Novelty Assessment Report

Paper: In-Place Test-Time Training
PDF URL: https://openreview.net/pdf?id=dTWfCLSoyl
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

The static "train then deploy" paradigm fundamentally limits Large Language Models (LLMs) from dynamically adapting their weights in response to continuous streams of new information inherent in real-world tasks. Test-Time Training (TTT) offers a compelling alternative by updating a subset of model parameters (fast weights) at inference time, yet its potential in the current LLM ecosystem is hindered by critical barriers including architectural incompatibility, computational inefficiency and misaligned fast weight objectives for language modeling. In this work, we introduce **In-Place Test-Time Training (In-Place TTT)**, a framework that seamlessly endows LLMs with Test-Time Training ability. In-Place TTT treats the final projection matrix of the ubiquitous MLP blocks as its adaptable fast weights, enabling a ``drop-in" enhancement for LLMs without costly retraining from scratch. Furthermore, we replace TTT's generic reconstruction objective with a tailored, theoretically-grounded objective explicitly aligned with the Next-Token-Prediction task governing autoregressive language modeling. This principled objective, combined with an efficient chunk-wise update mechanism, results in a highly scalable algorithm compatible with context parallelism. Extensive experiments validate our framework's effectiveness: as an in-place enhancement, it enables a 4B-parameter model to achieve superior performance on tasks with contexts up to 128k, and when pretrained from scratch, it consistently outperforms competitive TTT-related approaches. Ablation study results further provide deeper insights on our design choices. Collectively, our results establish In-Place TTT as a promising step towards a paradigm of continual learning in LLMs.

Core Task Landscape

This paper addresses: **Test-Time Training for Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **Test-Time Adaptation Mechanisms**
- **Inference-Time Optimization Strategies**
- **Training-Based Test-Time Scaling**
- **Post-Training Paradigms and Frameworks**
- **Domain-Specific Applications**
- **Computational Efficiency and System Design**
- **Analysis and Evaluation Frameworks**

Complete Taxonomy Tree

- Test-Time Training for Large Language Models Survey Taxonomy
- Test-Time Adaptation Mechanisms
  - Parameter Update Approaches ★ (6 papers)
  - [0] In-Place Test-Time Training (Anon et al., 2026) View paper
  - [1] Test-Time Learning for Large Language Models (Hu Jinwu, 2025) View paper
  - [7] Medadapter: Efficient test-time adaptation of large language models towards medical reasoning (Shi, 2024) View paper
  - [10] Revisiting dynamic evaluation: Online adaptation for large language models (Rannen-Triki, 2024) View paper
  - [35] The surprising effectiveness of test-time training for few-shot learning (AkyÃ¼rek, 2024) View paper
  - [50] Evaluating Test-Time Training for Conceptual Reasoning in Large Language Models (Derksen, 2025) View paper
  - Activation-Based Intervention (3 papers)
  - [3] Inference-time intervention: Eliciting truthful answers from a language model (Li Kenneth, 2023) View paper
  - [21] ControlMLLM: Training-Free Visual Prompt Learning for Multimodal Large Language Models (Xinyue Cai, 2024) View paper
  - [49] Bridging the language gaps in large language models with inference-time cross-lingual intervention (Birch, 2025) View paper
  - Auxiliary Model Integration (3 papers)
  - [30] Efficient Uncertainty Estimation via Distillation of Bayesian Large Language Models (Shi, 2025) View paper
  - [32] Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning (Ximing Lu, 2023) View paper
  - [33] Test-time Backdoor Mitigation for Black-Box Large Language Models with Defensive Demonstrations (Askari, 2023) View paper
- Inference-Time Optimization Strategies
  - Search and Sampling Methods (6 papers)
  - [14] Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters (Snell, 2024) View paper
  - [17] First Finish Search: Efficient Test-Time Scaling in Large Language Models (Sengupta Ayan, 2025) View paper
  - [19] Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning (CV Snell, 2025) View paper
  - [29] Enhancing Test-Time Scaling of Large Language Models with Hierarchical Retrieval-Augmented MCTS (Wan, 2025) View paper

