Novelty Assessment Report

Paper: Relative Scaling Laws for LLMs

PDF URL: https://openreview.net/pdf?id=RpX0k5RR6q

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Scaling laws describe how language models improve with additional data, parameters, and compute. While widely used, they are typically measured on aggregate test sets. Aggregate evaluations yield clean trends but average over heterogeneous subpopulations, obscuring performance disparities. We introduce relative scaling laws, which track how performance gaps between test distributions evolve with scale rather than focusing solely on absolute error. Using 255 decoder-only Transformers trained under matched-compute (IsoFLOP) budgets from 10^{18}--10^{20} FLOPs on standard pretraining datasets, we find diverse trajectories: academic domains on MMLU converge toward parity; regional English dialects shift depending on population size; and clusters of AI risk behaviours split, with capability- and influence-related risks increasing during pretraining while adversarial risks do not. These results show that although scaling improves overall performance, it is not a universal equalizer. To support further study, we release all model checkpoints from this work to enable practitioners to measure relative alongside traditional scaling laws, in order to better prioritize robustness challenges in light of the bitter lesson

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: **Tracking Performance Gaps Between Test Distributions with Scale**

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

Taxonomy Overview

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

- **Scaling Laws and Distribution-Dependent Performance**
- **Distribution Shift Characterization and Benchmarking**
- **Test-Time Adaptation Methods**
- **Training-Time Robustness and Data Strategies**
- **Robustness and Generalization Theory**
- **Cross-Domain and Multi-Modal Generalization**
- **Performance Evaluation Methodology and Reporting**
- **Application-Specific Performance Studies**

Complete Taxonomy Tree

- Tracking Performance Gaps Between Test Distributions with Scale Survey Taxonomy
- Scaling Laws and Distribution-Dependent Performance
  - Relative and Heterogeneous Scaling Dynamics ★ (3 papers)
  - [0] Relative Scaling Laws for LLMs (Anon et al., 2026) View paper
  - [2] Beyond neural scaling laws: beating power law scaling via data pruning (Sorscher, 2022) View paper
  - [15] Bias as a Virtue: Rethinking Generalization under Distribution Shifts (Chen, 2025) View paper
  - Aggregate Scaling Law Studies (3 papers)
  - [11] Mechanistic Design and Scaling of Hybrid Architectures (Poli, 2024) View paper
  - [38] An evaluation of methods for modelling species distributions (Pedro Segurado, 2004) View paper
  - [39] Scaling Laws for Multilingual Neural Machine Translation (Fernandes, 2023) View paper
  - Test-Time Compute Scaling (1 papers)
  - [3] Towards Thinking-Optimal Scaling of Test-Time Compute for LLM Reasoning (Yang, 2025) View paper
- Distribution Shift Characterization and Benchmarking
  - Temporal and Natural Distribution Shift Benchmarks (3 papers)
  - [14] DivShift: Exploring Domain-Specific Distribution Shifts in Large-Scale, Volunteer-Collected Biodiversity Datasets (Exposito-Alonso, 2025) View paper
  - [30] The Effect of Natural Distribution Shift on Question Answering Models (J. J. Miller, 2020) View paper
  - [44] Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time (Yao, 2022) View paper
  - Controlled and Domain-Specific Shift Benchmarks (4 papers)
  - [16] RobustCLEVR: A benchmark and framework for evaluating robustness in object-centric learning (Nathan Drenkow, 2024) View paper
  - [17] Ava-bench: Atomic visual ability benchmark for vision foundation models (Mai, 2025) View paper
  - [31] Characterizing generalization under out-of-distribution shifts in deep metric learning (Milbich, 2021) View paper
  - [47] Do-GOOD: Towards Distribution Shift Evaluation for Pre-Trained Visual Document Understanding Models (He, 2023) View paper
  - Fine-Grained Evaluation Toolkits and Metrics (2 papers)
  - [9] Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification (Borkan, 2019) View paper

