

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

Kaushal Kumar Maurya* KV Aditya Srivatsa* Ekaterina Kochmar
Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{kaushal.maurya, vaibhav.kuchibhotla, ekaterina.kochmar}@mbzuai.ac.ae

Abstract

Large language models (LLMs) have been widely adopted due to their remarkable performance across various applications, driving the accelerated development of a large number of diverse models. However, these individual LLMs show limitations in generalization and performance on complex tasks due to inherent training biases, model size constraints, and the quality or diversity of pre-training datasets. A promising direction is to efficiently harness the diverse capabilities of LLMs to overcome these individual limitations. To address these limitations, we introduce a novel LLM selection algorithm called SELECTLLM, which efficiently directs input queries to the most suitable subset of LLMs from a large pool, ensuring that the selected models collectively provide accurate responses. SELECTLLM employs a multi-label classifier and policy based on the classifier’s predictions and confidence scores in selecting an optimal, query-aware, and lightweight subset of LLMs. Our findings indicate that the proposed model outperforms existing ensemble-based baselines and achieves competitive performance with similarly sized top-performing LLMs while maintaining efficiency. Specifically, it achieves a huge reduction in inference latency on two challenging reasoning benchmarks: 13% on GSM8K and 70% on MMLU, compared to the top-performing baseline. Also, we establish a theoretical upper bound by an Oracle with LLMs and perform an in-depth linguistic analysis to understand the performance gap between the Oracle and SELECTLLM.

github.com/kaushal0494/SelectLLM

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in solving a wide range of core NLP tasks (Bommasani et al., 2021; Chang et al., 2023). Despite these

*Equal contribution

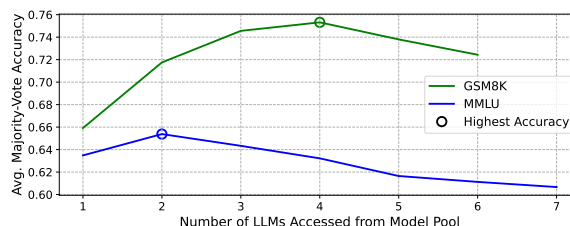


Figure 1: Accuracy (using majority voting) versus the number of LLMs plot for the GSM8K and MMLU test sets. Models are added in descending order of their performance on their corresponding training sets.

advances, existing LLMs still struggle with complex tasks such as factually-grounded reasoning and planning (Wei et al., 2022; Kojima et al., 2022; Minaee et al., 2024). Moreover, the wide range of LLMs available seem to **exhibit diverse capabilities** (Jiang et al., 2023), resulting in no single (especially open-source) LLM being effective across all benchmarks and datasets. Aligning with recent research trends (DeepSeek-AI et al., 2025; Mirzadeh et al., 2024), we focus on natural language understanding tasks grounded in reasoning, specifically those with discrete output values, and do not consider general natural language generation tasks.

Although newer and more powerful models are constantly introduced, an alternative and cost-effective approach involves harnessing the diverse capabilities of existing models to improve the overall response quality using ensembling (Wang et al., 2022, 2023; Li et al., 2024) and collaborative frameworks (Wu et al., 2023; Li et al., 2023). However, these approaches often require access to the responses from all models in the pool to choose the optimal response(s), which greatly increases the overall computational cost for such ensembles.

The individual LLMs exhibit diverse capabilities, i.e., *not all models may be suited for all kinds of tasks*. Fig. 1 reports the accuracy of a LLM model pool (spanning up to 7 diverse LLMs) on two challenging reasoning benchmarks – GSM8K

(Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021). As the plot demonstrates, utilizing more LLMs initially improves performance, which is supported by previous research towards employing more LLMs and more responses per model (Li et al., 2024). However, note that using more (or even all) models in the pool does not necessarily result in the best scores overall. Thus, *selectively abstaining* from querying unsuitable LLMs for a given task may help **improve the overall response quality** of such ensembles. Additionally, such an approach would **implicitly save computational resources** by accessing fewer models per query.

In this paper, we propose the novel SELECTLLM algorithm to explore this idea. Our approach first employs a *multi-label classifier* to learn the LLM-specific capabilities using a dataset of diverse queries. When running inference for a unseen query, this knowledge is utilized to predict confidence scores for each LLM in the model pool, reflecting *their likelihood of successfully solving the task*. Next, we develop various *selection policies* to determine the optimal subset of LLMs for each query based on these confidence scores and predictions. Additionally, we establish a theoretical Oracle model’s upper bound that can be achieved collectively by all LLMs and perform a *qualitative and quantitative linguistic analysis* of the inputs to understand the performance gap between the Oracle and SELECTLLM. The contributions of our work are as follows:

- We introduce the novel SELECTLLM algorithm, which is based on a multi-label classifier and an optimal confidence-based policy. This approach efficiently navigates input queries to the ideal subset of LLMs from a larger pool to improve response quality and simultaneously reduce computational costs.
- The efficacy of the proposed SELECTLLM algorithm is evaluated on two challenging reasoning benchmarks. We report an improvement of 1.90 points on GSM8K and 4.89 points in terms of accuracy on MMLU compared to the existing strong ensemble-based baselines and competitive performance with a model, which consists of similarly sized top-performing LLM subsets. Additionally, we observe significantly lower inference latencies, with reductions of 13% for GSM8K and 70% for MMLU, compared to the top-performing baseline.
- We present a theoretical upper bound es-

tablished by an Oracle model, representing the maximum performance achievable collectively by all LLMs in our pool. Furthermore, we conduct a linguistic feature-based analysis of the inputs to understand the gains achieved by SELECTLLM and the performance gap relative to the Oracle model.

2 Related Work

LLM Diversity and Capabilities: LLMs exhibit emergent capabilities, enabling them to perform beyond their explicit training objectives, with diversity in training data fostering broad domain expertise (Bommasani et al., 2021; Minaee et al., 2024). This diversity extends to architectural variations, multilingual proficiency, and adaptability across domains, making them highly effective for a wide range of tasks. Additionally, LLMs demonstrate strong generalization and reasoning abilities, enabling them to tackle complex challenges such as question answering, summarization, and classification (Hendrycks et al., 2021; Cobbe et al., 2021; Joshi et al., 2017; Tam et al., 2023; Zhang et al., 2023). Despite their strengths, no single open-source LLM consistently outperforms others across benchmarks (Jiang et al., 2023), highlighting the necessity of ensemble methods to leverage diverse model strengths. This work focuses on leveraging the diverse capabilities of LLMs in SELECTLLM to address individual LLM limitations.

LLM Ensembling Previous attempts at ensembling and routing LLMs typically fall into three categories: (1) Selecting the best response from multiple LLM generations, as seen in Liu and Liu (2021), Ravaut et al. (2022), and Jiang et al. (2023). However, this approach requires querying all LLMs in the model pool for each query during inference, which can be computationally expensive with a large number of LLMs. (2) Minimizing the number of queries to larger LLMs to reduce latency and computational costs, as demonstrated by Shnitzer et al. (2023) and Ding et al. (2024), who redirect simpler queries to the smallest model capable of handling the task. This minimizes querying costs with minimal performance drop. Routing to the single-best LLM while balancing both accuracy and efficiency has proven to be challenging (Srivatsa et al., 2024). (3) Using multiple LLMs in a multi-agent collaboration setting (Tran et al., 2025), where specialized LLMs work together to solve complex problems. However, it remains unclear

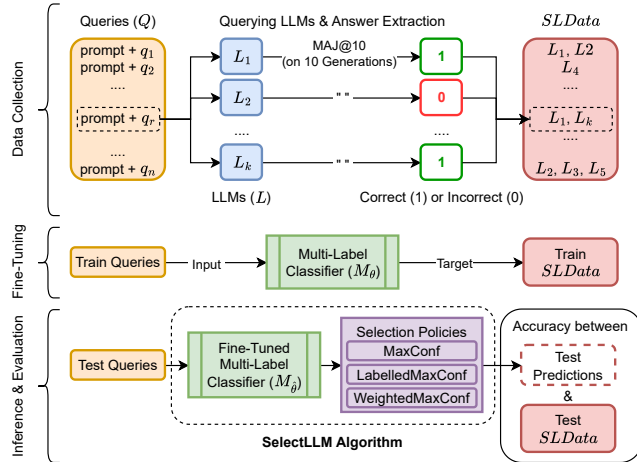


Figure 2: Overview of the proposed workflow.

how collaboration functions when non-specialized models are involved. The proposed model lays the foundation for addressing this gap. Moreover, we aim to develop an algorithm that improves response accuracy beyond individual LLMs and their combinations by querying only a subset of LLMs expected to be capable of solving the given query. This, in turn, reduces computational costs and latency by avoiding unnecessary queries to unsuitable LLMs.