- [41] A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods (Isha Puri, 2025) View paper
- [47] Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models (Zhichen Dong, 2024) View paper
- Verification and Reward-Based Guidance (4 papers)
- [8] Step-level Verifier-guided Hybrid Test-Time Scaling for Large Language Models (Chang, 2025) View paper
- [24] Ttrl: Test-time reinforcement learning (Zuo Yuxin, 2025) View paper
- [26] Inference-time alignment in continuous space (Yuan, 2025) View paper
- [46] Scaling Test-Time Compute Without Verification or RL is Suboptimal (Setlur, 2025) View paper
- Prompt-Based Methods (3 papers)
- [9] Boosted Prompt Ensembles for Large Language Models (Pitis, 2023) View paper
- [11] Automatic Prompt Selection for Large Language Models (Viet-Tung Do, 2024) View paper
- [40] Test-time Prompt Intervention (Yang Chenxu, 2025) View paper
- Reasoning Chain Optimization (2 papers)
- [34] Patience Is The Key to Large Language Model Reasoning (Yu, 2024) View paper
- [48] Towards thinking-optimal scaling of test-time compute for llm reasoning (Yang, 2025) View paper

- Training-Based Test-Time Scaling (3 papers)
- [22] EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test (Li, 2025) View paper
- [23] OpenR: An Open Source Framework for Advanced Reasoning with Large Language Models (Wang Jun, 2024) View paper
- [37] Optimizing Test-Time Compute via Meta Reinforcement Fine-Tuning (Qu Yuxiao, 2025) View paper

- Post-Training Paradigms and Frameworks (3 papers)
- [2] A survey of post-training scaling in large language models (Cheng, 2025) View paper
- [20] Almost surely safe alignment of large language models at inference-time (Xiaotong Ji, 2025) View paper
- [36] Sailing AI by the Stars: A Survey of Learning from Rewards in Post-Training and Test-Time Scaling of Large Language Models (Wu, 2025) View paper

- Domain-Specific Applications
- Medical and Scientific Reasoning (1 papers)
- [5] m1: Unleash the potential of test-time scaling for medical reasoning with large language models (Huang Xiaoke, 2025) View paper
- Multimodal and Vision-Language Models (4 papers)
- [4] InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models (Zhu JinGuo, 2025) View paper
- [27] Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling (Chen Zhe, 2024) View paper
- [38] Test-Time Warmup for Multimodal Large Language Models (Zemel, 2025) View paper
- [44] On the Test-Time Zero-Shot Generalization of Vision-Language Models: Do we Really need Prompt Learning? (Maxime Zanella, 2024) View paper
- Specialized Task Domains (4 papers)
- [6] Reimagining self-adaptation in the age of large language models (Raghav Donakanti, 2024) View paper
- [16] Generating Symbolic World Models via Test-time Scaling of Large Language Models (Yu, 2025) View paper
- [28] Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis (Ye Zhen, 2025) View paper
- [42] Evaluating large language model adaptation strategies for geospatial code generation (Kaiyuan, 2025) View paper

- Computational Efficiency and System Design (5 papers)
- [15] Distributed inference and fine-tuning of large language models over the internet (Borzunov, 2023) View paper
- [25] Tabi: An efficient multi-level inference system for large language models (Yiding Wang, 2023) View paper
- [31] An adaptive compute approach to optimize inference efficiency in large language models (James Lesatod, 2024) View paper
- [39] A review on edge large language models: Design, execution, and applications (Yue Zheng, 2025) View paper
- [43] Balcony: A Lightweight Approach to Dynamic Inference of Generative Language Models (Benyamin Jamialahmadi, 2025) View paper

- Analysis and Evaluation Frameworks (4 papers)
- [12] A Survey on Test-Time Scaling in Large Language Models: What, How, Where, and How Well? (Zhang, 2025) View paper
- [13] Harnessing the Reasoning Economy: A Survey of Efficient Reasoning for Large Language Models (Wang Rui, 2025) View paper
- [18] Inference-time computations for llm reasoning and planning: A benchmark and insights (Khurana, 2025) View paper
- [45] Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling (Liu Run-ze, 2025) View paper

Narrative

Core task: test-time training for large language models. This field explores how models can adapt or improve their performance during inference rather than relying solely on pre-training and fine-tuning. The taxonomy organizes the landscape into seven main branches. Test-Time Adaptation Mechanisms focuses on parameter update approaches and dynamic adjustments that modify model weights or internal states at inference, as seen in works like InPlace TestTime Training[0] and TestTime Learning LLMs[1]. Inference-Time Optimization Strategies emphasizes prompt engineering, search methods, and iterative refinement without parameter changes. Training-Based Test-Time Scaling investigates reinforcement learning and self-improvement loops that leverage test-time compute for reasoning tasks, exemplified by approaches such as Scaling TestTime Compute[14] and OpenR[23]. Post-Training Paradigms and Frameworks examine broader methodologies like continual learning and meta-learning that prepare models for test-time adaptation, while Domain-Specific Applications and Computational Efficiency branches address practical deployment concerns. Analysis and Evaluation Frameworks provide benchmarks and theoretical insights, including surveys like PostTraining Scaling Survey[2] and TestTime Scaling Survey[12].