    - [37] Cast: Conditional attribute subsampling toolkit for fine-grained evaluation (Wes Robbins, 2023) View paper
- Test-Time Adaptation Methods
    - Single-Model Test-Time Adaptation (4 papers)
    - [5] Efficient test-time model adaptation without forgetting (Niu, 2022) View paper
    - [22] Unsupervised Detection and Correction of Model Calibration Shift at Test-Time (Supreeth P. Shashikumar, 2023) View paper
    - [26] Test-time adaptation for optical flow estimation using motion vectors (Seyed Mehdi Ayyoubzadeh, 2023) View paper
    - [33] Calibration of Time-Series Forecasting: Detecting and Adapting Context-Driven Distribution Shift (Mouxiang Chen, 2023) View paper
    - Federated and Collaborative Test-Time Adaptation (2 papers)
    - [6] Test-time robust personalization for federated learning (Jiang, 2022) View paper
    - [28] Enabling collaborative test-time adaptation in dynamic environment via federated learning (Jiayuan Zhang, 2024) View paper
    - Domain-Specific Test-Time Adaptation (2 papers)
    - [12] Dual Test-Time Training for Out-of-Distribution Recommender System (Xihong Yang, 2024) View paper
    - [46] Scale-Aware Domain Adaptation for Robust UAV Tracking (Chang-hong Fu, 2023) View paper
- Training-Time Robustness and Data Strategies
    - Data Augmentation and Sampling Techniques (3 papers)
    - [7] A comprehensive evaluation of oversampling techniques for enhancing text classification performance (Salimkan Fatma Taskiran, 2025) View paper
    - [27] Disparity Distribution Equalization: An Effective Data Enhancement for Stereo Matching (Jiaxuan Zheng, 2024) View paper
    - [36] Adaptive sampling for stochastic risk-averse learning (Curi, 2020) View paper
    - Robust Training Objectives and Architectures (2 papers)
    - [42] Deep Model-Based Architectures for Inverse Problems Under Mismatched Priors (Shirin Shoushtari, 2022) View paper
    - [50] Learning to Retain while Acquiring: Combating Distribution-Shift in Adversarial Data-Free Knowledge Distillation (Gaurav Patel, 2023) View paper
    - Synthetic-to-Real Transfer and Pseudo-Labeling (2 papers)
    - [1] CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos (Wang, 2024) View paper
    - [4] Scaling laws of synthetic images for model training... for now (Lijie Fan, 2024) View paper
- Robustness and Generalization Theory
    - Spurious Correlation and Causal Representation Learning (1 papers)
    - [18] Examining and combating spurious features under distribution shift (Zhou, 2021) View paper
    - Holistic Reliability and Evaluation Frameworks (1 papers)
    - [8] A holistic assessment of the reliability of machine learning systems (Corso, 2023) View paper
- Cross-Domain and Multi-Modal Generalization
    - Pretraining-Downstream Task Alignment (2 papers)
    - [25] Data similarity is not enough to explain language model performance (Gregory Yauney, 2023) View paper
    - [41] Overestimation in LLM Evaluation: A Controlled Large-Scale Study on Data Contamination's Impact on Machine Translation (Kocyigit, 2025) View paper
    - Cross-Domain Adaptation and Transfer (2 papers)
    - [34] Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets (Andrea Roncoli, 2023) View paper
    - [45] Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing (Darestani, 2022) View paper
- Performance Evaluation Methodology and Reporting
    - Stochasticity and Reproducibility in Evaluation (2 papers)
    - [13] Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging (Reimers, 2017) View paper
    - [40] No Evidence of Reliability Across 36 Variations of the Emotional Dot-Probe Task in 9,600 Participants (I. Xu, 2024) View paper
    - Domain-Specific Evaluation Studies (3 papers)
    - [20] Testing the predictive performance of distribution models (Volker Bahn, 2013) View paper
    - [23] Industrial Anomaly Detection with Domain Shift: A Real-world Dataset and Masked Multi-scale Reconstruction (Zhang Zilong, 2023) View paper
    - [29] Scaling laws for language encoding models in fMRI (Antonello, 2023) View paper
    - Evaluation Infrastructure and Benchmarking Platforms (1 papers)
    - [24] BenchCloud: A Platform for Scalable Performance Benchmarking (Dirk Beyer, 2024) View paper
- Application-Specific Performance Studies
    - Software and System Performance (2 papers)
    - [21] A large-scale empirical study on mobile performance: energy, run-time and memory (Rua, 2023) View paper
    - [49] Performance Evaluation of NewSQL Databases in a Distributed Architecture (Zhiyao Zhang, 2025) View paper
    - Physical and Environmental System Evaluation (2 papers)
    - [10] Evaluation of multiple time scale rainfall erosivity models: A case study of subtropical regions in Central China (Yaodong Ping, 2025) View paper
    - [35] Performance evaluation of computed tomography systems: summary of AAPM Task Group 233 (Ehsan Samei, 2019) View paper
    - Social and Educational Assessment (4 papers)
    - [19] Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation (LA Rivera, 2019) View paper
    - [32] The Thermal Tolerances, Distributions, and Performances of Tropical Montane Tree Species (Kenneth J. Feeley, 2020) View paper
    - [43] Measuring economic competence of secondary school students in Germany (Kaiser, 2020) View paper
    - [48] Large scale real-world multi-person tracking (Bing Shuai, 2022) View paper