3 Problem Setting

We propose an ensembling-based LLM inference algorithm – SELECTLLM – to *efficiently* select the *most suitable* query-aware few LLMs from a large pool of available LLMs. The algorithm harnesses the diverse capabilities of different LLMs and selects a subset of models for the input query, jointly leading to the correct response, and this selection of a small subset leads to the reduction in latency.

Formally, for a given set of input queries $\mathcal{Q} = \{q_1, q_2, \dots, q_n\}$ and a pool of LLMs $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$, the objective is to learn a *selection model* \mathcal{M} that selects a subset of LLMs $\mathcal{L}_s \subseteq \mathcal{L}$, which jointly produce the correct answer for a given input query $q_i \in \mathcal{Q}$, such that the cumulative latency satisfies $\text{latency}(\mathcal{L}_s) < \text{latency}(\mathcal{L})$. In the best-case scenario, a query q_i is processed by a single LLM, while in the worst-case scenario, q_i is processed by all LLMs in \mathcal{L} .

4 Methodology

4.1 LLM Sampling

4.1.1 Selection of Benchmarks and LLMs

As discussed in Section 1, LLMs often struggle with reasoning tasks. To advance modeling and

performance in this domain, we have selected two challenging benchmarks. The GSM8K dataset, introduced by Cobbe et al. (2021), contains 8,792 grade-school level math word problems (MWP) in English, focusing on mathematical reasoning. The second benchmark, MMLU, proposed by Hendrycks et al. (2021), includes 14,572 multiple-choice questions across 57 subjects, assessing multi-domain natural language understanding and reasoning. See Appendix Table 5 for detailed dataset statistics.

We have selected a *diverse* and *sparse* set of LLMs based on *explicit* and *implicit* criteria. The explicit criteria includes performance on benchmarks, training methodologies, model specialization, and modes of operation (chat vs. non-chat), among others. Some of these diverse attributes are presented in the Appendix Table 6. The implicit criteria include factors such as diverse inference latencies (refer to Table 7) and prompting types (i.e., zero-shot vs. few-shot), among others. Further, we consider relatively small open-source LLMs (yet representative of exiting open-source LLM space) because: (i) Experiments with these LLMs are suitable for an academic lab setup, and (ii) This aligns with the research trend towards developing LLMs suitable for small mobile devices (Abdin et al., 2024). Furthermore, since these selection criteria ensure wide representativeness of the LLM pool, we hypothesize that if the proposed approach works for this pool of LLMs, it should be LLM pool-agnostic, although we leave the proof of this to future work. See Appendix A for more details.

4.1.2 Data Preparation for the SELECTLLM Model: SLData

In this study, we evaluate the performance of each LLM by generating M responses from it for each

LLM	GSM8K		MMLU	
	Train	Test	Train	Test
llama2-7b-1m	22.21	24.01	46.85	48.10
gemma-7b-1m	73.49	71.27	66.10	63.73
mistral-7b-1m	59.33	60.50	61.69	61.57
metamath-7b-1m	92.35	67.25	42.16	41.46
*gemma-7b-it	41.62	42.23	50.61	50.72
llama2-13b-chat	50.37	49.20	54.40	52.94
*mistral-7b-it	55.92	56.71	53.39	53.92

Table 1: Accuracy with majority voting (MAJ@M) for considered LLMs on GSM8K and MMLU datasets with train and test splits. All scores were calculated over M response generations for each LLM. Here we use $M=10$ inspired by Li et al. (2024).

input query to ensure reliable and replicable behavior of the proposed model. We employ *Majority Voting* (Li et al., 2024) to assess whether a query is correctly answered by the LLM or not. Specifically, *Majority Voting*(MAJ@M $\in \{0, 1\}$) determines whether the most frequent answer from an LLM matches (using *exact string match*) the gold answer or not. The accuracy with MAJ@M across all input prompts is reported in Table 1. In the rest of this paper, we consider only those LLMs for which the viable extracted answers are above 90% (see more details in Appendix Section A) to ensure response reliability, resulting in 6 acceptable LLMs for the GSM8K dataset and 7 for the MMLU dataset. We prepare the training dataset for the multi-label classification module of SELECTLLM (detailed in the next section) by associating each input query with the LLM(s) whose majority vote answer (across M samples) matches the gold answer, i.e., MAJ@M = 1. Formally, the target label for a query prompt $q_i \in \mathcal{Q}$ is given by $label(q_i) = \{l \mid l \in \mathcal{L}, maj@M(q_i, l) = 1\}$. We denote this dataset as SLDATA which is separately prepared for both GSM8K and MMLU.

4.2 The Proposed SELECTLLM Algorithm

Building on the discussion in Section 1, prior research (Ding et al., 2024) indicates that easy queries are correctly solved by smaller or general-purpose LLMs, whereas complex queries necessitate the use of specialized or larger LLMs. Conversely, there are rare queries that are incorrectly responded to by large LLMs but correctly answered by smaller LLMs (Nezhurina et al., 2024). Due to a lack of widely established query-to-LLM mappings, brute-force approaches are typically employed, querying every available LLM to obtain correct answers. As LLMs continue to advance rapidly in large numbers, such approaches become computationally inefficient and sometimes infeasible. One promising approach involves identifying and directing input queries to the most suitable subset of LLMs from a large pool, which jointly respond correctly. This ensures that the query is accurately addressed while maintaining lower latency compared to running inference on all LLMs in the pool.

Towards this objective, we introduce the query-aware SELECTLLM algorithm, designed to select a tailored subset of LLMs, taking into account the nature of the query and enabling them to collaboratively provide correct responses efficiently. The

SELECTLLM algorithm comprises two primary components: (i) A *multi-label classifier* (MLC) which is fine-tuned with SLDATA dataset, and (ii) A *selection policy*, which utilizes MLC’s prediction and confidence scores (i.e., the likelihood of an LLM responding correctly to the query) to determine the suitable subset of LLMs.

Multi-label Classifier (MLC) As indicated by prior studies (Hu et al., 2024), lightweight language models such as BERT, RoBERTa, and T5 exhibit negligible lower query inference latency compared to LLMs. So, we developed a multi-label classifier (MLC) based on these models. The models were fine-tuned using SLDATA with a multi-label classification objective, incorporating label imbalance techniques (Zhang et al., 2020) for both the GSM8K and MMLU datasets. The results are provided in Table 8. Among the tested models, the RoBERTa-based MLC outperformed the others, achieving weighted F1 scores of 0.71 and 0.68 for the GSM8K and MMLU test datasets, respectively. The fine-tuned model predicts a subset of LLMs (i.e., LLM identities) best suited to address the input query and confidence scores C for each model in the pool.

Confidence-Based Policies In this section, we discuss how the confidence scores C are utilized to select a suitable subset of LLMs \mathcal{L}_s for each input query. C is defined as $\{c_1, c_2, \dots, c_i, \dots, c_n\}$ where c_i is the confidence scores for the i^{th} input query. Each c_i is further represented as $\{c_i^{l_1}, c_i^{l_2}, \dots, c_i^{l_j}, \dots, c_i^{l_k}\}$ where $c_i^{l_j}$ is the confidence score of the j^{th} LLM for the i^{th} query.

The performance of SELECTLLM is determined by the selection policy used. For example, a greedy policy that always selects the LLM (or set of LLMs) with the highest confidence may be suboptimal. This is because another LLM (or set of LLMs) in the pool might have higher accuracy but may not be chosen due to slightly lower confidence. Additionally, when there are two subsets with similar confidence and accuracy, it is more efficient to select the subset with lower cumulative latency. Considering these aspects, we propose the following three optimal policies:

1. **LABELLEDMAXCONF:** This policy selects the top- s LLMs (L_s) for an input query q_i based on two constraints: (i) the LLMs should be present in the MLC predictions, and (ii) Only those LLMs that have confidence scores within the top- s from c_i are considered.

