



MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE



NAACL 2024

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

KV Aditya Srivatsa* and Kaushal Kumar Maurya* and Ekaterina Kochmar

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

`vaibhav.kuchibhotla@mbzuai.ac.ae`

at Workshop on Insights from Negative Results in NLP
(Co-located with NAACL 2024)



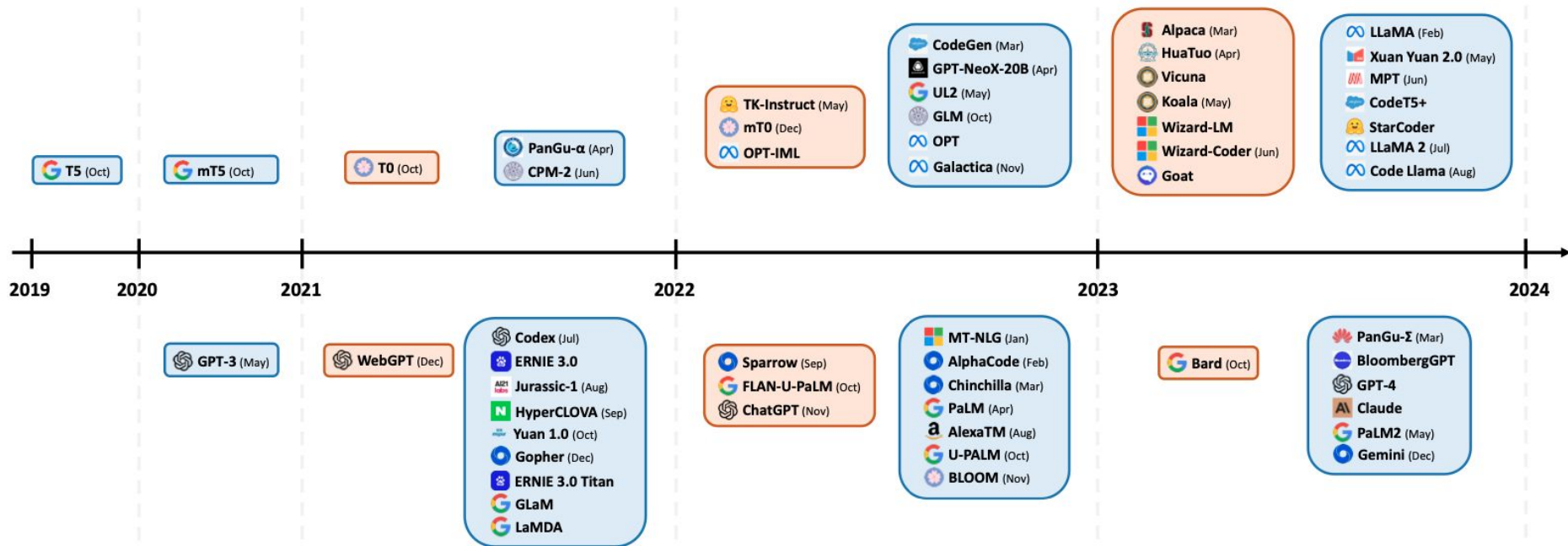
Outline

- ❑ Introduction and Motivation
- ❑ Research Hypothesis
- ❑ Methodology
- ❑ Experimental Setup
- ❑ Results and Learned Lessons
- ❑ Conclusion and Future Work

