The limited search scope means exhaustive coverage cannot be claimed, but within the examined set, the contributions appear distinct from existing concurrent inference and reasoning methods.

Given the sparse taxonomy leaf and absence of refutable candidates in the limited search, the work appears to occupy a relatively novel position within streaming LLM reasoning. However, the analysis covers only 29 candidates from semantic search, leaving open the possibility of relevant work in adjacent communities or under different terminology. The taxonomy's structure suggests the field is still coalescing around core abstractions for concurrent reasoning, and StreamingThinker's integration of streaming constraints with reasoning depth adjustment may represent an early exploration of this design space rather than an incremental refinement of established methods.

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

Contribution 1: Streaming thinking paradigm for LLMs

Description: The authors introduce a new reasoning paradigm where LLMs perform reasoning concurrently with input reception rather than waiting for complete input. This paradigm mirrors human cognition of thinking while reading and allows adaptive reasoning depth adjustment after input completion.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Depth-Breadth Synergy in RLVR: Unlocking LLM Reasoning Gains with Adaptive Exploration

URL: View paper

Brief Assessment

Depth-Breadth Synergy[52] focuses on reinforcement learning optimization for reasoning (RLVR with adaptive rollout sampling and batch scaling), not on concurrent input-reasoning paradigms or streaming processing architectures.

---

2. Scaling latent reasoning via looped language models

URL: View paper

Brief Assessment

Looped Language Models[55] focuses on iterative computation in latent space through recurrent parameter sharing for parameter efficiency, not on streaming input processing with concurrent reasoning during input reception as described in the original paper.

---

3. LARES: Latent Reasoning for Sequential Recommendation

URL: View paper

Brief Assessment

LARES[51] focuses on latent reasoning for sequential recommendation systems, not on streaming input processing paradigms for LLMs. The candidate addresses recommendation tasks through depth-recurrent reasoning over static sequences, whereas the original paper introduces concurrent reasoning during incremental input reception for language models.

---

4. A Multi-Layered AI-Driven Cybersecurity Architecture: Integrating Entropy Analytics, Fuzzy Reasoning, Game Theory, and Multi-Agent Reinforcement Learning for Adaptive Threat Defense

URL: View paper

Brief Assessment

Multi-Layered Cybersecurity[59] focuses on cybersecurity defense mechanisms integrating entropy analytics, fuzzy reasoning, game theory, and multi-agent reinforcement learning. It does not address LLM reasoning paradigms or sequential input processing with adaptive depth adjustment.

---

5. Dynamic chain-of-thought: Towards adaptive deep reasoning

URL: View paper

Brief Assessment

Dynamic Chain-of-Thought[53] focuses on adaptive reasoning depth through pruning and reward optimization to reduce computational costs, not on concurrent input processing with reasoning. The candidate addresses computational efficiency in reasoning steps, while the original introduces a paradigm where reasoning unfolds during input reception rather than after complete input availability.

---

6. Toward adaptive reasoning in large language models with thought rollback

URL: View paper

Brief Assessment

Thought Rollback[58] focuses on adaptive reasoning through backward rollback to previous thoughts for error correction, not on concurrent processing of input with reasoning. The candidate's core mechanism is 'rolling back thoughts' to revise errors, whereas the original paper's paradigm is about thinking while reading during input reception.

---

7. Think Before Recommend: Unleashing the Latent Reasoning Power for Sequential Recommendation

URL: View paper

Brief Assessment

Think Before Recommend[60] focuses on inference-time computing for sequential recommendation systems, not LLM reasoning paradigms. It addresses item prediction through multi-step reasoning in recommender systems, which is a fundamentally different domain from the streaming thinking paradigm for processing text input in LLMs.

---

8. RL for Reasoning by Adaptively Revealing Rationales

URL: View paper

Brief Assessment

Adaptive Rationale Revealing[57] focuses on curriculum learning through partial expert demonstrations in RL, not on concurrent input-reasoning processing. The candidate addresses when to reveal rationales during training, while the original addresses how to process input streams concurrently with reasoning generation.