A particularly active line of work contrasts parameter-updating methods with parameter-free inference strategies. Parameter update approaches, such as InPlace TestTime Training[0] and Medadapter[7], directly modify model weights using test examples or domain-specific data, trading computational cost for potentially stronger adaptation. In contrast, methods like Dynamic Evaluation Online[10] and TestTime Training FewShot[35] explore lighter-weight adjustments or prompt-based interventions that preserve the original model. InPlace TestTime Training[0] sits squarely within the parameter update cluster, emphasizing efficient in-place weight modifications during inference. Compared to TestTime Learning LLMs[1], which may explore broader learning signals, and Medadapter[7], which targets medical domain adaptation, InPlace TestTime Training[0] appears to prioritize computational efficiency and minimal overhead while still enabling meaningful model updates. This positioning reflects ongoing debates about the trade-offs between adaptation strength, inference latency, and resource constraints in real-world deployments.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Test-Time Learning for Large Language Models

Authors: Hu Jinwu, Zhang Zhi-tian, Chen Guohao, Wen Xutao, Shuai Chao, et al. (11 authors total) | **Year/Venue**: 2025 • International Conference on Machine Learning | **URL**: View paper

Abstract

While Large Language Models (LLMs) have exhibited remarkable emergent capabilities through extensive pre-training, they still face critical limitations in generalizing to specialized domains and handling diverse linguistic variations, known as distribution shifts. In this paper, we propose a Test-Time Learning (TTL) paradigm for LLMs, namely TLM, which dynamically adapts LLMs to target domains using only unlabeled test data during testing. Specifically, we first provide empirical evidence and th...

Relationship Analysis

Both papers belong to the Parameter Update Approaches category, focusing on gradient-based optimization of model parameters at test time. They overlap in their core mechanism of adapting LLM weights during inference to handle distribution shifts and long-context scenarios. However, the original paper (In-Place TTT) updates MLP projection matrices using chunk-wise mechanisms with a next-token prediction objective, while the candidate paper (TLM) performs test-time learning by minimizing input perplexity on unlabeled test data using LoRA for lightweight adaptation, representing different optimization targets and parameter selection strategies.

---

2. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning

Authors: Shi, Wenqi, Sun Haotian, Wu Hang, Xu Ran, et al. (10 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Despite their improved capabilities in generation and reasoning, adapting large language models (LLMs) to the biomedical domain remains challenging due to their immense size and privacy concerns. In this study, we propose MedAdapter, a unified post-hoc adapter for test-time adaptation of LLMs towards biomedical applications. Instead of fine-tuning the entire LLM, MedAdapter effectively adapts the original model by fine-tuning only a small BERT-sized adapter to rank candidate solutions generated ...

Relationship Analysis

Both papers belong to the Parameter Update Approaches category, employing gradient-based optimization to update model components at test time. MedAdapter focuses on training a separate lightweight BERT-sized adapter (110M parameters) to rank candidate solutions generated by LLMs for biomedical reasoning tasks, while In-Place TTT directly updates the final projection matrix of MLP blocks within the LLM architecture itself using a chunk-wise mechanism with an NTP-aligned objective. The key distinction is that MedAdapter uses an external verifier model for post-hoc adaptation without modifying the LLM's internal parameters, whereas In-Place TTT performs in-place updates of specific weight matrices within the LLM during inference.

---

3. Revisiting dynamic evaluation: Online adaptation for large language models

Authors: Rannen-Triki, Amal, Bornschein, Jorg, Pascanu, et al. (16 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We consider the problem of online fine tuning the parameters of a language model at test time, also known as dynamic evaluation. While it is generally known that this approach improves the overall predictive performance, especially when considering distributional shift between training and evaluation data, we here emphasize the perspective that online adaptation turns parameters into temporally changing states and provides a form of context-length extension with memory in weights, more in line w...

