Novelty Assessment Report

Paper: Universal Model Routing for Efficient LLM Inference
PDF URL: https://openreview.net/pdf?id=ka82fvJ5f1
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Model routing is a simple technique for reducing the inference cost of large language models (LLMs), wherein one maintains a pool of candidate LLMs, and learns to route each prompt to the smallest feasible LLM. Existing works focus on learning a router for a fixed pool of LLMs. In this paper, we consider the problem of dynamic routing, where new, previously unobserved LLMs are available at test time. We propose UniRoute, a new approach to this problem that relies on representing each LLM as afeature vector, derived based on predictions on a set of representative prompts. Based on this, we detail two effective instantiations of UniRoute, relying on cluster-based routing and a learned cluster map respectively. We show that these are estimates of a theoretically optimal routing rule, and quantify their errors via an excess risk bound. Experiments on a range of public benchmarks show the effectiveness of UniRoute in routing amongst more than 30 unseen LLMs.

Core Task Landscape

This paper addresses: **Dynamic Model Routing for Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **30 categories**.

Taxonomy Overview

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

- **Query-Adaptive Model Selection and Routing**
- **Internal Model Architecture and Execution Optimization**
- **Model Merging and Multi-Model Integration**
- **Inference Scheduling and Resource Management**
- **Multi-Agent Coordination and Collaboration**
- **Continual Learning and Model Adaptation**
- **Auxiliary Optimization and Supporting Techniques**
- **Application-Specific Routing and Control**
- **Survey and Taxonomy Studies**

Complete Taxonomy Tree

- Dynamic Model Routing for Large Language Models Survey Taxonomy
- Query-Adaptive Model Selection and Routing
  - Universal and Cross-Model Routing Frameworks ★ (3 papers)
  - [0] Universal Model Routing for Efficient LLM Inference (Anon et al., 2026) View paper
  - [10] Tryage: Real-time, intelligent routing of user prompts to large language models (Surya N. Hari, 2023) View paper
  - [46] Universal LLM Routing with Correctness-Based Representation (W Jitkrittum, 2025) View paper
  - Confidence and Category-Aware Routing (2 papers)
  - [9] CARGO: A Framework for Confidence-Aware Routing of Large Language Models (Barrak, 2025) View paper
  - [26] LLM-based evaluation for dynamic routing among large language models (T Tsuchiya, 2025) View paper
  - Lookahead and Response-Informed Routing (1 papers)
  - [11] Lookahead Routing for Large Language Models (Huang Canbin, 2025) View paper
  - Multi-Objective and Cost-Aware Routing (4 papers)
  - [2] MixLLM: Dynamic Routing in Mixed Large Language Models (Wang Xin-yuan, 2025) View paper
  - [22] Dynamic Quality-Latency Aware Routing for LLM Inference in Wireless Edge-Device Networks (Rui Bao, 2025) View paper
  - [30] Tensoropera router: A multi-model router for efficient llm inference (Dimitris Stripelis, 2024) View paper
  - [32] Cost-effective online multi-llm selection with versatile reward models (Li Jin, 2024) View paper
  - Online Learning and Adaptive Selection (3 papers)
  - [24] Dynamically Learned Test-Time Model Routing in Language Model Zoos with Service Level Guarantees (H WoisetschlÃ¤ger, 2025) View paper
  - [43] Online Multi-LLM Selection via Contextual Bandits under Unstructured Context Evolution (Manhin Poon, 2025) View paper
  - [50] Convergence-aware online model selection with time-increasing bandits (Y Xia, 2024) View paper
  - Reasoning-Aware and Complexity-Based Routing (4 papers)
  - [13] SynapseRoute: An Auto-Route Switching Framework on Dual-State Large Language Model (Zhang Wen-cheng, 2025) View paper
  - [29] Route to Reason: Adaptive Routing for LLM and Reasoning Strategy Selection (Pan Zhihong, 2025) View paper
  - [31] Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router (Shao, 2025) View paper
  - [36] Adaptivellm: A framework for selecting optimal cost-efficient llm for code-generation based on cot length (Liu Fang, 2025) View paper

