Novelty Assessment Report

Paper: Train-before-Test Harmonizes Language Model Rankings
PDF URL: https://openreview.net/pdf?id=ORv3SAzus1
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Existing language model benchmarks provide contradictory model rankings, even for benchmarks that aim to capture similar skills. This dilemma of conflicting rankings hampers model selection, clouds model comparisons, and adds confusion to a growing ecosystem of competing models. In this paper, we take a different perspective on model comparison: instead of relying on out-of-the-box performance via direct evaluation, we compare model potential by providing each model with identical benchmark-specific fine-tuning before evaluation. We call this approach train-before-test. Our primary contribution is a comprehensive empirical evaluation of model potential across 24 benchmarks and 61 models. First, we demonstrate that model potential rankings obtained through train-before-test exhibit remarkable consistency across all benchmarks. Whereas traditional rankings demonstrate little external validity under direct evaluation, they enjoy a significant degree of external validity when applying train-before-test: model potential rankings transfer gracefully from one benchmark to another. Second, train-before-test restores the connection between perplexity and downstream task performance, lost under direct evaluation. Remarkably, even pre-finetuning perplexity of a base model predicts post-finetuning downstream performance, suggesting that ranking consistency reflects inherent model potential rather than fine-tuning artifacts. Finally, train-before-test reduces the model-score matrix to essentially rank one, indicating that model potential is dominated by one latent factor, uncovered by train-before-test. While direct evaluation remains useful for assessing deployment-ready performance, train-before-test provides a complementary lens for understanding achievable performance of models after adaptation.

Core Task Landscape

This paper addresses: **Harmonizing Language Model Rankings Through Fine-Tuning Before Evaluation**

A total of **50 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

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

- **Alignment and Preference Learning**
- **Ranking Architectures and Fine-Tuning Strategies**
- **Parameter-Efficient Fine-Tuning**
- **Domain Applications**
- **Training Data and Active Learning**
- **Reasoning and Chain-of-Thought Enhancement**

Complete Taxonomy Tree

- Harmonizing Language Model Rankings Through Fine-Tuning Before Evaluation Survey Taxonomy
- Alignment and Preference Learning
  - Reinforcement Learning from Human Feedback ★ (6 papers)
  - [0] Train-before-Test Harmonizes Language Model Rankings (Anon et al., 2026) View paper
  - [1] Preference ranking optimization for human alignment (Huang Fei, 2024) View paper
  - [10] Tuning for LLM alignment (Uday Kamath, 2024) View paper
  - [16] Rrhf: Rank responses to align language models with human feedback (H Yuan, 2023) View paper
  - [17] Fine-tuning language models from human preferences (Ziegler, 2019) View paper
  - [48] RRHF: Rank Responses to Align Language Models with Human Feedback without tears (YUAN Zheng, 2023) View paper
  - Direct Preference Optimization (2 papers)
  - [14] Is Crowdsourcing Breaking Your Bank? Cost-Effective Fine-Tuning of Pre-trained Language Models with Proximal Policy Optimization (Yang Shuo, 2024) View paper
  - [22] Finetuning Large Language Model for Personalized Ranking (Wu Ning, 2024) View paper
  - Ranking Feedback (2 papers)
  - [19] Pangu-coder2: Boosting large language models for code with ranking feedback (Shen Bo, 2023) View paper
  - [29] Oracle-RLAIF: An Improved Fine-Tuning Framework for Multi-modal Video Models through Reinforcement Learning from Ranking Feedback (Glatt, 2025) View paper
  - Diverse and Personalized Alignment (2 papers)
  - [4] Fine-tuning language models to find agreement among humans with diverse preferences (Bakker, 2022) View paper
  - [9] Towards an End-to-End Personal Fine-Tuning Framework for AI Value Alignment (Eleanor Watson, 2024) View paper
  - Alignment Surveys and Fairness (2 papers)
  - [2] Aligning large language models with human: A survey (Wang Yu-Fei, 2023) View paper
  - [15] A study of implicit ranking unfairness in large language models (Chua, 2024) View paper
- Ranking Architectures and Fine-Tuning Strategies
  - Pre-trained Model Ranking Adaptation (3 papers)

