Novelty Assessment Report

Paper: Scaling Laws Meet Model Architecture: Toward Inference-Efficient LLMs

PDF URL: https://openreview.net/pdf?id=0TmVqOpBbK

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Scaling the number of parameters and the size of training data has proven to be an effective strategy for improving large language model (LLM) performance. Yet, as these models grow increasingly powerful and widely deployed, the cost of inference has become a pressing concern. Despite its importance, the trade-off between model accuracy and inference efficiency remains underexplored. In this work, we examine how key architectural factors, hidden size, the allocation of parameters between MLP and attention (mlp-to-attention ratio), and grouped-query attention (GQA), influence both inference cost and accuracy. We introduce a conditional scaling law that augments the Chinchilla framework with architectural information, along with a search framework for identifying architectures that are simultaneously inference-efficient and accurate. To validate our approach, we train more than 200 models spanning 80M to 3B parameters and 8B to 100B training tokens, and fit the proposed conditional scaling law. Our results show that the conditional scaling law reliably predicts optimal architectural choices and that the resulting models outperform existing open-source baselines. Under the same training budget, optimized architectures achieve up to 2.1\% higher accuracy and 42\% greater inference throughput compared to LLaMA-3.2.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Architecture-Aware Scaling Laws for Inference-Efficient Language Models**

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:

- **Inference-Time Compute Scaling**
- **Model Architecture Design for Efficiency**
- **Scaling Laws and Predictive Modeling**
- **Training-Time Scaling and Optimization**
- **Quantization and Low-Bit Models**
- **System Infrastructure and Deployment**
- **Multimodal and Domain-Specific Efficiency**
- **Empirical Analysis and Benchmarking**

Complete Taxonomy Tree

- Architecture-Aware Scaling Laws for Inference-Efficient Language Models Survey Taxonomy
- Inference-Time Compute Scaling
  - Search-Based Inference Scaling (5 papers)
  - [1] Scaling llm test-time compute optimally can be more effective than scaling model parameters (Snell, 2024) View paper
  - [3] Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning (CV Snell, 2025) View paper
  - [7] Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving (Y Wu, 2025) View paper
  - [15] A*-Decoding: Token-Efficient Inference Scaling (Chatziveroglou, 2025) View paper
  - [25] ETS: Efficient Tree Search for Inference-Time Scaling (Hooper, 2025) View paper
  - Sampling-Based Inference Scaling (2 papers)
  - [21] An empirical analysis of compute-optimal inference for problem-solving with language models (Yangzhen Wu, 2024) View paper
  - [34] Rollout Roulette: A Probabilistic Inference Approach to Inference-Time Scaling of LLMs using Particle-Based Monte Carlo Methods (Puri, 2025) View paper
  - Multi-Agent and Collaborative Inference (1 papers)
  - [29] Multi-Agent Sampling: Scaling Inference Compute for Data Synthesis with Tree Search-Based Agentic Collaboration (Ye Hai, 2024) View paper
  - Inference Scaling for Specialized Tasks (3 papers)
  - [9] Inference-time scaling for complex tasks: Where we stand and what lies ahead (Balachandran, 2025) View paper
  - [42] Inference Scaling for Long-Context Retrieval Augmented Generation (Yue, 2024) View paper
  - [43] Inference-Time Scaling for Generalist Reward Modeling (Liu, 2025) View paper
  - Reinforcement Learning for Inference Scaling (2 papers)
  - [24] Advancing language model reasoning through reinforcement learning and inference scaling (Hou Zhenyu, 2025) View paper
  - [30] Think Deep, Think Fast: Investigating Efficiency of Verifier-free Inference-time-scaling Methods (Wang Junlin, 2025) View paper
- Model Architecture Design for Efficiency
  - Mixture-of-Experts Architectures (6 papers)
  - [2] Towards greater leverage: Scaling laws for efficient mixture-of-experts language models (Tian, 2025) View paper