    - Domain-Specific and Application-Oriented Routing (2 papers)
    - [17] Dynamicroutegpt: A real-time multi-vehicle dynamic navigation framework based on large language models (Ziai Zhou, 2024) View paper
    - [47] CarbonCall: Sustainability-Aware Function Calling for Large Language Models on Edge Devices (Karatzas, 2025) View paper
- Internal Model Architecture and Execution Optimization
    - Dynamic Layer and Depth Routing (2 papers)
    - [3] Fast Thinking for Large Language Models (Zheng Haoyu, 2025) View paper
    - [7] Radial Networks: Dynamic Layer Routing for High-Performance Large Language Models (Dotzel, 2024) View paper
    - Mixture-of-Experts and Expert Routing (3 papers)
    - [6] Routing experts: Learning to route dynamic experts in existing multi-modal large language models (Wu Qiong, 2025) View paper
    - [45] D2MoE: Dual Routing and Dynamic Scheduling for Efficient On-Device MoE-based LLM Serving (Wang Haodong, 2025) View paper
    - [48] T-REX: Mixture-of-Rank-One-Experts with Semantic-aware Intuition for Multi-task Large Language Model Finetuning (Zhang Rong-yu, 2024) View paper
    - Quantization and Precision Adaptation (2 papers)
    - [39] DP-LLM: Runtime Model Adaptation with Dynamic Layer-wise Precision Assignment (Kwon Sang-Woo, 2025) View paper
    - [41] QLLMS: Quantization-Adaptive LLM Scheduling for Partially Informed Edge Serving Systems (Miao Hu, 2025) View paper
- Model Merging and Multi-Model Integration
    - Task-Specific Model Merging and Quantized Integration (1 papers)
    - [1] 1bit-Merging: Dynamic Quantized Merging for Large Language Models (Liu, 2025) View paper
    - Expert Switching and Memory Management (2 papers)
    - [8] Me-switch: A memory-efficient expert switching framework for large language models (Liu Jing, 2024) View paper
    - [33] Engine-Agnostic Model Hot-Swapping for Cost-Effective LLM Inference (Radostin Stoyanov, 2025) View paper
- Inference Scheduling and Resource Management
    - Dynamic Scheduling and Request Management (1 papers)
    - [5] Llumnix: Dynamic Scheduling for Large Language Model Serving (Sun Biao, 2024) View paper
    - Parallelism Switching and Training Optimization (2 papers)
    - [14] A learning rate path switching training paradigm for version updates of large language models (Wang Zhihao (ç□å¿□æµ©), 2024) View paper
    - [16] Enabling Parallelism Hot Switching for Efficient Training of Large Language Models (Hao Ge, 2024) View paper
    - Data Selection and Sampling (2 papers)
    - [15] Lead: Iterative data selection for efficient llm instruction tuning (Lin, 2025) View paper
    - [44] Reinforce-ada: An adaptive sampling framework for reinforce-style llm training (Xiong Wei, 2025) View paper
    - Serverless and Deployment Challenges (1 papers)
    - [18] Illuminating the Hidden Challenges of Serverless LLM Systems (A Samanta, 2025) View paper
- Multi-Agent Coordination and Collaboration
    - Dynamic Task Routing and Agent Collaboration (2 papers)
    - [20] Parallelism Meets Adaptiveness: Scalable Documents Understanding in Multi-Agent LLM Systems (Wu, 2025) View paper
    - [37] Rcr-router: Efficient role-aware context routing for multi-agent llm systems with structured memory (Liu Jun, 2025) View paper
    - Modular Task Decomposition and Hierarchical Coordination (1 papers)
    - [21] Modular Task Decomposition and Dynamic Collaboration in Multi-Agent Systems Driven by Large Language Models (Wu Di, 2025) View paper
    - Knowledge-Aware Orchestration and Cognitive Synergy (1 papers)
    - [28] OSC: Cognitive Orchestration through Dynamic Knowledge Alignment in Multi-Agent LLM Collaboration (Jusheng Zhang, 2025) View paper
- Continual Learning and Model Adaptation
    - Continual Instruction Tuning and Task Switching (1 papers)
    - [23] SwitchCIT: Switching for Continual Instruction Tuning of Large Language Models (Wu Xinbo, 2024) View paper
    - Multimodal Continual Learning (1 papers)
    - [4] MLLM-CL: Continual Learning for Multimodal Large Language Models (Zhao Hongbo, 2025) View paper
    - Few-Shot Adaptation of Multi-Modal Models (1 papers)
    - [27] Few-shot adaptation of multi-modal foundation models: A survey (Fan Liu, 2024) View paper
- Auxiliary Optimization and Supporting Techniques
    - Retrieval-Augmented Generation and Adaptive Retrieval (1 papers)
    - [42] PAIRS: Parametric-Verified Adaptive Information Retrieval and Selection for Efficient RAG (Chen Wang, 2025) View paper
    - Hierarchical Representation and Query-Adaptive Structures (1 papers)
    - [12] VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos (Zi-yang Wang, 2024) View paper
    - Kernel Design and Bayesian Optimization (1 papers)
    - [38] Adaptive Kernel Design for Bayesian Optimization Is a Piece of CAKE with LLMs (Yin Feng, 2025) View paper
    - Knowledge Distillation and Model Compression (1 papers)
    - [34] GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model (Shicheng Tan, 2023) View paper
    - Performance Optimization and Deployment Strategies (1 papers)
    - [35] Performance Optimization of Large Language Models (LLMs) in Web Applications (Sivakumar, 2024) View paper
- Application-Specific Routing and Control
    - Robotics and Real-Time Control (2 papers)
    - [19] Real-Time Anomaly Detection and Reactive Planning with Large Language Models (Rohan Sinha, 2024) View paper
    - [40] LAMS: LLM-Driven Automatic Mode Switching for Assistive Teleoperation (Yiran Tao, 2025) View paper
    - Wireless and Edge Network Routing (1 papers)
    - [49] LLM-Driven Adaptive 6G-Ready Wireless Body Area Networks: Survey and Framework (Mohammad Jalili Torkamani, 2025) View paper

