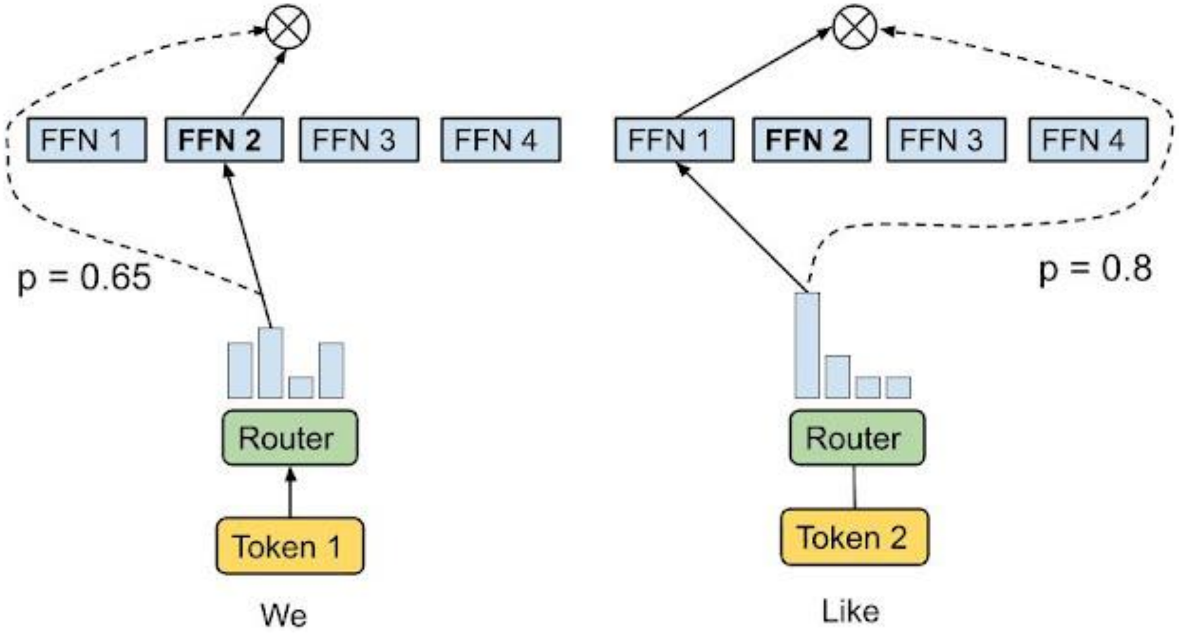
# RouterEval: A Comprehensive Benchmark for Routing LLMs to Explore Model-level Scaling Up in LLMs

Zhongzhan Huang
Sun Yat-sen University

**Project Page**: https://github.com/MilkThink-Lab/RouterEval
[Data, Code, Paper, Baselines and Tutorial]

2025.03.09

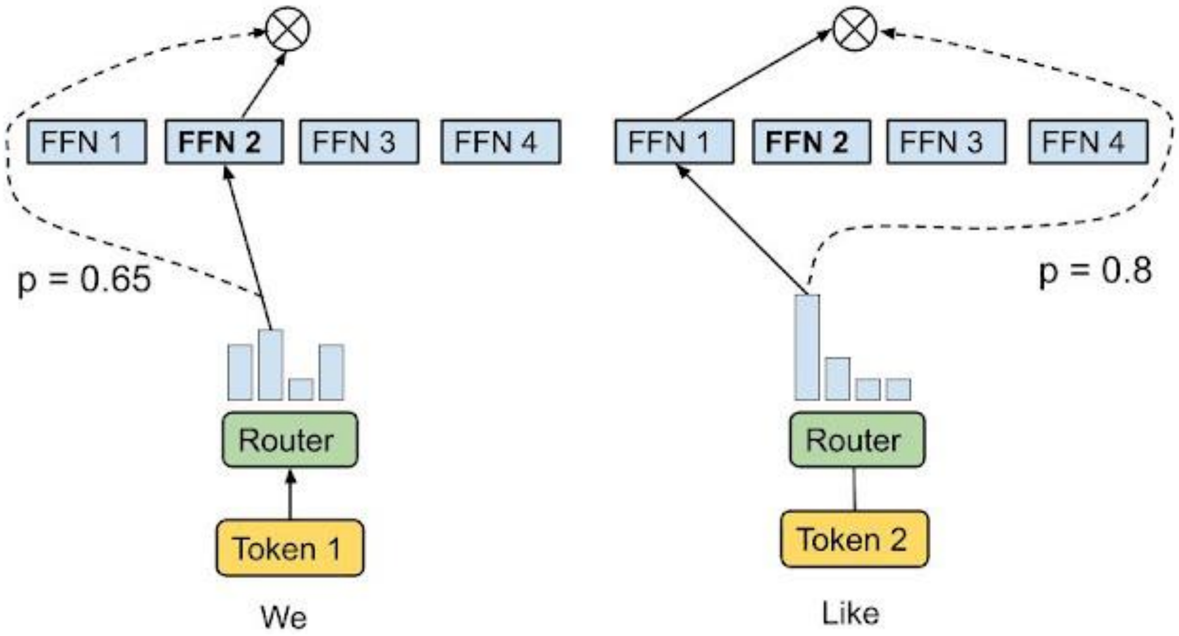
# What’s your first impression of Mixture-of-Experts (**MoE)**? When you hear this term?