    - [6] Glam: Efficient scaling of language models with mixture-of-experts (Du, 2022) View paper
    - [16] Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing (Julian BÃ¼chel, 2025) View paper
    - [18] Toward Inference-optimal Mixture-of-Expert Large Language Models (Yun, 2024) View paper
    - [28] Uni-MoE: Scaling Unified Multimodal LLMs With Mixture of Experts (Yunxin Li, 2024) View paper
    - [40] Joint MoE Scaling Laws: Mixture of Experts Can Be Memory Efficient (Pioro, 2025) View paper
    - Hybrid and Alternative Architectures (3 papers)
    - [20] Llamba: Scaling Distilled Recurrent Models for Efficient Language Processing (Aviv Bick, 2025) View paper
    - [39] Hybrid Architectures for Language Models: Systematic Analysis and Design Insights (Bae, 2025) View paper
    - [41] Mechanistic Design and Scaling of Hybrid Architectures (Poli, 2024) View paper
    - Attention Mechanism Optimization (1 papers)
    - [38] Mixture of Attention Spans: Optimizing LLM Inference Efficiency with Heterogeneous Sliding-Window Lengths (Fu Tian-Yu, 2024) View paper
    - Activation Sparsity and Pruning (2 papers)
    - [48] NeuroPrune: A Neuro-inspired Topological Sparse Training Algorithm for Large Language Models (Dhurandhar, 2024) View paper
    - [50] Q-Sparse: All Large Language Models can be Fully Sparsely-Activated (Wang Hongyu, 2024) View paper
- Scaling Laws and Predictive Modeling
    - Architecture-Conditional Scaling Laws ★ (2 papers)
    - [0] Scaling Laws Meet Model Architecture: Toward Inference-Efficient LLMs (Anon et al., 2026) View paper
    - [8] Scaling Inference-Efficient Language Models (Bian, 2025) View paper
    - Training Compute-Optimal Scaling Laws (2 papers)
    - [10] Scaling laws for predicting downstream performance in llms (Chen Yangyi, 2024) View paper
    - [13] Beyond chinchilla-optimal: Accounting for inference in language model scaling laws (Sardana, 2023) View paper
    - Inference-Aware Scaling Laws (2 papers)
    - [31] Inference economics of language models (Erdil, 2025) View paper
    - [32] A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search (Nayak, 2025) View paper
    - Parallel and Alternative Scaling Paradigms (1 papers)
    - [26] Parallel scaling law for language models (Hui, 2025) View paper
- Training-Time Scaling and Optimization
    - Efficient Training Strategies (1 papers)
    - [14] SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer (Xie, 2025) View paper
    - Parameter-Efficient Fine-Tuning (1 papers)
    - [37] A Semantic-Aware Layer-Freezing Approach to Computation-Efficient Fine-Tuning of Language Models (Aleti, 2024) View paper
    - Model Compression and Pruning (1 papers)
    - [12] Training Language Models to Reason Efficiently (Arora, 2025) View paper
- Quantization and Low-Bit Models (2 papers)
    - [35] Spectra 1.1: Scaling Laws and Efficient Inference for Ternary Language Models (Vaidhya, 2025) View paper
    - [49] Scaling Laws and Efficient Inference for Ternary Language Models (Jain Vineet, 2025) View paper
- System Infrastructure and Deployment
    - Distributed Inference Systems (3 papers)
    - [4] Efficiently scaling transformer inference (Pope, 2023) View paper
    - [45] AIBrix: Towards Scalable, Cost-Effective Large Language Model Inference Infrastructure (Gupta, 2025) View paper
    - [46] UELLM: A Unified and Efficient Approach for Large Language Model Inference Serving (Yiyuan He, 2024) View paper
    - Hardware Co-Design and Accelerators (3 papers)
    - [22] Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures (Zhao Cheng-gang, 2025) View paper
    - [23] Memory Is All You Need: An Overview of Compute-in-Memory Architectures for Accelerating Large Language Model Inference (Wolters, 2024) View paper
    - [33] System-performance and cost modeling of Large Language Model training and inference (Guo WenZhe, 2025) View paper
- Multimodal and Domain-Specific Efficiency (2 papers)
    - [5] Inference Compute-Optimal Video Vision Language Models (Huang, 2025) View paper
    - [19] LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via a Hybrid Architecture (Xidong Wang, 2024) View paper
- Empirical Analysis and Benchmarking (6 papers)
    - [11] Efficiency in language understanding and generation: An evaluation of four open-source large language models (Siu Ming Wong, 2024) View paper
    - [17] Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings (Ilager, 2025) View paper
    - [27] Sprout: Green Generative AI with Carbon-Efficient LLM Inference (Baolin Li, 2024) View paper
    - [36] How to Upscale Neural Networks with Scaling Law? A Survey and Practical Guidelines (Sengupta Ayan, 2025) View paper
    - [44] EfficientLLM: Efficiency in Large Language Models (Yuan, 2025) View paper
    - [47] The Efficiency Spectrum of Large Language Models: An Algorithmic Survey (DING Tianyu, 2023) View paper