---

9. PATS: Process-Level Adaptive Thinking Mode Switching

URL: View paper

Brief Assessment

PATS[56] focuses on switching between different reasoning complexity modes (simple/medium/complex) at the process level based on step difficulty, not on concurrent input processing with reasoning. The original paper's streaming paradigm enables reasoning during input reception, while PATS[56] addresses adaptive depth adjustment after complete input is available.

Contribution 2: StreamingThinker framework

Description: The authors develop a complete framework implementing the streaming thinking paradigm. It integrates three components: a generation pipeline for streaming chain-of-thought traces, training mechanisms with streaming attention masks and position encoding, and parallel KV cache inference that decouples input encoding from reasoning generation.

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. Visual Structures Helps Visual Reasoning: Addressing the Binding Problem in VLMs

URL: View paper

Brief Assessment

Visual Structures Reasoning[64] addresses visual reasoning in vision-language models through spatial input structuring, not streaming text processing or concurrent reasoning during sequential input reception. The candidate focuses on binding visual features to referents, while the original contribution concerns temporal streaming of textual/reasoning tokens with parallel KV caches.

---

2. Parallel-r1: Towards parallel thinking via reinforcement learning

URL: View paper

Brief Assessment

Parallel-r1[61] focuses on parallel thinking via reinforcement learning for mathematical reasoning, where multiple reasoning paths are explored concurrently and then summarized. This differs fundamentally from StreamingThinker's streaming paradigm, which enables reasoning to unfold alongside incremental input reception with streaming attention masks and parallel KV caches for concurrent input encoding and reasoning generation.

---

3. Adaptive Termination for Multi-round Parallel Reasoning: An Universal Semantic Entropy-Guided Framework

URL: View paper

Brief Assessment

Adaptive Termination Framework[68] focuses on test-time collaborative inference combining sequential and parallel reasoning with semantic entropy-based termination control. It does not address concurrent reasoning during input reception or streaming paradigms.

---

4. A survey on parallel reasoning

URL: View paper

Brief Assessment

Parallel Reasoning Survey[65] focuses on parallel reasoning paradigms where multiple reasoning paths are explored concurrently and then aggregated, rather than streaming thinking where reasoning unfolds alongside incremental input reception. The survey does not address frameworks that enable reasoning during input streaming with parallel KV cache inference.

---

5. Generalizable Reasoning through Compositional Energy Minimization

URL: View paper

Brief Assessment

Compositional Energy Minimization[70] focuses on compositional reasoning through energy landscape optimization over solution spaces, not on streaming/concurrent processing during input reception or parallel KV cache inference mechanisms.

---

6. Learning adaptive parallel reasoning with language models

URL: View paper

Brief Assessment

Adaptive Parallel Reasoning[62] focuses on parallelizing reasoning computations across multiple inference threads to improve efficiency and reduce latency, rather than enabling concurrent reasoning during sequential input reception as in StreamingThinker's streaming paradigm.

---

7. Distributional reasoning in LLMs: Parallel reasoning processes in multi-hop reasoning

URL: View paper

Brief Assessment

Distributional Reasoning[67] analyzes internal multi-hop reasoning processes through linear transformations and parallel reasoning paths in hidden layers, rather than proposing a framework for streaming input processing with concurrent reasoning generation.

---

8. Dynamic Parallel Tree Search for Efficient LLM Reasoning

URL: View paper

Brief Assessment

Dynamic Parallel Tree[66] focuses on parallelizing tree-of-thought search algorithms for efficiency in reasoning path exploration, not on enabling concurrent reasoning during incremental input reception as StreamingThinker does.

---

9. Instilling parallel reasoning into language models

URL: View paper

Brief Assessment

Instilling Parallel Reasoning[63] focuses on parallel reasoning across multiple independent threads to explore diverse strategies simultaneously, not on streaming reasoning during input reception. The candidate's parallel threads execute different solution approaches concurrently, while StreamingThinker enables reasoning to unfold alongside incremental input arrival with streaming attention masks and position encoding.