Narrative

Core task: tracking performance gaps between test distributions with scale. The field examines how model performance evolves across different test distributions as models and datasets grow, organizing research into several major branches. Scaling Laws and Distribution-Dependent Performance investigates how accuracy improvements vary heterogeneously across in-distribution versus out-of-distribution settings, with works like Relative Scaling Laws[0] and Beyond Neural Scaling[2] exploring whether all test scenarios benefit equally from increased compute. Distribution Shift Characterization and Benchmarking focuses on measuring and taxonomizing the types of shifts that occur in practice, while Test-Time Adaptation Methods and Training-Time Robustness strategies offer complementary approaches to closing performance gaps. Additional branches address theoretical foundations of generalization, cross-domain transfer, and rigorous evaluation methodology, reflecting the community's recognition that a single accuracy number often masks important disparities.

Particularly active lines of work reveal tensions between scaling optimism and robustness challenges. Some studies suggest that larger models naturally improve worst-case performance, yet others document persistent or even widening gaps on certain distribution shifts, raising questions about whether scale alone suffices or whether targeted interventions remain necessary. Relative Scaling Laws[0] sits within the branch examining heterogeneous scaling dynamics, emphasizing how different test distributions respond differently to model scale—a perspective closely aligned with Beyond Neural Scaling[2], which questions uniform scaling benefits, and contrasting with works like Thinking Optimal Scaling[3] that explore allocation strategies. Nearby efforts such as Efficient Test Time Adaptation[5] and Test Time Robust Personalization[6] offer adaptive mechanisms to address gaps that scaling does not fully resolve, highlighting an ongoing dialogue about whether robustness emerges automatically or requires explicit design choices at training or test time.

Related Works in Same Category

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

1. Beyond neural scaling laws: beating power law scaling via data pruning

Authors: Sorscher, Ben, Ben Sorscher, Geirhos, Robert, et al. (15 authors total) | **Year/Venue**: 2022 • Neural Information Processing Systems | **URL**: View paper

Abstract

Widely observed neural scaling laws, in which error falls off as a power of the training set size, model size, or both, have driven substantial performance improvements in deep learning. However, these improvements through scaling alone require considerable costs in compute and energy. Here we focus on the scaling of error with dataset size and show how in theory we can break beyond power law scaling and potentially even reduce it to exponential scaling instead if we have access to a high-qualit...

Relationship Analysis

Both papers belong to the category of Relative and Heterogeneous Scaling Dynamics, examining how performance gaps evolve with scale rather than aggregate performance. The original paper introduces relative scaling laws to track performance disparities between test distributions (e.g., MMLU domains, regional English dialects, AI risks) as compute increases, while the candidate paper focuses on beating power law scaling through data pruning strategies that selectively retain training examples based on difficulty metrics. The key difference is that the original paper analyzes distribution-dependent performance gaps during standard scaling, whereas the candidate paper proposes methods to improve scaling efficiency by pruning training data rather than studying naturally occurring performance gaps across test distributions.

2. Bias as a Virtue: Rethinking Generalization under Distribution Shifts

Authors: Chen, Ruixuan, Li Wentao, Xiao, Jiahui, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Machine learning models often degrade when deployed on data distributions different from their training data. Challenging conventional validation paradigms, we demonstrate that higher in-distribution (ID) bias can lead to better out-of-distribution (OOD) generalization. Our Adaptive Distribution Bridge (ADB) framework implements this insight by introducing controlled statistical diversity during training, enabling models to develop bias profiles that effectively generalize across distributions. ...