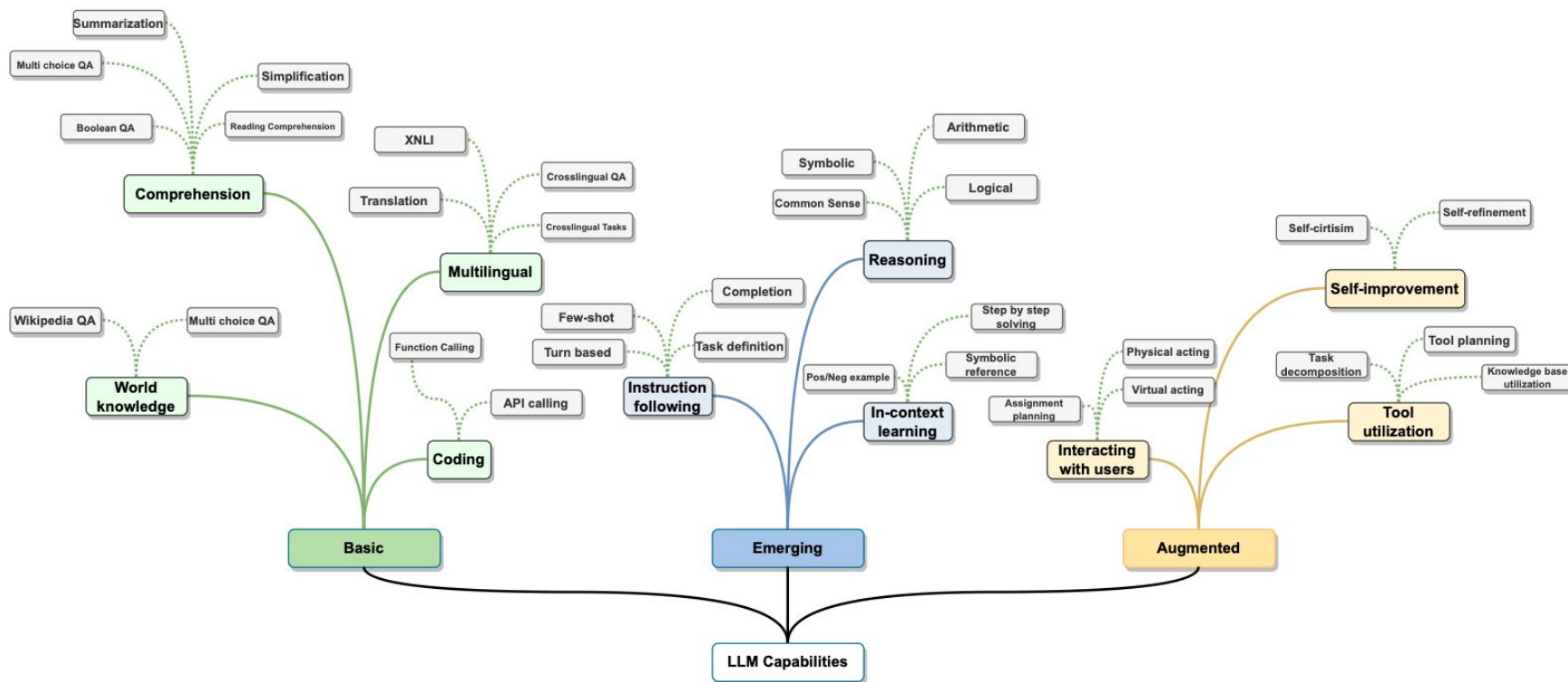
Outline

- ❏ Introduction and Motivation
- ❏ Research Hypothesis
- ❏ Methodology
- ❏ Experimental Setup
- ❏ Results and Learned Lessons
- ❏ Conclusion and Future Work

Landscape of Large Languages Models



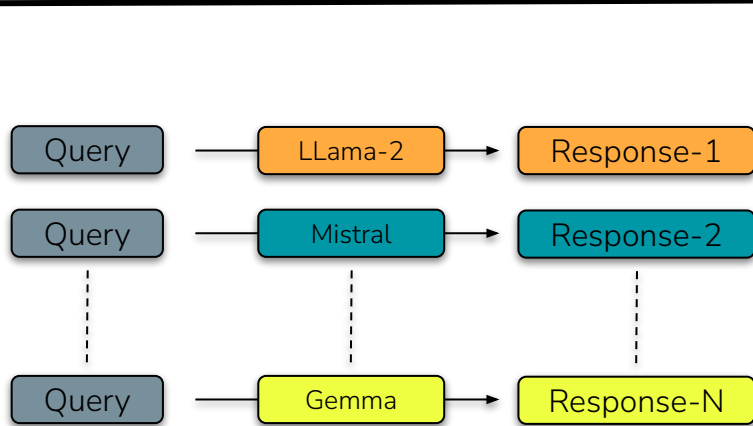
Diversity Capabilities of Large Languages Models