Traditional Mixture-of-Experts (MoE)

https://research.google/blog/mixture-of-experts-with-expert-choice-routing/

# What's your first impression of Mixture-of-Experts (**MoE**)? When you hear this term?

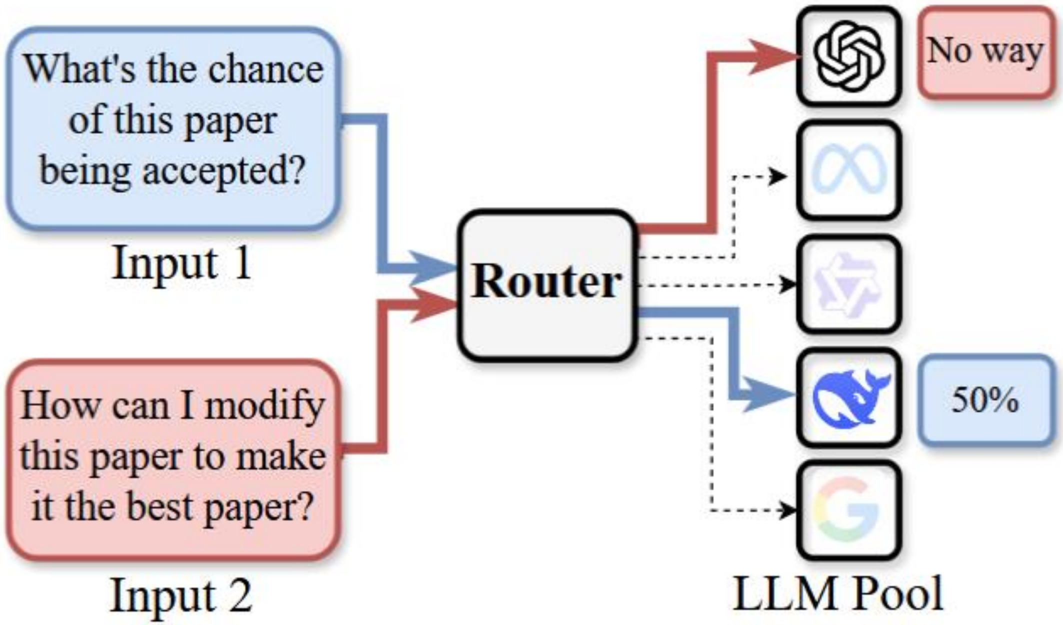


Traditional Mixture-of-Experts (MoE)

https://research.google/blog/mixture-of-experts-with-expert-choice-routing/

My First Impression

# Routing LLMs —— Model-level “MoE” ?


What's the chance of this paper being accepted?
Input 1
How can I modify this paper to make it the best paper?
Input 2
Router
No way
50%
LLM Pool

