ROBUST IN-CONTEXT LEARNING VIA MULTI-ARMED BANDIT-BASED PARTITION SELECTION

Varul Srivastava
IIIT Hyderabad, India
{varul,srivastava}@gmail.com

Sankarshan Damle
EPFL Lausanne, Switzerland
{sankarshan,damle}@epfl.ch

Manisha Padala
IIT Gandhinagar, India
{manisha,padala}@iitgn.ac.in

ABSTRACT

In-context learning (ICL) enables Large Language Models (LLMs) to adapt to new tasks without parameter updates, relying solely on exemplar selection. However, in real-world scenarios, data partitions may contain corrupted labels, degrading ICL performance. We address this challenge by formulating partition selection as a multi-armed bandit (MAB) problem, where each evaluation sample serves as a pull, allowing the model to identify the most reliable partitions iteratively. Using an Upper Confidence Bound (UCB) strategy, we progressively refine exemplar selection to mitigate the impact of noisy data. Empirical results demonstrate that UCB-based partition selection recovers performance comparable to settings without label noise, highlighting its effectiveness in improving ICL robustness.

1 INTRODUCTION

In-context Learning (ICL) (Brown et al., 2020; Min et al., 2022) enables Large Language Models (LLMs) (Achiam et al., 2023) to perform well on downstream tasks (e.g., sentiment or text classification) by adapting to new inputs without parameter updates. ICL only requires black-box access, making it an increasingly favored approach for efficiently incorporating *private data* without fine-tuning (Van Veen et al., 2023). For example, a hospital chain can use patient notes as in-context exemplars to query a proprietary LLM (e.g., OpenAI (OpenAI, 2024)) for diagnoses or treatment recommendations, leveraging its capabilities without exposing sensitive data or retraining the model.

The effectiveness of ICL in such a setup depends heavily on the quality and relevance of patient notes provided by the hospitals, as better exemplars lead to more accurate and reliable model outputs (Ye et al., 2023; Wang et al., 2024a). This scenario can be viewed as a case where the data universe is partitioned across hospitals, and the performance of in-context learning may be sensitive to the quality and representativeness of exemplars drawn from each partition (Cheng et al., 2024). Abstracting out this scenario, we have a data universe $\mathcal{D}$ being partitioned into subsets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$, where each $\mathcal{D}_i$ represents the data held by hospital i. The exemplar set $\mathcal{E}$, used for in-context learning, is drawn from these partitions.

However, not all partitions may contain high-quality data; some may introduce noise, errors, or *adversarial* manipulations that degrade ICL performance. For instance, certain hospitals might have inconsistent/outdated record-keeping practices, leading to mislabeled or incomplete patient notes. In more severe cases, an adversarial partition $\mathcal{D}_j$ could intentionally inject misleading exemplars, influencing the LLM's outputs in undesirable ways. Recent ICL literature discusses ICL's performance relative to the quality of in-context examples, with noisy or mislabeled data affecting reliability and accuracy (Cheng et al., 2024; Gao et al., 2024). This paper addresses the problem of selecting data partitions for drawing the exemplar set $\mathcal{E}$ in ICL, where some partitions may contain corrupted labels, by modeling partition selection as a *multi-armed bandit* (MAB) problem (Thompson, 1933; Bubeck et al., 2012) (refer to Figure 1).