---

10. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning

URL: View paper

Brief Assessment

Mechanistic Chain-of-Thought[69] investigates the internal neural mechanisms of how LLMs generate chain-of-thought reasoning, focusing on attention heads and layer-wise functional components. It does not address streaming processing, concurrent input-output generation, or parallel KV cache inference architectures.

Contribution 3: Streaming CoT generation pipeline with quality control

Description: The authors design a data generation method that produces streaming-compatible reasoning traces. It employs boundary tokens to define reasoning units, uses teacher model reconstruction for alignment, and includes quality metrics (granularity and sequential consistency scores) with depth-controlled reasoning variants.

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. Vchain: Chain-of-visual-thought for reasoning in video generation

URL: View paper

Brief Assessment

VChain[77] focuses on visual reasoning for video generation using keyframes from multimodal models, not on streaming text-based CoT generation with sequential alignment to input context order.

---

2. Internal Chain-of-Thought: Empirical Evidence for Layer-wise Subtask Scheduling in LLMs

URL: View paper

Brief Assessment

Internal Chain-of-Thought[74] investigates how LLMs internally decompose tasks across layers during inference, not data generation pipelines. The candidate focuses on analyzing existing model behavior through layer-wise masking and patching, while the original designs a method to produce streaming-compatible reasoning traces with boundary tokens and quality metrics.

---

3. Dissecting chain-of-thought: Compositionality through in-context filtering and learning

URL: View paper

Brief Assessment

Dissecting Chain-of-Thought[76] focuses on theoretical analysis of compositional function learning through filtering and in-context learning phases, not on streaming-compatible reasoning trace generation with quality metrics and depth control as described in the original paper.

---

4. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing

URL: View paper

Brief Assessment

ThinkSound[79] focuses on audio generation from video using chain-of-thought reasoning for multimodal tasks, not on streaming-compatible reasoning traces aligned with sequential input context order for general LLM reasoning tasks.

---

5. Context-preserving logical drift confinement for large language model reasoning through recursive constraint projection

URL: View paper

Brief Assessment

Logical Drift Confinement[80] focuses on reasoning stability and trajectory discretization during multistep inference, not on streaming-compatible CoT generation with boundary tokens, teacher model reconstruction, or sequential consistency metrics as proposed in the original paper.

---

6. MCoT-RE: Multi-Faceted Chain-of-Thought and Re-Ranking for Training-Free Zero-Shot Composed Image Retrieval

URL: View paper

Brief Assessment

MCoT-RE[78] focuses on composed image retrieval using multi-faceted chain-of-thought for balancing visual-textual context in a two-stage retrieval framework. It does not address streaming-compatible reasoning traces, boundary tokens, teacher model reconstruction, or sequential consistency metrics for language model reasoning.

---

7. Why think step by step? reasoning emerges from the locality of experience

URL: View paper

Brief Assessment

Locality of Experience[75] focuses on why chain-of-thought reasoning works based on local statistical structure in training data, not on designing generation pipelines with boundary tokens, teacher reconstruction, or quality metrics for streaming-compatible traces.

---

8. Enhancing Long Chain-of-Thought Reasoning through Multi-Path Plan Aggregation

URL: View paper

Brief Assessment

Multi-Path Plan Aggregation[73] focuses on multi-path exploration and aggregation of planning steps in long CoT reasoning, not on streaming-compatible generation aligned with sequential input order. The candidate addresses CoT derailment through plan exploration rather than streaming processing during input reception.

---

9. Chain of thought empowers transformers to solve inherently serial problems

URL: View paper

Brief Assessment

Chain-of-Thought Empowers[71] focuses on theoretical expressiveness of CoT for decoder-only transformers on inherently serial problems, not on streaming-compatible data generation pipelines with quality metrics and sequential alignment.