- [8] Pre-trained language model-based retrieval and ranking for web search (Lixin Zou, 2022) View paper
- [23] Modeling relevance ranking under the pre-training and fine-tuning paradigm (Bo Lin, 2021) View paper
- [24] P3 Ranker: Mitigating the Gaps between Pre-training and Ranking Fine-tuning with Prompt-based Learning and Pre-finetuning (Hu, 2022) View paper
- Prompt-Based and Zero-Shot Ranking (4 papers)
- [3] Large language models are effective text rankers with pairwise ranking prompting (Bendersky, 2024) View paper
- [11] Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models (Crystina Zhang, 2025) View paper
- [33] Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning (Zhuang, 2025) View paper
- [47] RaCT: Ranking-aware Chain-of-Thought Optimization for LLMs (Liu Haowei, 2024) View paper
- Encoder-Decoder Ranking (2 papers)
- [45] RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses (Honglei Zhuang, 2023) View paper
- [50] Fine-tuning large language models for text ranking with listwise constraints (J Song, n.d.) View paper
- Bi-Encoder Architectures (1 papers)
- [49] Semi-Siamese Bi-encoder Neural Ranking Model Using Lightweight Fine-Tuning (Jung, 2022) View paper
- Large-Scale Deployment (2 papers)
- [5] Make large language model a better ranker (Zhi Zheng, 2024) View paper
- [26] Pre-trained language model based ranking in Baidu search (Zou LiXin, 2021) View paper
- Distillation and Compression (2 papers)
- [13] Efficient re-ranking with cross-encoders via early exit (Francesco Busolin, 2025) View paper
- [20] Best Practices for Distilling Large Language Models into BERT for Web Search Ranking (Ye DeZhi, 2025) View paper
- Two-Stage and Hybrid Ranking (2 papers)
- [21] A two-stage adaptation of large language models for text ranking (Long, 2024) View paper
- [27] Fine-Tuning a rerank Model Using Negative Sample Mining and Cross-Encoder for Enhanced Text Ranking Optimization (Jing Zhang, 2025) View paper

• Parameter-Efficient Fine-Tuning
- Low-Rank Adaptation (3 papers)
- [30] GLR: Graph Chain-of-Thought with LoRA Fine-Tuning and Confidence Ranking for Knowledge Graph Completion (Yifei Chen, 2025) View paper
- [32] Text to Trust: Evaluating Fine-Tuning and LoRA Trade-offs in Language Models for Unfair Terms of Service Detection (Noshitha Padma Pratyusha Juttu, 2025) View paper
- [41] Dynamic Adaptive Rank Space Exploration for Efficient Sentiment Analysis with Large Language Models (Ding Hongcheng, 2024) View paper
- Adapter and Prompt-Based Tuning (1 papers)
- [37] PARAMETER EFFICIENT FINE-TUNING AND OVERFITTING IN GPT LARGE LANGUAGE MODELS: A METRIC-BASED COMPARISON (Bohdan Pavlyshenko, 2025) View paper
- Hybrid Efficient Methods (1 papers)
- [35] Edinburgh Clinical NLP at SemEval-2024 Task 2: Fine-tune your model unless you have access to GPT-4 (Giwon Hong, 2024) View paper

• Domain Applications
- Automated Essay Scoring (3 papers)
- [7] Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking (Ruosong Yang, 2020) View paper
- [12] Rank-then-score: Enhancing large language models for automated essay scoring (Cai YiDa, 2025) View paper
- [18] Leveraging Large Language Models for Automated Chinese Essay Scoring (Haiyue Feng, 2024) View paper
- Clinical and Specialized NLP (3 papers)
- [28] CogBench: A Large Language Model Benchmark for Multilingual Speech-Based Cognitive Impairment Assessment (Feng Rui, 2025) View paper
- [36] Fine-tuning Large Language Model (LLM) Artificial Intelligence Chatbots in Ophthalmology and LLM-based evaluation using GPT-4 (Tan, 2024) View paper
- [43] Assertion Detection Large Language Model In-context Learning LoRA Fine-tuning (Ji, 2024) View paper
- Code Generation and Cybersecurity (1 papers)
- [46] HackMentor: Fine-Tuning Large Language Models for Cybersecurity (Jie Zhang, 2023) View paper
- Multimodal and Vision-Language Tasks (2 papers)
- [31] PreResQ-R1: Towards Fine-Grained Rank-and-Score Reinforcement Learning for Visual Quality Assessment via Preference-Response Disentangled Policy Optimization (Feng ZeHui, 2025) View paper
- [34] MotIF: Motion Instruction Fine-tuning (Minyoung Hwang, 2024) View paper
- Structured Prediction and Knowledge Graphs (2 papers)
- [25] Sort by Structure: Language Model Ranking as Dependency Probing (MÃ¼ller-Eberstein, 2022) View paper
- [44] Simplify Prompt Template and Optimum Label-Select for Few-Shot NER (Xiao Qin, 2024) View paper