Relationship Analysis

Both papers belong to the 'Relative and Heterogeneous Scaling Dynamics' category, analyzing how performance gaps between distributions change with scale rather than aggregate performance alone. The original paper introduces relative scaling laws to track performance disparities across test distributions (MMLU domains, regional English dialects, AI risks) as compute increases from 10^18 to 10^20 FLOPs, finding diverse trajectories where some gaps converge while others diverge. The candidate paper focuses on a different aspect: how in-distribution bias affects out-of-distribution generalization under distribution shifts, proposing the ADB framework to strategically increase training bias for better OOD performance, rather than studying how existing performance gaps evolve with model scale.

Contributions Analysis

Overall novelty summary. The paper introduces relative scaling laws, which track how performance gaps between test distributions evolve with scale rather than measuring aggregate error alone. It sits in the 'Relative and Heterogeneous Scaling Dynamics' leaf, which contains only three papers total, indicating a relatively sparse research direction. The sibling papers examine related questions about heterogeneous scaling behavior, but the taxonomy shows this subpopulation-focused perspective remains less explored than aggregate scaling law studies, which form a separate, more established leaf.

The taxonomy places this work within 'Scaling Laws and Distribution-Dependent Performance,' adjacent to branches on test-time compute scaling and aggregate scaling studies. Neighboring leaves address distribution shift characterization and test-time adaptation methods, reflecting the field's broader concern with robustness. The scope note for the paper's leaf explicitly excludes aggregate-only trends, positioning relative scaling laws as a complementary lens that examines whether scale acts as a universal equalizer or produces divergent trajectories across subpopulations.

Among 24 candidates examined across three contributions, none were flagged as clearly refuting the work. The relative scaling laws framework examined 10 candidates with zero refutations; the open-source IsoFLOP suite examined 4 with zero refutations; and the empirical case studies examined 10 with zero refutations. This suggests that within the limited search scope, the specific combination of tracking performance gaps across diverse test distributions under matched-compute budgets appears relatively unexplored, though the analysis does not claim exhaustive coverage of all prior scaling law research.

Based on the limited literature search, the work appears to occupy a distinct position within scaling law research by systematically measuring relative rather than absolute performance trends. The sparse population of its taxonomy leaf and absence of refuting candidates among those examined suggest novelty, though the search scope of 24 papers leaves open the possibility of relevant work outside the top semantic matches. The release of 255 model checkpoints may enable future comparative studies that were not captured in this analysis.

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

Contribution 1: Relative scaling laws framework

Description: The authors formalize a framework for measuring how performance disparities between different test distributions change as models scale, separating initial gaps from differences in improvement rates. This is formulated as a power law that indicates whether gaps narrow, persist, or widen with increased compute.

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. Unveiling Downstream Performance Scaling of LLMs: A Clustering-Based Perspective

URL: View paper

Brief Assessment

Downstream Performance Clustering[65] focuses on predicting downstream task performance through clustering tasks by difficulty, not on measuring how performance gaps between distributions evolve with scale as formulated in the original paper's relative scaling framework.

2. Quartet: Native FP4 Training Can Be Optimal for Large Language Models

URL: View paper

Brief Assessment

Quartet FP4 Training[66] focuses on low-precision training optimization and introduces scaling laws relating evaluation loss to forward/backward precision (Equation 1), not on measuring performance disparities between test distributions as models scale. The candidate's scaling law models precision-accuracy trade-offs, while the original models how gaps between different test distributions evolve with compute.

3. Unlocking high-accuracy differentially private image classification through scale

URL: View paper

Brief Assessment

Differentially Private Scale[62] focuses on differential privacy techniques for image classification, not on measuring performance disparities across test distributions with model scale. The paper addresses privacy-utility tradeoffs in deep learning rather than relative scaling behavior across subpopulations.

4. Delving deep into the generalization of vision transformers under distribution shifts

URL: View paper

Brief Assessment

Vision Transformers Generalization[63] focuses on out-of-distribution generalization of vision transformers across different types of distribution shifts (background, corruption, texture, style), not on how performance gaps between test distributions evolve with model scale in language models.

5. The evolution of the Black-White test score gap in Grades Kâ□□3: The fragility of results

URL: View paper

Brief Assessment

Test Score Gap Evolution[70] examines how Black-White test score gaps evolve across grades K-3 using different scale transformations of test scores. This focuses on educational assessment scaling issues, not on how performance disparities between test distributions change with model scale in machine learning systems.

