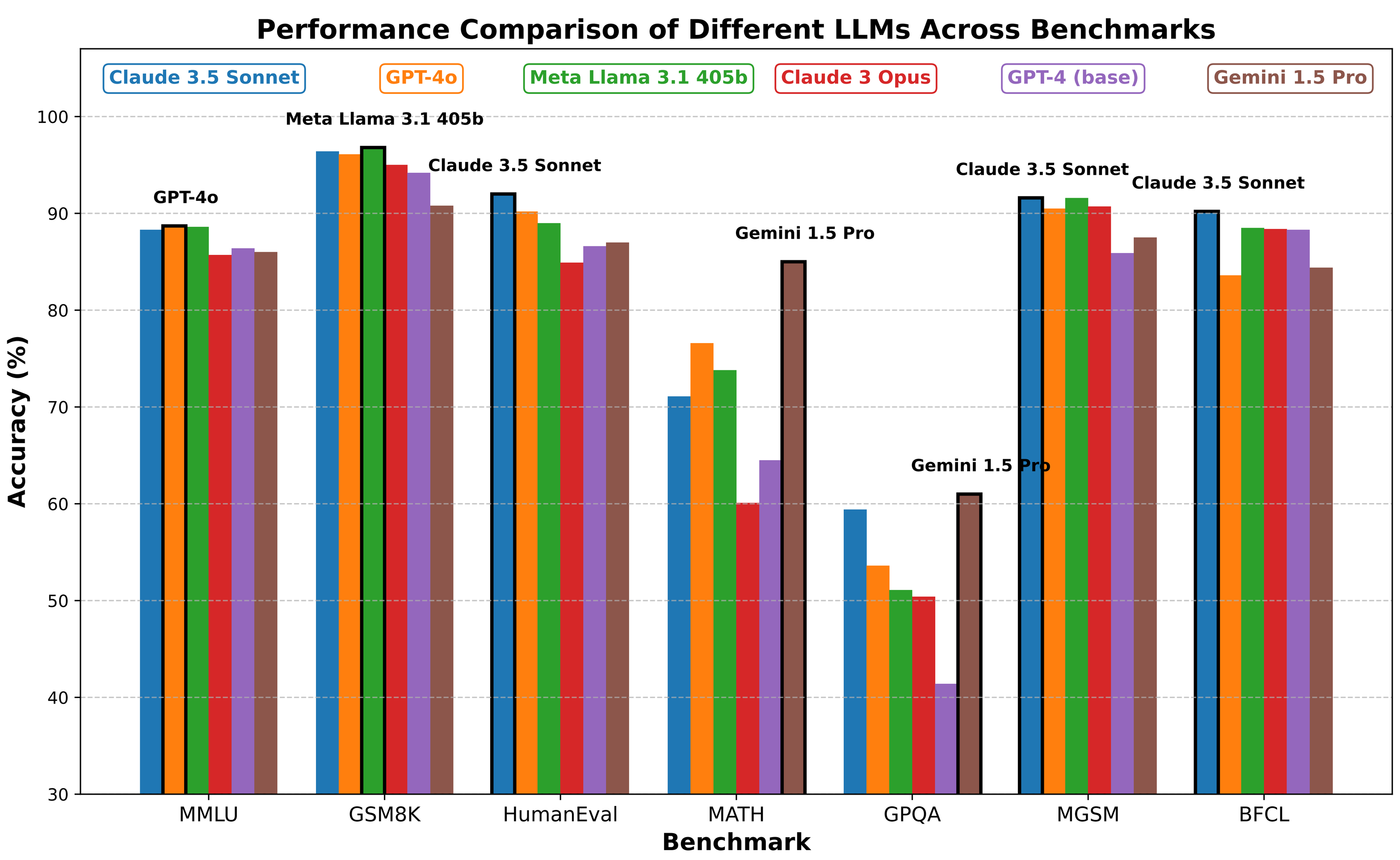
SELECTLLM: Query-Aware Efficient Selection Algorithm for Large Language Models