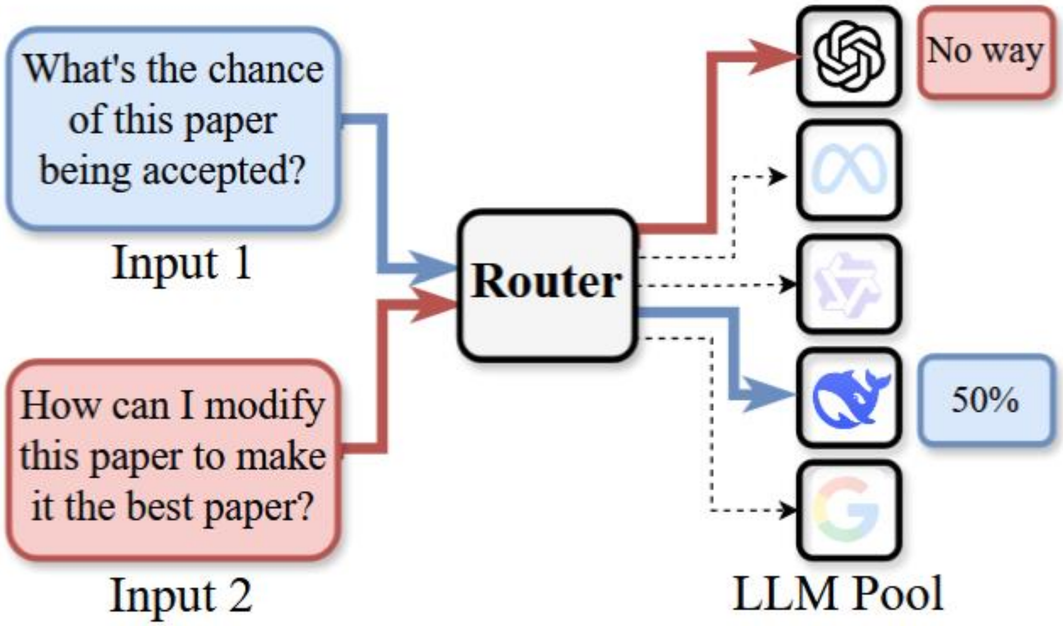
The Overview of Routing LLMs

Given an input, a capable router assigns it to the appropriate LLM for processing.

- Goal:
  - **High** accuracy, ✓
  - **Low** computational cost, ⚠
  - **Low** hallucinations, etc. ⚠

- Task Type:
  - ~Classification
  - ~Agents Scheduling
  - ~Recommender system
  - ~Retrieval problem
    - Retrieval data for LLM ✗
    - Retrieval LLM for data ✓

# Routing LLMs —— Model-level “MoE” ?


What's the chance of this paper being accepted?
Input 1
How can I modify this paper to make it the best paper?
Input 2
Router
No way
50%
LLM Pool

The Overview of Routing LLMs

Given an input, a capable router assigns it to the appropriate LLM for processing.