- Survey and Taxonomy Studies (1 papers)
  - [25] Doing More with Less--Implementing Routing Strategies in Large Language Model-Based Systems: An Extended Survey (Bouvard, 2025) View paper

Narrative

Core task: dynamic model routing for large language models. The field addresses how to intelligently select, switch, or coordinate among multiple LLMs or model variants to balance quality, latency, cost, and resource constraints. The taxonomy reveals several major branches: Query-Adaptive Model Selection and Routing focuses on matching individual queries to appropriate models based on difficulty or domain; Internal Model Architecture and Execution Optimization examines within-model mechanisms such as mixture-of-experts and layer-skipping; Model Merging and Multi-Model Integration explores combining parameters or outputs from diverse models; Inference Scheduling and Resource Management tackles system-level orchestration and load balancing; Multi-Agent Coordination and Collaboration studies how multiple LLM agents can work together; Continual Learning and Model Adaptation considers evolving model capabilities over time; and Application-Specific Routing tailors routing strategies to domains like code generation or video understanding. Works such as Tryage[10] and MixLLM[2] illustrate early query-adaptive approaches, while Llumnix[5] exemplifies scheduling and resource management at scale.

Particularly active lines of work center on learning universal routing policies that generalize across diverse model pools and query distributions, trading off the need for task-specific tuning against the desire for broad applicability. Universal Model Routing[0] sits squarely in this universal and cross-model routing cluster, aiming to develop routing frameworks that adapt to heterogeneous LLM ensembles without extensive retraining. Nearby efforts like Universal LLM Routing[46] share this ambition of generality, while Tryage[10] represents an earlier, more heuristic approach to cascading models by difficulty. A central open question is how to efficiently learn routing policies that remain robust as new models enter the pool or as query distributions shift, with some works exploring online learning (e.g., contextual bandits) and others leveraging distillation or meta-learning. Universal Model Routing[0] contributes to this landscape by emphasizing cross-model generalization, positioning itself as a step toward routing systems that require minimal per-model customization.

Related Works in Same Category

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

1. Tryage: Real-time, intelligent routing of user prompts to large language models

Authors: Surya N. Hari, Thomson, Matt, Matt Thomson, S. N. Hari | **Year/Venue**: 2023 | **URL**: View paper

Abstract

The introduction of the transformer architecture and the self-attention mechanism has led to an explosive production of language models trained on specific downstream tasks and data domains. With over 200, 000 models in the Hugging Face ecosystem, users grapple with selecting and optimizing models to suit multifaceted workflows and data domains while addressing computational, security, and recency concerns. There is an urgent need for machine learning frameworks that can eliminate the burden of ...