• Training Data and Active Learning
- Data Augmentation and Synthesis (1 papers)
- [39] Augmented Relevance Datasets with Fine-Tuned Small LLMs (Deveaud, 2025) View paper
- Active Learning Strategies (1 papers)
- [40] Annotating Data for Fine-Tuning a Neural Ranker? Current Active Learning Strategies are not Better than Random Selection (Sophia Althammer, 2023) View paper
- Retrieval-Augmented Methods (1 papers)
- [6] Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning and Their Synergistic Fusion for Enhanced â¦ (G Budakoglu, 2025) View paper
- Academic Search and Citation Ranking (1 papers)
- [38] Learning to Research: Learning to Ranking the Similar Papers via BERT Fine-Tuning (Jiaxin Ye, 2023) View paper

• Reasoning and Chain-of-Thought Enhancement (1 papers)
- [42] Making Large Language Models Better Reasoners with Alignment (Wang Peiyi, 2023) View paper

Narrative

Core task: Harmonizing language model rankings through fine-tuning before evaluation. The field addresses how to align and optimize language models so that their outputs better reflect human preferences and task-specific quality criteria. The taxonomy reveals six main branches: Alignment and Preference Learning focuses on methods such as reinforcement learning from human feedback (RLHF) and preference ranking optimization to steer models toward desired behaviors; Ranking Architectures and Fine-Tuning Strategies explores diverse model designs and training regimes for scoring or reranking candidate outputs; Parameter-Efficient Fine-Tuning investigates lightweight adaptation techniques that reduce computational overhead; Domain Applications examines specialized use cases in areas like clinical NLP, essay scoring, and code generation; Training Data and Active Learning considers how to curate high-quality datasets and iteratively select informative examples; and Reasoning and Chain-of-Thought Enhancement targets improvements in multi-step inference and logical consistency. Representative works such as Aligning LLMs Survey[2] and RRHF[16] illustrate foundational alignment strategies, while Pre-trained Retrieval Ranking[8] and RankT5[45] exemplify architectural innovations in ranking.

Several active lines of work highlight key trade-offs and open questions. One central tension lies between sample efficiency and alignment quality: methods like Preference Ranking Optimization[1] and Tuning for Alignment[10] seek to maximize preference learning from limited human feedback, whereas Cost-Effective PPO[14] and RRHF without Tears[48] aim to reduce the computational burden of reinforcement learning. Another contrast emerges between pointwise scoring and listwise or pairwise ranking, with approaches such as Pairwise Ranking Prompting[3] and Better Ranker[5] exploring how to best capture relative quality judgments. Train before Test[0] sits within the Alignment and Preference Learning branch, emphasizing the importance of fine-tuning models on preference data prior to evaluation—a perspective closely aligned with Fine-tuning Human Preferences[17] and RRHF[16]. Compared to Better Ranker[5], which focuses on architectural refinements for ranking, Train before Test[0] underscores the procedural step of harmonizing rankings through targeted pre-evaluation tuning, thereby bridging alignment objectives with practical evaluation protocols.

Related Works in Same Category

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

1. Preference ranking optimization for human alignment

Authors: Huang Fei, Li, Yongbin, Li Ming-hao, Song, et al. (11 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large language models (LLMs) often contain misleading content, emphasizing the need to align them with human values to ensure secure AI systems. Reinforcement learning from human feedback (RLHF) has been employed to achieve this alignment. However, it encompasses two main drawbacks: (1) RLHF exhibits complexity, instability, and sensitivity to hyperparameters in contrast to SFT. (2) Despite massive trial-and-error, multiple sampling is reduced to pair-wise contrast, thus lacking contrasts from a...

Relationship Analysis

Both papers belong to the Reinforcement Learning from Human Feedback category, focusing on aligning language models with human preferences through reward-based optimization. The original paper (Train-before-Test) harmonizes model rankings by providing identical benchmark-specific fine-tuning before evaluation to assess model potential, while the candidate paper (PRO) proposes a supervised fine-tuning algorithm that directly optimizes language models using preference rankings of arbitrary length, extending beyond pairwise contrasts to multi-positional comparisons. The key difference is that the original paper focuses on evaluation methodology and ranking consistency across benchmarks, whereas the candidate paper focuses on a novel training algorithm (PRO) that replaces RLHF's trial-and-error process with direct optimization on ranked preference sequences.

---

2. Tuning for LLM alignment

Authors: Uday Kamath, Kevin Keenan, Garrett Somers, Sarah Sorenson | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â It is thus an essential part of LLM fine-tuning to align the model â are trained using annotated data to assess response safety â is supplied to both the original LM and the one that we tuned â

Relationship Analysis

Both papers belong to the Reinforcement Learning from Human Feedback category, focusing on aligning language models with human preferences. The candidate paper appears to address general LLM alignment through tuning with annotated safety data and reward modeling, overlapping with the original paper's concern for model evaluation and comparison. However, the original paper specifically investigates how fine-tuning before evaluation harmonizes model rankings across benchmarks, while the candidate paper focuses on the alignment tuning process itself rather than evaluation methodology or ranking consistency.