- Goal:
  - **High** accuracy, ✓
  - **Low** computational cost, ⚠
  - **Low** hallucinations, etc. ⚠

- Classical Example:
  - Machine-Human Chatting Handoff
    - Intelligent customer service
  - Large/Small LLM switch
    - GPT-4~GPT-3 switch for trade-off between Perf. and cost.
    - Sometimes we might feel that LLMs seem to be acting a bit silly

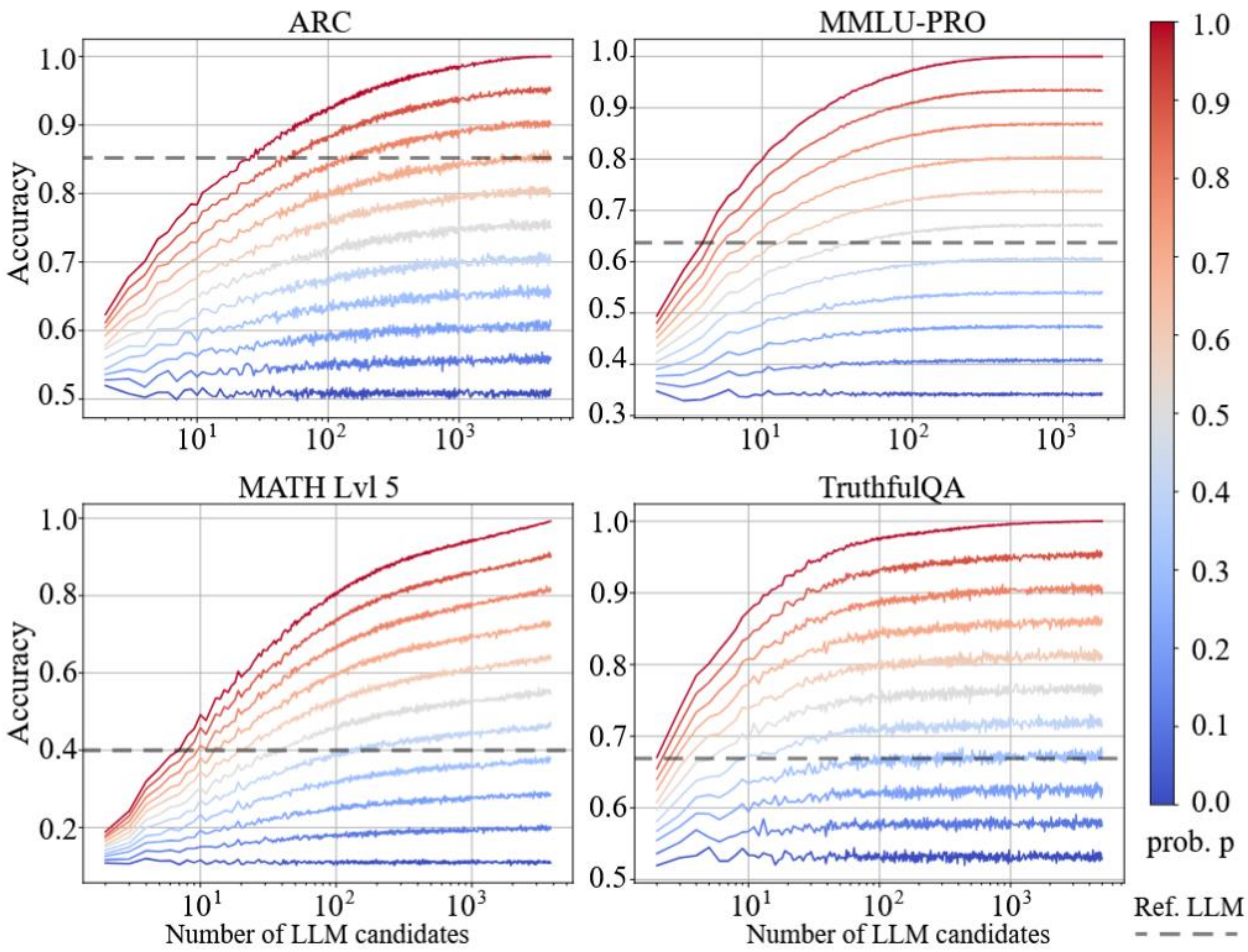
- Potential of Routing LLM
    - Prompt-to-Leaderboard @ UCB (2025.02.20, 10+days ago) ~ Arena Rank 1 ≈ Grok3 (with 200000+GPU)
- Current Shortcomings of Routing LLM
    - Without unified benchmark (everyone did their own thing)
    - Existing benchmarks (Limited LLMs/evaluations and ~~Open-source~~ Proprietary )