6. Data Contamination or Genuine Generalization? Disentangling LLM Performance on Benchmarks

URL: View paper

Brief Assessment

Data Contamination Disentangling[64] focuses on distinguishing genuine generalization from data contamination in LLMs through n-gram alignment and perturbation testing. It does not address how performance gaps between test distributions evolve with model scale, which is the core focus of the relative scaling laws framework.

7. Genrep for first-shot unsupervised anomalous sound detection of dcase 2025 challenge

URL: View paper

Brief Assessment

Genrep Anomalous Sound[61] focuses on anomalous sound detection using frozen audio embeddings and domain shift handling. It does not address scaling laws, performance gap evolution across test distributions, or how model scale affects performance disparities.

8. Navigating the Accuracy-Size Trade-Off with Flexible Model Merging

URL: View paper

Brief Assessment

Flexible Model Merging[68] focuses on model merging techniques and accuracy-size trade-offs in combining fine-tuned models, not on measuring performance disparities across test distributions as models scale with compute.

9. Scaling up Masked Diffusion Models on Text

URL: View paper

Brief Assessment

Masked Diffusion Text[69] focuses on establishing scaling laws for masked diffusion models in language tasks, not on measuring performance disparities between test distributions or how gaps evolve with scale across subpopulations.

10. FairTune: A Bias-Aware Fine-Tuning Framework Towards Fair Heart Rate Prediction from PPG

URL: View paper

Brief Assessment

FairTune[67] focuses on bias-aware fine-tuning of physiological foundation models for heart rate prediction across demographic groups, not on measuring performance gap evolution across test distributions with model scale or formulating power laws for scaling behavior.

Contribution 2: Open-source IsoFLOP scaling suite of 255 models

Description: The authors train and publicly release 255 decoder-only Transformer models under matched-compute budgets spanning three orders of magnitude across three distinct pretraining datasets. This resource enables reproducible study of both traditional and relative scaling laws.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and â¦

URL: View paper

Brief Assessment

Encoder Decoder Comparison[71] focuses on comparing encoder-only (RoBERTa) and decoder-only (LLaMA2) models for intent classification and sentiment analysis tasks, not on training models under matched-compute budgets or releasing scaling suites for studying scaling laws.

2. Scaling Sparse and Dense Retrieval in Decoder-Only LLMs

URL: View paper

Brief Assessment

Sparse Dense Retrieval Scaling[73] focuses on retrieval models (sparse vs. dense) trained on MSMarco with different optimization objectives, not on general-purpose decoder-only Transformers across diverse pretraining datasets under IsoFLOP budgets.

3. Decoder-only architecture for streaming end-to-end speech recognition

URL: View paper

Brief Assessment

Decoder Only Streaming[72] focuses on decoder-only architectures for streaming speech recognition, not on training scaling suites across matched compute budgets for language models.

4. Towards Neural Scaling Laws for Time Series Foundation Models

URL: View paper

Brief Assessment

Time Series Scaling Laws[74] focuses on time series foundation models with encoder/decoder transformers trained on time series data, not general language models. The candidate trains models on time series forecasting tasks rather than language modeling with matched-compute budgets across diverse pretraining corpora.

Contribution 3: Empirical case studies demonstrating diverse relative scaling trajectories

Description: The authors demonstrate the application of relative scaling laws across three distinct domains, revealing diverse trajectories including convergence of academic domains, mixed effects for regional English dialects, and divergence between capability-related and adversarial AI risks. These studies show that scale has non-uniform impacts on distributional robustness.

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. On the robustness of chatgpt: An adversarial and out-of-distribution perspective

URL: View paper

Brief Assessment

ChatGPT Robustness[51] focuses on adversarial and out-of-distribution robustness evaluation of ChatGPT on classification and translation tasks, not on relative scaling laws or trajectories across different distributions during model training.

2. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges

URL: View paper

Brief Assessment

Alignment Safety LLMs[58] is a comprehensive survey on alignment and safety mechanisms in large language models, focusing on training paradigms, evaluation frameworks, and ethical considerations. It does not present empirical scaling law studies or analyze distributional robustness trajectories across domains, making it fundamentally different from the original paper's contribution.