---

3. Rrhf: Rank responses to align language models with human feedback

Authors: H Yuan, Z Yuan, C Tan, W Wang | **Year/Venue**: 2023 | **URL**: View paper

Abstract

â Besides, fine-tuning language models with PPO needs to â named RRHF for large language models that can leverage â initial language model policy as Ï, the online language model â

Relationship Analysis

Both papers belong to the Reinforcement Learning from Human Feedback category, using RL techniques to align language models with human preferences. The original paper (Train-before-Test) focuses on harmonizing model rankings across benchmarks by providing identical fine-tuning before evaluation, while the candidate paper (RRHF) proposes a novel training paradigm that ranks sampled responses using conditional probabilities and ranking loss as an alternative to PPO for alignment. The key difference is that the original paper addresses evaluation methodology and ranking consistency, whereas RRHF addresses the training methodology for alignment itself.

---

4. Fine-tuning language models from human preferences

Authors: Ziegler, Daniel M., Daniel M. Ziegler, Stiennon, Nisan, et al. (23 authors total) | **Year/Venue**: 2019 | **URL**: View paper

Abstract

Reward learning enables the application of reinforcement learning (RL) to tasks where reward is defined by human judgment, building a model of reward by asking humans questions. Most work on reward learning has used simulated environments, but complex information about values is often expressed in natural language, and we believe reward learning for language is a key to making RL practical and safe for real-world tasks. In this paper, we build on advances in generative pretraining of language mo...

Relationship Analysis

Both papers belong to the Reinforcement Learning from Human Feedback category, using RL techniques to align language models with human preferences. The original paper focuses on harmonizing model rankings through standardized fine-tuning before evaluation across diverse benchmarks, while the candidate paper focuses on fine-tuning language models using reward models trained from human

preference comparisons for specific tasks like sentiment continuation and summarization. The key difference is that the original paper uses fine-tuning as an evaluation methodology to compare model potential, whereas the candidate paper uses RLHF as a training methodology to optimize models for human-preferred outputs on specific tasks.

---

5. RRHF: Rank Responses to Align Language Models with Human Feedback without tears

Authors: YUAN Zheng, Yuan, Hongyi, Tan Chuanqi, Wang Wei, et al. (8 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (RLHF) facilitates the alignment of large language models with human preferences, significantly enhancing the quality of interactions between humans and models. InstructGPT implements RLHF through several stages, including Supervised Fine-Tuning (SFT), reward model training, and Proximal Policy Optimization (PPO). However, PPO is sensitive to hyperparameters and requires multiple models in its standard implementation, making it hard to train and scale u...

Relationship Analysis

Both papers belong to the Reinforcement Learning from Human Feedback category, focusing on aligning language models with human preferences through reward-based optimization. While the original paper (Train-before-Test) harmonizes model rankings by applying standardized fine-tuning before evaluation across diverse benchmarks, the candidate paper (RRHF) proposes a simplified alternative to PPO that ranks multiple sampled responses using log probabilities and ranking loss to align with human preferences. The key difference is that the original paper addresses evaluation methodology and ranking consistency across benchmarks, whereas RRHF focuses on developing a more efficient training paradigm for alignment that avoids the complexity of PPO while achieving comparable performance.

Contributions Analysis

Overall novelty summary. The paper proposes a train-before-test methodology to harmonize language model rankings by applying identical benchmark-specific fine-tuning before evaluation. It resides in the Reinforcement Learning from Human Feedback leaf, which contains six papers addressing alignment through reward modeling and policy optimization. This leaf sits within the broader Alignment and Preference Learning branch, indicating a moderately populated research direction focused on steering models toward human preferences. The taxonomy shows five sibling leaves in alignment (Direct Preference Optimization, Ranking Feedback, Diverse Alignment, Surveys/Fairness), suggesting the field is actively exploring multiple paradigms for preference learning.

The taxonomy reveals neighboring branches addressing complementary concerns: Ranking Architectures and Fine-Tuning Strategies (seven leaves, covering prompt-based ranking, encoder-decoder designs, and deployment) focuses on architectural innovations, while Parameter-Efficient Fine-Tuning (three leaves) targets low-rank adaptation and adapter methods to reduce tuning costs. The paper's emphasis on fine-tuning before evaluation bridges alignment objectives with practical ranking protocols, connecting to both the alignment branch (where it resides) and the ranking architectures branch (which addresses model designs for scoring). The scope_note for RLHF explicitly excludes ranking-only approaches, clarifying that this work's contribution lies in evaluation methodology rather than novel ranking architectures.