- Potential of Routing LLM
    - Prompt-to-Leaderboard @ UCB (2025.02.20, 10+days ago) ~ Arena Rank 1 ≈ Grok3 (with 200000+GPU)
- Current Shortcomings of Routing LLM
    - Without unified benchmark (everyone did their own thing)
    - Existing benchmarks (Limited LLMs/evaluations and ~~Open-source~~ Proprietary )

- Our Contribution

  [Collect/ Organize /Open-source] **8,567** LLMs ~ **12** benchmarks ~ **201,715,850** performance records

    - We find the model-level scaling up phenomenon in LLMs (With capable router: Performance $\propto$ #LLMs)
    - Constructing the RouterEval benchmark tailored for router design



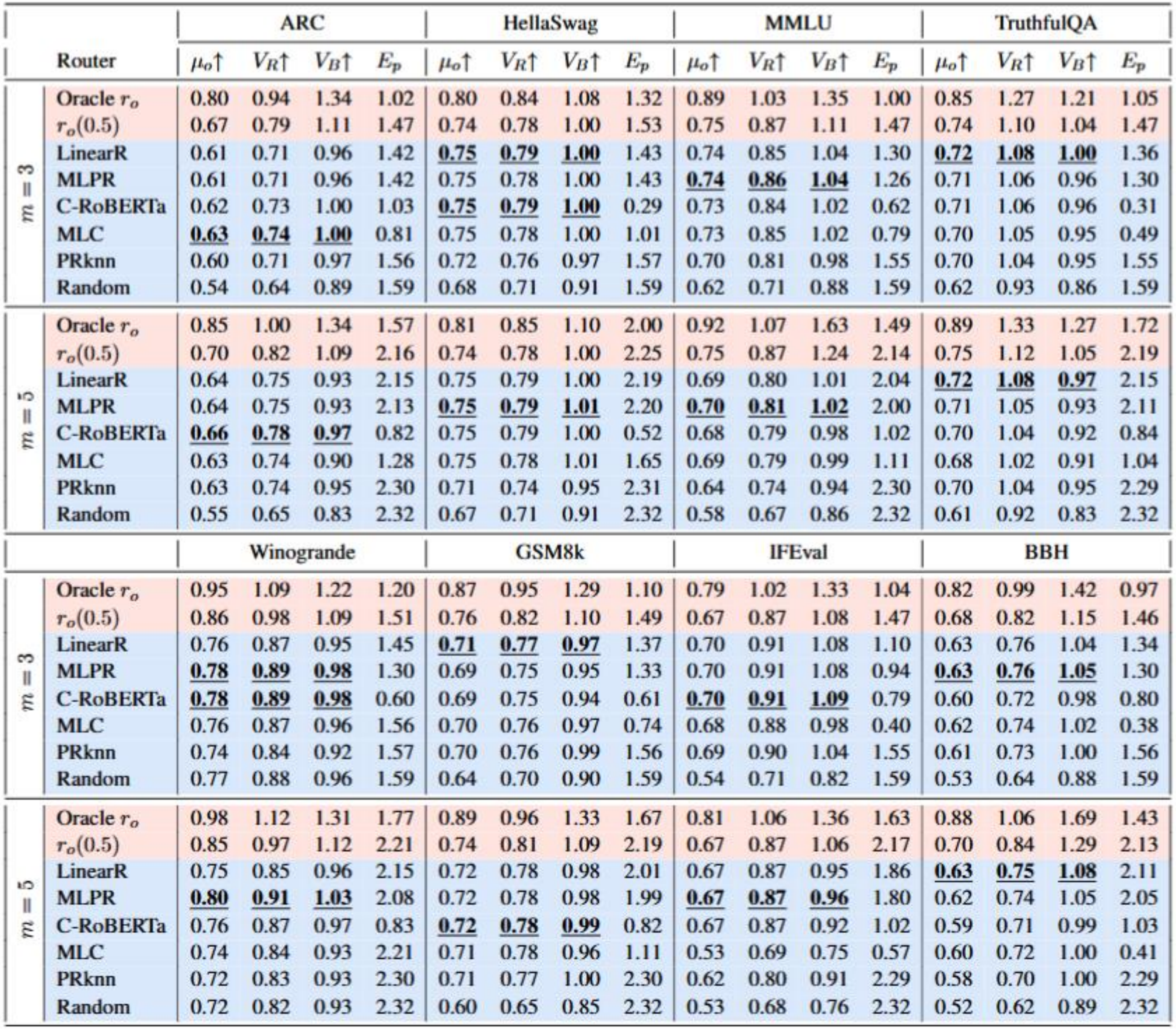
p → 1: oracle router

P → 0: random router

- **Model-level Scaling up**: With capable router, Performance ∝ #LLMs

- **Weak Candidates Can Also be Promising:** 5 weak LLMs (perf. ≤ 0.3) can achieve 0.95 ≥ GPT-4 on MMLU

- **Small Number of Candidates is Enough:** 3~10 candidates seems most cost-effective