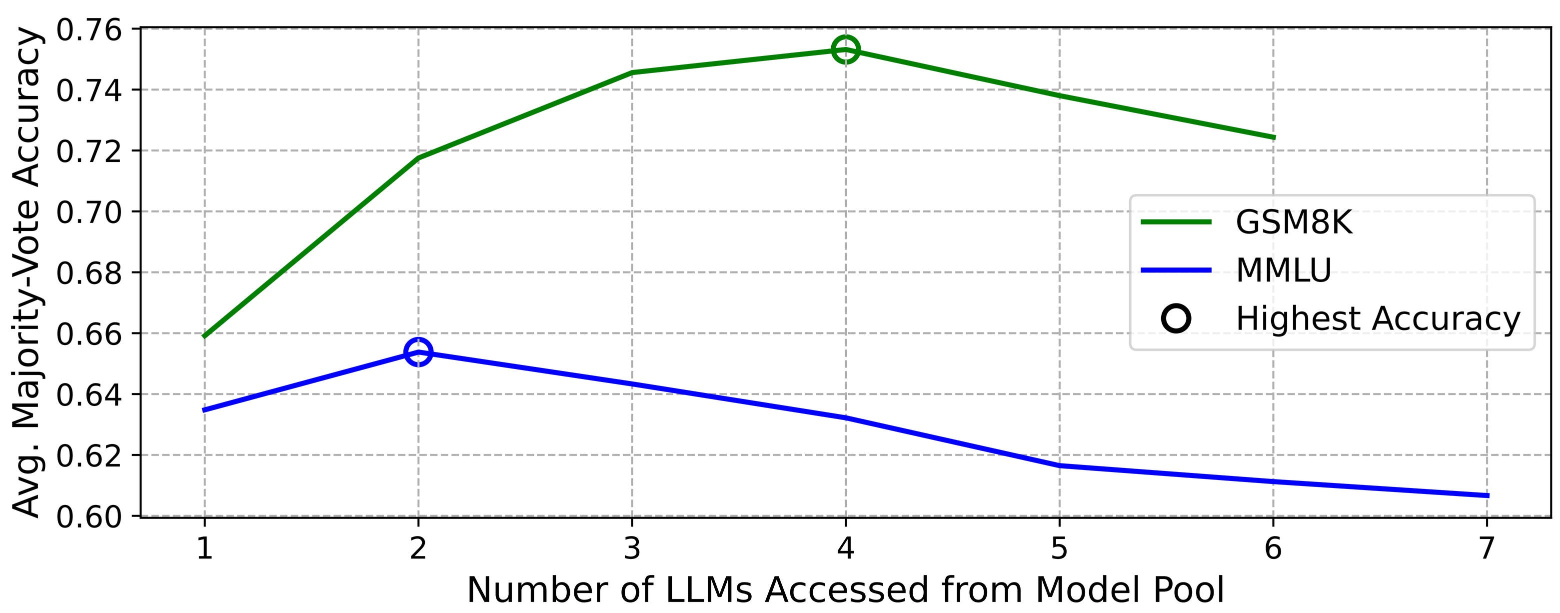
Kaushal Kumar Maurya*, KV Aditya Srivatsa*, and Ekaterina Kochmar
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
kaushal.maurya@mbzuai.ac.ae
*Equal Contributions

Mohamed bin Zayed University of Artificial Intelligence
ACL 2025 VIENNA

Introduction

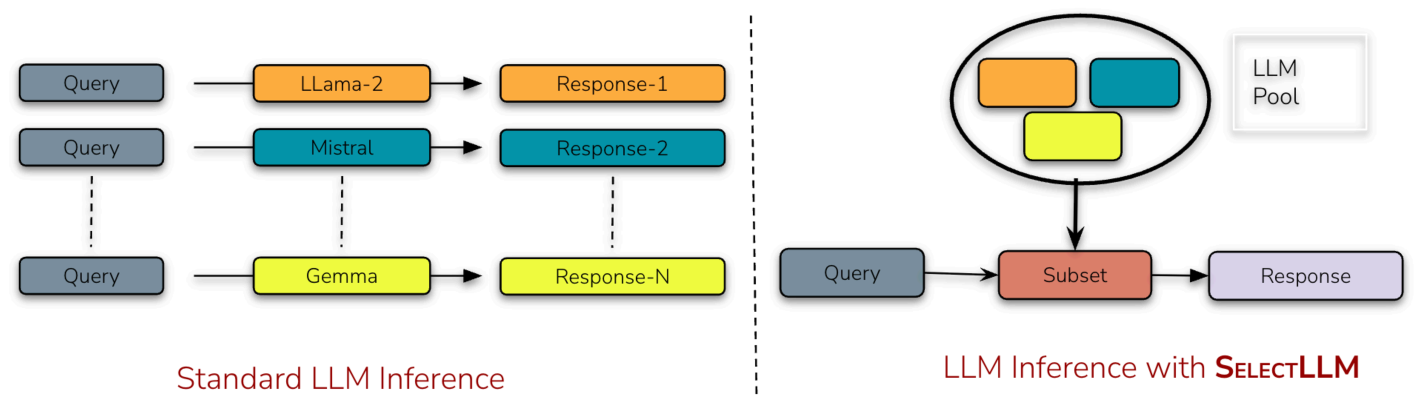


- Different LLMs exhibit diverse capabilities [1].
- It is natural to ask: How can we harness the diverse capabilities of LLMs? Can we simply use all available LLMs?