Narrative

Core task: architecture-aware scaling laws for inference-efficient language models. The field has evolved to address the dual challenge of predicting model performance while accounting for deployment costs and architectural choices. The taxonomy reveals eight major branches spanning the full lifecycle from design to deployment. Scaling Laws and Predictive Modeling focuses on mathematical frameworks that relate model size, data, and compute to performance, with recent extensions to architecture-conditional settings and inference budgets (Scaling Laws Architecture[0], Inference Scaling Laws[7]). Model Architecture Design for Efficiency explores structural innovations like mixture-of-experts and hybrid architectures (MoE Scaling Laws[2], Hybrid Architecture Analysis[39]), while Inference-Time Compute Scaling examines how test-time computation can be traded for accuracy (Test-Time Compute Scaling[1], Test-Time Reasoning Scaling[3]). Training-Time Scaling, Quantization, and System Infrastructure branches address optimization strategies, low-bit representations (Ternary Scaling Laws[49]), and practical deployment concerns (Cost Modeling LLMs[33]), with additional branches covering multimodal extensions and empirical benchmarking (LLM Efficiency Evaluation[11]).

A central tension emerges between predictive accuracy and practical deployment constraints. Works like Inference-Efficient Models[8] and Beyond Chinchilla[13] challenge traditional compute-optimal training by incorporating inference costs into scaling decisions, while Inference Economics[31] explicitly models the economic trade-offs. Scaling Laws Architecture[0] sits within the Architecture-Conditional Scaling Laws cluster, emphasizing how different architectural families—dense transformers, MoE variants, or hybrid designs—exhibit distinct scaling behaviors that must be captured for accurate performance prediction under inference budgets. This contrasts with earlier work that treated architecture as fixed, and complements neighboring efforts like Inference-Efficient Models[8] which focus more on empirical comparisons across architectures. The original paper's emphasis on architecture-aware prediction bridges the gap between theoretical scaling frameworks and the practical reality that deployment efficiency depends critically on structural choices, a theme echoed across multiple branches but rarely integrated into unified predictive models.

Related Works in Same Category

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

1. Scaling Inference-Efficient Language Models

Authors: Bian, Song, Yan Minghao, Venkataraman, Shivaram | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Scaling laws are powerful tools to predict the performance of large language models. However, current scaling laws fall short of accounting for inference costs. In this work, we first show that model architecture affects inference latency, where models of the same size can have up to 3.5x difference in latency. To tackle this challenge, we modify the Chinchilla scaling laws to co-optimize the model parameter count, the number of training tokens, and the model architecture. Due to the reason that...

Relationship Analysis

Both papers belong to the Architecture-Conditional Scaling Laws category, incorporating architectural parameters into scaling law formulations to predict model performance. They overlap in addressing the trade-off between inference efficiency and accuracy by extending traditional scaling laws (Chinchilla) with architectural factors like hidden size and layer configuration. The key difference is that the original paper focuses on a broader set of architectural factors (hidden size, MLP-to-attention ratio, and GQA) with a multiplicative/additive calibration approach, while the candidate paper emphasizes the aspect ratio (dmodel/nlayers) with a multiplicative term (1 + εRγ) and includes a novel model selection methodology for ranking candidates based on predicted loss.

Contributions Analysis

Overall novelty summary. The paper proposes a conditional scaling law that extends the Chinchilla framework by incorporating architectural parameters—hidden size, MLP-to-attention ratio, and grouped-query attention—to predict both accuracy and inference cost. It resides in the Architecture-Conditional Scaling Laws leaf, which contains only two papers including this one. This represents a relatively sparse research direction within the broader Scaling Laws and Predictive Modeling branch, suggesting the integration of architectural factors into scaling law frameworks remains an emerging area despite the maturity of compute-optimal scaling research.