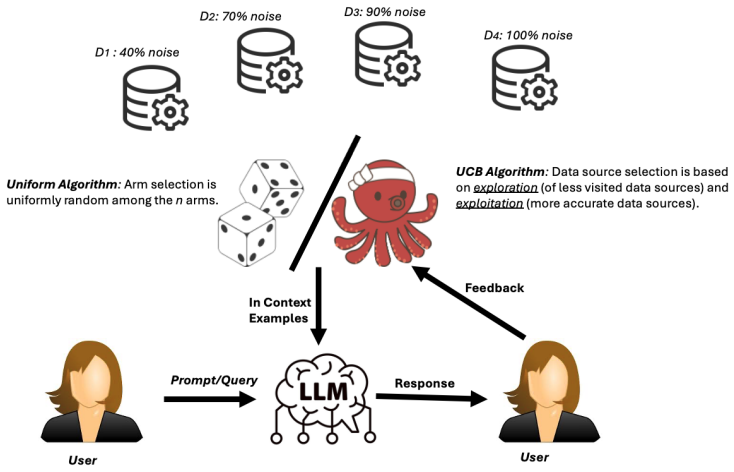

Figure 1: Uniform (random) and UCB (MAB) algorithms for the selection of data source from $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4$. User feedback (correct/wrong prediction) is used to update UCB values.

Our Approach & Contributions. We focus on open-source instruction fine-tuned autoregressive LLMs, including the `LLaMA3` family (Dubey et al., 2024): (i) `LLaMA-3.2-3B`, (ii) `LLaMA-3.1-8B`, and (iii) `LLaMA-3.1-70B` as well as (iv) `phi-3.5-mini` (Abdin et al., 2024). **First**, on ICL benchmarks like `AGNews` (Zhang et al., 2015), `MMLU` (Hendrycks et al., 2020) and `MMLU-Pro` Wang et al. (2024b) as well as synthetic dataset the introduction of data partitions with corrupted labels significantly reduces the ICL performance across LLMs. **Second**, to address this, we reformulate the problem of data partition selection in ICL as a MAB problem, where each evaluation sample serves as a pull, allowing the model to iteratively refine its choice of the most reliable partition. Specifically, we employ an Upper Confidence Bound (UCB) (Auer et al., 2002) strategy to progressively identify the partition that yields the most reliable ICL performance as evaluation progresses. **Third**, we show that UCB-based partition selection enables ICL to recover performance comparable to the baseline setting without corrupted labels.

2 Related Work

In-Context Learning (ICL) using Noisy labels. Recent research explores performance and robustness of ICL when faced with noisy examples. Gao et al. (2024) investigates performance drop due to noisy annotation on ICL in generation tasks and propose Local Perplexity Ranking (LPR) method to replace noisy labels. Cheng et al. (2024) introduces ICL robustness against noisy labels by introducing noisy data during training and Pan et al. (2024) studies robustness against noise in machine translation. Wang et al. (2023) compares performance of LLMs with supervised learning (SL) showing robustness improves with model size. We formulate partition selection as a multi-armed bandit problem, allowing dynamic adaptation to noise during evaluation.

Large Language Models (LLMs) and Multi-Arm Bandits (MABs). Most work in the intersection of LLMs and MABs has been to enhance MABs performance using LLMs (Baheri & Alm; Sun et al., 2025; de Curtò et al., 2023). Felicioni et al. (2024) studies the role of epistemic uncertainty estimation in decision-making tasks that use natural language as input. Contrary to existing work, which utilizes LLMs to improve MABs; this paper utilizes MABs to improve ICL performance.

3 ICL via Multi-Armed Bandit-based Partition Selection

We define the ICL setting and introduce our optimal partition selection policy based on MAB.

3.1 IN-CONTEXT LEARNING (ICL)

Given an LLM[1] $f_\theta(\cdot)$ with parameters θ, the few-shot prompting set can be defined as: $\mathbb{C} := \{I, (x_1, y_1), \ldots, (x_k, y_k)\}$ where I is the *system prompt* (explaining the evaluation task), and $(x_1, y_1), \ldots, (x_k, y_k)$ are the demonstration exemplars. Here, $k \in \mathbb{Z}_{\geq 0}$ is the number of exemplars provided. With $k = 0$, this is zero-shot prompting, with $k \geq 1$, it is few-shot.

Consider a task with a training and validation set, $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$. The exemplars for training are distributed across N i.i.d. partitions, i.e., $\mathcal{D}_{train} = \bigcup_{i \in [N]} \mathcal{D}_i$. We consider three ICL evaluation tasks: `AGNews`, `MMLU`, `MMLU-Pro`, and a synthetic task. For our experiments, we use $k \in \{0, 10\}$.