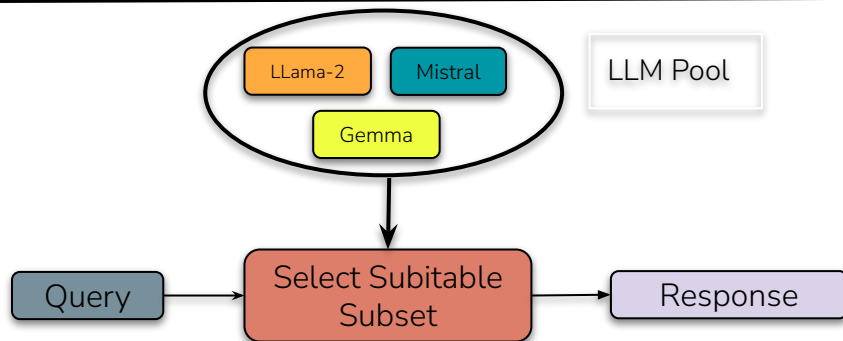
Outline

- ❏ Introduction and Motivation
- ❏ **Research Hypothesis**
- ❏ Methodology
- ❏ Experimental Setup
- ❏ Results and Learned Lessons
- ❏ Conclusion and Future Work

Research Hypothesis



Standard LLM Inference



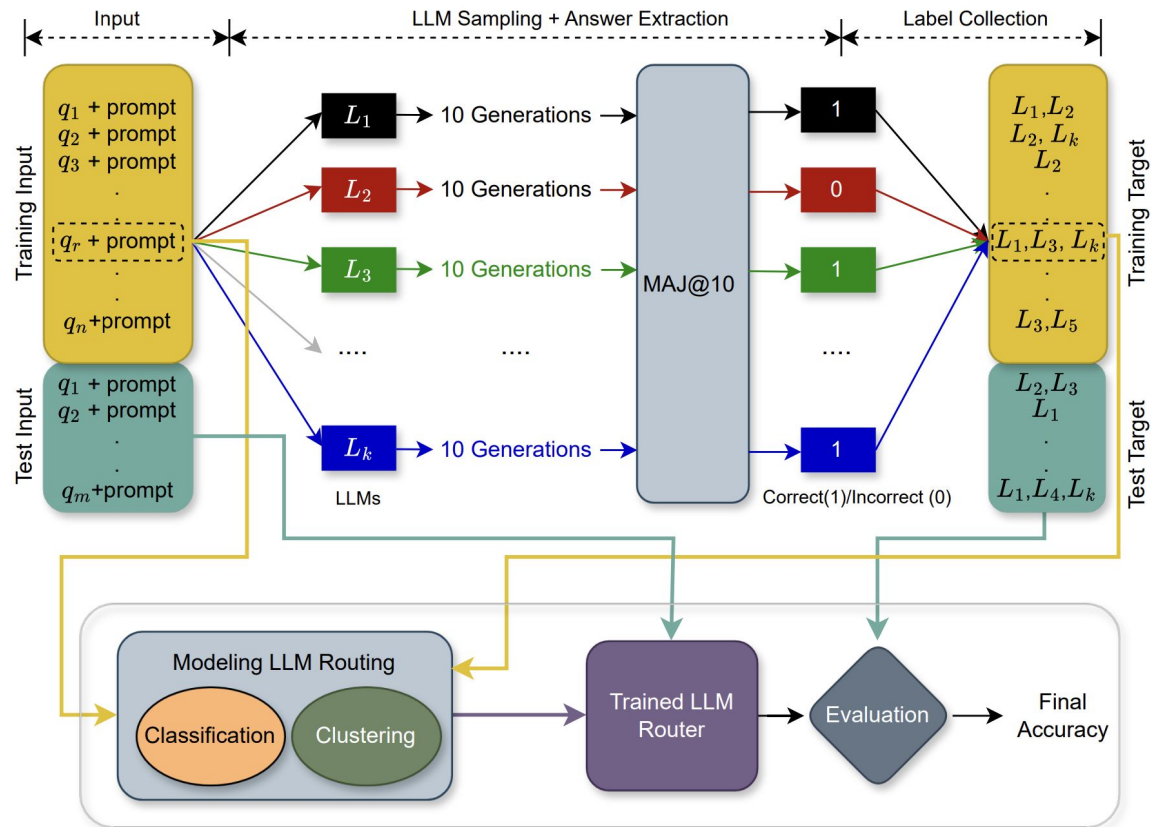
LLM Inference with **SELECTLLM**

Whether directing an input query to the most *suitable subset of LLM* from a large pool of diverse LLMs *lead to better performance in terms of accuracy and latency?*