3. Bridging Distribution Shift and AI Safety: Conceptual and Methodological Synergies

URL: View paper

Brief Assessment

Distribution Shift Safety[53] focuses on conceptual and methodological connections between distribution shift and AI safety across domains like fairness, security, and trustworthiness. It does not examine relative scaling trajectories across academic domains, linguistic variation, or AI risk behaviors as the original paper does.

4. Navigating the safety landscape: Measuring risks in finetuning large language models

URL: View paper

Brief Assessment

Safety Landscape Navigation[54] focuses on measuring safety risks in finetuning LLMs through weight perturbations and safety basins, not on distributional robustness trajectories across different domains during pretraining scaling.

5. JailbreaksOverTime: Detecting Jailbreak Attacks Under Distribution Shift

URL: View paper

Brief Assessment

JailbreaksOverTime[56] focuses on jailbreak detection under distribution shift in AI safety systems, not on relative scaling laws or distributional robustness trajectories across different domains like academic knowledge, linguistic variation, or AI risks.

6. ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models

URL: View paper

Brief Assessment

ASSERT[59] focuses on robustness evaluation through adversarial prompt generation in safety domains, not on scaling law trajectories across different distributions or compute budgets.

7. Uncertainty-Aware Trajectory Prediction via Rule-Regularized Heteroscedastic Deep Classification

URL: View paper

Brief Assessment

Uncertainty Aware Trajectory[55] focuses on trajectory prediction for autonomous driving with uncertainty quantification and traffic rule integration. It does not address scaling laws, distributional robustness across knowledge domains, linguistic variation, or AI risk trajectories.

8. Towards Robust Machine Learning under Distribution Shifts: From Causal Guarantees to Robust Federated Learning

URL: View paper

Brief Assessment

Robust Federated Learning[60] focuses on distribution shifts in federated learning and causal robustness guarantees, not on scaling law trajectories across academic domains, linguistic variation, or AI risks as studied in the original paper.

9. Robust LLM Alignment via Distributionally Robust Direct Preference Optimization

URL: View paper

Brief Assessment

Robust LLM Alignment[52] focuses on preference distribution shift in alignment algorithms using distributionally robust optimization, not on scaling law trajectories across knowledge domains, linguistic variation, or AI risks as studied in the original paper.

10. Evaluating model robustness and stability to dataset shift

URL: View paper

Brief Assessment

Model Robustness Evaluation[57] focuses on evaluating model stability to dataset shifts using distributional robustness optimization, not on scaling laws or how performance gaps evolve with compute/scale across different domains or tasks.

Appendix: Text Similarity Detection

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

References

- [0] Relative Scaling Laws for LLMs View paper
- [1] CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos View paper
- [2] Beyond neural scaling laws: beating power law scaling via data pruning View paper
- [3] Towards Thinking-Optimal Scaling of Test-Time Compute for LLM Reasoning View paper
- [4] Scaling laws of synthetic images for model training... for now View paper
- [5] Efficient test-time model adaptation without forgetting View paper
- [6] Test-time robust personalization for federated learning View paper
- [7] A comprehensive evaluation of oversampling techniques for enhancing text classification performance View paper
- [8] A holistic assessment of the reliability of machine learning systems View paper
- [9] Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification View paper
- [10] Evaluation of multiple time scale rainfall erosivity models: A case study of subtropical regions in Central China View paper
- [11] Mechanistic Design and Scaling of Hybrid Architectures View paper
- [12] Dual Test-Time Training for Out-of-Distribution Recommender System View paper
- [13] Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging View paper
- [14] DivShift: Exploring Domain-Specific Distribution Shifts in Large-Scale, Volunteer-Collected Biodiversity Datasets View paper
- [15] Bias as a Virtue: Rethinking Generalization under Distribution Shifts View paper
- [16] RobustCLEVR: A benchmark and framework for evaluating robustness in object-centric learning View paper
- [17] Ava-bench: Atomic visual ability benchmark for vision foundation models View paper
- [18] Examining and combating spurious features under distribution shift View paper
- [19] Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation View paper
- [20] Testing the predictive performance of distribution models View paper
- [21] A large-scale empirical study on mobile performance: energy, run-time and memory View paper
- [22] Unsupervised Detection and Correction of Model Calibration Shift at Test-Time View paper
- [23] Industrial Anomaly Detection with Domain Shift: A Real-world Dataset and Masked Multi-scale Reconstruction View paper
- [24] BenchCloud: A Platform for Scalable Performance Benchmarking View paper
- [25] Data similarity is not enough to explain language model performance View paper
- [26] Test-time adaptation for optical flow estimation using motion vectors View paper
- [27] Disparity Distribution Equalization: An Effective Data Enhancement for Stereo Matching View paper
- [28] Enabling collaborative test-time adaptation in dynamic environment via federated learning View paper
- [29] Scaling laws for language encoding models in fMRI View paper
- [30] The Effect of Natural Distribution Shift on Question Answering Models View paper
- [31] Characterizing generalization under out-of-distribution shifts in deep metric learning View paper
- [32] The Thermal Tolerances, Distributions, and Performances of Tropical Montane Tree Species View paper
- [33] Calibration of Time-Series Forecasting: Detecting and Adapting Context-Driven Distribution Shift View paper
- [34] Domain Adaptive Graph Neural Networks for Constraining Cosmological Parameters Across Multiple Data Sets View paper
- [35] Performance evaluation of computed tomography systems: summary of AAPM Task Group 233 View paper
- [36] Adaptive sampling for stochastic risk-averse learning View paper
- [37] Cast: Conditional attribute subsampling toolkit for fine-grained evaluation View paper
- [38] An evaluation of methods for modelling species distributions View paper
- [39] Scaling Laws for Multilingual Neural Machine Translation View paper
- [40] No Evidence of Reliability Across 36 Variations of the Emotional Dot-Probe Task in 9,600 Participants View paper
- [41] Overestimation in LLM Evaluation: A Controlled Large-Scale Study on Data Contamination's Impact on Machine Translation View paper
- [42] Deep Model-Based Architectures for Inverse Problems Under Mismatched Priors View paper
- [43] Measuring economic competence of secondary school students in Germany View paper
- [44] Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time View paper