| | Router | ARC | | | | HellaSwag | | | | MMLU | | | | TruthfulQA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mu_o\uparrow$ | $V_R\uparrow$ | $V_B\uparrow$ | $E_p$ | $\mu_o\uparrow$ | $V_R\uparrow$ | $V_B\uparrow$ | $E_p$ | $\mu_o\uparrow$ | $V_R\uparrow$ | $V_B\uparrow$ | $E_p$ | $\mu_o\uparrow$ | $V_R\uparrow$ | $V_B\uparrow$ | $E_p$ |
| $m = 3$ | Oracle $r_o$ | 0.80 | 0.94 | 1.34 | 1.02 | 0.80 | 0.84 | 1.08 | 1.32 | 0.89 | 1.03 | 1.35 | 1.00 | 0.85 | 1.27 | 1.21 | 1.05 |
| | $r_o(0.5)$ | 0.67 | 0.79 | 1.11 | 1.47 | 0.74 | 0.78 | 1.00 | 1.53 | 0.75 | 0.87 | 1.11 | 1.47 | 0.74 | 1.10 | 1.04 | 1.47 |
| | LinearR | 0.61 | 0.71 | 0.96 | 1.42 | **0.75** | **0.79** | **1.00** | 1.43 | 0.74 | 0.85 | 1.04 | 1.30 | **0.72** | **1.08** | **1.00** | 1.36 |
| | MLPR | 0.61 | 0.71 | 0.96 | 1.42 | 0.75 | 0.78 | 1.00 | 1.43 | **0.74** | **0.86** | **1.04** | 1.26 | 0.71 | 1.06 | 0.96 | 1.30 |
| | C-RoBERTa | 0.62 | 0.73 | 1.00 | 1.03 | **0.75** | **0.79** | **1.00** | 0.29 | 0.73 | 0.84 | 1.02 | 0.62 | 0.71 | 1.06 | 0.96 | 0.31 |
| | MLC | **0.63** | **0.74** | **1.00** | 0.81 | 0.75 | 0.78 | 1.00 | 1.01 | 0.73 | 0.85 | 1.02 | 0.79 | 0.70 | 1.05 | 0.95 | 0.49 |
| | PRknn | 0.60 | 0.71 | 0.97 | 1.56 | 0.72 | 0.76 | 0.97 | 1.57 | 0.70 | 0.81 | 0.98 | 1.55 | 0.70 | 1.04 | 0.95 | 1.55 |
| | Random | 0.54 | 0.64 | 0.89 | 1.59 | 0.68 | 0.71 | 0.91 | 1.59 | 0.62 | 0.71 | 0.88 | 1.59 | 0.62 | 0.93 | 0.86 | 1.59 |
| $m = 5$ | Oracle $r_o$ | 0.85 | 1.00 | 1.34 | 1.57 | 0.81 | 0.85 | 1.10 | 2.00 | 0.92 | 1.07 | 1.63 | 1.49 | 0.89 | 1.33 | 1.27 | 1.72 |
| | $r_o(0.5)$ | 0.70 | 0.82 | 1.09 | 2.16 | 0.74 | 0.78 | 1.00 | 2.25 | 0.75 | 0.87 | 1.24 | 2.14 | 0.75 | 1.12 | 1.05 | 2.19 |
| | LinearR | 0.64 | 0.75 | 0.93 | 2.15 | 0.75 | 0.79 | 1.00 | 2.19 | 0.69 | 0.80 | 1.01 | 2.04 | **0.72** | **1.08** | **0.97** | 2.15 |
| | MLPR | 0.64 | 0.75 | 0.93 | 2.13 | **0.75** | **0.79** | **1.01** | 2.20 | **0.70** | **0.81** | **1.02** | 2.00 | 0.71 | 1.05 | 0.93 | 2.11 |
| | C-RoBERTa | **0.66** | **0.78** | **0.97** | 0.82 | 0.75 | 0.79 | 1.00 | 0.52 | 0.68 | 0.79 | 0.98 | 1.02 | 0.70 | 1.04 | 0.92 | 0.84 |
| | MLC | 0.63 | 0.74 | 0.90 | 1.28 | 0.75 | 0.78 | 1.01 | 1.65 | 0.69 | 0.79 | 0.99 | 1.11 | 0.68 | 1.02 | 0.91 | 1.04 |
| | PRknn | 0.63 | 0.74 | 0.95 | 2.30 | 0.71 | 0.74 | 0.95 | 2.31 | 0.64 | 0.74 | 0.94 | 2.30 | 0.70 | 1.04 | 0.95 | 2.29 |
| | Random | 0.55 | 0.65 | 0.83 | 2.32 | 0.67 | 0.71 | 0.91 | 2.32 | 0.58 | 0.67 | 0.86 | 2.32 | 0.61 | 0.92 | 0.83 | 2.32 |