---

10. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification

URL: View paper

Brief Assessment

SpecInfer[72] focuses on accelerating LLM inference through speculative decoding with token trees, not on generating chain-of-thought reasoning traces aligned with sequential input order. The systems address fundamentally different problems: streaming reasoning generation versus inference acceleration.

Appendix: Text Similarity Detection

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

References

- [0] StreamingThinker: Large Language Models Can Think While Reading View paper
- [1] Inference-Time Scaling for Generalist Reward Modeling View paper
- [2] Online Scheduling of Edge Multiple-Model Inference with DAG Structure and Retraining View paper
- [3] Streaming video understanding and multi-round interaction with memory-enhanced knowledge View paper
- [4] Scout Sketch+: Finding Both Promising and Damping Items Simultaneously in Data Streams View paper

- [5] Optimizing CNN inference speed over big social data through efficient model parallelism for sustainable web of things View paper
- [6] High Confidence Level Inference is Almost Free using Parallel Stochastic Optimization View paper
- [7] SimulS2S-LLM: Unlocking Simultaneous Inference of Speech LLMs for Speech-to-Speech Translation View paper
- [8] A Survey of AI Inference Technologies for On-Device Systems View paper
- [9] Privacy Preserving Publish/Subscribe for Geo-Textual Data Streams View paper
- [10] A Computing-in-Memory Engine Supporting One-Shot Floating-Point NN Inference and On-Device Fine-Tuning for Edge AI View paper
- [11] Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation View paper
- [12] Real-Time Semantic Data Integration and Reasoning in Life- and Time-Critical Decision Support Systems View paper
- [13] InSS: An Intelligent Scheduling Orchestrator for Multi-GPU Inference With Spatio-Temporal Sharing View paper
- [14] Securely and Efficiently Outsourcing Neural Network Inference via Parallel MSB Extraction View paper
- [15] Streammlops: Online learning in practice from big data streams & real-time applications View paper
- [16] Real-time photonic blind interference cancellation View paper
- [17] Helios: Efficient Distributed Dynamic Graph Sampling for Online GNN Inference View paper
- [18] V2X-UniPool: Unifying Multimodal Perception and Knowledge Reasoning for Autonomous Driving View paper
- [19] Evidential Ensemble Preference-Guided Learning Approach for Real-Time Multimode Fault Diagnosis View paper
- [20] Real-time suspicious detection framework for financial data streams View paper
- [21] StreamMLOps: Operationalizing online learning for big data streaming & real-time applications View paper
- [22] Real-Time Decompression and Rasterization of Massive Point Clouds View paper
- [23] Real Time Complex Event Processing And Stream Reasoning for Low-Cost IoT Systems View paper
- [24] Parallel Control With Event-Based Adaptive Critic Implementation for Robust Optimal Tracking of Uncertain Nonlinear Systems View paper
- [25] LLMEasyQuant: Scalable Quantization for Parallel and Distributed LLM Inference View paper
- [26] Stream reasoning playground View paper
- [27] Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model View paper
- [28] A review on big data real-time stream processing and its scheduling techniques View paper
- [29] Homomorphic data compression for real time photon correlation analysis. View paper
- [30] Top: task-based operator parallelism for asynchronous deep learning inference on gpu View paper
- [31] Liger: Interleaving Intra- and Inter-Operator Parallelism for Distributed Large Model Inference View paper
- [32] Query-Driven Multimodal GraphRAG: Dynamic Local Knowledge Graph Construction for Online Reasoning View paper
- [33] ParaFIS:A new online fuzzy inference system based on parallel drift anticipation View paper
- [34] SMART-BARN: Scalable multimodal arena for real-time tracking behavior of animals in large numbers View paper
- [35] Fast online inference for nonlinear contextual bandit based on Generative Adversarial Network View paper