Relationship Analysis

Both papers belong to the Parameter Update Approaches category, focusing on gradient-based weight updates at test time for language models. They overlap in their core mechanism of adapting model parameters during inference to handle distribution shifts and long contexts. However, the original paper introduces an in-place framework that repurposes existing MLP blocks as fast weights with a novel next-token prediction objective, while the candidate paper revisits dynamic evaluation with emphasis on online SGD adaptation trade-offs, memory efficiency through LoRA, and empirical analysis of compute-performance Pareto fronts across different context sizes and model scales.

---

4. The surprising effectiveness of test-time training for few-shot learning

Authors: AkyÃ¼rek, Ekin, Damani, Mehul, Qiu, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Language models (LMs) have shown impressive performance on tasks within their training distribution, but often struggle with structurally novel tasks even when given a small number of in-context task examples. We investigate the effectiveness of test-time training (TTT) -- temporarily updating model parameters during inference using a loss derived from input data -- as a mechanism for improving LMs' reasoning and few-shot learning capabilities. On the Abstraction and Reasoning Corpus (ARC), perf...

Relationship Analysis

Both papers belong to the Parameter Update Approaches category, employing gradient-based optimization to update model parameters at test time. The original paper (In-Place TTT) focuses on continual adaptation for long-context language modeling by updating MLP projection matrices in-place during inference with a next-token prediction objective, while the candidate paper investigates test-time training for few-shot learning on novel reasoning tasks (ARC, BBH) by temporarily updating LoRA adapters using in-context examples. The key distinction is that the original targets streaming context adaptation in autoregressive LMs, whereas the candidate addresses transductive learning for structurally novel tasks with limited demonstrations.

---

5. Evaluating Test-Time Training for Conceptual Reasoning in Large Language Models

Authors: F Derksen | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Many high-performing ARC solvers use test-time training (TTT): they add Low-Rank Adaptation (LoRA) â¦ These results show that a data-eï¬cient TTT step can more than quadruple LLM â¦

Relationship Analysis

Both papers belong to the Parameter Update Approaches category, employing gradient-based optimization to update model weights at test time. The original paper introduces In-Place TTT, which repurposes MLP blocks' final projection matrices as fast weights and uses a Next-Token Prediction-aligned objective for language modeling tasks with contexts up to 128k tokens. The candidate paper evaluates test-time training with LoRA adapters for conceptual reasoning on the ConceptARC benchmark, focusing on abstract reasoning tasks

rather than long-context language modeling, and demonstrates that TTT can quadruple performance on aggregated reasoning while doubling per-concept accuracy.

Contributions Analysis

Overall novelty summary. The paper introduces In-Place Test-Time Training, a framework enabling LLMs to update parameters during inference by treating MLP projection matrices as adaptable fast weights. It resides in the Parameter Update Approaches leaf, which contains six papers including the original work. This leaf sits within Test-Time Adaptation Mechanisms, one of seven major branches in a taxonomy spanning fifty papers across fourteen leaf nodes. The Parameter Update Approaches direction represents a moderately populated research area, focusing on gradient-based weight modifications at test time rather than activation manipulation or external model integration.

The taxonomy reveals neighboring leaves such as Activation-Based Intervention and Auxiliary Model Integration, both under Test-Time Adaptation Mechanisms. Activation-Based Intervention manipulates internal states without weight updates, while Auxiliary Model Integration employs separate lightweight models to guide inference. The paper's approach diverges from these by directly modifying base model parameters in-place. Sibling papers in the same leaf include works like TestTime Learning LLMs and Medadapter, which also update weights but may differ in architectural targets or domain focus. The broader Inference-Time Optimization Strategies branch explores parameter-free methods like prompt engineering and search algorithms, highlighting a fundamental methodological split in the field.

Among twenty-one candidates examined through semantic search and citation expansion, the three contributions show no clear refutation. The In-Place TTT framework examined nine candidates with zero refutable matches, suggesting limited direct overlap in the specific architectural approach of using MLP projection matrices as fast weights. The LM-aligned objective contribution examined ten candidates without refutation, indicating the tailored next-token-prediction objective may represent a novel formulation within the limited search scope. The chunk-wise update mechanism examined only two candidates, reflecting either a sparse research direction or narrow search coverage for this efficiency-focused component.