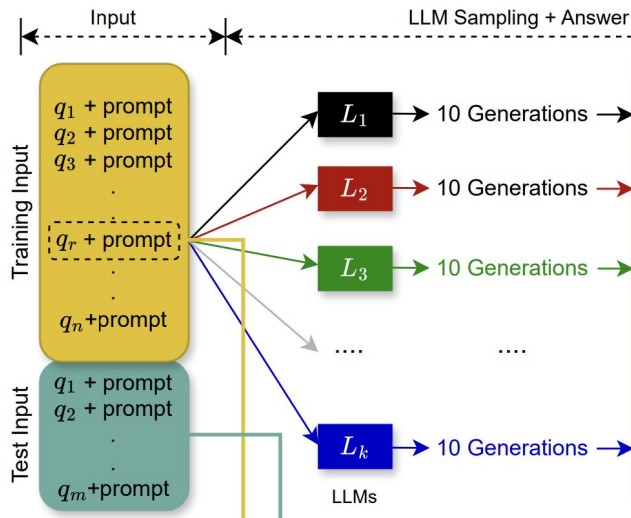
Outline

- ❏ Introduction and Motivation
- ❏ Research Hypothesis
- ❏ **Methodology**
- ❏ Experimental Setup
- ❏ Results and Learned Lessons
- ❏ Conclusion and Future Work

Methodology: Overview



Methodology: LLM Sampling



LLMs	Chat?	Specialized?	#Parameters
llama2-7b	×	×	7B
llama2-13b-chat	✓	×	13B
mistral-7b	×	×	7B
mistral-7b-it	✓	×	7B
gemma-7b	×	×	7B
gemma-7b-it	✓	×	7B
metamath-7b	×	✓	7B

Zero-Shot CoT Prompt

Answer the following question step by step to the best of your ability.

Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

Model Response

Sure, I'd be happy to help! Let's analyze the information given:
 In April, Natalia sold clips to 48 of her friends.
 ...
 So, Natalia sold 24 clips in May.
 Now, let's add the number of clips she sold in April and May to find out how many clips she sold altogether:
 48 clips + 24 clips = 72 clips
 Therefore, Natalia sold a total of **72** clips in April and May.

Few-Shot CoT Prompt

□ Few-Shot Exemplars

<START_QUESTION> Question: An auditor traces the serial numbers ... is supported by this test?
 (A) Valuation and allocation ... (D) Presentation and disclosure <END_QUESTION>
 <START_ANSWER> Answer: Let's think step by step. ... The answer is (B). <END_ANSWER>

<START_QUESTION> Question: One hundred years ago, ... what is the investment worth today?
 (A) \$13,000 (B) \$600 (C) \$15,000 (D) \$28,000 <END_QUESTION>
 <START_ANSWER> Answer: Let's think step by step. ... The answer is (A). <END_ANSWER>

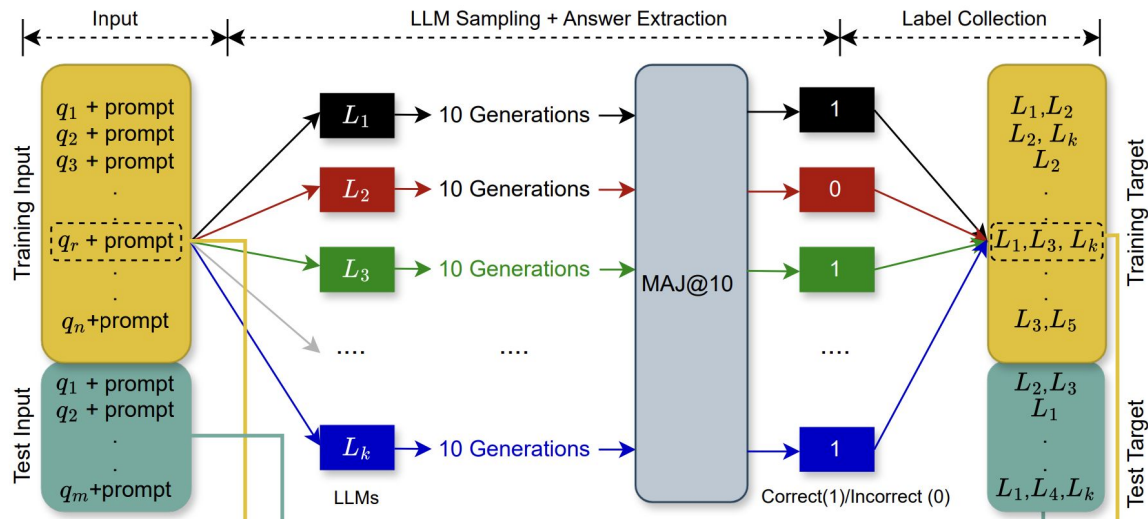
...