Relationship Analysis

Both papers belong to the Universal and Cross-Model Routing Frameworks category, focusing on routing systems that generalize across heterogeneous LLM pools. They overlap in addressing dynamic model selection where new LLMs may be encountered at test time, using learned representations to route queries to appropriate models. The key difference is that the original paper (UniRoute) uses cluster-based routing with per-cluster error vectors and provides theoretical excess risk bounds, while the candidate paper (Tryage) employs a thalamic-inspired perceptive router that predicts downstream model performance and incorporates multi-objective optimization with user-defined constraints (model size, recency, security) through a Pareto front exploration framework.

2. Universal LLM Routing with Correctness-Based Representation

Authors: W Jitkrittum, H Narasimhan, AS Rawat | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Large language modelsâ significant advances in capabilities are accompanied by significant â¦ We now introduce the dynamic model routing problem. Suppose Hall denotes the set of all â¦

⚠ Similarity Notice

These papers appear to be highly similar or variants of the same work. Both propose 'UniRoute' for universal LLM routing using correctness-based representations and cluster-based strategies, with nearly identical technical approaches including K-means clustering on training sets, per-cluster error vectors for LLM representation, and the same routing formulation. The core methodology, experimental setup, and theoretical framework (Proposition 1 on optimal routing) are essentially identical, suggesting these are likely different versions of the same paper.

Contributions Analysis

Overall novelty summary. The paper introduces UniRoute, a framework for dynamic model routing that handles previously unseen LLMs at test time by representing each model as a feature vector derived from predictions on representative prompts. This work sits within the 'Universal and Cross-Model Routing Frameworks' leaf of the taxonomy, which contains only three papers total including this one. The leaf focuses specifically on routing systems designed to generalize across heterogeneous or unseen LLM pools, distinguishing it from confidence-based or category-specific routing methods. This represents a relatively sparse research direction within the broader query-adaptive model selection landscape.

The taxonomy reveals that UniRoute's immediate neighbors include confidence-aware routing methods and lookahead-based approaches, which rely on different signals for routing decisions. The broader 'Query-Adaptive Model Selection and Routing' branch encompasses seven distinct sub-areas, from multi-objective optimization to reasoning-aware routing, suggesting a fragmented field with multiple competing paradigms. UniRoute's emphasis on cross-model generalization through learned representations positions it at the intersection of universal routing and representation learning, diverging from methods that require per-model training or task-specific tuning. The scope note explicitly excludes confidence-based methods, clarifying that UniRoute's feature-vector approach represents a distinct technical strategy.

Among the three contributions analyzed, the formalization of the dynamic LLM pool routing problem shows the most substantial prior work overlap: one refutable candidate was identified among ten examined papers. The UniRoute framework itself and the cluster-based instantiations appear more novel, with zero refutable candidates found among four and ten examined papers respectively. However, the literature search examined only 24 total candidates through top-K semantic search and citation expansion, representing a limited sample of the field. The single refutable case suggests that aspects of the problem formalization may have been explored previously, though the specific instantiations and theoretical guarantees appear less anticipated by prior work.

Based on this limited search scope covering 24 candidates across three contributions, UniRoute appears to occupy a relatively under-explored niche within dynamic model routing. The sparse population of its taxonomy leaf and the low refutation rate suggest meaningful novelty in the cross-model generalization approach, though the analysis cannot rule out relevant work outside the top-K semantic

matches examined. The field structure indicates active parallel development in related but distinct routing paradigms, positioning UniRoute as one of several competing frameworks rather than a definitive solution.

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

Contribution 1: UniRoute framework for dynamic model routing

Description: The authors introduce UniRoute, a novel routing framework that represents each LLM as a feature vector based on its prediction errors on representative prompts. This enables routing among previously unseen LLMs without retraining the router.

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. Incoherence as Oracle-less Measure of Error in LLM-Based Code Generation

URL: View paper

Brief Assessment

Incoherence Oracle-less Measure[67] addresses error detection in LLM-generated code using behavioral disagreement between samples, not model routing or feature vector representations for routing decisions.

---

2. HIPPD: Brain-Inspired Hierarchical Information Processing for Personality Detection

URL: View paper

Brief Assessment

HIPPD[69] focuses on personality detection using brain-inspired hierarchical processing with dynamic routing among specialized lightweight models for personality pattern recognition. This is fundamentally different from UniRoute's LLM routing framework based on prediction error feature vectors for efficient inference across unseen language models.