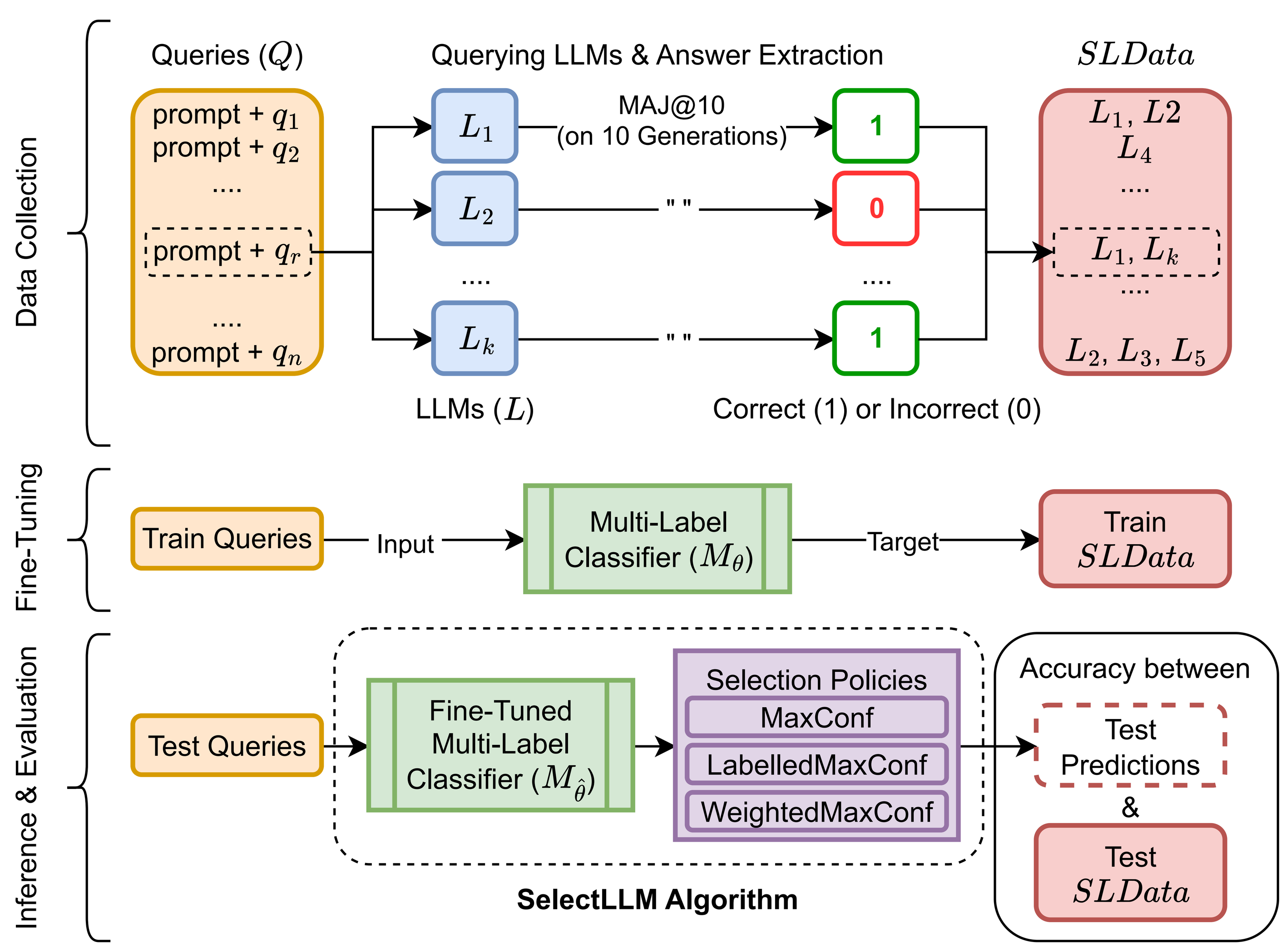
We propose SELECTLLM, an approach that **dynamically** selects the **most suitable, query-aware** subset of LLMs from a large pool.

