# Harnessing the Power of Multiple Minds: Lessons Learned from LLM Routing

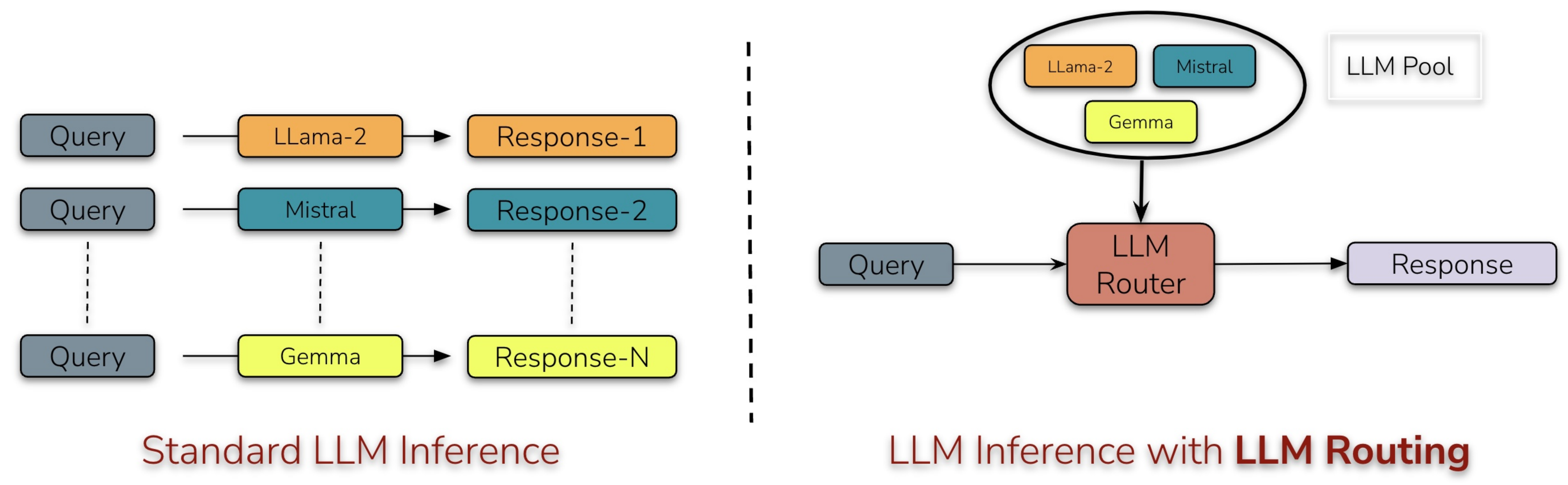
KV Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
vaibhav.kuchibhotla@mbzuai.ac.ae

NAACL 2024

## Introduction

- Large language models (LLMs) demonstrate remarkable capabilities in many natural language generation and understanding tasks.
- Different LLMs exhibit diverse capabilities [2].
- It is natural to ask how to harness these diverse capabilities of LLMs efficiently.
- Towards this end, we investigate the feasibility of developing an **LLM Routing** model, *which efficiently directs an input query to the most suitable single LLM from a pool of LLMs.*