Algorithm 1 SELECTLLM Inference Algorithm with WEIGHTEDMAXCONF Policy

Require: Queries \mathcal{Q} , LLMs \mathcal{L} , SLDATA $(\mathcal{X}, \mathcal{Y}) \in \mathcal{D}$, pre-trained model \mathcal{M}_θ with parameters θ

- 1: Fine-tune the model: $\mathcal{M}_{\hat{\theta}} = \arg \min_{\theta} \sum_{(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})} \text{Loss}(\mathcal{M}_\theta(x_i), y_i)$ ▷ Fine-tuning \mathcal{M}_θ with \mathcal{D}
 - 2: **for** each query q_i in \mathcal{Q} **do**
 - 3: Perform inference: $q_i^{\text{logits}}, q_i^{\text{labels}} = \mathcal{M}_{\hat{\theta}}(q_i)$ ▷ Using fine-tuned model $\mathcal{M}_{\hat{\theta}}$
 - 4: Calculate confidence scores: $c_i = \sigma(q_i^{\text{logits}})$ ▷ σ is the sigmoid activation function
 - 5: Select top- s confidence scores: $c_i^{(s)} = \max(c_i, s)$ and associated LLMs: $L_s = L(c_i^{(s)})$
 - 6: Initialize an empty set for answer set $A_i \leftarrow \phi$
 - 7: **for** each l_j in L_s **do**
 - 8: Generate M responses: $a_i^M = l_j(q_i)$ ▷ Generate M responses with LLM l_j
 - 9: Find answer frequency $a_i^f = \{a_i^k : \text{countof}(a_i^k) / \sqrt{c_i^{(s)j}} \mid \text{for } a_i^k \text{ in } \text{unique}(a_i^k)\}$
 - 10: $A_i \leftarrow A_i \cup a_i^f$
 - 11: **end for**
 - 12: Return most frequent answer in A_i
 - 13: **end for**
-

2. **MAXCONF:** This is a more flexible policy than LABELLEDMAXCONF as it only takes into account the second constraint, i.e., it selects the top- s LLMs corresponding to the top- s confidence scores from c_i .
3. **WEIGHTEDMAXCONF:** This policy begins by selecting the top- s LLMs based on their high confidence scores, i.e., for a given query q_i , we denote the selected LLMs as $L_s^{q_i}$. Subsequently, we modify the frequency of answer values extracted from the responses of each selected LLM, which involves dividing the frequency of each value by the square root of the confidence score associated with the respective LLM. Finally, we collect all response values and their modified frequencies across the selected LLMs (frequencies are added for the same value). The value with the highest frequency after the modification is selected as the final response. The formal steps are presented in Algorithm 1. Intuitively, dividing by the square root of the confidence score aims to mitigate biased selection of LLM by the policy similar to (Wu et al., 2024). This adjustment ensures fairer opportunities for each selected LLM to contribute to the majority voting.

Across all three policies, in case of a conflict where two LLMs have similar confidence, the *light-weight* LLM (i.e., the lower latency) is preferred.

5 Experimental Setup

5.1 Baseline Models

Based on recent literature, the following baseline models are included for comparison:

1. **Oracle:** The maximum performance is assumed under the premise that an Oracle always selects the lowest latency subset of LLMs that generates the correct majority vote answer for each question (if possible; otherwise, the question attempt is marked as incorrect). Empirically, this is obtained by evaluating all subsets of LLMs, i.e., $(2^k - 1)$, where k is the total number of LLMs.
2. **Random:** This represents the mean performance of uniformly randomly selecting an LLM subset from all possible $(2^k - 1)$ subsets for each query. We report mean scores across 1,000 independent runs to avoid biases.
3. **All LLMs:** This baseline reports the mean accuracy of MAJ@ $(M \times |L|)$ based on the combined pool of M generations from each LLM and is similar to Li et al. (2024).
4. **LLM-Blender** (Jiang et al., 2023): An ensembling framework was developed to utilize the diverse strengths of multiple open-source LLMs. Specifically, it employs PAIRRANKER, which utilizes a cross-attention-based method for pairwise comparison of different LLM responses to determine the superior one. We use the officially released model checkpoint in our setting.
5. **Top- s LLMs:** For this baseline, we consider the responses of the top- s scoring LLMs using a majority-vote strategy. The top-performing models are determined by comparing the overall accuracies of the LLMs.

For *All LLMs*, *LLM-Blender* and *Top- s LLMs*, the latency remains constant since they need inference with all LLMs to determine the performance.

5.2 Evaluation Metrics

We evaluate the performance of all models with the *accuracy* (Acc) metric using majority voting (see Section 4.1.2). Additionally, we report the *latency per query* (Lat) to estimate efficiency. The exact costs of model execution, including factors like latency, FLOPs, and energy consumption, may vary and are influenced by factors such as prompt templates, hardware capabilities, and network connectivity, especially in LLM inference scenarios. To ensure a fair comparison, we record the inference latency of each LLM under uniform conditions using single A100 GPUs. The individual latencies for each LLM are detailed in Appendix Table 7.

6 Results and Discussion

Table 2 presents the performance results for the Oracle, baselines, and the proposed SELECTLLM models across both GSM8K and MMLU datasets. We have also reported respective inference latencies to analyze the efficiency of different models. We make the following major observations:

Performance of baselines: We evaluated four baseline models – *Random*, *All LLMs*, *LLM-Blender*, and *Top-s LLMs* – to assess the impact of ensembling LLMs. The performance of the *Random* baseline model surpasses that of several individual LLMs reported in Fig. 1, demonstrating the potential of utilizing multiple LLMs. Li et al. (2024) reports that increasing the number of LLMs in the model pool generally increases ensemble performance. Interestingly, however, selecting only the top few LLMs in our experiments performs better than including all LLMs (in *all LLMs*). Thus, the optimal number of LLMs may depend on the diversity of the model pool (i.e., individual capabilities and overall performance on benchmarks). The *Top-s LLMs* model proves to be the most effective baseline, outperforming even robust approaches such as LLM-Blender, indicating the benefit of prioritizing the best-performing LLMs. However, the latency of most baseline models is high (as they must utilize the entire model pool at inference), limiting their real-world practicality.

The effect of different policies with the SELECTLLM algorithm: It can be observed that the LABELLEDMAXCONF policy yields the lowest performance on the GSM8K dataset. This suboptimal performance may be attributed to the policy’s dependence on both the MLC predictions and

confidence scores, wherein the classifier predominantly assigns the label *metamath-7b-1m*. Given that *metamath-7b-1m* is a specialized model for mathematics and 88% of the SLDATA training data is associated with this label, this reliance skews the predictions. However, this effect is minimal for the MMLU dataset, where the label distribution across LLMs is more balanced. This limitation is addressed by the MAXCONF and WEIGHTEDMAXCONF, which relax the constraint on MLC label prediction and only operate on confidence scores. This allows models to incorporate other LLMs and push the performance, particularly for the GSM8K data. Mathematical tuning in WEIGHTEDMAXCONF allows policies to select LLMs more effectively and improve the scores. Overall, the WEIGHTEDMAXCONF policy emerged as the best performing, with a slight edge over its closest competitor, the MAXCONF policy.

SELECTLLM vs. baseline models: The proposed SELECTLLM model, utilizing the WEIGHTEDMAXCONF policy, demonstrates superior performance compared to the *Random*, *All LLMs*, and *LLM-Blender* baselines. Additionally, SELECTLLM maintains competitive accuracy relative to the *Top-s LLMs* baseline while achieving significantly lower latency. Specifically, it reduces latency by 13% for the GSM8K dataset and by 70% for the MMLU dataset compared to all baselines (except *Random*). This substantial efficiency gain underscores the effectiveness of SELECTLLM in balancing performance and operational efficiency.

Query awareness analysis of SELECTLLM: With this analysis, we aim to understand how the distribution of selected LLMs changes as more LLMs are selected (with increasing values of s) in SELECTLLM algorithm. Figure 6 in the Appendix presents such distribution. For both datasets, in the top-1 and top-2 subsets, most of the queries are directed to the best-performing LLMs. However, as the subset size increases, the dominance of the top-performing models diminishes, leading to a more uniform distribution where queries are routed towards more LLMs to boost the performance. This indicates the input query awareness of the SELECTLLM model, which is adept at assigning a suitable set of LLMs for the input query.