Based on the limited search scope of twenty-one candidates, the work appears to occupy a distinct position within parameter update approaches, particularly in its architectural integration strategy and objective design. However, the analysis does not cover exhaustive prior work in test-time adaptation or continual learning, and the moderate size of the Parameter Update Approaches leaf suggests active but not overcrowded research activity. The absence of refutable candidates may reflect genuine novelty or limitations in semantic search coverage for this specific combination of techniques.

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

Contribution 1: In-Place Test-Time Training framework for LLMs

Description: The authors propose a framework that enables Large Language Models to dynamically update their weights at inference time by repurposing existing MLP blocks as adaptable fast weights. This drop-in enhancement requires no architectural modifications or costly retraining from scratch, addressing the architectural incompatibility barrier of previous TTT methods.

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. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

URL: View paper

Brief Assessment

Scaling TestTime Compute[14] focuses on optimizing inference-time computation through search and adaptive distribution updates, not on dynamic weight updates or architectural modifications to LLMs during test time.

---

2. Steering language models with activation engineering

URL: View paper

Brief Assessment

Activation Engineering[69] focuses on inference-time modification of activations through steering vectors to control model outputs (sentiment, toxicity), not on dynamic weight updates or test-time training mechanisms for continual learning.

---

3. Test-Time Learning for Large Language Models

URL: View paper

Brief Assessment

TestTime Learning LLMs[1] focuses on test-time adaptation through input perplexity minimization for domain shifts, using LoRA updates. The original paper's in-place TTT repurposes MLP blocks as fast weights with chunk-wise updates for dynamic context adaptation—fundamentally different mechanisms and objectives.

---

4. Sensitivity-lora: Low-load sensitivity-based fine-tuning for large language models

URL: View paper

Brief Assessment

SensitivityLoRA[67] focuses on parameter-efficient fine-tuning through low-rank adaptation with sensitivity-based rank allocation, not on test-time training with dynamic weight updates during inference.

---

5. Test-Time Training Done Right

URL: View paper

Brief Assessment

TestTime Training Right[65] focuses on large chunk updates (2k-1m tokens) across diverse modalities including images and videos, while the original paper specifically addresses architectural compatibility for LLMs by repurposing MLP blocks as fast weights with a language modeling-aligned objective.

---

6. Efficient test-time adaptation of vision-language models

URL: View paper

Brief Assessment

Efficient TestTime VisionLanguage[63] focuses on test-time adaptation for vision-language models (CLIP) using a training-free dynamic cache for image classification, not on test-time training for large language models with dynamic weight updates in MLP blocks.

---

7. Dual prototype evolving for test-time generalization of vision-language models

URL: View paper

Brief Assessment

Dual Prototype Evolving[70] focuses on test-time adaptation for vision-language models (VLMs) like CLIP, not large language models. The candidate addresses visual-textual prototype evolution for image classification tasks, while the original proposes dynamic weight updates for autoregressive language modeling in LLMs.

---

8. Grounded Test-Time Adaptation for LLM Agents

URL: View paper

Brief Assessment

Grounded TestTime Agents[66] focuses on test-time adaptation for LLM agents in interactive environments (web navigation, function calling) through distributional adaptation and dynamics grounding, not on architectural modifications enabling dynamic weight updates during inference for language modeling tasks.

---

9. Towards Stable Test-Time Adaptation in Dynamic Wild World

URL: View paper

Brief Assessment

Stable TestTime Adaptation[64] focuses on test-time adaptation for image classification under distribution shifts, using batch normalization analysis and entropy minimization. It does not address test-time training for large language models or dynamic weight updates in the context of autoregressive language modeling.

Contribution 2: LM-aligned objective for TTT

Description: The authors introduce a novel learning objective that aligns with the Next-Token Prediction goal of language models, replacing the generic reconstruction targets used in prior TTT work. This objective is designed to encourage fast weights to store predictively useful information for autoregressive language modeling.

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. Autotimes: Autoregressive time series forecasters via large language models

URL: View paper

Brief Assessment

Autotimes[56] focuses on adapting LLMs for time series forecasting through autoregressive token prediction, not on test-time training objectives for language models. The paper addresses time series modality alignment rather than TTT fast weight objectives.