- [36] GhostShell: Streaming LLM Function Calls for Concurrent Embodied Programming View paper
- [37] Real-Time Go-Around Prediction: A case study of JFK airport View paper
- [38] EdgeLoc: A Communication-Adaptive Parallel System for Real-Time Localization in Infrastructure-Assisted Autonomous Driving View paper
- [39] EndoMamba: An Efficient Foundation Model for Endoscopic Videos View paper
- [40] Kraken: Inherently Parallel Transformers For Efficient Multi-Device Inference View paper
- [41] Real-time processing of streaming big data View paper
- [42] â□□1,2-Norm and CUR Decomposition based Sparse Online Active Learning for Data Streams with Streaming Features View paper
- [43] ParaTra: A parallel transformer inference framework for concurrent service provision in edge computing View paper
- [44] Grounding stream reasoning research View paper
- [45] Real-time forecasting of data revisions in epidemic surveillance streams. View paper
- [46] Stream reasoning in temporal datalog View paper
- [47] Context-aware query derivation for IoT data streams with DIVIDE enabling privacy by design View paper
- [48] Real-Time Semantic Indexing for High-Volume Data Streams View paper
- [49] A novel online real-time classifier for multi-label data streams View paper
- [50] Parallel CPU-GPU Execution for LLM Inference on Constrained GPUs View paper
- [51] LARES: Latent Reasoning for Sequential Recommendation View paper
- [52] Depth-Breadth Synergy in RLVR: Unlocking LLM Reasoning Gains with Adaptive Exploration View paper
- [53] Dynamic chain-of-thought: Towards adaptive deep reasoning View paper
- [54] Adaptivestep: Automatically dividing reasoning step through model confidence View paper
- [55] Scaling latent reasoning via looped language models View paper
- [56] PATS: Process-Level Adaptive Thinking Mode Switching View paper
- [57] RL for Reasoning by Adaptively Revealing Rationales View paper
- [58] Toward adaptive reasoning in large language models with thought rollback View paper
- [59] A Multi-Layered AI-Driven Cybersecurity Architecture: Integrating Entropy Analytics, Fuzzy Reasoning, Game Theory, and Multi-Agent Reinforcement Learning for Adaptive Threat Defense View paper
- [60] Think Before Recommend: Unleashing the Latent Reasoning Power for Sequential Recommendation View paper
- [61] Parallel-r1: Towards parallel thinking via reinforcement learning View paper
- [62] Learning adaptive parallel reasoning with language models View paper
- [63] Instilling parallel reasoning into language models View paper
- [64] Visual Structures Helps Visual Reasoning: Addressing the Binding Problem in VLMs View paper
- [65] A survey on parallel reasoning View paper
- [66] Dynamic Parallel Tree Search for Efficient LLM Reasoning View paper
- [67] Distributional reasoning in LLMs: Parallel reasoning processes in multi-hop reasoning View paper
- [68] Adaptive Termination for Multi-round Parallel Reasoning: An Universal Semantic Entropy-Guided Framework View paper
- [69] How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning View paper
- [70] Generalizable Reasoning through Compositional Energy Minimization View paper
- [71] Chain of thought empowers transformers to solve inherently serial problems View paper

- [72] Specinfer: Accelerating large language model serving with tree-based speculative inference and verification View paper
- [73] Enhancing Long Chain-of-Thought Reasoning through Multi-Path Plan Aggregation View paper
- [74] Internal Chain-of-Thought: Empirical Evidence for Layer-wise Subtask Scheduling in LLMs View paper
- [75] Why think step by step? reasoning emerges from the locality of experience View paper
- [76] Dissecting chain-of-thought: Compositionality through in-context filtering and learning View paper
- [77] Vchain: Chain-of-visual-thought for reasoning in video generation View paper
- [78] MCoT-RE: Multi-Faceted Chain-of-Thought and Re-Ranking for Training-Free Zero-Shot Composed Image Retrieval View paper
- [79] Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing View paper
- [80] Context-preserving logical drift confinement for large language model reasoning through recursive constraint projection View paper