Prediction distribution analysis: Appendix Figure 7 presents the distribution of the number of input queries with correct answer predictions us-

Models / Setups		GSM8K		MMLU	
		Acc (↑)	Lat (↓)	Acc (↑)	Lat (↓)
Oracle		90.52	3.24	90.46	1.75
Baseline	Random	69.49	9.65	58.20	8.27
	LLM-Blender (Jiang et al., 2023)	75.28	19.00	60.27	16.40
	All LLMs (Li et al., 2024)	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
SELECTLLM	MLC + WEIGHTEDMAXCONF ($s = 1$)	67.24	4.70	63.52	2.97

Table 2: Performance of different models on the GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021) test sets. By default, $s = 4$ is used for GSM8K and $s = 2$ for MMLU. The value of s is selected based on experiments on the validation set (see Appendix Section 6). Acc: with MAJ@($M \times L_s$) scores reported in percentage (%); Lat: runtime of M generations for a single query (in seconds); MLC: multi-label classifier; s : the number of LLMs considered.

ing the top-3 individual LLMs and models with the SELECTLLM algorithm across both datasets. It can be observed (count from the right column) that the proposed model is able to correctly provide answers to input queries compared to other individual LLMs, which supports the performance gain reported in Table 2. The distribution also indicates (count from the bottom) that the proposed model utilizes the capabilities of multiple individual LLMs to extract the correct answers. Additionally, Figure 9 illustrates the subject-wise performance differences between SELECTLLM and the best performing individual LLM (i.e., gemma-7b-1m) for MMLU. It can be observed that the proposed model shows substantial gains in the majority of subjects, while performing slightly worse for a few of them, showing scope for improvement.

Upper-bound and performance gap with SELECTLLM: We measure the maximum performance achievable by SELECTLLM with the WEIGHTEDMAXCONF policy by establishing an upper-bound performance. Specifically, we compute potential subsets based on the labels predicted by the MLC for each query q_i . A query is considered solved if at least one subset yields a correct answer. The upper bound scores for GSM8K and MMLU are 78.77 and 76.20, respectively. This reveals the following insights: (1) the classifier’s performance is constrained, with weighted F1 scores of 0.71 for GSM8K and 0.68 for MMLU, due to limited training data (approximately 7K for GSM8K and 14K for MMLU), likely contributing to the performance gap between SELECTLLM and Oracle. *Therefore, augmenting the training data or improving the classifier model could enhance scores.* (2) A performance disparity between the best policy and the upper bound suggests the potential for develop-

ing better policies (specifically for MMLU). *However, since policies rely on the classifier’s confidence scores, enhancing the classifier could also bridge this gap.* We will investigate this performance gap from a linguistic perspective in Section 6.1.

Ablation studies: *Is the proposed algorithm effective with different LLMs pool sizes (i.e., value of k)?* To investigate this question, we conduct ablation studies considering various LLM pool sizes, i.e., $k = 1, \dots, 6$, for the GSM8K dataset. We examine two extreme settings: pools with top- k and bottom- k LLMs based on individual LLM performance. This encompasses many configurations as LLMs with similar or different performances may be present in the LLM pool. We also compare this with closet strong Top- s LLM baseline. The results are shown in the Appendix Figure 8. We observe that even with different k values across both top- k and bottom- k setups, the proposed SELECTLLM outperforms (in terms of accuracy) the Top- s LLM baseline. *This indicates that the proposed approach is LLM pool-agnostic.* Moreover, as the number of k values increases, latency becomes a factor: for larger pool sizes, the latency for SELECTLLM is much lower than for Top- s LLM. Similar results are observed with the MMLU dataset.

Latency vs. Accuracy vs. s -Value: Figure 4 shows the relationship between latency, accuracy, and s -values for different SELECTLLM policies across both datasets along with the corresponding development sets. WEIGHTEDMAXCONF consistently outperforms both MAXCONF and RANDOM in terms of accuracy and latency for most s -values (except for 5 and 6 on GSM8K). This highlights the superiority of WEIGHTEDMAXCONF, maintaining its effectiveness even with a small number of

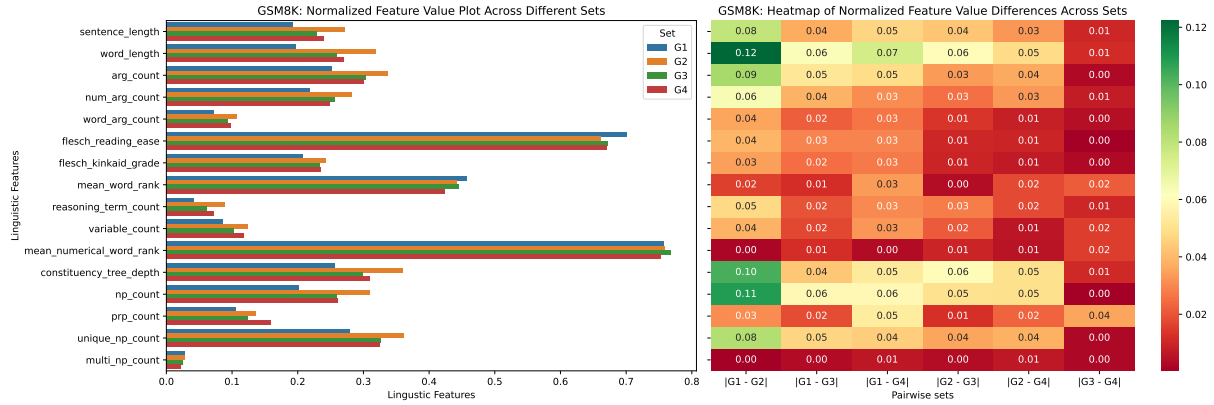


Figure 3: Quantitative Analysis (GSM8K): Distribution of normalized linguistic feature values (left) and the pairwise absolute difference in normalized feature values between different sets (right).

LLMs. A larger number of LLMs and higher latency are required for GSM8K, whereas a lower s -value and latency suffice for MMLU to achieve high accuracy. *Selecting 4 LLMs for GSM8K and 2 for MMLU yields optimal performance, demonstrating the effectiveness and efficiency of SELECTLLM with WEIGHTEDMAXCONF.* Using this optimal value of s results on the test set are reported in Table 2.

Additionally, we also report the performance of SELECTLLM with WEIGHTEDMAXCONF for $s = 1$ in Table 2 to understand the impact of selecting the best query-aware LLM. The results indicate that there is a large performance gap when selecting multiple LLMs, underscoring that ensembling is a promising direction.

6.1 Interpretation of Performance Gaps

We investigate the underlying factors contributing to the performance gap between SELECTLLM and the Oracle. To gain insights, we perform both linguistic feature-based quantitative analysis and a SHAP values-based qualitative analysis, examining four subsets (i.e., group) of the test data (numbers are specific to the GSM8K dataset): **G1**: The set of questions correctly solved by all individual LLMs (949 questions), **G2**: The set of *additional* questions solved only by SELECTLLM and the Oracle (46 questions), **G3**: The set of *additional* questions only solved by the Oracle (168 questions) and **G4**: The set of questions that could not be solved by any LLM (125 questions).

Quantitative Analysis: We extracted 16 linguistic features specific to input questions, as proposed by Srivatsa and Kochmar (2024) for the GSM8K dataset. In Figure 3, we present the feature value

distribution (bar plot) and pair-wise value differences (heatmap) across all four subsets. Major insights are: (1) The feature value distributions for G3 and G4 are notably similar, suggesting that questions only solvable by the Oracle exhibit comparable complexity to those not solved by any LLMs. (2) Features such as *length*, *tree depth*, *NP count*, and *argument count* emerge as influential across the different subsets, which aligns with the findings of Rabinovich et al. (2023) and Srivatsa and Kochmar (2024). A similar observation holds for the MMLU dataset. Future models should incorporate these features to address the performance gap between SELECTLLM and the Oracle.