The taxonomy reveals neighboring leaves focused on Training Compute-Optimal Scaling Laws (architecture-agnostic Chinchilla-style analyses) and Inference-Aware Scaling Laws (optimizing for deployment costs without explicit architectural conditioning). The paper bridges these directions by making architectural choices explicit predictors of performance under inference constraints. Nearby branches like Model Architecture Design for Efficiency explore structural innovations (MoE, hybrid models, attention mechanisms) but typically lack unified predictive frameworks. The scope_note for Architecture-Conditional Scaling Laws explicitly excludes architecture-agnostic approaches, positioning this work as addressing a gap between theoretical scaling and practical architectural diversity.

Among 20 candidates examined across three contributions, the conditional scaling law itself shows one refutable candidate among 10 examined, indicating some prior work on architecture-aware prediction exists within the limited search scope. The search framework for identifying inference-efficient architectures found no refutable candidates among 10 examined, suggesting this systematic optimization approach may be less explored. The characterization of architectural factors' impact was not examined against candidates. These statistics reflect a targeted literature search rather than exhaustive coverage, and the single refutable pair for the core contribution suggests the specific formulation may overlap with existing architecture-conditional frameworks.

Based on the limited search scope of 20 candidates, the work appears to occupy a moderately novel position by unifying architectural conditioning with inference-aware optimization in a single predictive framework. The sparse population of its taxonomy leaf and the absence of refutable candidates for the search framework component suggest potential novelty in the systematic approach, though the core scaling law formulation shows some overlap with prior architecture-aware prediction efforts. The analysis does not cover exhaustive comparison with all Chinchilla extensions or architecture search methods beyond the top-K semantic matches examined.

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

Contribution 1: Conditional scaling law augmenting Chinchilla with architectural factors

Description: The authors propose a conditional extension of the Chinchilla scaling laws that incorporates architectural parameters such as hidden size, mlp-to-attention ratio, and grouped-query attention. This framework enables predicting model performance while accounting for architectural design choices.

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. Physics of language models: Part 3.3, knowledge capacity scaling laws

URL: View paper

Brief Assessment

Knowledge Capacity Scaling[52] focuses on knowledge storage capacity (bits per parameter) in language models, not on predicting model performance via architectural parameters as the original paper does. The candidate examines how models store factual knowledge tuples, while the original work predicts training loss and optimal architectures for inference efficiency.

2. CodeGen2: Lessons for Training LLMs on Programming and Natural Languages

URL: View paper

Brief Assessment

CodeGen2[53] focuses on training efficiency for code generation models through unified learning objectives and data mixtures, not on scaling laws incorporating architectural parameters for performance prediction.

3. AI and Memory Wall

URL: View paper

Brief Assessment

Memory Wall[55] focuses on memory bandwidth bottlenecks in AI systems and argues for architectural redesign to overcome memory limitations, not on scaling laws that predict model performance based on architectural parameters like hidden size or mlp-to-attention ratio.

4. Slamming: Training a Speech Language Model on One GPU in a Day

URL: View paper

Brief Assessment

Slamming Speech[54] focuses on training speech language models (SLMs) with emphasis on model initialization, synthetic data, and preference optimization. It does not propose scaling laws that incorporate architectural parameters for performance prediction.

5. Scaling Inference-Efficient Language Models

URL: View paper

Prior Art Analysis

Inference-Efficient Models[8] demonstrates prior work that extends Chinchilla scaling laws by incorporating architectural parameters. The candidate paper explicitly proposes 'inference-efficient scaling laws' that augment the Chinchilla framework with architectural factors including hidden size (dmodel), number of layers (nlayers), and aspect ratio (dmodel/nlayers). Both papers modify the Chinchilla loss function to include architectural terms, with the candidate using a multiplicative factor (1 + εrγ) where r represents aspect ratio. This shows that incorporating architectural parameters into Chinchilla-style scaling laws was already explored in Inference-Efficient Models[8], challenging the novelty claim that the original paper was first to propose such conditional extensions.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly state they are augmenting/extending the Chinchilla framework with additional constraints and architectural considerations. - **Original**: we introduce a conditional scaling law that augments the chinchilla framework with architectural information - **Candidate**: we propose rewriting eq. (1) as below to meet practical requirements: arg min n,d l(n, d) s.t. n ≤ nc, d≤ dc, tinf ≤ tc