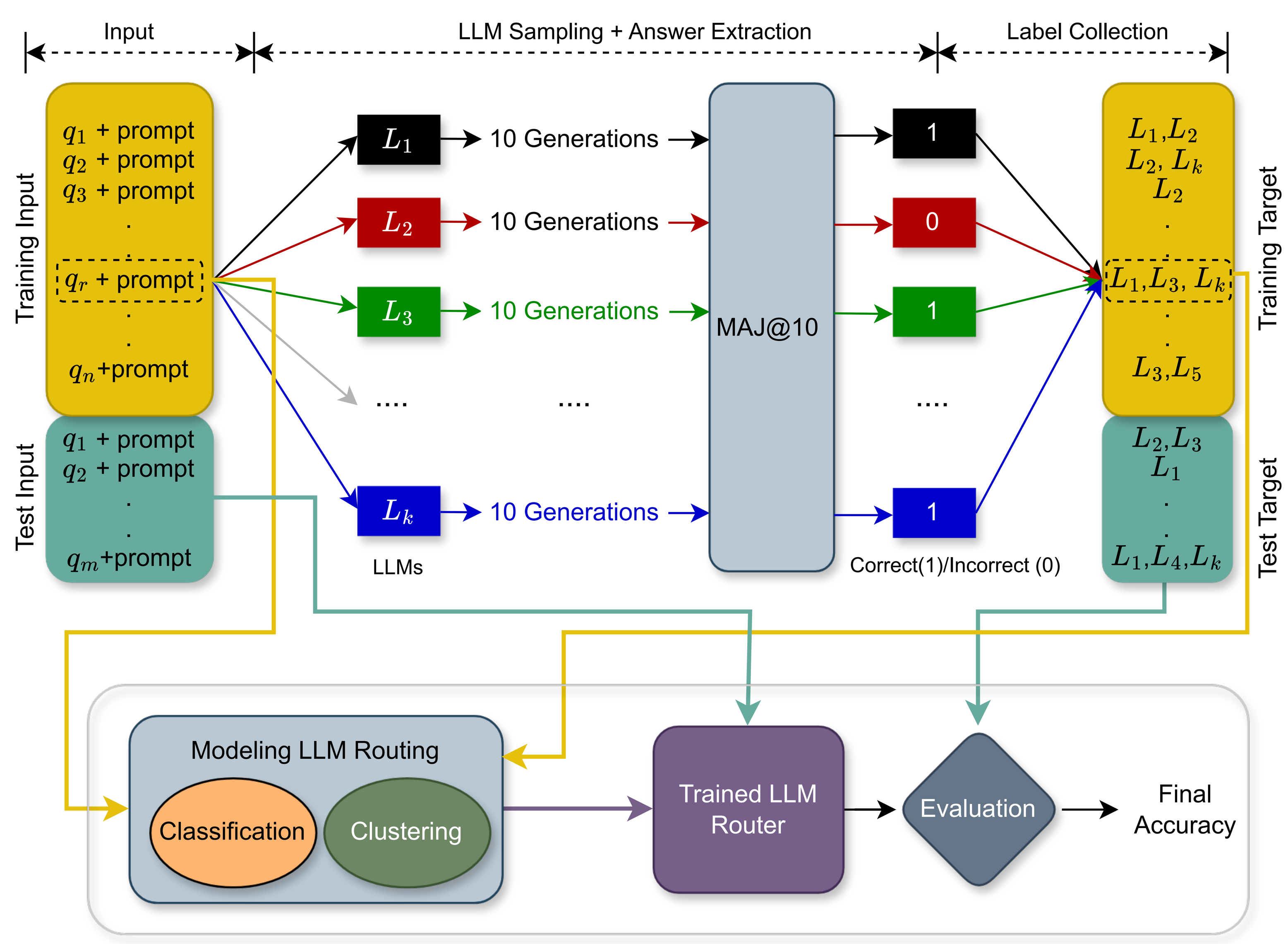
## Research Statement




**LLM Routing**

Whether directing an input query to the most *suitable single LLM* from a pool of diverse LLMs improves performance compared to individual LLMs while maintaining reasonable latency (e.g., similar to a single LLM)?