---

3. Benchmarking and Improving LLM Robustness for Personalized Generation

URL: View paper

Brief Assessment

Personalized Generation Robustness[71] focuses on evaluating LLM robustness for personalized generation tasks using error-based metrics, not on dynamic model routing using feature vector representations from prediction errors.

---

4. Peering Inside the Black Box: Uncovering LLM Errors in Optimization Modelling through Component-Level Evaluation

URL: View paper

Brief Assessment

Component-Level Evaluation[70] focuses on evaluating LLM-generated optimization formulations through component-level metrics (precision, recall, RMSE), not on routing among multiple LLMs. The paper addresses optimization modeling evaluation, not dynamic model selection or routing frameworks.

Contribution 2: Cluster-based routing instantiations with theoretical guarantees

Description: The authors propose two concrete implementations of UniRoute using unsupervised and supervised prompt clustering. They provide theoretical analysis showing these methods estimate the optimal routing rule and derive an excess risk bound quantifying their approximation error.

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. Geographical cluster-based routing in sensing-covered networks

URL: View paper

Brief Assessment

The candidate paper focuses on geographical cluster-based routing in sensor networks with sensing coverage constraints, not general reinforcement learning frameworks for language model routing with excess risk bounds.

---

2. Multiply Robust Estimation for Local Distribution Shifts with Multiple Domains

URL: View paper

Brief Assessment

Multiply Robust Estimation[55] focuses on tabular data analysis with distribution shifts across population segments, not LLM routing. The clustering approach serves a fundamentally different purpose (handling local distribution shifts) rather than routing queries to language models.

---

3. Adv-SSL: Adversarial Self-Supervised Representation Learning with Theoretical Guarantees

URL: View paper

Brief Assessment

Adv-SSL[60] focuses on self-supervised representation learning for transfer learning, not on routing mechanisms for LLM inference. The theoretical guarantees in Adv-SSL[60] concern clustering in embedding spaces for classification tasks, which is fundamentally different from routing rules and excess risk bounds for model selection.

---

4. Group Distributionally Robust Dataset Distillation with Risk Minimization

URL: View paper

Brief Assessment

Group Distributionally Robust[54] focuses on dataset distillation with clustering for robustness across subgroups, not on routing mechanisms or excess risk bounds for model selection systems.

---

5. Uneven Clustering Routing Protocols for Multi-Hop Cognitive Radio Sensor Networks: General Design Principles and an Illustrative Example

URL: View paper

Brief Assessment

Uneven Clustering Routing[51] focuses on cognitive radio sensor networks with physical cluster formation for energy efficiency, not LLM routing with theoretical excess risk bounds for model selection.

---

6. Improving Approximate and Exact Approaches Based on Decision Diagrams and Dynamic Programming for Combinatorial Optimization

URL: View paper

Brief Assessment

Decision Diagrams Optimization[59] focuses on combinatorial optimization using decision diagrams and dynamic programming for problems like scheduling and resource allocation, not on LLM routing with cluster-based methods and excess risk bounds.

---

7. Provable pathways: Learning multiple tasks over multiple paths

URL: View paper

Brief Assessment

Provable Pathways[56] focuses on multitask learning with compositional representations over pathways in a supernet, not on LLM routing. The clustering here groups tasks for representation learning, not routing decisions among LLMs based on prompt clusters.

---

8. Compressive statistical learning with random feature moments

URL: View paper

Brief Assessment

Compressive Statistical Learning[52] focuses on compressive clustering for data compression and dimensionality reduction in statistical learning, not on routing decisions among language models with theoretical excess risk bounds.

---

9. Can Evolutionary Clustering Have Theoretical Guarantees?

URL: View paper

Brief Assessment

Evolutionary Clustering Guarantees[57] focuses on clustering algorithms with theoretical guarantees for partitioning data sets, not on routing among language models. The technical domains are fundamentally different.

---

10. Clustering-based Meta Bayesian Optimization with Theoretical Guarantee

URL: View paper

Brief Assessment

Clustering-based Meta Bayesian[53] focuses on meta-learning for Bayesian optimization with function clustering, not LLM routing. The technical domains are fundamentally different - one addresses black-box optimization with surrogate models, the other addresses prompt-to-model routing decisions.