Evidence 2 - **Rationale**: Both papers present frameworks that augment Chinchilla with architectural considerations to identify inference-efficient architectures, demonstrating that this approach existed prior to the original paper's submission. - **Original**: a conditional scaling law that augments the chinchilla framework with architectural information, along with a search framework for identifying architectures that are simultaneously inference-efficient and accurate - **Candidate**: in this work, we aim to address the following question: given dataset and parameter constraints, can we train an inference-efficient and accurate model for downstream tasks? we first show that the number of parameters is not the exclusive factor affecting inference efficiency. as illustrated in figu...

6. Farseer: A Refined Scaling Law in Large Language Models

URL: View paper

Brief Assessment

Farseer Scaling[57] focuses on refining the Chinchilla scaling law by modeling the loss surface L(N,D) with improved mathematical formulations for data-model interactions, but does not incorporate architectural parameters like hidden size, MLP-to-attention ratio, or grouped-query attention into its scaling law framework.

7. Observational scaling laws and the predictability of langauge model performance

URL: View paper

Brief Assessment

Observational Scaling Laws[51] focuses on predicting performance across multiple model families using observable benchmark metrics rather than architectural parameters. The candidate does not address architectural factors like hidden size, mlp-to-attention ratio, or grouped-query attention in scaling laws.

8. Collaborative performance prediction for large language models

URL: View paper

Brief Assessment

Collaborative Performance Prediction[59] focuses on predicting LLM performance across downstream tasks using collaborative filtering methods and historical performance data, not on extending scaling laws with architectural parameters like hidden size or MLP-to-attention ratio for pre-training loss prediction.

9. Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling?

URL: View paper

Brief Assessment

Inductive Bias Scaling[58] studies how different model architectures (e.g., Transformers, Performers, MLP-Mixers) scale differently across compute regions, but does not propose conditional extensions of Chinchilla scaling laws incorporating specific architectural parameters like hidden size, mlp-to-attention ratio, or grouped-query attention for performance prediction.

10. Scaling laws with vocabulary: Larger models deserve larger vocabularies

URL: View paper

Brief Assessment

Vocabulary Scaling Laws[56] focuses on vocabulary size as an architectural parameter for scaling laws, not the architectural factors (hidden size, mlp-to-attention ratio, grouped-query attention) studied in the original paper. The candidate extends Chinchilla laws with vocabulary considerations, while the original extends them with model architecture parameters like hidden size and attention mechanisms.

Contribution 2: Search framework for inference-efficient and accurate architectures

Description: The authors develop a systematic framework (Algorithm 1) that uses the conditional scaling law to identify model architectures optimizing both inference efficiency and accuracy under fixed parameter and token budgets, including a local search procedure for grouped-query attention.

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. Neural architecture search as multiobjective optimization benchmarks: Problem formulation and performance assessment

URL: View paper

Brief Assessment

Multiobjective NAS Benchmarks[67] focuses on formulating NAS as multiobjective optimization problems and creating benchmark test suites for evaluating EMO algorithms. It does not present a search framework that uses conditional scaling laws to identify optimal architectures under fixed parameter and token budgets, which is the core novelty of the original paper's contribution.

2. Faststereonet: A fast neural architecture search for improving the inference of disparity estimation on resource-limited platforms

URL: View paper

Brief Assessment

FastStereoNet[69] focuses on neural architecture search for disparity estimation in computer vision, using LAHC and simulated annealing. The original paper addresses LLM architecture optimization under parameter/token budgets with conditional scaling laws, which is a fundamentally different domain and methodology.

3. MARCO: Hardware-Aware Neural Architecture Search for Edge Devices with Multi-Agent Reinforcement Learning and Conformal Prediction Filtering

URL: View paper

Brief Assessment

MARCO Edge NAS[61] focuses on hardware-aware neural architecture search for edge devices using multi-agent RL and conformal prediction filtering, targeting resource-constrained deployment. The original paper addresses scaling laws for LLM architectures under parameter/token budgets, optimizing hidden size, MLP-to-attention ratio, and GQA for inference efficiency in large language models—a fundamentally different problem domain and methodology.