- [45] Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing View paper
- [46] Scale-Aware Domain Adaptation for Robust UAV Tracking View paper
- [47] Do-GOOD: Towards Distribution Shift Evaluation for Pre-Trained Visual Document Understanding Models View paper
- [48] Large scale real-world multi-person tracking View paper
- [49] Performance Evaluation of NewSQL Databases in a Distributed Architecture View paper
- [50] Learning to Retain while Acquiring: Combating Distribution-Shift in Adversarial Data-Free Knowledge Distillation View paper
- [51] On the robustness of chatgpt: An adversarial and out-of-distribution perspective View paper
- [52] Robust LLM Alignment via Distributionally Robust Direct Preference Optimization View paper
- [53] Bridging Distribution Shift and AI Safety: Conceptual and Methodological Synergies View paper
- [54] Navigating the safety landscape: Measuring risks in finetuning large language models View paper
- [55] Uncertainty-Aware Trajectory Prediction via Rule-Regularized Heteroscedastic Deep Classification View paper
- [56] JailbreaksOverTime: Detecting Jailbreak Attacks Under Distribution Shift View paper
- [57] Evaluating model robustness and stability to dataset shift View paper
- [58] Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges View paper
- [59] ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models View paper
- [60] Towards Robust Machine Learning under Distribution Shifts: From Causal Guarantees to Robust Federated Learning View paper
- [61] Genrep for first-shot unsupervised anomalous sound detection of dcase 2025 challenge View paper
- [62] Unlocking high-accuracy differentially private image classification through scale View paper
- [63] Delving deep into the generalization of vision transformers under distribution shifts View paper
- [64] Data Contamination or Genuine Generalization? Disentangling LLM Performance on Benchmarks View paper
- [65] Unveiling Downstream Performance Scaling of LLMs: A Clustering-Based Perspective View paper
- [66] Quartet: Native FP4 Training Can Be Optimal for Large Language Models View paper
- [67] FairTune: A Bias-Aware Fine-Tuning Framework Towards Fair Heart Rate Prediction from PPG View paper
- [68] Navigating the Accuracy-Size Trade-Off with Flexible Model Merging View paper
- [69] Scaling up Masked Diffusion Models on Text View paper
- [70] The evolution of the Black-White test score gap in Grades Kâ□□3: The fragility of results View paper
- [71] A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and â□¦ View paper
- [72] Decoder-only architecture for streaming end-to-end speech recognition View paper
- [73] Scaling Sparse and Dense Retrieval in Decoder-Only LLMs View paper
- [74] Towards Neural Scaling Laws for Time Series Foundation Models View paper