---

2. Out-of-Distribution Detection and Selective Generation for Conditional Language Models

URL: View paper

Brief Assessment

OOD Selective Generation[58] focuses on out-of-distribution detection and quality estimation for conditional language models using embedding-based methods, not on test-time training objectives for autoregressive language modeling.

---

3. PARM: Multi-Objective Test-Time Alignment via Preference-Aware Autoregressive Reward Model

URL: View paper

Brief Assessment

PARM[60] focuses on multi-objective test-time alignment using autoregressive reward models for preference-based generation, not on test-time training objectives aligned with next-token prediction for language model adaptation.

---

4. Long-context autoregressive video modeling with next-frame prediction

URL: View paper

Brief Assessment

LongContext Video Modeling[52] focuses on video generation using frame-level autoregressive modeling with flow matching objectives, not language model next-token prediction objectives for test-time training.

---

5. NEP: Autoregressive Image Editing via Next Editing Token Prediction

URL: View paper

Brief Assessment

NEP[57] focuses on autoregressive image editing via next editing-token prediction for visual content, not on test-time training objectives for language models. The domains and technical approaches are fundamentally different.

---

6. ICRT: In-Context Imitation Learning via Next-Token Prediction

URL: View paper

Brief Assessment

ICRT[54] focuses on in-context imitation learning for robotics using autoregressive prediction on sensorimotor trajectories. It does not address test-time training objectives for language models or next-token prediction alignment.

---

7. In-context imitation learning via next-token prediction

URL: View paper

Brief Assessment

InContext Imitation NextToken[53] focuses on in-context imitation learning for robotics using sensorimotor trajectories, not on test-time training objectives for language models. The candidate addresses robot control via next-token prediction on robot actions, while the original contribution concerns aligning TTT objectives with autoregressive language modeling goals.

---

8. Generative Verifiers: Reward Modeling as Next-Token Prediction

URL: View paper

Brief Assessment

Generative Verifiers[51] focuses on training verifiers using next-token prediction for solution verification and ranking, not on test-time training objectives for dynamic weight adaptation in language models.

---

9. Go with Your Gut: Scaling Confidence for Autoregressive Image Generation

URL: View paper

Brief Assessment

Confidence Scaling Autoregressive[55] focuses on test-time scaling for next-token prediction in autoregressive image generation using token entropy signals, not on designing learning objectives for test-time training in language models.

---

10. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval

URL: View paper

Brief Assessment

Tranception[59] focuses on protein fitness prediction using autoregressive transformers with inference-time retrieval of homologous sequences, not on test-time training objectives for general language models. The paper does not address TTT learning objectives aligned with next-token prediction in the context of language modeling.

Contribution 3: Efficient chunk-wise update mechanism with context parallelism

Description: The authors develop an efficient chunk-wise update strategy that leverages parallel scan algorithms to enable context parallelism while maintaining strict causal semantics. This design addresses the computational inefficiency of per-token updates in previous TTT methods and enables high throughput on modern accelerators.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Exploring Diffusion with Test-Time Training on Efficient Image Restoration

URL: View paper

Brief Assessment

Diffusion TestTime Restoration[62] focuses on image restoration tasks using diffusion models with chunk-optimized flash processing for 2D spatial data, not on language modeling with causal chunk-wise updates for sequential token processing as in the original paper.

---

2. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing

URL: View paper

Brief Assessment

Diffusion Faster Inference[61] focuses on discrete diffusion models for text generation with block-wise autoregressive decoding, not on test-time training mechanisms for LLMs. The technical approaches are fundamentally different.