Qualitative Analysis: We extracted word-level SHAP values¹ from the SELECTLLM classifier for GSM8K. After mapping key content words to WordNet (Miller, 1994) synsets, we identified synset groups most detrimental to the solvability of each LLM. Table 3 shows the corresponding SHAP values for GSM8K. See Appendix K for more details. We find that: (1) Not all synset groups affect LLMs uniformly; some are detrimental to one model but beneficial to another (e.g., *small numbers & quantifiers*). *Frequency terms* (e.g., “twice”, “thrice”) consistently reduce solvability for most LLMs, but MetaMath remains resilient. (2) *Gain using SELECTLLM*: SELECTLLM solves questions with more challenging *fractional values* and *rates and ratios* than individual LLMs. (3) *Gap between SELECTLLM and Oracle*: *quantifiers*, *age units*, *other units*, and questions about groups like “family” or “team” relatively increase difficulty in G3 compared to G2. (4) *Solved by None*:

¹Some samples of interactive SHAP plots are hosted at <https://anonymous.4open.science/r/SelectLLM-3621>.

Set	LLM	frequency terms	time duration units	age units	small numbers	quantifiers	ordinals	fractional values	rates and ratios	named entities	other units	groups
(G1) Solved by All	gemma-7b-it	-0.938	-0.845	-0.167	-1.0	-0.982	-0.084	-0.138	0.014	1.0	0.073	0.296
	gemma-7b-lm	-0.936	0.624	0.214	1.0	1.0	0.152	0.743	-0.172	1.0	0.699	0.669
	llama2-13b-chat	-0.988	-0.786	-0.095	-0.999	-0.837	-0.038	-0.264	-0.107	1.0	0.135	0.555
	metamath-7b-lm	0.065	0.962	0.33	1.0	1.0	0.279	0.989	0.372	1.0	0.753	0.687
	mistral-7b-it	-0.993	-0.531	-0.058	-0.842	0.98	-0.023	0.195	0.049	1.0	0.364	0.519
mistral-7b-lm	-0.994	-0.452	-0.009	-1.0	0.857	-0.009	0.085	-0.002	1.0	0.327	0.472	
(G2) Solved by SelectLLM	gemma-7b-it	-1.0	-0.998	-0.949	-1.0	-1.0	-0.312	-0.996	-0.607	1.0	-0.225	0.911
	gemma-7b-lm	-1.0	0.775	0.999	1.0	1.0	0.854	-0.268	0.472	1.0	0.913	0.988
	llama2-13b-chat	-1.0	-0.994	-0.068	-1.0	-1.0	0.021	-0.998	-0.717	1.0	0.081	0.977
	metamath-7b-lm	0.45	1.0	1.0	1.0	1.0	0.945	0.995	0.897	1.0	0.964	0.968
	mistral-7b-it	-1.0	-0.988	0.07	-1.0	1.0	0.34	-0.996	-0.314	1.0	0.504	0.983
mistral-7b-lm	-1.0	-0.984	0.448	-1.0	0.989	0.322	-0.994	-0.513	1.0	0.472	0.966	
(G3) Solved by Oracle	gemma-7b-it	-0.674	-0.77	-0.334	-1.0	-0.977	-0.272	-0.176	0.05	-1.0	-0.281	0.037
	gemma-7b-lm	-0.707	0.185	0.248	0.999	0.502	0.357	0.351	0.332	1.0	0.183	0.129
	llama2-13b-chat	-0.835	-0.689	-0.183	-1.0	-0.944	-0.116	-0.043	0.042	0.988	-0.241	0.12
	metamath-7b-lm	0.102	0.818	0.466	1.0	0.999	0.599	0.62	0.367	1.0	0.475	0.196
	mistral-7b-it	-0.864	-0.635	-0.194	-1.0	-0.542	-0.027	-0.01	0.195	0.998	-0.197	0.102
mistral-7b-lm	-0.88	-0.681	-0.189	-1.0	-0.734	-0.013	0.013	0.115	1.0	-0.166	0.09	
(G4) Solved by None	gemma-7b-it	-0.646	-0.684	-0.817	-1.0	-0.948	-0.279	-0.713	-0.317	-1.0	-0.192	-0.002
	gemma-7b-lm	-0.629	0.246	0.129	0.997	0.802	-0.07	-0.245	0.171	1.0	0.112	0.198
	llama2-13b-chat	-0.776	-0.53	-0.742	-1.0	-0.872	-0.259	-0.75	-0.332	-0.981	-0.165	0.113
	metamath-7b-lm	0.145	0.763	0.857	1.0	0.992	0.264	0.525	0.513	1.0	0.304	0.304
	mistral-7b-it	-0.858	-0.508	-0.744	-1.0	-0.463	-0.265	-0.735	-0.132	-0.884	-0.127	0.111
mistral-7b-lm	-0.849	-0.47	-0.753	-1.0	-0.606	-0.223	-0.712	-0.237	-0.06	-0.117	0.103	

Table 3: Normalized SHAP values for the SelectLLM classifier (GSM8K) over questions from the four subsets of the Test split (see Section 6.1). Each column represents a related group of Synsets. A negative SHAP value (in red) is detrimental to the predicted solvability of corresponding questions, and a positive value (in green) is beneficial.

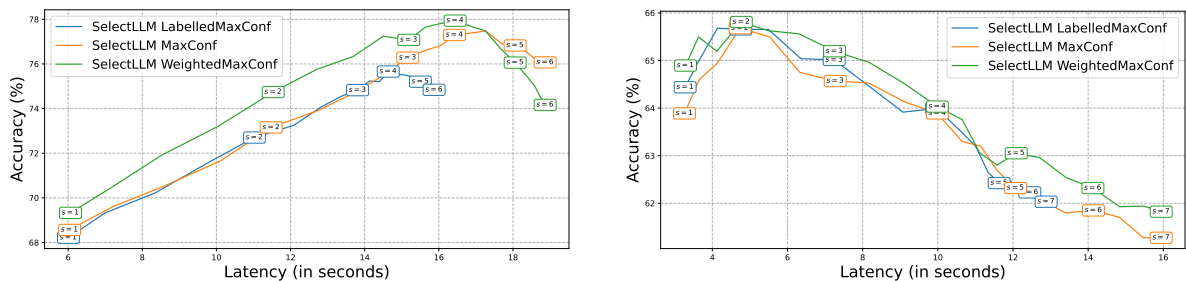


Figure 4: Latency vs. accuracy vs. s -values plots for different SELECTLLM policies on GSM8K (left) and MMLU (right) development sets. s -value: total number of LLMs selected with SELECTLLM model.

G3/G4 rows show that unsolvable questions contain challenging *age units*, *ordinals* (e.g., "first", "third"), and *named entities*.

6.2 Out-of-domain (OOD) Generalization

This section presents a preliminary experiment to validate out-of-domain (OOD) generalization of SELECTLLM using the MMLU dataset. As described in Section 4.1, the MMLU dataset comprises 57 category-specific subsets based on grade level (high school and college), subject, domain, etc. We conducted two OOD experiments: one at the grade level and another at the subject level. For the grade-level experiment, the model was trained on high school data and evaluated on college-level data. For the subject-level experiment, we created two non-overlapping splits for training and testing. The results, presented in Table 4, indicate that OOD performance is comparable to in-domain performance, with subject-level generalization yielding even higher accuracy.

7 Conclusions and Future Directions

In this work, we introduce an efficient and novel LLM selection algorithm, SELECTLLM, to nav-

Setup	Acc (\uparrow)	Lat (\downarrow)
OOD (Grade-level)	58.97	4.80
OOD (Subject-level)	79.75	4.89
In-domain	65.81	4.78

Table 4: OOD experiment results with MMLU dataset for SELECTLLM with WEIGHTEDMAXCONF policy and $s=2$.

igate input queries to the most suitable subset of LLMs from a large pool. SELECTLLM employs a multi-label classifier and confidence-based optimal policies to select a lightweight subset of LLMs. The model is evaluated on two challenging reasoning datasets and compared against several strong baseline models. SELECTLLM demonstrates superior performance compared to these baseline models, achieving competitive accuracy with a similar subset size of top-performing LLMs, while maintaining significantly lower latency. Despite the promising results, we recognize that further advancements in modeling could bring performance closer to the Oracle benchmark. Our findings provide a robust foundation for future research. A potential avenue for improvement could involve integrating LLM-related and query-related features to enhance the model's query and LLM awareness.

Limitations

We believe that this study provides a useful starting point for optimal subset selection of LLMs from a large pool. However, we acknowledge that there are certain limitations of this work, and addressing these limitations in the future is an important task.