Contribution 3: Formalization of dynamic LLM pool routing problem

Description: The authors formally define the problem of routing when the set of available LLMs can change dynamically at test time, extending beyond the standard static pool assumption in prior work. This includes a meta-distribution over LLM pools and characterization of the Bayes-optimal routing rule.

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. CARGO: A Framework for Confidence-Aware Routing of Large Language Models

URL: View paper

Brief Assessment

CARGO[9] focuses on confidence-aware routing with category-specific regressors for model selection, not on formalizing the dynamic pool routing problem with meta-distributions over LLM pools or characterizing Bayes-optimal routing rules as in the original paper.

---

2. Tryage: Real-time, intelligent routing of user prompts to large language models

URL: View paper

Brief Assessment

Tryage[10] focuses on context-aware routing for model selection from a library based on prompt analysis and user constraints, but does not formalize the dynamic pool routing problem with meta-distributions over LLM pools or characterize Bayes-optimal routing rules as done in the original paper.

---

3. Routing experts: Learning to route dynamic experts in existing multi-modal large language models

URL: View paper

Brief Assessment

Routing Experts[6] focuses on dynamic expert routing within multi-modal LLMs by treating each layer as an expert, not on routing across a dynamic pool of different LLM models as in the original paper.

---

4. RadialRouter: Structured Representation for Efficient and Robust Large Language Models Routing

URL: View paper

Brief Assessment

RadialRouter[63] focuses on improving routing effectiveness through a novel transformer-based architecture and query-LLM relationship modeling, but does not formalize the dynamic pool routing problem with meta-distributions or characterize Bayes-optimal rules as the original paper does.

---

5. Efficient dynamic ensembling for multiple LLM experts

URL: View paper

Brief Assessment

Dynamic Ensembling[65] focuses on sequential knowledge transfer through MDP-based routing where models build upon previous responses, rather than formalizing the meta-distribution framework and Bayes-optimal routing characterization for dynamic pools.

---

6. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms

URL: View paper

Brief Assessment

RouterEval[66] focuses on benchmarking router performance across large pools of LLMs but does not formalize the dynamic routing problem with meta-distributions over LLM pools or characterize Bayes-optimal routing rules as the original paper does.

---

7. Mixture of experts in large language models

URL: View paper

Brief Assessment

Mixture of Experts[61] is a general survey paper on MoE architectures in LLMs, focusing on expert gating mechanisms and model capacity scaling. It does not address the specific problem of routing when the set of available LLMs can change dynamically at test time, which is the core novelty claim of the original paper.

---

8. Routellm: Learning to route llms with preference data

URL: View paper

Brief Assessment

RouteLLM[64] focuses on routing between two models (strong/weak) using preference data, not on the general problem of dynamic pools where the set of available LLMs changes at test time with a meta-distribution over pools.

---

9. Routing experts: Learning to route dynamic experts in multi-modal large language models

URL: View paper

Brief Assessment

The candidate paper focuses on dynamic expert routing within multi-modal large language models (mixture of experts architecture), not on the problem of routing across a dynamically changing pool of independent LLMs as formalized in the original paper.

---

10. Universal LLM Routing with Correctness-Based Representation

URL: View paper

Prior Art Analysis

Universal LLM Routing[46] demonstrates that the dynamic routing problem was formalized prior to the original paper. The candidate paper explicitly defines the problem where 'new, previously unobserved llms are available at test time' and formalizes it with a meta-distribution over LLM pools. The candidate provides the same mathematical framework including the constrained optimization problem and characterization of the Bayes-optimal routing rule (Proposition 1), which matches the original paper's formulation. Both papers use identical notation and problem setup, with the candidate paper being published at a workshop (SCOPE - ICLR 2025) that would have preceded or coincided with the original submission timeline.

Evidence

Evidence 1 - **Rationale**: Both papers claim to formalize the dynamic routing problem where new LLMs appear at test time, using nearly identical terminology. - **Original**: we formalise the problem setting of model routing with adynamicpool of llms - **Candidate**: in this paper, we consider the problem of dynamic routing, where new, previously unobserved llms are available at test time

Evidence 2 - **Rationale**: The mathematical formalization is virtually identical between the two papers, including notation, set definitions, and problem structure, demonstrating prior work exists. - **Original**: let hall denote the set of all possible llm predictors, where for simplicity we assume |hall|<+∞ . let h . = 2hall denote the set of all subsets of hall. let htr ={h (1) tr , . . . , h(m) tr } ∈hdenote the set of llm predictors observed during training. during evaluation, we seek to route amongst ht... - **Candidate**: suppose hall denotes the set of all possible feasible llm predictors, where we assume |hall|< +∞. let h . = 2 hall denote the set of all subsets of h. let htr = {h(1),...,h (m)} ∈ h denote the set of m llm predictors observed during training. during evaluation, we seek to route amongst the llm predic...