The test instance, x', is sampled from the validation set $\mathcal{D}_{test}$. The LLM uses the setup and exemplars to predict the test label y', i.e., the LLM outputs $y' = f_\theta(x'; (x_1, y_1), \ldots, (x_k, y_k), I)$.

3.2 DISTRIBUTION OF DATA SOURCES

We consider a noisy ICL setting, where the i.i.d data distributed across the partitions may be corrupted. We assume that the noise is with respect to the label. For each partition $i \in [N]$, $\mathcal{D}_i$, the given label is *correct* with probability q_i and *noisy* with probability $1 - q_i$. Formally, for a classification task with c classes, we define a noisy set $\mathcal{D}'_i$, $\mathcal{D}'_i = \{(x_i, y'_i)\}_{i \in |\mathcal{D}_i|}$, where for each feature-label pair $(x_i, y_i) \sim \mathcal{D}_i$,

$$y'_i = \begin{cases} y_i & \text{w.p. } q_i \\ \hat{y}_i \sim \text{Uniform}[\{1, \ldots, c\} \setminus \{y_i\}] & \text{w.p. } (1 - q_i) \end{cases}$$

That is, we consider a noisy ICL setting with $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ data partitions, where given the noisy probabilities $\{q_i\}_{i \in [N]}$, the exemplars are sampled from $\mathcal{D}_{train} = \bigcup_{i \in N} \mathcal{D}'_i$.

3.3 ICL VIA MULTI-ARMED BANDIT-BASED PARTITION SELECTION

At time t we sample $x'_t \sim \mathcal{D}_{test}$, we select a partition $i \in [N]$ and sample k exemplars from $\mathcal{D}'_i$. To improve performance we must select a partition with least expected noise. *Each of the partitions is considered to be an arm.* At time t an arm pulled using our algorithm is denoted by $\mathbb{I}_t$ i.e., $\mathbb{I}_t \in [N]$. Pulling an arm, $\mathbb{I}_t = i$ corresponds to a reward of 1 if LLM outputs the correct label with the exemplar sampled from i otherwise 0,

$$r_{\mathbb{I}_t} = \begin{cases} 1 & f_\theta(x'_t; (x_1, y_1), \ldots, (x_k, y_k), I) = y' \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Each arm is associated with a expected reward taken over the k exemplars selected

$$\mu_i = \mathbb{E}_{\{(x_1, y_1), \ldots, (x_k, y_k)\} \sim \mathcal{D}'_i}[r_i]$$

The best arm $i^\star$ corresponds to the partition with highest expected reward denoted by $\mu^\star$. The expected cumulative regret is given by,

$$R_T = \sum_{t=1}^{T} [\mu^\star - \mathbb{E}[r_{\mathbb{I}_t}]]$$

The goal is to minimize the above regret. We use the Upper Confidence Bound (UCB) (Auer et al., 2002) algorithm with sub-linear regret guarantees (Algorithm 1). Figure 1 depicts our setting.

In the UCB, the score at every round $t \in [T]$ for each arm or partition $i \in [N]$ is calculated as:

$$UCB_{i,t} = \hat{r}_{i,t} + c \cdot \sqrt{\frac{\ln(t)}{n_{i,t}}} \tag{2}$$

where $\hat{r}_{i,t}$ is the expected reward from arm i till round t and $n_{i,t}$ is the number of times arm i is pulled till round t. Formally, $\hat{r}_{i,t} = \frac{1}{n_{i,t}} \sum_{t'=1}^{t} r_{\mathbb{I}_t} \cdot \mathbb{I}_{i,t}$, where $\mathbb{I}_{i,t} = 1$ when arm i is pulled at round t. Here c is the exploration constant set to 0.5 in our experiments.

[1]We refer to LLMs generically without discussing their size, as size is not central to our focus.