Among thirty candidates examined, none clearly refute the three core contributions. The train-before-test methodology examined ten candidates with zero refutable overlaps, as did the empirical demonstration of ranking consistency and the perplexity-performance restoration findings. This suggests that within the limited search scope, the procedural innovation of harmonizing rankings through pre-evaluation fine-tuning appears relatively unexplored. However, the sibling papers in the RLHF leaf (e.g., Preference Ranking Optimization, RRHF) address related alignment objectives, indicating that while the specific evaluation protocol may be novel, the underlying fine-tuning paradigm is well-established in the alignment literature.

Based on the top-thirty semantic matches and taxonomy structure, the work introduces a methodological contribution to evaluation practices within a moderately active alignment subfield. The analysis does not cover exhaustive literature on benchmark design or meta-evaluation frameworks outside the alignment and ranking domains. The absence of refutable candidates reflects the limited search scope rather than definitive novelty, and a broader survey of evaluation methodology papers might reveal closer precedents.

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

Contribution 1: Train-before-test evaluation methodology

Description: The authors introduce train-before-test, a novel evaluation methodology that compares language models by fine-tuning each model on identical task-specific data before testing, rather than evaluating out-of-the-box performance. This approach aims to equalize model preparation and reveal inherent model potential.

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. Adapting large language models via reading comprehension

URL: View paper

Brief Assessment

Reading Comprehension Adaptation[68] focuses on continued pre-training with reading comprehension texts to adapt models to specific domains, not on fine-tuning models on task-specific data before benchmark evaluation as a methodology for comparing model potential.

---

2. Fine-tuning large language models for domain-specific machine translation

URL: View paper

Brief Assessment

Domain-Specific Translation[61] focuses on fine-tuning LLMs for domain-specific machine translation tasks, not on benchmark evaluation methodology. The candidate does not address train-before-test as an evaluation framework for comparing model potential across benchmarks.

---

3. Large language models: A survey

URL: View paper

Brief Assessment

LLM Survey[62] discusses pre-training and fine-tuning paradigms for language models but does not propose comparing models by fine-tuning them on identical task-specific data before evaluation. The survey focuses on general training approaches rather than evaluation methodologies for model comparison.

---

4. Scaling instruction-finetuned language models

URL: View paper

Brief Assessment

Scaling Instruction-Finetuned[60] focuses on instruction finetuning to improve model performance on unseen tasks, not on comparing models by fine-tuning each on identical task-specific data before testing. The candidate's methodology is about improving model capabilities through instruction tuning, not about equalizing model preparation for fair comparison.

---

5. Fine-tuning large language models with sequential instructions

URL: View paper

Brief Assessment

Sequential Instructions[66] focuses on fine-tuning models with multi-step sequential instructions to improve instruction-following behavior, not on equalizing model preparation for benchmark evaluation comparisons.

---

6. How abilities in large language models are affected by supervised fine-tuning data composition

URL: View paper

Brief Assessment

Fine-tuning Data Composition[63] focuses on supervised fine-tuning data composition for activating multiple abilities (math, code, general alignment) in LLMs, not on benchmark evaluation methodology or model comparison frameworks.

---

7. Dynamic adaptation of lora fine-tuning for efficient and task-specific optimization of large language models

URL: View paper

Brief Assessment

Dynamic LoRA Adaptation[67] focuses on optimizing LoRA hyperparameters during fine-tuning for task-specific adaptation, not on establishing a comparative evaluation methodology across models and benchmarks as the original paper does.

---

8. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and â¦

URL: View paper

Brief Assessment

Ultimate Fine-Tuning Guide[69] is a comprehensive technical review of fine-tuning methodologies and best practices. It does not propose train-before-test as an evaluation methodology for comparing models. The guide focuses on practical implementation of fine-tuning techniques rather than novel evaluation frameworks for model comparison.

---

9. Fine-tuning protein language models boosts predictions across diverse tasks

URL: View paper

Brief Assessment

Protein Language Models[64] focuses on fine-tuning protein language models for specific protein prediction tasks (e.g., secondary structure, disorder, subcellular location), not on evaluating language models by fine-tuning them on benchmark-specific data before testing to harmonize model rankings across diverse benchmarks.

---

10. Pixiu: A large language model, instruction data and evaluation benchmark for finance

URL: View paper

Brief Assessment

Pixiu[65] focuses on instruction-tuning LLMs for financial tasks using task-specific training data, but does not propose a comparative evaluation methodology for ranking models. The paper's approach is domain-specific fine-tuning for deployment, not a benchmarking methodology for comparing model potential across diverse tasks.