4. Automated UAV Object Detector Design Using Large Language Model-Guided Architecture Search

URL: View paper

Brief Assessment

UAV Detector LLM[66] focuses on automated object detector design for UAV applications using LLM-guided architecture search, not on general LLM architecture optimization under parameter and token budgets as in the original paper.

5. Efficient neural architecture search via parameters sharing

URL: View paper

Brief Assessment

Parameter Sharing NAS[60] focuses on efficient architecture search through parameter sharing across child models during search, not on optimizing inference efficiency under fixed parameter/token budgets. The original paper's framework explicitly targets inference throughput and accuracy tradeoffs using conditional scaling laws, which is a different objective.

6. Sparse: Sparse architecture search for cnns on resource-constrained microcontrollers

URL: View paper

Brief Assessment

Sparse Microcontroller[63] focuses on CNN architecture search for resource-constrained microcontrollers with RAM/ROM constraints, not LLM architecture search under parameter/token budgets. The technical domains and optimization objectives differ fundamentally.

7. Semi-supervised neural architecture search

URL: View paper

Brief Assessment

Semi-Supervised NAS[65] focuses on reducing the number of architecture evaluations needed during search by using semi-supervised learning to train an accuracy predictor. The original paper's framework optimizes architectural choices (hidden size, MLP-to-attention ratio, GQA) under fixed parameter and token budgets to balance inference efficiency and accuracy. These are fundamentally different objectives and methodologies.

8. Hao: Hardware-aware neural architecture optimization for efficient inference

URL: View paper

Brief Assessment

HAO Architecture Optimization[68] focuses on hardware-aware neural architecture search for FPGA deployment with integer programming, not on scaling laws for LLM architectures under parameter/token budgets.

9. Mixed precision neural architecture search for energy efficient deep learning

URL: View paper

Brief Assessment

Mixed Precision NAS[64] focuses on joint optimization of neural architecture and quantization (bit-widths) for energy efficiency in edge devices, while the original paper develops a conditional scaling law framework for LLM architectures optimizing inference throughput and accuracy under parameter/token budgets. These are fundamentally different problem domains and methodologies.

10. Neural architecture search for resource constrained hardware devices: A survey

URL: View paper

Brief Assessment

Resource-Constrained NAS Survey[62] is a survey paper reviewing existing NAS methods for hardware devices, not proposing a novel search framework. The original paper develops a specific systematic framework (Algorithm 1) using conditional scaling laws for LLMs, which is distinct from the survey's scope of reviewing general hardware-aware NAS approaches.

Contribution 3: Characterization of architectural factors' impact on inference and accuracy

Description: The authors systematically study how hidden size, mlp-to-attention ratio, and GQA affect both inference throughput and model accuracy by training over 200 models ranging from 80M to 3B parameters, revealing U-shaped relationships between these factors and training loss.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