Algorithm 1 Upper Confidence Bound (UCB) Algorithm for ICL

Input: Data paritions $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$, validation set $\mathcal{D}_{test}$, number of few-shots k, model f_θ
Initialize:
$t \leftarrow 0$
for each arm $i \in 1 \ldots n$ **do**
    $n_{i,t} \leftarrow 1$ ▷ Number of times arm i was played
    $\hat{r}_{i,t} \leftarrow 0$ ▷ Initial reward sample
end for
for Query $x'_t \in \mathcal{D}_{test}$ **do**
    $t \leftarrow t + 1$
    Compute UCB for each arm:
    for each arm $i = 1 \ldots n$ **do**
        $\text{UCB}_{i,t} \leftarrow \hat{r}_{i,t} + c \cdot \sqrt{\frac{\ln t}{n_{i,t}}}$
    end for
    Select arm $a = \arg\max_{i \in [n]} \text{UCB}_{i,t}$
    Sample k examples $\{(x_r, y_r)\}_{r=1}^{k}$ from D_a
    Observe reward r_t according to Equation 1 using f_θ and x'_t
    $n_{a,t} \leftarrow n_{a,t} + 1$
    $\hat{r}_{a,t} \leftarrow \frac{1}{n_{a,t}} \left((n_{a,t} - 1) \cdot \hat{r}_{a,t-1} + r_t \right)$
end for

4 Experimental Setup

4.1 Noisy In-context Learning (ICL)

We consider $N = 4$ data partitions and evaluate performance across multiple noisy strategies. Except for one, all strategies use $k = 10$-shot prompting. Appendix A presents other details.

1. **Zero**: Zero-shot prompting. A baseline.
2. **Worst-Case**: 10-shot prompting with $q_i = 0$ for each partition $\mathcal{D}_i$. That is, all exemplars are noisy. We uniformly select a partition from N partitions.
3. **Uniform**: 10-shot prompting with $q_i = 0.25$ for each partition $\mathcal{D}_i$. We uniformly select a partition from N partitions.
4. **No-Noise**: 10-shot prompting with $q_i = 1$ for each partition $\mathcal{D}_i$. That is the non-noisy (baseline) ICL setting. (no noise). We uniformly select a partition from N partitions.
5. **UCB-d1**: 10-shot prompting with $\{q_1, q_2, q_3, q_4\} = \{1, 0, 0, 0\}$ (i.e., on expectation 75% of exemplars are noisy). Partition sampling is performed using UCB (Algorithm 1).
6. **UCB-d2**: 10-shot prompting with $\{q_1, q_2, q_3, q_4\} = \{0.6, 0.3, 0.1, 0\}$ (i.e., on expectation 75% of exemplars are noisy). Partition sampling is performed using UCB (Algorithm 1).

With **UCB-d1** and **UCB-d2**, we consider differences between two types of distributions. While **UCB-d1** and **UCB-d2** have 75% of noisy data. However, the best data partition is of higher quality in **UCB-d1**. When the UCB algorithm converges to the optimal partition, it exploits the better quality partition in the case of **UCB-d1** (as we see later in Table 1).

4.2 Evaluation Tasks and Models

We run experiments using `LLaMA-3.2-3B` on 4 datasets: AG-News Zhang et al. (2015), MMLU Hendrycks et al. (2020), MMLU-Pro Wang et al. (2024b) and synthetic data.

Evaluation Tasks.

- `AGNews` (Zhang et al., 2015). A text classification benchmark consisting of news articles categorized into four topics: World, Sports, Business, and Science/Technology. `AGNews` evaluates language models on topic classification.

- `MMLU` (Hendrycks et al., 2020) & `MMLU-Pro` (Wang et al., 2024b). The Massive Multitask Language Understanding (`MMLU`) benchmark assesses models on diverse subjects spanning humanities, sciences, and mathematics. `MMLU-Pro` (Wang et al., 2024b) is an extended version with additional challenges and improved question quality. `MMLU` is an instruction-following task.
- `Synthetic`. We generate d-dimensional synthetic classification data with c classes. This is done by randomly generating several points and c random targets. Each point is mapped with the label of the closest target. For experiments, we use $(d, c) = (2, 5)$ to generate data.