Contribution 2: Comprehensive empirical demonstration of ranking consistency

Description: The authors conduct extensive experiments showing that train-before-test produces remarkably consistent model rankings across diverse benchmarks, with average Kendall's tau increasing from 0.52 to 0.76, demonstrating that model potential rankings transfer gracefully across tasks.

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. NevIR: Negation in Neural Information Retrieval

URL: View paper

Brief Assessment

NevIR[59] focuses on negation understanding in neural information retrieval systems, not on model ranking consistency across benchmarks after fine-tuning.

---

2. How to Benchmark Vision Foundation Models for Semantic Segmentation?

URL: View paper

Brief Assessment

Vision Foundation Benchmarking[57] focuses on semantic segmentation model rankings under different fine-tuning settings, not general language model rankings across diverse benchmarks. The domains and methodologies are fundamentally different.

---

3. Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation

URL: View paper

Brief Assessment

Online Self-Training Stability[53] focuses on semantic segmentation model adaptation using online self-training methods, not on language model ranking consistency across benchmarks. The paper addresses stability issues in domain adaptation rather than benchmark evaluation methodology.

---

4. RaCT: Ranking-aware Chain-of-Thought Optimization for LLMs

URL: View paper

Brief Assessment

RaCT[47] focuses on text reranking using chain-of-thought prompting for information retrieval tasks, not on demonstrating consistent model rankings across diverse benchmarks after fine-tuning.

---

5. A comparative study on large language models' accuracy in cross-lingual professional terminology processing: An evaluation across multiple domains

URL: View paper

Brief Assessment

Cross-lingual Terminology Accuracy[52] focuses on evaluating LLM accuracy in translating professional terminology across domains and languages, not on model ranking consistency across benchmarks after fine-tuning.

---

6. A linearized framework and a new benchmark for model selection for fine-tuning

URL: View paper

Brief Assessment

Linearized Model Selection[55] focuses on model selection from a pre-trained model zoo for fine-tuning tasks, not on demonstrating ranking consistency across diverse benchmarks. The candidate addresses a different problem (selecting which pre-trained model to fine-tune) rather than evaluating ranking consistency of models across multiple benchmarks.

---

7. Merging models on the fly without retraining: A sequential approach to scalable continual model merging

URL: View paper

Brief Assessment

Merging Models Fly[58] focuses on sequential model merging techniques for combining fine-tuned models, not on evaluating model rankings across benchmarks or demonstrating ranking consistency after fine-tuning.

---

8. Die SuperGLEBer at GermEval 2025 shared tasks: Growing pains-when more isn't always better

URL: View paper

Brief Assessment

SuperGLEBer Growing Pains[54] focuses on analyzing German language understanding tasks and model rankings within a specific benchmark framework, not on demonstrating that train-before-test produces consistent model rankings across diverse benchmarks with improved Kendall's tau values.

---

9. Curriculum Direct Preference Optimization for Diffusion and Consistency Models

URL: View paper

Brief Assessment

Curriculum DPO[56] focuses on curriculum learning for text-to-image generation using difficulty-based pair sampling, not on demonstrating consistent model rankings across diverse benchmarks after fine-tuning.

---

10. CURLoRA: Stable LLM Continual Fine-Tuning and Catastrophic Forgetting Mitigation

URL: View paper

Brief Assessment

CURLoRA[51] focuses on continual fine-tuning methods for LLMs to mitigate catastrophic forgetting, not on evaluating model ranking consistency across benchmarks. The papers address entirely different research questions.

Contribution 3: Restoration of perplexity-performance alignment

Description: The authors show that train-before-test re-establishes the fundamental relationship between perplexity and downstream performance. Notably, pre-fine-tuning perplexity of base models predicts post-fine-tuning downstream performance, suggesting ranking consistency reflects inherent model potential rather than fine-tuning artifacts.

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. Bitnet: 1-bit pre-training for large language models

URL: View paper

Brief Assessment

BitNet[71] focuses on 1-bit quantization for LLMs and does not address perplexity-performance alignment or train-before-test methodology. The paper's contributions center on efficient model architecture design, not evaluation methodology or ranking consistency across benchmarks.

---

2. Quantifying the importance of data alignment in downstream model performance

URL: View paper

Brief Assessment

Data Alignment Importance[73] focuses on alignment between training and evaluation data predicting perplexity, not on the train-before-test methodology that restores perplexity-performance relationships across diverse benchmarks after fine-tuning.

---

3. Demystifying prompts in language models via perplexity estimation

URL: View paper

Brief Assessment

Demystifying Prompts Perplexity[79] examines perplexity as a predictor of prompt effectiveness in zero-shot settings, not as a predictor of post-fine-tuning downstream performance after task-specific adaptation. The original paper's contribution focuses on how train-before-test restores perplexity-performance alignment and shows pre-fine-tuning perplexity predicts post-fine-tuning performance, which is a different evaluation paradigm.