Research Statement



How can we select the **most effective query-aware** subset of LLMs from a large pool to simultaneously improve accuracy and minimize computational cost?

Methodology

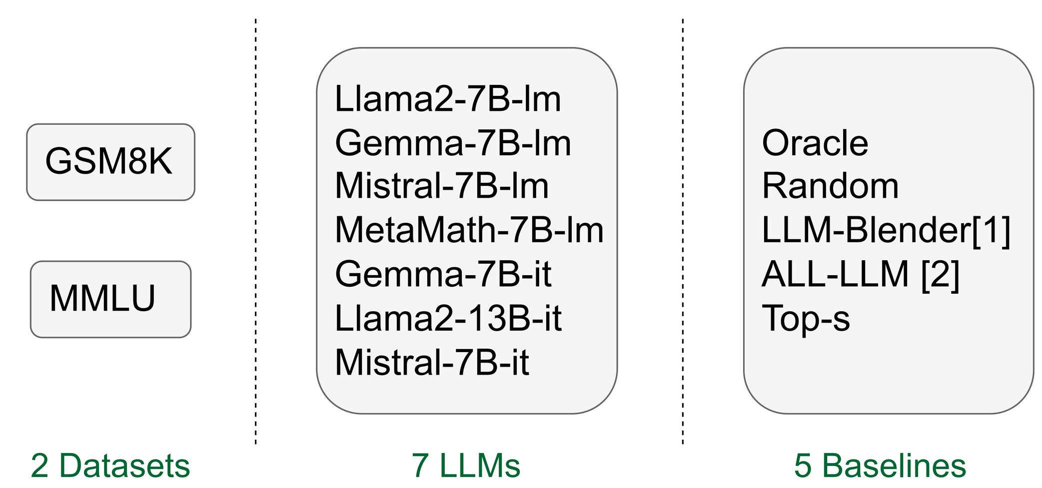


- We use *Majority Voting* (MAJ@K $\in 0, 1$) to determine whether the most frequent answer among the model outputs matches the gold reference.
- **SLData Data Preparation:** Each input query is paired with candidate LLM(s) L for which the MAJ@10 score equals 1. Formally, for a query $q \in Q$, the target label is defined as follows:

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

- **LLM Selection Strategy:**
  - *Multi-Label Classifier:* We train lightweight language models (e.g., BERT, RoBERTa, T5) with SLDATA to predict the optimal subset of LLMs. Among them, RoBERTa-based classifiers achieved the best performance.
  - *Selection Policies Based on Classifier Predictions:* We propose three inference-time selection strategies: (1) LABELLEDMAXCONF, (2) MAXCONF, and (3) WEIGHTEDMAXCONF.

Experimental Setup



Results and Observations

Models / Setups		GSM8K Acc (↑)	GSM8K Lat (↓)	MMLU Acc (↑)	MMLU Lat (↓)
Oracle		90.52	3.24	90.46	1.75
Baseline	Random	69.49	9.65	58.20	8.27
	LLM-Blender [2]	75.28	19.00	60.27	16.40
	All LLMs [1]	76.04	19.00	60.92	16.40
	Top-s LLMs	77.48	19.00	65.75	16.40
SELECTLLM	MLC + LABELLEDMAXCONF	75.66	14.69	65.68	4.78
	MLC + MAXCONF	77.48	16.50	65.68	4.78
	MLC + WEIGHTEDMAXCONF	**77.94**	16.50	**65.81**	4.78

- Quantitative analysis identifies length, tree depth, NP count, and argument count as the most influential among the 16 linguistic features explored.
- Out-of-distribution evaluation on the MMLU dataset demonstrates that SELECTLLM generalizes effectively across both grades and subjects.
- The performance gap with the Oracle is primarily due to limited training data for the classifier.

Conclusions and Future Directions

- We propose SELECTLLM, an efficient LLM selection algorithm based on multi-label classification with confidence-based policies.
- It achieves higher accuracy and lower latency than strong baselines on both GSM8K and MMLU datasets.
- Future work should include LLM-specific and query-level linguistic and structural features to bridge the performance gap with the Oracle.

References

[1] Zhengbao Jiang, Frank F. Xu, Suchin Gururangan, and Noah A. Smith. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *EMNLP 2023*.

[2] Qin Zhang, Yangbin Yu, Qiang Fu, Deheng Ye, et al. More agents is all you need. *TMLR 2024*.