<START_QUESTION> Question: A company whose stock ... stock dividend? <END_QUESTION>
 (A) Treasury stock is debited for \$300. ... (D) Common stock is debited for \$3,000.
 <START_ANSWER> Answer: Let's Think step by step.

Model Response

The stock dividend is calculated based on the number of shares outstanding, not the current market price. Hence the adjustment for the stock dividend is $\$1 \times 1,000 \times 30\% = 300$, which is in Retained earnings. The answer is (C). <END_ANSWER>

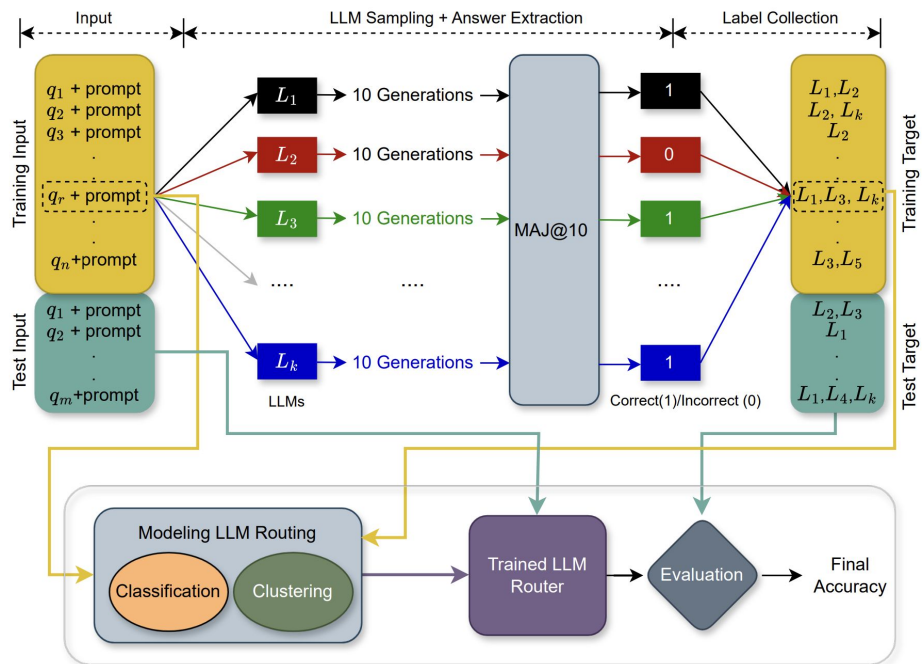
Methodology: Answer Extraction and Data Preparation



Majority Voting (MAJ@K $\in \{0, 1\}$) determines whether the most frequent answer matches the gold answer or not.