Evidence 3 - **Rationale**: Both papers introduce the meta-distribution concept in identical terms, showing the original paper's claimed novelty of this formalization was already present in prior work. - **Original**: we assume that the set of training llms htr is itself drawn from somemeta-distribution h over h. rather than perform well on thespecificset htr, we would like to generalise toanyset of llms drawn fromh - **Candidate**: we assume that the set of llms observed during training is itself drawn from some meta-distribution h over h. rather than perform well on the specific set of training llms, we would like to generalise to any set of llms drawn from h

Evidence 4 - **Rationale**: The characterization of the Bayes-optimal routing rule is presented identically in both papers, including the proposition statement, mathematical formulation, and Lagrange multiplier approach, demonstrating this theoretical contribution existed prior to the original paper. - **Original**: proposition 1(optimal dynamic routing).under a mild regularity condition on p, for any input x∈x , llm pool h∈h , and budget b >0, there exists a lagrange multiplier λh ≥0 such that the optimal dynamic routerr ∗ for the constrained optimization in(5)is r∗(x,h) = argmin m∈[|h|] h ey|x h ℓ(x,y, h(m)) ... - **Candidate**: proposition 1 (optimal dynamic routing). under a mild regularity condition on p, for any input x ∈x, llm candidate set h ∈h, and budget b >0, the optimal dynamic router r∗for the constrained optimization in (2) is r∗(x,h) = argmin m∈[|h|] [ey|x [ℓ(x,y,h(m))] + λh ·c(h(m))]

Appendix: Text Similarity Detection

Textual similarity detection checked 25 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Universal LLM Routing with Correctness-Based Representation

Detected in: Core Task (sibling), Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Universal Model Routing for Efficient LLM Inference View paper
- [1] 1bit-Merging: Dynamic Quantized Merging for Large Language Models View paper
- [2] MixLLM: Dynamic Routing in Mixed Large Language Models View paper
- [3] Fast Thinking for Large Language Models View paper
- [4] MLLM-CL: Continual Learning for Multimodal Large Language Models View paper
- [5] Llumnix: Dynamic Scheduling for Large Language Model Serving View paper
- [6] Routing experts: Learning to route dynamic experts in existing multi-modal large language models View paper
- [7] Radial Networks: Dynamic Layer Routing for High-Performance Large Language Models View paper
- [8] Me-switch: A memory-efficient expert switching framework for large language models View paper
- [9] CARGO: A Framework for Confidence-Aware Routing of Large Language Models View paper
- [10] Tryage: Real-time, intelligent routing of user prompts to large language models View paper
- [11] Lookahead Routing for Large Language Models View paper
- [12] VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos View paper