Appendix: Text Similarity Detection

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

References

- [0] In-Place Test-Time Training View paper
- [1] Test-Time Learning for Large Language Models View paper
- [2] A survey of post-training scaling in large language models View paper
- [3] Inference-time intervention: Eliciting truthful answers from a language model View paper
- [4] InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models View paper
- [5] m1: Unleash the potential of test-time scaling for medical reasoning with large language models View paper
- [6] Reimagining self-adaptation in the age of large language models View paper
- [7] Medadapter: Efficient test-time adaptation of large language models towards medical reasoning View paper
- [8] Step-level Verifier-guided Hybrid Test-Time Scaling for Large Language Models View paper
- [9] Boosted Prompt Ensembles for Large Language Models View paper
- [10] Revisiting dynamic evaluation: Online adaptation for large language models View paper
- [11] Automatic Prompt Selection for Large Language Models View paper
- [12] A Survey on Test-Time Scaling in Large Language Models: What, How, Where, and How Well? View paper
- [13] Harnessing the Reasoning Economy: A Survey of Efficient Reasoning for Large Language Models View paper
- [14] Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters View paper
- [15] Distributed inference and fine-tuning of large language models over the internet View paper
- [16] Generating Symbolic World Models via Test-time Scaling of Large Language Models View paper
- [17] First Finish Search: Efficient Test-Time Scaling in Large Language Models View paper
- [18] Inference-time computations for llm reasoning and planning: A benchmark and insights View paper
- [19] Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning View paper
- [20] Almost surely safe alignment of large language models at inference-time View paper
- [21] ControlMLLM: Training-Free Visual Prompt Learning for Multimodal Large Language Models View paper
- [22] EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test View paper
- [23] OpenR: An Open Source Framework for Advanced Reasoning with Large Language Models View paper
- [24] Ttrl: Test-time reinforcement learning View paper
- [25] Tabi: An efficient multi-level inference system for large language models View paper
- [26] Inference-time alignment in continuous space View paper
- [27] Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling View paper
- [28] Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis View paper
- [29] Enhancing Test-Time Scaling of Large Language Models with Hierarchical Retrieval-Augmented MCTS View paper
- [30] Efficient Uncertainty Estimation via Distillation of Bayesian Large Language Models View paper
- [31] An adaptive compute approach to optimize inference efficiency in large language models View paper
- [32] Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning View paper
- [33] Test-time Backdoor Mitigation for Black-Box Large Language Models with Defensive Demonstrations View paper
- [34] Patience Is The Key to Large Language Model Reasoning View paper
- [35] The surprising effectiveness of test-time training for few-shot learning View paper
- [36] Sailing AI by the Stars: A Survey of Learning from Rewards in Post-Training and Test-Time Scaling of Large Language Models View paper
- [37] Optimizing Test-Time Compute via Meta Reinforcement Fine-Tuning View paper
- [38] Test-Time Warmup for Multimodal Large Language Models View paper
- [39] A review on edge large language models: Design, execution, and applications View paper

- [40] Test-time Prompt Intervention View paper
- [41] A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods View paper
- [42] Evaluating large language model adaptation strategies for geospatial code generation View paper
- [43] Balcony: A Lightweight Approach to Dynamic Inference of Generative Language Models View paper
- [44] On the Test-Time Zero-Shot Generalization of Vision-Language Models: Do we Really need Prompt Learning? View paper
- [45] Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling View paper
- [46] Scaling Test-Time Compute Without Verification or RL is Suboptimal View paper
- [47] Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models View paper
- [48] Towards thinking-optimal scaling of test-time compute for llm reasoning View paper
- [49] Bridging the language gaps in large language models with inference-time cross-lingual intervention View paper
- [50] Evaluating Test-Time Training for Conceptual Reasoning in Large Language Models View paper
- [51] Generative Verifiers: Reward Modeling as Next-Token Prediction View paper
- [52] Long-context autoregressive video modeling with next-frame prediction View paper
- [53] In-context imitation learning via next-token prediction View paper
- [54] ICRT: In-Context Imitation Learning via Next-Token Prediction View paper
- [55] Go with Your Gut: Scaling Confidence for Autoregressive Image Generation View paper
- [56] Autotimes: Autoregressive time series forecasters via large language models View paper
- [57] NEP: Autoregressive Image Editing via Next Editing Token Prediction View paper
- [58] Out-of-Distribution Detection and Selective Generation for Conditional Language Models View paper
- [59] Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval View paper
- [60] PARM: Multi-Objective Test-Time Alignment via Preference-Aware Autoregressive Reward Model View paper
- [61] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing View paper
- [62] Exploring Diffusion with Test-Time Training on Efficient Image Restoration View paper
- [63] Efficient test-time adaptation of vision-language models View paper
- [64] Towards Stable Test-Time Adaptation in Dynamic Wild World View paper
- [65] Test-Time Training Done Right View paper
- [66] Grounded Test-Time Adaptation for LLM Agents View paper
- [67] Sensitivity-lora: Low-load sensitivity-based fine-tuning for large language models View paper
- [68] Depth-Aware Test-Time Training for Zero-Shot Video Object Segmentation View paper
- [69] Steering language models with activation engineering View paper
- [70] Dual prototype evolving for test-time generalization of vision-language models View paper