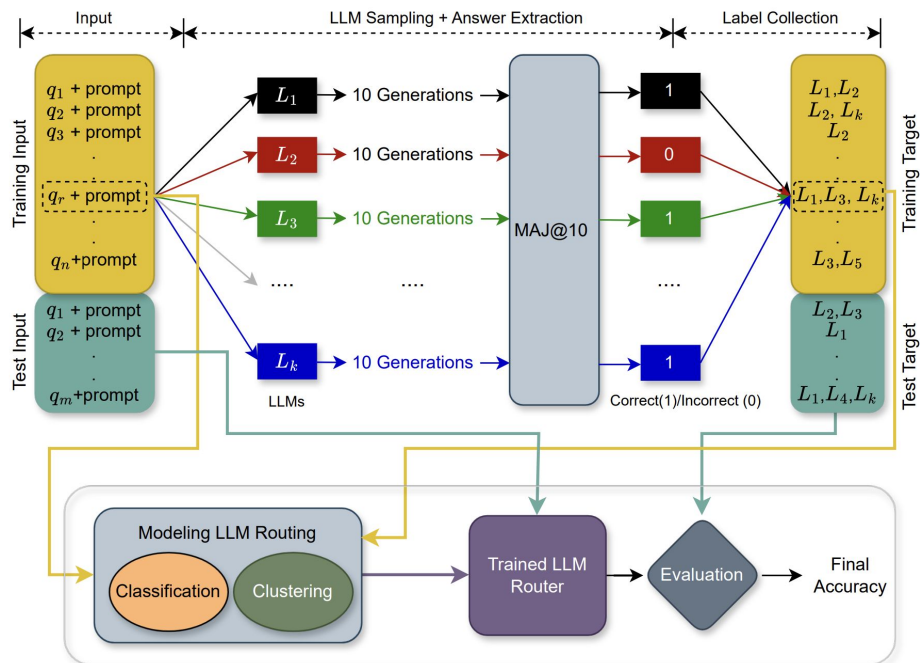
The target label for an input query $q \in Q$ is given by label $(q) = \{l \mid l \in L, \text{maj@10}(q, l) = 1\}$

Methodology: LLM Routing Models



- Multi-label Classifier
 - BERT, DistilBERT, RoBERTa, and T5
- Separate Classifier
 - BERT, DistilBERT, RoBERTa, and T5
- Clustering-Based Routing
 - TF-IDF, RoBERTa
- Smaller models, utilizing only a few layers of PLMs, Random Forests, etc.

Methodology: LLM Routing Policies



With Classifier Predicted Confidence Score

1. ArgMax
2. Random
3. Prediction with random forest
4. Sorted Prediction (similar to 3)

Outline

- ❏ Introduction and Motivation
- ❏ Research Hypothesis
- ❏ Methodology
- ❏ **Experimental Setup**
- ❏ Results and Learned Lessons
- ❏ Conclusion and Future Work

Experimental Setup

GSM8K

Mathematical Reasoning Task

MMLU

Natural Language
Reasoning/Understanding Task

Datasets/Benchmarks

Accuracy

Latency

Evaluation Metrics

1. Oracle
2. Random
3. Individual Models
4. All LLMs
5. Classifier Upper Bound
6. LLM Routing Models + Policies

Baselines/Other Models

Outline

- ❏ Introduction and Motivation
- ❏ Research Hypothesis
- ❏ Methodology
- ❏ Experimental Setup
- ❏ **Results and Learned Lessons**
- ❏ Conclusion and Future Work

Evaluation Results

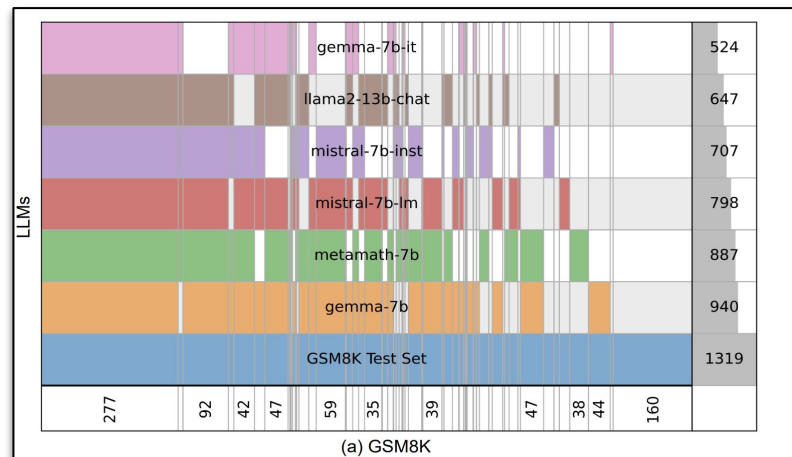
Models		GSM8K		MMLU	
		ACC	LAT (sec)	ACC	LAT (sec)
Oracle		87.18	3.46	89.15	1.89
Random		55.37	3.52	52.50	2.35
gemma-7b		<u>71.11</u>	7.10	<u>63.85</u>	3.00
metamath-7b		67.55	4.70	42.28	2.40
mistral-7b		59.74	3.70	62.09	1.80
*mistral-7b-it		50.41	1.00	51.63	1.10
llama2-13b-chat		46.70	1.80	50.52	4.80
*gemma-7b-it		36.84	0.70	49.28	1.00
llama2-7b		–	–	48.36	2.30
All LLMs		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