| | | Winogrande | | | | GSM8k | | | | IFEval | | | | BBH | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m = 3$ | Oracle $r_o$ | 0.95 | 1.09 | 1.22 | 1.20 | 0.87 | 0.95 | 1.29 | 1.10 | 0.79 | 1.02 | 1.33 | 1.04 | 0.82 | 0.99 | 1.42 | 0.97 |
| | $r_o(0.5)$ | 0.86 | 0.98 | 1.09 | 1.51 | 0.76 | 0.82 | 1.10 | 1.49 | 0.67 | 0.87 | 1.08 | 1.47 | 0.68 | 0.82 | 1.15 | 1.46 |
| | LinearR | 0.76 | 0.87 | 0.95 | 1.45 | **0.71** | **0.77** | **0.97** | 1.37 | 0.70 | 0.91 | 1.08 | 1.10 | 0.63 | 0.76 | 1.04 | 1.34 |
| | MLPR | **0.78** | **0.89** | **0.98** | 1.30 | 0.69 | 0.75 | 0.95 | 1.33 | 0.70 | 0.91 | 1.08 | 0.94 | **0.63** | **0.76** | **1.05** | 1.30 |
| | C-RoBERTa | **0.78** | **0.89** | **0.98** | 0.60 | 0.69 | 0.75 | 0.94 | 0.61 | **0.70** | **0.91** | **1.09** | 0.79 | 0.60 | 0.72 | 0.98 | 0.80 |
| | MLC | 0.76 | 0.87 | 0.96 | 1.56 | 0.70 | 0.76 | 0.97 | 0.74 | 0.68 | 0.88 | 0.98 | 0.40 | 0.62 | 0.74 | 1.02 | 0.38 |
| | PRknn | 0.74 | 0.84 | 0.92 | 1.57 | 0.70 | 0.76 | 0.99 | 1.56 | 0.69 | 0.90 | 1.04 | 1.55 | 0.61 | 0.73 | 1.00 | 1.56 |
| | Random | 0.77 | 0.88 | 0.96 | 1.59 | 0.64 | 0.70 | 0.90 | 1.59 | 0.54 | 0.71 | 0.82 | 1.59 | 0.53 | 0.64 | 0.88 | 1.59 |
| $m = 5$ | Oracle $r_o$ | 0.98 | 1.12 | 1.31 | 1.77 | 0.89 | 0.96 | 1.33 | 1.67 | 0.81 | 1.06 | 1.36 | 1.63 | 0.88 | 1.06 | 1.69 | 1.43 |
| | $r_o(0.5)$ | 0.85 | 0.97 | 1.12 | 2.21 | 0.74 | 0.81 | 1.09 | 2.19 | 0.67 | 0.87 | 1.06 | 2.17 | 0.70 | 0.84 | 1.29 | 2.13 |
| | LinearR | 0.75 | 0.85 | 0.96 | 2.15 | 0.72 | 0.78 | 0.98 | 2.01 | 0.67 | 0.87 | 0.95 | 1.86 | **0.63** | **0.75** | **1.08** | 2.11 |
| | MLPR | **0.80** | **0.91** | **1.03** | 2.08 | 0.72 | 0.78 | 0.98 | 1.99 | **0.67** | **0.87** | **0.96** | 1.80 | 0.62 | 0.74 | 1.05 | 2.05 |
| | C-RoBERTa | 0.76 | 0.87 | 0.97 | 0.83 | **0.72** | **0.78** | **0.99** | 0.82 | 0.67 | 0.87 | 0.92 | 1.02 | 0.59 | 0.71 | 0.99 | 1.03 |
| | MLC | 0.74 | 0.84 | 0.93 | 2.21 | 0.71 | 0.78 | 0.96 | 1.11 | 0.53 | 0.69 | 0.75 | 0.57 | 0.60 | 0.72 | 1.00 | 0.41 |
| | PRknn | 0.72 | 0.83 | 0.93 | 2.30 | 0.71 | 0.77 | 1.00 | 2.30 | 0.62 | 0.80 | 0.91 | 2.29 | 0.58 | 0.70 | 1.00 | 2.29 |
| | Random | 0.72 | 0.82 | 0.93 | 2.32 | 0.60 | 0.65 | 0.85 | 2.32 | 0.53 | 0.68 | 0.76 | 2.32 | 0.52 | 0.62 | 0.89 | 2.32 |

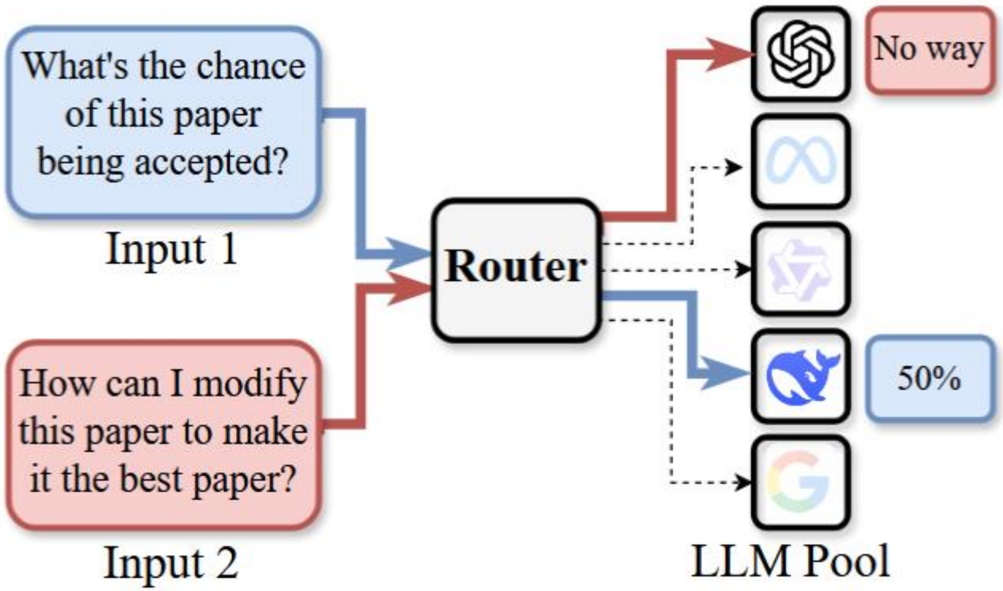
🟦 Existing routers still have significant room for improvement.

🟦 Classification bias is a major issue.