Models. We run experiments on different models `LLaMA-3.2-3B`, `LLaMA-3.1-8B`, `LLaMA-3.1-70B` (Dubey et al., 2024) and `phi-3.5-mini` (Abdin et al., 2024). We choose open-source LLMs, focusing on model size and architecture.

5 Results and Discussion

Table 1 presents the mean and standard deviation for accuracy of `LLaMA-3.2-3B` over different evaluation tasks, across three independent runs. Table 2 compares performance across model sizes on `AGNews`.

Dataset	Zero-Shot	Worst-Case	No-Noise	Uniform	UCB-d1	UCB-d2
AG-News	$58.38_{\pm 1.9}$	$61.68_{\pm 3.8}$	$74.25_{\pm 1.39}$	$64.67_{\pm 3.86}$	$73.45_{\pm 1.64}$	$69.93_{\pm 1.82}$
MMLU	$34.13_{\pm 2.22}$	$45.84_{\pm 0.11}$	$49.23_{\pm 1.26}$	$46.31_{\pm 0.53}$	$48.77_{\pm 1.17}$	$48.70_{\pm 0.34}$
MMLU-Pro	$19.36_{\pm 0.72}$	$27.41_{\pm 1.02}$	$29.21_{\pm 1.02}$	$29.67_{\pm 0.5}$	$30.27_{\pm 0.41}$	$26.95_{\pm 0.91}$
Synthetic	$20.95_{\pm 1.69}$	$25.75_{\pm 0.28}$	$39.02_{\pm 2.11}$	$29.84_{\pm 2.40}$	$39.22_{\pm 1.83}$	$31.74_{2.26}$

Table 1: Results for `LLaMA-3.2-3B`

Accuracy Decreases with Noisy In-Context Exemplars. From Table 1, we first establish that the introduction of noise in ICL exemplars hurts accuracy. While the baseline "No-Noise" achieves the highest accuracy, the worst-case with 100% noisy samples consistently performs poorly.

Models	Zero-Shot	Worst-Case	No-Noise	Uniform	UCB-d1	UCB-d2
`phi-3.5-mini`	47.70	48.70	69.86	53.69	69.36	61.88
`LLaMA-3.2-3B`	58.38	61.68	74.25	64.67	73.45	69.93
`LLaMA-3.1-8B`	80.08	63.38	84.11	75.78	83.92	81.31
`LLaMA-3.1-70B`	86.69	86.65	87.76	87.76	87.55	87.37

Table 2: Results for different model sizes and architecture on `AG-News`

Effectiveness of MAB Approach. From Table 1, for the same number of noisy samples (75%) overall, performance is better when data partition selection uses UCB (Algorithm 1). Further, UCB-d1 performs better than UCB-d2. This is because of the convergence of UCB to the optimal data partition. UCB-d1 has distribution $\{q_1, q_2, q_3, q_4\} = \{1, 0, 0, 0\}$ and UCB-d2 has $\{0.6, 0.3, 0.1, 0\}$. When UCB converges to the optimal partition, it samples from a partition with no noise while in UCB-d2 it samples from a partition that adds noisy labels with probability $1 - q_1 = 0.4$.

Increased Robustness with LLM size. We observe through Table 2 that LLMs become increasingly robust against noisy perturbations to example labels for larger model sizes. This trend is attributed to improved generalization and representation capabilities of LMs with higher number of parameters; also aligning with the findings of Wang et al. (2023).

6 Conclusion & Future Work

In this paper, we explored the setting where ICL examples are drawn from multiple data sources, some of which may be noisy. We demonstrated that in such scenarios, a bandit-based approach generally outperforms uniform random selection. Additionally, we observed that as the parameter

size of LLMs increases, they become more generalized and, consequently, more robust to noisy in-context examples. Therefore, UCB-based selection (Algorithm 1) for sourcing in-context examples proves to be an effective strategy for ICL in LLMs.