Evaluation Results: Learned Lessons-I

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		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

Ques: Does including multiple LLMs solve all questions in a given dataset?

Ans: 10% of questions cannot be solved by all LLMs combined.



Evaluation Results: Learned Lessons-II

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		71.11	7.10	63.85	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		74.37	19.00	60.39	16.40
	Upper bound	79.68	5.16	77.18	1.94
MLC	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

Ques: Can the upper bound performance of the classifier/clustering be equal to the Oracle model performance?

Ans: NO, due to small training data (~9k in GSM8K and 15k in MMLU)

Split/Criteria	GSM8K	MMLU
Training	6,816	13,757
Validation	359	285
Test	1,319	1,530
#examples for few-shot CoT	5	5

Evaluation Results: Learned Lessons-III

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		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

Ques: Does router modeling with multi-label classifiers exhibit better performance than individual LLMs?

Ans: **NOT** with **ALL**, better than weaker LLMs but lower or similar to the best single LLM

Evaluation Results: Learned Lessons-IV

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		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

Ques: What is the impact of different policies on LLM router modeling?

Ans: The predictions-based policy is better than other policies; however, the classifier performance presents a **serious bottleneck**.

Evaluation Results: Learned Lessons-V

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		71.11	7.10	63.85	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		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

Ques: What is the impact of LLM routing on inference latency?

Ans: The proposed LLM routing model consistently maintains a latency score **equal to or lower** than any individual LLM.

Prompt Type	LLM	GSM8K	MMLU
FCoT	llama2-7b	4.21	2.30
	gemma-7b	7.10	3.00
	mistral-7b	3.70	1.80
	metamath-7b	4.70	2.40
	gemma-7b-it	0.70	1.00
ZCoT	llama2-13b-chat	1.80	4.80
	mistral-7b-it	1.00	1.10

Inference latency (in second) with A100 GPU

Outline

- ❏ Introduction and Motivation
- ❏ Research Hypothesis
- ❏ Methodology
- ❏ Experimental Setup
- ❏ Results and Learned Lessons
- ❑ Conclusion and Future Work

Conclusion

- The theoretical bounds of LLM routing is **much higher** than individual models' performance.
- **LLMs routing is a feasible direction** that works best with equally capable LLMs.
- If a **few** LLMs dominate, the router's performance **degrades**, even though it **still outperforms weak LLMs**.
- The inference latency of the routing model is **at least at the same level** as that of single LLMs.

Future Work

- Larger datasets for LLM routing
- Novel models for LLM routing
- Better routing policies
- Incorporating LLM-specific features
- Scaling to diverse LLMs and benchmarks

Acknowledgement

- Jad Doughman and Prof. Ted Briscoe for insightful discussions and initial guidance
- Campus Super Computing Center at MBZUAI
- Anonymous reviewers for their valuable feedback

References

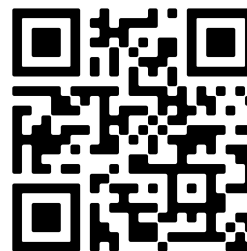
1. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, Barnes N, Mian A. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435. 2023 Jul 12.
2. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J. Large language models: A survey. arXiv preprint arXiv:2402.06196. 2024 Feb 9.
3. Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2080–2094, Online. Association for Computational Linguistics.
4. Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
5. Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More Agents Is All You Need. arXiv preprint arXiv:2402.05120.
6. Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, MarieAnne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv preprint arXiv:2401.04088.
7. Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor R hle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. 2024. Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing. In The Twelfth International Conference on Learning Representation
8. Noam Shazeer, Azalia Mirhoseini, Krzysztof Mazi r, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv preprint arXiv:1701.06538.



Thank you!!



Paper



GitHub