Extension to general generation tasks The scope of the current study is limited to two challenging reasoning benchmarks: GSM8K and MMLU. These tasks are framed as generative tasks, where LLMs are required to produce discrete final answers. The study does not focus on general generation tasks such as machine translation, question generation, etc.

It is well known that the *one model fits all* approach is not universally applicable (Huang et al., 2024). For instance, advanced LLMs such as DeepSeek-R1 (DeepSeek-AI et al., 2025) demonstrate superior performance in reasoning tasks but exhibit comparatively lower effectiveness in general generation tasks (Mercer et al., 2025). Similar observations hold for OpenAI’s and Google’s reasoning models. Building on this observation, we posit that SelectLLM serves as a pivotal step toward the efficient selection of an optimal subset of LLMs from a larger pool, balancing both latency and accuracy, with a focus on reasoning tasks. While the current study does not cover general generation tasks, which are left for future work, the primary challenge for such adaptation lies in implementing robust voting mechanisms. These can be modeled in a simpler manner, as proposed in the All LLMs (Li et al., 2023) baseline, to ensure feasibility. Similarly, the idea proposed by Odu-making et al. (2024) could be adapted for general generation tasks.

Collection of larger training data for multi-label classifier Another limitation of the proposed SELECTLLM algorithm is the limited availability of training data for the multi-label classifier, with only 7K instances for GSM8K and 14K for MMLU. This limitation can potentially lead to biased learning. Despite several measures to address this issue, such as weighing labels to counteract label imbalance, conducting extensive optimal hyperparameter searches, experimenting with different sizes of probabilistic and LLM-based models (with RoBERTa performing the best), and obtaining the best checkpoint with the validation set, the performance remains suboptimal. The algorithm achieves a weighted F1

score of 0.71 for GSM8K and 0.68 for MMLU.

Reduction in invalid answers We are able to extract viable answers for 92% to 95% of queries across different LLMs, while the remaining queries resulted in invalid/incorrect outputs from the extraction algorithm. These invalid responses can be attributed to two primary factors: (i) limitations in the LLMs’ response generation, where outputs are not structured in an extractable format, and (ii) limitations in the extraction algorithm, which fails to accurately parse the generated text. Similar limitations have also been reported in previous studies (Singh et al., 2023; Chen et al., 2022), particularly for tasks requiring discrete outputs from LLMs. Since the extraction algorithm is the same across all LLMs, the findings are expected to remain the same even if a more effective extraction algorithm is proposed. Future work should prioritize the development of advanced prompting methodologies and optimized LLM architectures that facilitate the generation of discrete, extraction-friendly outputs. Additionally, refining extraction algorithms to improve robustness and adaptability can further reduce the rate of invalid responses.

Ethics Statement

This paper introduces the SELECTLLM algorithm, a novel approach designed to leverage the diverse capabilities of various LLMs. While SELECTLLM utilizes LLMs, it is crucial to recognize that, independent of this study, LLMs inherently present risks. These models may generate outputs that, despite being plausible, are factually inaccurate or nonsensical. Such *hallucinations* can lead to misguided decision-making and the propagation of biases, particularly in high-stakes contexts where accuracy is paramount. In the absence of appropriate safeguards, the broad deployment of LLMs could exacerbate these issues. Thus, it is imperative to develop mechanisms that mitigate the risks of hallucinations to ensure the responsible and effective application of these models.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.
- 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*.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *Preprint*, arXiv:2501.12948.
- Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, 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 Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Mengya Hu, Rui Xu, Deren Lei, Yaxi Li, Mingyu Wang, Emily Ching, Eslam Kamal, and Alex Deng. 2024. SLM Meets LLM: Balancing Latency, Interpretability and Consistency in Hallucination Detection. *arXiv preprint arXiv:2408.12748*.
- Wenhao Huang, Qianyu He, Zhixu Li, Jiaqing Liang, and Yanghua Xiao. 2024. [Is There a One-Model-Fits-All Approach to Information Extraction? Revisiting Task Definition Biases](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10274–10287, Miami, Florida, USA. Association for Computational Linguistics.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023.

- Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.
- Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. [More Agents Is All You Need](#). *ArXiv*, abs/2402.05120.
- Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.
- Sarah Mercer, Samuel Spillard, and Daniel P Martin. 2025. Brief analysis of DeepSeek R1 and it's implications for Generative AI. *arXiv preprint arXiv:2502.02523*.
- George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncl Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.
- Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. *arXiv preprint arXiv:2406.02061*.
- Ayomide Odumakinde, Daniel D'souza, Pat Verga, Beyza Ermis, and Sara Hooker. 2024. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress. *arXiv preprint arXiv:2408.14960*.
- Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. 2023. [Predicting Question-Answering Performance of Large Language Models through Semantic Consistency](#). *Preprint*, arXiv:2311.01152.
- Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.
- Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. 2023. [Large Language Model Routing with Benchmark Datasets](#). *Preprint*, arXiv:2309.15789.
- Chandan Singh, John Morris, Alexander M Rush, Jianfeng Gao, and Yuntian Deng. 2023. Tree prompting: Efficient task adaptation without fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6267.
- KV Aditya Srivatsa and Ekaterina Kochmar. 2024. What Makes Math Word Problems Challenging for LLMs?
- KV Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2024. Harnessing the Power of Multiple Minds: Lessons Learned from LLM Routing.
- Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the Factual Consistency of Large Language Models Through News Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.
- Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint arXiv:2501.06322*.
- Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. 2023. Fusing Models with Complementary Expertise. *arXiv preprint arXiv:2310.01542*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *Advances in neural information processing systems*, 35:22199–22213.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Auto-gen: Enabling next-gen LLM applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yu-Chang Wu, Shen-Huan Lyu, Haopu Shang, Xiangyu Wang, and Chao Qian. 2024. Confidence-aware contrastive learning for selective classification. *arXiv preprint arXiv:2406.04745*.

Min-Ling Zhang, Yu-Kun Li, Hao Yang, and Xu-Ying Liu. 2020. Towards class-imbalance aware multi-label learning. *IEEE Transactions on Cybernetics*, 52(6):4459–4471.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. *Sentiment Analysis in the Era of Large Language Models: A Reality Check*. Preprint, arXiv:2305.15005.

A Prompting Templates and Answer Extraction

Considering diverse LLMs and benchmarks adds challenges to prompting, as no single uniform prompting approach fits all LLMs (Sclar et al., 2023). Based on insights from recent work on the appropriate usage of prompts (Sahoo et al., 2024) and our own experiments, we make the following observations about prompting trends:

1. For non-chat LLMs, few-shot chain-of-thought (COT) prompting (Wei et al., 2022) works better than zero-shot (Kojima et al., 2022) across both datasets. Therefore, we use five few-shot random examples obtained from the development set. The few-shot prompting results in over 95% *viable* answers (except for the llama2-7b-1m LLM, which has a viability score of 83%). An answer is considered *viable* if it is represented by a single numeric/alphabetic string that can be extracted from the generated solution to compare with the reference answer. Viability is estimated using an automated regular expressions based-script and verified through manual inspection.
2. For chat LLMs, few-shot COT distracts the generator, leading to unexpected outputs, so zero-shot COT works best. To ensure correctness, we utilize different models’ chat templates from HuggingFace.² The viability of answer extraction for chat models is approximately 92%. Examples of zero-shot and few-shot prompting are presented in Appendix Figure 5.

The limitation of answer extraction from 92% to 95% (except for llama2-7b-1m) can be attributed

²https://huggingface.co/docs/transformers/en/chat_templating

to two factors: (i) limitations in the LLMs’ response generation, where outputs are not structured in an extractable format, and (ii) limitations in the extraction algorithm, which fails to accurately parse the generated text. For more details, see the limitations section.

The adapted prompting approaches used in our LLM queries are designed to instruct LLMs to specify that their final answers should be provided at the very end of each of their responses. We thus use a simple answer extraction policy of selecting the last mentioned numerical value (for GSM8K) and multiple-choice option (for MMLU) from the generated responses. Figure 5 in the Appendix shows a sample generation example. Responses failing to include any final answer are considered non-viable or invalid (‘INVALID’) and counted as incorrect responses. For MMLU, we evaluate the extracted options directly against the annotated correct answers (‘A’, ‘B’, ‘C’, and ‘D’) from the dataset. For GSM8K, questions where the absolute difference between the ground truth and predicted numerical answers is less than $\epsilon = 0.1$ are evaluated as solved correctly. This threshold was set to accommodate instances where model-generated real-valued answers differ slightly from the expected answers, e.g., due to rounding errors.