- [13] SynapseRoute: An Auto-Route Switching Framework on Dual-State Large Language Model View paper
- [14] A learning rate path switching training paradigm for version updates of large language models View paper
- [15] Lead: Iterative data selection for efficient llm instruction tuning View paper
- [16] Enabling Parallelism Hot Switching for Efficient Training of Large Language Models View paper
- [17] Dynamicroutegpt: A real-time multi-vehicle dynamic navigation framework based on large language models View paper
- [18] Illuminating the Hidden Challenges of Serverless LLM Systems View paper
- [19] Real-Time Anomaly Detection and Reactive Planning with Large Language Models View paper
- [20] Parallelism Meets Adaptiveness: Scalable Documents Understanding in Multi-Agent LLM Systems View paper
- [21] Modular Task Decomposition and Dynamic Collaboration in Multi-Agent Systems Driven by Large Language Models View paper
- [22] Dynamic Quality-Latency Aware Routing for LLM Inference in Wireless Edge-Device Networks View paper
- [23] SwitchCIT: Switching for Continual Instruction Tuning of Large Language Models View paper
- [24] Dynamically Learned Test-Time Model Routing in Language Model Zoos with Service Level Guarantees View paper
- [25] Doing More with Less--Implementing Routing Strategies in Large Language Model-Based Systems: An Extended Survey View paper
- [26] LLM-based evaluation for dynamic routing among large language models View paper
- [27] Few-shot adaptation of multi-modal foundation models: A survey View paper
- [28] OSC: Cognitive Orchestration through Dynamic Knowledge Alignment in Multi-Agent LLM Collaboration View paper
- [29] Route to Reason: Adaptive Routing for LLM and Reasoning Strategy Selection View paper
- [30] Tensoropera router: A multi-model router for efficient llm inference View paper
- [31] Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router View paper
- [32] Cost-effective online multi-llm selection with versatile reward models View paper
- [33] Engine-Agnostic Model Hot-Swapping for Cost-Effective LLM Inference View paper
- [34] GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model View paper
- [35] Performance Optimization of Large Language Models (LLMs) in Web Applications View paper
- [36] Adaptivellm: A framework for selecting optimal cost-efficient llm for code-generation based on cot length View paper
- [37] Rcr-router: Efficient role-aware context routing for multi-agent llm systems with structured memory View paper
- [38] Adaptive Kernel Design for Bayesian Optimization Is a Piece of CAKE with LLMs View paper
- [39] DP-LLM: Runtime Model Adaptation with Dynamic Layer-wise Precision Assignment View paper
- [40] LAMS: LLM-Driven Automatic Mode Switching for Assistive Teleoperation View paper
- [41] QLLMS: Quantization-Adaptive LLM Scheduling for Partially Informed Edge Serving Systems View paper
- [42] PAIRS: Parametric-Verified Adaptive Information Retrieval and Selection for Efficient RAG View paper
- [43] Online Multi-LLM Selection via Contextual Bandits under Unstructured Context Evolution View paper
- [44] Reinforce-ada: An adaptive sampling framework for reinforce-style llm training View paper
- [45] D2MoE: Dual Routing and Dynamic Scheduling for Efficient On-Device MoE-based LLM Serving View paper
- [46] Universal LLM Routing with Correctness-Based Representation View paper
- [47] CarbonCall: Sustainability-Aware Function Calling for Large Language Models on Edge Devices View paper
- [48] T-REX: Mixture-of-Rank-One-Experts with Semantic-aware Intuition for Multi-task Large Language Model Finetuning View paper
- [49] LLM-Driven Adaptive 6G-Ready Wireless Body Area Networks: Survey and Framework View paper
- [50] Convergence-aware online model selection with time-increasing bandits View paper
- [51] Uneven Clustering Routing Protocols for Multi-Hop Cognitive Radio Sensor Networks: General Design Principles and an Illustrative Example View paper
- [52] Compressive statistical learning with random feature moments View paper
- [53] Clustering-based Meta Bayesian Optimization with Theoretical Guarantee View paper
- [54] Group Distributionally Robust Dataset Distillation with Risk Minimization View paper
- [55] Multiply Robust Estimation for Local Distribution Shifts with Multiple Domains View paper
- [56] Provable pathways: Learning multiple tasks over multiple paths View paper
- [57] Can Evolutionary Clustering Have Theoretical Guarantees? View paper
- [58] Geographical cluster-based routing in sensing-covered networks View paper
- [59] Improving Approximate and Exact Approaches Based on Decision Diagrams and Dynamic Programming for Combinatorial Optimization View paper
- [60] Adv-SSL: Adversarial Self-Supervised Representation Learning with Theoretical Guarantees View paper
- [61] Mixture of experts in large language models View paper
- [62] Routing experts: Learning to route dynamic experts in multi-modal large language models View paper
- [63] RadialRouter: Structured Representation for Efficient and Robust Large Language Models Routing View paper
- [64] Routellm: Learning to route llms with preference data View paper
- [65] Efficient dynamic ensembling for multiple LLM experts View paper
- [66] Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms View paper
- [67] Incoherence as Oracle-less Measure of Error in LLM-Based Code Generation View paper
- [68] Enhancing Requirements Engineering with Large Language Models: From Elicitation and Classification to Traceability, Ambiguity Management and API â View paper
- [69] HIPPD: Brain-Inspired Hierarchical Information Processing for Personality Detection View paper
- [70] Peering Inside the Black Box: Uncovering LLM Errors in Optimization Modelling through Component-Level Evaluation View paper
- [71] Benchmarking and Improving LLM Robustness for Personalized Generation View paper