🟥 Fast Experiments (Even CPU only)

🟥 **Try** Few-shot learning, data augmentation, recommender systems, regularization methods, and pre-training, etc!

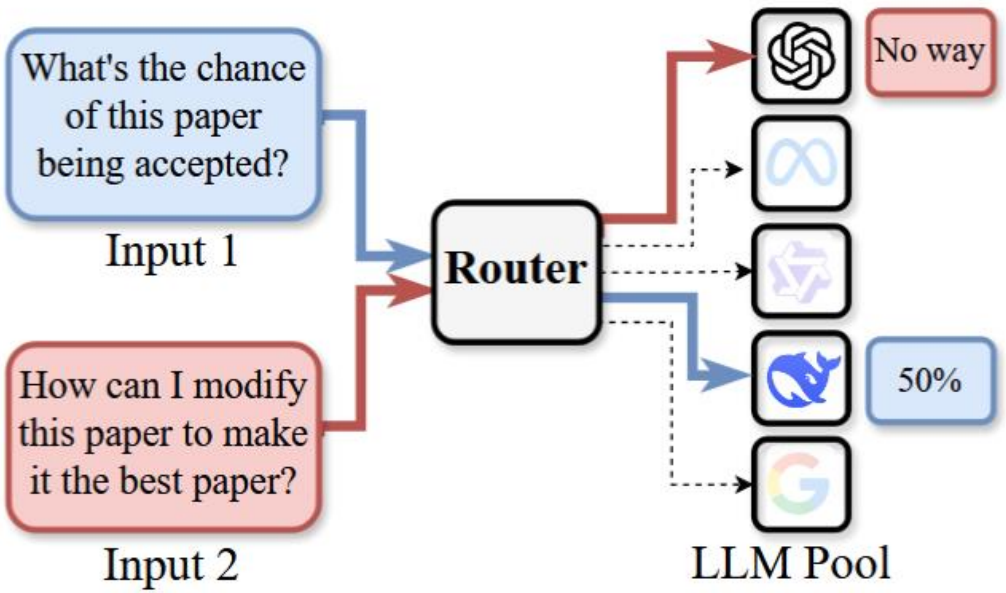
## Relationship and Differences with Existing Paradigms


What's the chance of this paper being accepted?
Input 1
How can I modify this paper to make it the best paper?
Input 2
Router
No way
50%
LLM Pool

The Overview of Routing LLMs

- **Recommender Systems**
  - Routing LLM is a specialized recommender system (input~user / LLM~item / Perf. Record~ history)
- **LLM Ensemble**
  - Typically post-decision, while Routing LLM ~ pre-decision
- **LLM Fusion**
  - Typically homogeneous LLMs, while Routing LLM can involve heterogeneous LLMs
- **Mixture-of-Experts (MoE)**
  - Routing LLM is a model-level Mixture of Experts

## Relationship and Differences with Existing Paradigms


What's the chance of this paper being accepted?
Input 1
How can I modify this paper to make it the best paper?
Input 2
Router
No way
50%
LLM Pool

The Overview of Routing LLMs

- **Recommender Systems**
  - Routing LLM is a specialized recommender system (input~user / LLM~item / Perf. Record~ history)
- **LLM Ensemble**
  - Typically post-decision, while Routing LLM ~ pre-decision
- **LLM Fusion**
  - Typically homogeneous LLMs, while Routing LLM can involve heterogeneous LLMs
- **Mixture-of-Experts (MoE)**
  - Routing LLM is a model-level Mixture of Experts

Routing LLM is compatible with all the above paradigms (viewed as LLMs in the candidate pool).

- Limitation and Challenge
  - **Severe lack of data**
    - performance records are typically proprietary and expensive
  - **How to maintain router performance with multiple candidates?**
    - lack of large data / multi-class issue
  - **RouterEval currently focuses only on performance**
    - But can easily expand to computational cost, hallucination rate, etc
    - However, performance alone is still far from sufficiently usable
  - **The challenge of deployment**
    - Large batch input for industrial deployment (avg. cost ↓)
    - 3~10 candidates + weak candidates seem “enough”

# RouterEval: A Comprehensive Benchmark for Routing LLMs to Explore Model-level Scaling Up in LLMs

Zhongzhan Huang[1], Guoming Ling[1], Vincent S. Liang[2], Yupei Lin[1], Yandong Chen[1], Shanshan Zhong[1], Hefeng Wu[1], Liang Lin[1]
[1]Sun Yat-sen University, [2]Purdue University

**Project Page**: https://github.com/MilkThink-Lab/RouterEval
[Data, Code, Paper, Baselines and Tutorial]