B Data Statistics and Selected LLMs

The statistics of the dataset and the list of considered diverse LLMs are presented in Table 5 and 6, respectively.

C LLM Latency Estimation

Runtime statistics of various LLMs are presented in Table 7.

D Sample Prompts

Prompting examples of few-shot COT and zero-shot COT are illustrated in Figure 5.

Split	GSM8K	MMLU
Train	6,816	13,757
Validation	359	285
Test	1,319	1,530

Table 5: Dataset statistics for GSM8K and MMLU benchmarks. For the MMLU, the officially released training and test splits have been swapped to align with the distribution of the GSM8K.

LLMs	Chat?	Spec?	#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

Table 6: List of diverse LLMs selected for this study. Spec: Specialized LLM

LLM	Prompt Type	GSM8K (prompt/sec)	MMLU (prompt/sec)
Few-shot COT	llama2-7b	4.21	2.30
	gemma-7b	7.10	3.00
	mistral-7b	3.70	1.10
	metamath-7b	4.70	2.40
Zero-shot COT	gemma-7b-it	0.70	1.00
	llama2-13b-chat	1.80	4.80
	mistral-7b-it	3.70	1.80

Table 7: Runtime statistics on various LLMs over M generations for each input query. The timings are recorded using a single A100 GPU. ‘sec’ denotes seconds, and COT denotes Chain-of-thought. For few-shot COT, we have considered 5 random examples from the development set.

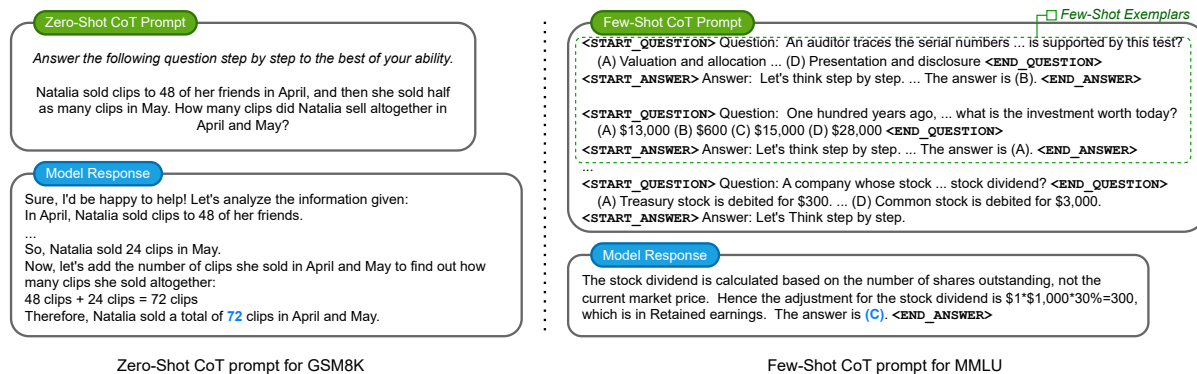


Figure 5: Sample zero-shot COT prompt for a chat (or instruction-tuned) LLM is shown for a GSM8K sample, and the few-shot COT prompt for non-chat LLMs is shown for an MMLU sample. Note that the type of prompting is associated with the LLM type rather than the dataset. Specifically, non-chat LLMs use few-shot COT, while chat models use zero-shot COT.

E Distribution of Different LLMs in Top-s Subset Selected with SELECTLLM

The details are shown in Figure 6.

F Query Response Distribution with Different Models

The details are presented in Figure 7.

G Implementation Details

Querying LLMs We use the vLLM³ package to query LLMs. All models were queried with a temperature of 0.8 and a max token length of 2000. Each question prompt was queried K times with different initialization seeds. We used a single NVIDIA A100 GPU for all runs. Querying each dataset once took approximately 1-2 hours.

³<https://github.com/vllm-project/vllm>

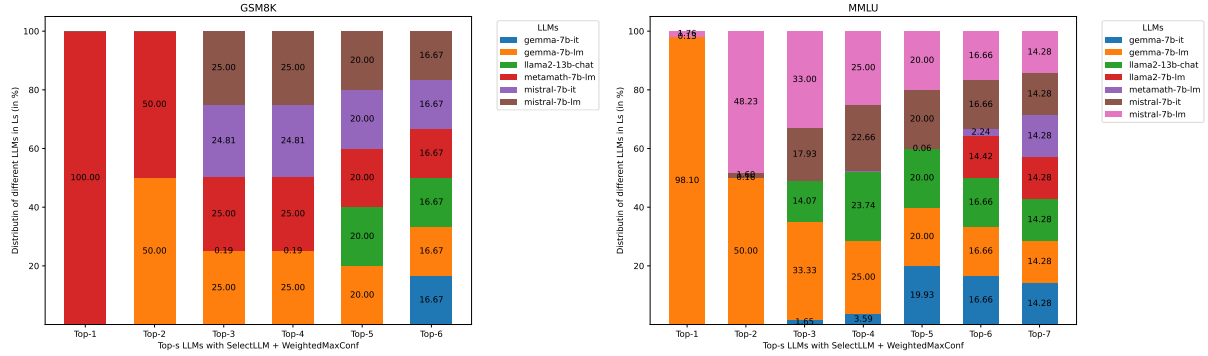


Figure 6: The distribution of different LLMs in the predicted subset of LLMs with SELECTLLM algorithm for both GSM8K (top) and MMLU (bottom) datasets.

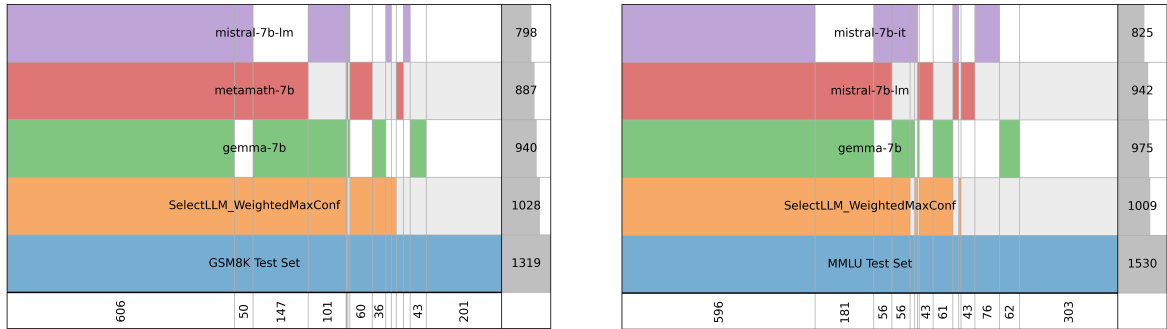


Figure 7: Distribution of the top-3 individual LLMs and proposed model responses to input queries for the test split of the GSM8K (left) and MMLU (right) datasets. The count in the rightmost column of each figure indicates the number of queries whose responses have been correctly answered by each LLM or the proposed model. The counts at the bottom denote the number of queries that have correct answers from one or more LLMs or the proposed model.

Model/Setup	GSM8K	MMLU
BERT (base-uncased)	0.63	0.59
BERT (base-uncased) + Weighted-label penalty	0.64	0.62
RoBERTa (base)	0.67	0.61
BRoBERTa (base) + Weighted-label penalty	0.71	0.68
T5 (base)	0.66	0.60
T5 (base) + Weighted-label penalty	0.69	0.65

Table 8: The F1 score for the MLC classifier is evaluated with different pre-trained language models. Additionally, a weighted-label penalty is applied to handle label imbalance, inspired by (Zhang et al., 2020). This evaluation is conducted across both GSM8K and MMLU datasets.

MLC Training We use the HuggingFace⁴ library for loading and tuning all pre-trained Transformer encoders in our experiments. Each model was trained for 10 epochs, with an initial learning rate of $1e-6$, a warmup ratio of 0.1, and class-balanced CrossEntropy loss. The training checkpoint with the lowest validation loss was selected for inference. In this study, we have considered $M=10$ inspired by Li et al. (2024).