---

4. What is Wrong with Perplexity for Long-context Language Modeling?

URL: View paper

Brief Assessment

Perplexity Long-Context[75] focuses on identifying key tokens in long-context scenarios to improve perplexity's correlation with downstream performance, while the original paper addresses perplexity-performance alignment through train-before-test methodology that equalizes model preparation across benchmarks. These are fundamentally different approaches to different problems.

---

5. Monotonic paraphrasing improves generalization of language model prompting

URL: View paper

Brief Assessment

Monotonic Paraphrasing[77] focuses on paraphrasing prompts to lower perplexity for improved task performance, not on establishing fundamental relationships between pre-training perplexity and post-fine-tuning performance across model rankings.

---

6. Contrastive perplexity for controlled generation: An application in detoxifying large language models

URL: View paper

Brief Assessment

Contrastive Perplexity[72] focuses on using perplexity as a contrastive objective for controlled generation (detoxification), not on establishing relationships between perplexity and downstream task performance across models or benchmarks.

---

7. Rethinking the role of text complexity in language model pretraining

URL: View paper

Brief Assessment

Text Complexity Pretraining[74] examines how text complexity affects language modeling and downstream performance, but does not investigate the relationship between perplexity and downstream task performance across different models or evaluation methodologies. The candidate focuses on simplifying text while preserving content, not on perplexity as a predictor of performance.

---

8. Perplexed by perplexity: Perplexity-based pruning with small reference models

URL: View paper

Brief Assessment

Perplexed by Perplexity[78] focuses on using perplexity from small reference models for data pruning to improve downstream task performance, not on establishing fundamental relationships between pre-fine-tuning perplexity and post-fine-tuning performance across model rankings.

---

9. On the worst prompt performance of large language models

URL: View paper

Brief Assessment

Worst Prompt Performance[70] focuses on prompt robustness and variability in LLM performance across semantically equivalent queries, not on perplexity predicting downstream task performance after fine-tuning.

---

10. Language models scale reliably with over-training and on downstream tasks

URL: View paper

Brief Assessment

Over-Training Scaling[76] focuses on predicting downstream task performance from perplexity in the context of over-training regimes and compute-optimal scaling laws. The original paper examines how train-before-test (fine-tuning) restores perplexity-performance alignment that is lost under direct evaluation, which is a different methodological approach and research question.

Appendix: Text Similarity Detection

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

References

- [0] Train-before-Test Harmonizes Language Model Rankings View paper
- [1] Preference ranking optimization for human alignment View paper
- [2] Aligning large language models with human: A survey View paper
- [3] Large language models are effective text rankers with pairwise ranking prompting View paper
- [4] Fine-tuning language models to find agreement among humans with diverse preferences View paper
- [5] Make large language model a better ranker View paper
- [6] Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning and Their Synergistic Fusion for Enhanced â View paper
- [7] Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking View paper
- [8] Pre-trained language model-based retrieval and ranking for web search View paper
- [9] Towards an End-to-End Personal Fine-Tuning Framework for AI Value Alignment View paper
- [10] Tuning for LLM alignment View paper
- [11] Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models View paper
- [12] Rank-then-score: Enhancing large language models for automated essay scoring View paper
- [13] Efficient re-ranking with cross-encoders via early exit View paper
- [14] Is Crowdsourcing Breaking Your Bank? Cost-Effective Fine-Tuning of Pre-trained Language Models with Proximal Policy Optimization View paper
- [15] A study of implicit ranking unfairness in large language models View paper
- [16] Rrhf: Rank responses to align language models with human feedback View paper
- [17] Fine-tuning language models from human preferences View paper
- [18] Leveraging Large Language Models for Automated Chinese Essay Scoring View paper
- [19] Pangu-coder2: Boosting large language models for code with ranking feedback View paper
- [20] Best Practices for Distilling Large Language Models into BERT for Web Search Ranking View paper
- [21] A two-stage adaptation of large language models for text ranking View paper
- [22] Finetuning Large Language Model for Personalized Ranking View paper
- [23] Modeling relevance ranking under the pre-training and fine-tuning paradigm View paper
- [24] P3 Ranker: Mitigating the Gaps between Pre-training and Ranking Fine-tuning with Prompt-based Learning and Pre-finetuning View paper
- [25] Sort by Structure: Language Model Ranking as Dependency Probing View paper
- [26] Pre-trained language model based ranking in Baidu search View paper
- [27] Fine-Tuning a rerank Model Using Negative Sample Mining and Cross-Encoder for Enhanced Text Ranking Optimization View paper
- [28] CogBench: A Large Language Model Benchmark for Multilingual Speech-Based Cognitive Impairment Assessment View paper