## Methodology



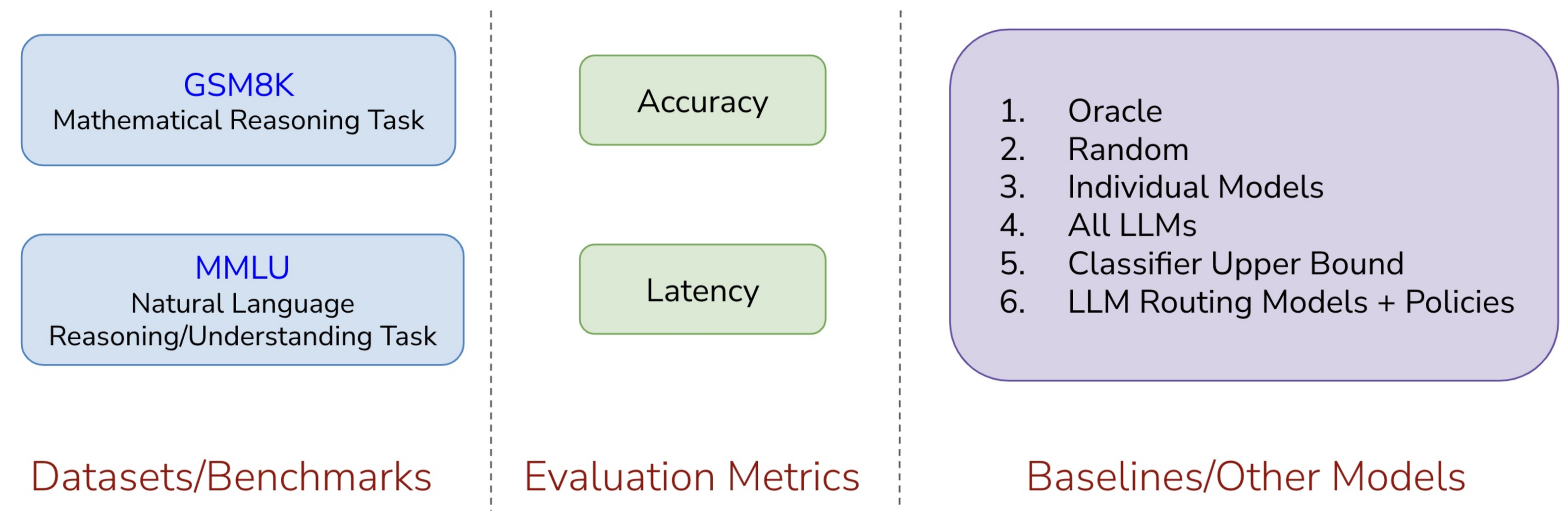

- **LLMs:** 3 chat LLMs and 4 non-chat (standard auto-regressive) LLMs
- **LLM Sampling:** *Zero-shot COT* for chat LLMs and *Few-shot COT* for non-chat LLMs
- 10 generations for each input query to improve *reproducibility*
- **Answer Extraction:** Using *Majority Voting* (MAJ@K $\in \{0, 1\}$) to determine whether the most frequent answer matches the gold answer or not
- **Data Preparation for LLM Routing:** Associate each input query with those viable LLM(s) that have a MAJ@10 score of 1. Formally, the target label for an input query $q \in Q$ is given by:

$$label\,(q) = \{l \mid l \in L, maj@10\,(q, l) = 1\}$$

where $L$ is the set of candidate LLMs and $Q$ is the set of query prompts.

- **LLM Routing Models:**
  - *Classification:* Developed on top of Pre-trained Language Models (PLMs) like BERT, DistilBERT, RoBERTa, and T5. RoBERTa works best.
    - Multi-Label Classifier (MLC)
    - Separate Classifier (SC)
  - *Clustering:* Feature extraction with TF-IDF and RoBERTa PLM
- **Predicted Confidence Score-based Policies:** (1) ArgMax, (2) Random, (3) Prediction with Random Forest, and (4) Sorted Prediction.

## Experimental Setup




## Results and Lessons Learned

| Models | | GSM8K | | MMLU | |
|---|---|---|---|---|---|
| | | Acc | Lat (sec) | Acc | Lat (sec) |
| Oracle | | 87.18 | 3.46 | 89.15 | 1.89 |
| Random | | 55.37 | 3.52 | 52.50 | 2.35 |
| gemma-7b | | <u>71.11</u> | 7.10 | <u>63.85</u> | 3.00 |
| metamath-7b | | 67.55 | 4.70 | 42.28 | 2.40 |
| mistral-7b | | 59.74 | 3.70 | 62.09 | 1.80 |
| mistral-7b-it | | 50.41 | 1.00 | 51.63 | 1.10 |
| llama2-13b-chat | | 46.70 | 1.80 | 50.52 | 4.80 |
| gemma-7b-it | | 36.84 | 0.70 | 49.28 | 1.00 |
| llama2-7b | | – | – | 48.36 | 2.30 |
| All LLMs [1] | | 74.37 | 19.00 | 60.39 | 16.40 |
| MLC | Upper bound | 79.68 | 5.16 | 77.18 | 1.94 |
| | ArgMax policy | 67.62 | 4.76 | 62.28 | 2.95 |
| | Random policy | 67.47 | 4.76 | 58.16 | 2.86 |
| | Prediction policy | **67.70** | 4.77 | **63.85** | 2.95 |
| | Sorted Pred policy | 59.90 | 4.77 | 48.36 | 2.92 |
| SC | ArgMax policy | 67.55 | 4.70 | 62.87 | 2.94 |
| Clustering | TF-IDF | 67.55 | 4.70 | 61.76 | 2.83 |
| | RoBERTa | 67.55 | 4.70 | 61.76 | 2.83 |

- ~10% of questions cannot be solved by all LLMs combined.
- Currently, the upper bound performance of the classifier/clustering model is not equal to the Oracle model due to the small size of the training data.
- The model with LLM routing performs better than weaker LLMs but worse or similar to the best single LLM.
- The predictions-based policy is slightly better than other policies; however, the classifier performance presents a serious bottleneck.
- The proposed LLM routing model consistently maintains a latency score equal to or lower than any individual LLM.

## Conclusions and Future Directions

- The theoretical upper bounds of LLM routing are much higher than individual models' performance.
- The proposed LLMs routing is a feasible direction that works best with equally capable LLMs.
- Future research should focus on generating more training data for router training.
- Future research should also incorporate LLM-specific features in router modeling.

## References

[1] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.

[2] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

## Acknowledgements

We thank *Jad Doughman* and *Ted Briscoe* for insightful discussions about this research. We are grateful to the Campus Super Computing Center at MBZUAI for supporting this work. We also thank the anonymous reviewers for their valuable feedback.