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

References

- [0] Scaling Laws Meet Model Architecture: Toward Inference-Efficient LLMs View paper
- [1] Scaling llm test-time compute optimally can be more effective than scaling model parameters View paper
- [2] Towards greater leverage: Scaling laws for efficient mixture-of-experts language models View paper
- [3] Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning View paper
- [4] Efficiently scaling transformer inference View paper
- [5] Inference Compute-Optimal Video Vision Language Models View paper
- [6] Glam: Efficient scaling of language models with mixture-of-experts View paper
- [7] Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving View paper
- [8] Scaling Inference-Efficient Language Models View paper
- [9] Inference-time scaling for complex tasks: Where we stand and what lies ahead View paper
- [10] Scaling laws for predicting downstream performance in llms View paper
- [11] Efficiency in language understanding and generation: An evaluation of four open-source large language models View paper
- [12] Training Language Models to Reason Efficiently View paper
- [13] Beyond chinchilla-optimal: Accounting for inference in language model scaling laws View paper
- [14] SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer View paper
- [15] A*-Decoding: Token-Efficient Inference Scaling View paper
- [16] Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing View paper
- [17] Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings View paper
- [18] Toward Inference-optimal Mixture-of-Expert Large Language Models View paper
- [19] LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via a Hybrid Architecture View paper
- [20] Llamba: Scaling Distilled Recurrent Models for Efficient Language Processing View paper
- [21] An empirical analysis of compute-optimal inference for problem-solving with language models View paper
- [22] Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures View paper
- [23] Memory Is All You Need: An Overview of Compute-in-Memory Architectures for Accelerating Large Language Model Inference View paper
- [24] Advancing language model reasoning through reinforcement learning and inference scaling View paper
- [25] ETS: Efficient Tree Search for Inference-Time Scaling View paper
- [26] Parallel scaling law for language models View paper
- [27] Sprout: Green Generative AI with Carbon-Efficient LLM Inference View paper
- [28] Uni-MoE: Scaling Unified Multimodal LLMs With Mixture of Experts View paper
- [29] Multi-Agent Sampling: Scaling Inference Compute for Data Synthesis with Tree Search-Based Agentic Collaboration View paper
- [30] Think Deep, Think Fast: Investigating Efficiency of Verifier-free Inference-time-scaling Methods View paper
- [31] Inference economics of language models View paper
- [32] A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search View paper
- [33] System-performance and cost modeling of Large Language Model training and inference View paper
- [34] Rollout Roulette: A Probabilistic Inference Approach to Inference-Time Scaling of LLMs using Particle-Based Monte Carlo Methods View paper
- [35] Spectra 1.1: Scaling Laws and Efficient Inference for Ternary Language Models View paper
- [36] How to Upscale Neural Networks with Scaling Law? A Survey and Practical Guidelines View paper
- [37] A Semantic-Aware Layer-Freezing Approach to Computation-Efficient Fine-Tuning of Language Models View paper
- [38] Mixture of Attention Spans: Optimizing LLM Inference Efficiency with Heterogeneous Sliding-Window Lengths View paper
- [39] Hybrid Architectures for Language Models: Systematic Analysis and Design Insights View paper
- [40] Joint MoE Scaling Laws: Mixture of Experts Can Be Memory Efficient View paper
- [41] Mechanistic Design and Scaling of Hybrid Architectures View paper
- [42] Inference Scaling for Long-Context Retrieval Augmented Generation View paper
- [43] Inference-Time Scaling for Generalist Reward Modeling View paper
- [44] EfficientLLM: Efficiency in Large Language Models View paper
- [45] AIBrix: Towards Scalable, Cost-Effective Large Language Model Inference Infrastructure View paper
- [46] UELLM: A Unified and Efficient Approach for Large Language Model Inference Serving View paper
- [47] The Efficiency Spectrum of Large Language Models: An Algorithmic Survey View paper
- [48] NeuroPrune: A Neuro-inspired Topological Sparse Training Algorithm for Large Language Models View paper
- [49] Scaling Laws and Efficient Inference for Ternary Language Models View paper
- [50] Q-Sparse: All Large Language Models can be Fully Sparsely-Activated View paper
- [51] Observational scaling laws and the predictability of langauge model performance View paper
- [52] Physics of language models: Part 3.3, knowledge capacity scaling laws View paper
- [53] CodeGen2: Lessons for Training LLMs on Programming and Natural Languages View paper
- [54] Slamming: Training a Speech Language Model on One GPU in a Day View paper
- [55] AI and Memory Wall View paper
- [56] Scaling laws with vocabulary: Larger models deserve larger vocabularies View paper
- [57] Farseer: A Refined Scaling Law in Large Language Models View paper
- [58] Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling? View paper
- [59] Collaborative performance prediction for large language models View paper
- [60] Efficient neural architecture search via parameters sharing View paper
- [61] MARCO: Hardware-Aware Neural Architecture Search for Edge Devices with Multi-Agent Reinforcement Learning and Conformal Prediction Filtering View paper
- [62] Neural architecture search for resource constrained hardware devices: A survey View paper
- [63] Sparse: Sparse architecture search for cnns on resource-constrained microcontrollers View paper
- [64] Mixed precision neural architecture search for energy efficient deep learning View paper
- [65] Semi-supervised neural architecture search View paper
- [66] Automated UAV Object Detector Design Using Large Language Model-Guided Architecture Search View paper

- [67] Neural architecture search as multiobjective optimization benchmarks: Problem formulation and performance assessment View paper
- [68] Hao: Hardware-aware neural architecture optimization for efficient inference View paper
- [69] Faststereonet: A fast neural architecture search for improving the inference of disparity estimation on resource-limited platforms View paper