- [29] Oracle-RLAIF: An Improved Fine-Tuning Framework for Multi-modal Video Models through Reinforcement Learning from Ranking Feedback View paper
- [30] GLR: Graph Chain-of-Thought with LoRA Fine-Tuning and Confidence Ranking for Knowledge Graph Completion View paper
- [31] PreResQ-R1: Towards Fine-Grained Rank-and-Score Reinforcement Learning for Visual Quality Assessment via Preference-Response Disentangled Policy Optimization View paper
- [32] Text to Trust: Evaluating Fine-Tuning and LoRA Trade-offs in Language Models for Unfair Terms of Service Detection View paper
- [33] Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning View paper
- [34] MotIF: Motion Instruction Fine-tuning View paper
- [35] Edinburgh Clinical NLP at SemEval-2024 Task 2: Fine-tune your model unless you have access to GPT-4 View paper
- [36] Fine-tuning Large Language Model (LLM) Artificial Intelligence Chatbots in Ophthalmology and LLM-based evaluation using GPT-4 View paper
- [37] PARAMETER EFFICIENT FINE-TUNING AND OVERFITTING IN GPT LARGE LANGUAGE MODELS: A METRIC-BASED COMPARISON View paper
- [38] Learning to Research: Learning to Ranking the Similar Papers via BERT Fine-Tuning View paper
- [39] Augmented Relevance Datasets with Fine-Tuned Small LLMs View paper
- [40] Annotating Data for Fine-Tuning a Neural Ranker? Current Active Learning Strategies are not Better than Random Selection View paper
- [41] Dynamic Adaptive Rank Space Exploration for Efficient Sentiment Analysis with Large Language Models View paper
- [42] Making Large Language Models Better Reasoners with Alignment View paper
- [43] Assertion Detection Large Language Model In-context Learning LoRA Fine-tuning View paper
- [44] Simplify Prompt Template and Optimum Label-Select for Few-Shot NER View paper
- [45] RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses View paper
- [46] HackMentor: Fine-Tuning Large Language Models for Cybersecurity View paper
- [47] RaCT: Ranking-aware Chain-of-Thought Optimization for LLMs View paper
- [48] RRHF: Rank Responses to Align Language Models with Human Feedback without tears View paper
- [49] Semi-Siamese Bi-encoder Neural Ranking Model Using Lightweight Fine-Tuning View paper
- [50] Fine-tuning large language models for text ranking with listwise constraints View paper
- [51] CURLoRA: Stable LLM Continual Fine-Tuning and Catastrophic Forgetting Mitigation View paper
- [52] A comparative study on large language models' accuracy in cross-lingual professional terminology processing: An evaluation across multiple domains View paper
- [53] Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation View paper
- [54] Die SuperGLEBer at GermEval 2025 shared tasks: Growing pains-when more isn't always better View paper
- [55] A linearized framework and a new benchmark for model selection for fine-tuning View paper
- [56] Curriculum Direct Preference Optimization for Diffusion and Consistency Models View paper
- [57] How to Benchmark Vision Foundation Models for Semantic Segmentation? View paper
- [58] Merging models on the fly without retraining: A sequential approach to scalable continual model merging View paper
- [59] NevIR: Negation in Neural Information Retrieval View paper
- [60] Scaling instruction-finetuned language models View paper
- [61] Fine-tuning large language models for domain-specific machine translation View paper
- [62] Large language models: A survey View paper
- [63] How abilities in large language models are affected by supervised fine-tuning data composition View paper
- [64] Fine-tuning protein language models boosts predictions across diverse tasks View paper
- [65] Pixiu: A large language model, instruction data and evaluation benchmark for finance View paper
- [66] Fine-tuning large language models with sequential instructions View paper
- [67] Dynamic adaptation of lora fine-tuning for efficient and task-specific optimization of large language models View paper
- [68] Adapting large language models via reading comprehension View paper
- [69] The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and â¦ View paper
- [70] On the worst prompt performance of large language models View paper
- [71] Bitnet: 1-bit pre-training for large language models View paper
- [72] Contrastive perplexity for controlled generation: An application in detoxifying large language models View paper
- [73] Quantifying the importance of data alignment in downstream model performance View paper
- [74] Rethinking the role of text complexity in language model pretraining View paper
- [75] What is Wrong with Perplexity for Long-context Language Modeling? View paper
- [76] Language models scale reliably with over-training and on downstream tasks View paper
- [77] Monotonic paraphrasing improves generalization of language model prompting View paper
- [78] Perplexed by perplexity: Perplexity-based pruning with small reference models View paper
- [79] Demystifying prompts in language models via perplexity estimation View paper