⁴<https://huggingface.co/>

H Performance of Individual LLMs

Table 9 presents the individual LLM performance for both the GSM8K and MMLU datasets. There is no clear best-performing model for both datasets. For instance, the gemma-7b-lm model performs the best in terms of accuracy for GSM8K, but its latency cost is also high. Similar trends are observed for the MMLU dataset. The performance of individual LLMs should not be directly compared with the SELECTLLM model. Selecting the best LLM with high accuracy from a large pool requires running inferences across all LLMs to identify the top performer, resulting in the cumulative latency of all models, rather than just the selected one. On the other hand, randomly selecting an LLM may not necessarily choose the best performer. In this context, comparing SELECTLLM with ensemble-based baselines is more appropriate, which is done in the main results (Table 2) where SELECTLLM outperforms all baseline models, such as strange LLM-Blender baseline, in terms of both latency and accuracy.

Models / Setups	GSM8K		MMLU	
	Acc ↑	Lat ↓	Acc ↑	Lat ↓
gemma-7b-1m	71.27	7.10	63.73	3.00
mistral-7b-1m	60.50	3.70	61.57	1.80
metamath-7b-1m	67.25	4.70	41.76	2.40
llama2-7b-1m	–	–	48.10	2.30
llama2-13b-chat	49.20	1.80	52.94	4.80
mistral-7b-it	56.71	1.00	53.92	1.10
gemma-7b-it	42.23	0.70	50.72	1.00

Table 9: Performance and latency scores for different LLMs on GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021) test sets. Acc: accuracy in percentage (%); Lat: latency in seconds; ‘–’: values are not available.

I Ablation Studies

The details are presented in Figure 8.

J Importance of Domain: A Case Study with MMLU

The MMLU dataset comprises 57 subjects. In this analysis, we evaluate the performance of the proposed SELECTLLM algorithm using the WEIGHTEDMAXCONF policy, employing the best-performing individual LLM (gemma-7b-1m) on a subject-wise basis. Figure 9 illustrates that while the proposed model’s performance may be sub-par for a few subjects, it demonstrates high performance for a significant portion of the subjects, indicating the effectiveness of the proposed model in general.

K SHAP Analysis

In this section, we provide the details of the SHAP-values-based qualitative analysis discussed in Section 6.1.

Extracting Shapley Values We first extract token-level Shapley values for the input question text to the fine-tuned MLC classifiers for each LLM using the PartitionExplainer module of the SHAP library⁵. The token-level SHAP values are then summed according to word boundaries to get word-level SHAP values. This is a valid operation as SHAP values are additive in nature.

Synset Mapping In order to interpret the impact of different semantic word categories on the predicted solvability of questions, we map all viable words to their corresponding synsets using NLTK’s WordNet module⁶. In this process, all instances of

a word are recorded as instances of its respective synset. Next, we merge entries for synsets that represent less than 0.3% (decided by observation) of the total number of tokens with their hypernyms. This helps reduce sparsity in the synsets to review. The mean SHAP values for each synset are then calculated across the test set of the datasets.

Synset Grouping Based on the top 15 most detrimental synsets for each LLM in each of the four sets (i.e., having the lowest non-positive average SHAP values), we manually identify groups of related and recurring synsets. The groups for GSM8K, as presented in Table 3, are as follows:

- **frequency terms:**
Synset(‘twice.r.01’),
Synset(‘thrice.r.01’),
Synset(‘times.n.01’), etc.
- **time duration units:**
Synset(‘hours.n.01’),
Synset(‘day.n.01’),
Synset(‘week.n.01’),
Synset(‘time_unit.n.01’),
Synset(‘time_period.n.01’), etc.
- **age units:** Though similar to time-units, these synsets correspond specifically to instances where the age of an entity is being described. These include: Synset(‘age.n.01’), Synset(‘year.n.01’), Synset(‘time_of_life.n.01’), etc.
- **small numbers:** Though some small integers and numbers possess their own individual synsets, we identify and filter all instances of small numbers (≤ 100) (in numerical and word form) before synset matching and group them as small numbers.
- **quantifiers:**
Synset(‘every.s.01’),
Synset(‘many.a.01’),
Synset(‘more.a.01’),
Synset(‘less.a.01’), etc.
- **ordinals:**
Synset(‘first.a.01’),
Synset(‘second.s.01’),
Synset(‘third.s.01’), etc.
- **fractional values:**
Synset(‘common_fraction.n.01’).

⁵<https://shap.readthedocs.io/en/latest/>

⁶<https://www.nltk.org/>

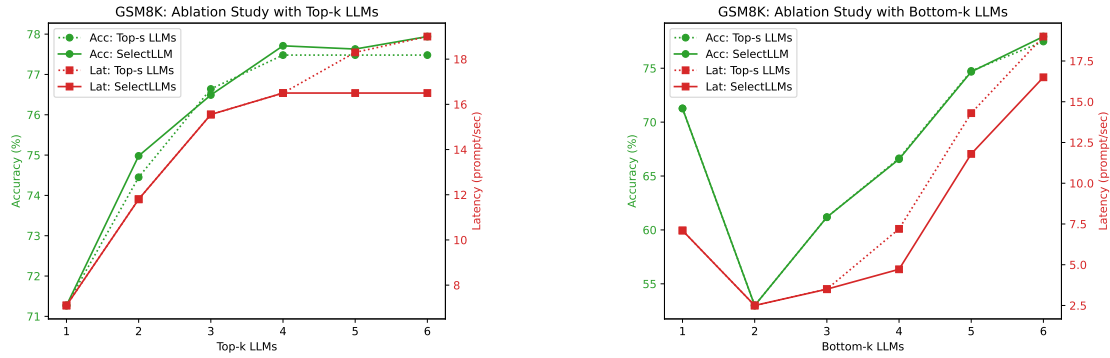


Figure 8: Ablation studies with top- k and bottom- k LLMs for GSM8K dataset (left). Similar observations are made on the MMLU dataset (right). For each LLM set, we have considered the values with an optimal s for which the top- s value is the highest.

- **rates and ratios:**

Synset('rate.n.01'),
 Synset('ratio.n.01'),
 Synset('proportion.n.01'),
 Synset('magnitude_relation.n.01'),
 etc.

- **named entities:** As WordNet does not record proper-nouns or named-entities like persons, organizations, locations, dates, and times, we identify such words before synset matching and designate them a separate group.

- **other units:**

Synset('monetary_unit.n.01'),
 Synset('linear_unit.n.01'),
 Synset('work_unit.n.01'),
 Synset('area_unit.n.01'),
 Synset('definite_quantity.n.01'),
 etc.

- **groups:**

Synset('family.n.01'),
 Synset('team.n.01'),
 Synset('unit.n.03'), etc.

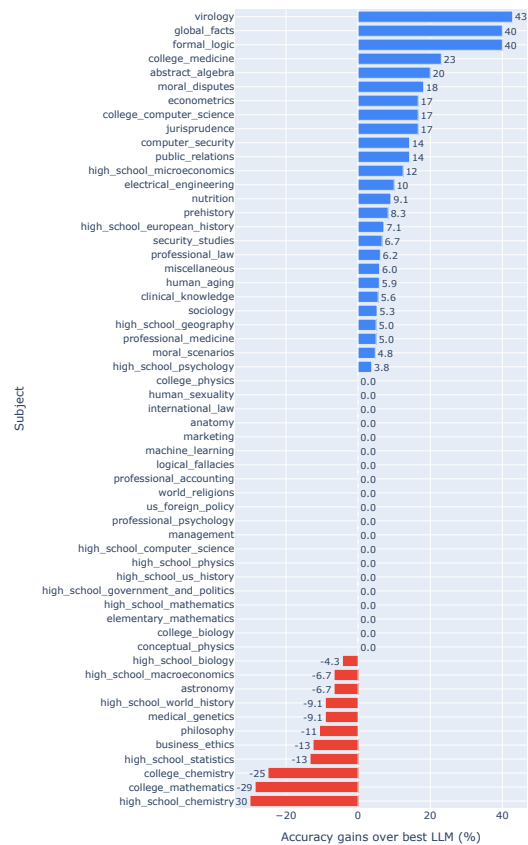


Figure 9: Subject-wise relative accuracy gain by SELECTLLM with WEIGHTEDMAXCONF policy over the performance of the best-performing individual LLM (gemma-7b-1m).