Future Work. While this study focuses on scenarios with noisy data sources, we conjecture that under non-IID data distributions across sources, employing combinatorial bandits (Chen et al., 2013; Bubeck et al., 2013) would be a more effective choice. A natural extension of this work is to generalize the threat model from noisy label corruption to adversarial corruption.

REFERENCES

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Ali Baheri and Cecilia Alm. Llms-augmented contextual bandit. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 258–265, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/bubeck13.html.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/chen13a.html.

Chen Cheng, Xinzhi Yu, Haodong Wen, Jingsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. Exploring the robustness of in-context learning with noisy labels. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. URL https://openreview.net/forum?id=ib4cAWZKXa.

J. de Curtò, I. de Zarzà, G. Roig, J. C. Cano, P. Manzoni, and C. T. Calafate. Llm-informed multi-armed bandit strategies for non-stationary environments. *Electronics*, 12(13):2814, 2023. doi: 10.3390/electronics12132814.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nicolò Felicioni, Lucas Maystre, Sina Ghiassian, and Kamil Ciosek. On the importance of uncertainty in decision-making with large language models. *arXiv preprint arXiv:2404.02649*, 2024.

Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. On the noise robustness of in-context learning for text generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=00uVk06eVK.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

OpenAI. Openai models documentation, 2024. URL https://platform.openai.com/docs/models.

Leiyu Pan, Yongqi Leng, and Deyi Xiong. Can large language models learn translation robustness from noisy-source in-context demonstrations? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2798–2808, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.249/.

Jiahang Sun, Zhiyong Wang, Runhan Yang, Chenjun Xiao, John Lui, and Zhongxiang Dai. Large language model-enhanced multi-armed bandits. *arXiv preprint arXiv:2502.01118*, 2025.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts. *Research Square*, 2023.

Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1752–1767, 2024a.

Xindi Wang, Yufei Wang, Can Xu, Xiubo Geng, Bowen Zhang, Chongyang Tao, Frank Rudzicz, Robert E Mercer, and Daxin Jiang. Investigating the learning behaviour of in-context learning: a comparison with supervised learning. In *ECAI 2023*, pp. 2543–2551. IOS Press, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

A TASK AND PROMPT DESCRIPTION

In this section, we describe tasks, performance evaluation metric and prompt used in each dataset.

A.1 AG-NEWS

The AG News dataset Zhang et al. (2015) is a widely used benchmark for text classification tasks in machine learning and natural language processing. It consists of 120,000 training samples and 7,600 test samples, each categorized into one of four news topics: World, Sports, Business, and Science/Technology. Prompt includes the news article, and the goal of LLM is to classify the category into one of these 4 categories labeled 0,1,2 and 3. We evaluate the accuracy of the LLM (correct if predicted category matches the actual category).

> Classify the news based on whether their type is a World, Sports, Business, or Technology. The type and label are as follows:
>
> World: 1
> Sports: 2
> Business: 3
> Technology: 4
>
> Your goal is to learn the correlation between prompt and label.
> ###
> Format: Article: ⟨news⟩
> Response is a single number between 1 to 4. Output only the number and nothing else. Just say the type label (1,2,3,4). Do not say anything more. No explanations needed. Don't respond anything except answer type.

Figure 2: System Prompt for AG-News

A.2 MMLU

The Massive Multitask Language Understanding (MMLU) Hendrycks et al. (2020) dataset is a comprehensive benchmark designed to evaluate the knowledge and reasoning abilities of language models across 57 diverse subjects, including STEM, humanities, social sciences, and more. It consists of multiple-choice questions, with four answer options per question, covering both high-school and professional-level topics. MMLU is particularly useful for assessing a model's generalization and factual knowledge beyond simple language understanding. We evaluate the accuracy of LLM based on it's ability to correctly answer the question (correct if predicted option is correct option).

MMLU-Pro Wang et al. (2024b) is an extension of MMLU that introduces more challenging and professionally relevant questions, often requiring deeper reasoning, domain-specific expertise, and multi-step problem-solving. This enhanced version serves as a more rigorous benchmark for evaluating advanced language models, distinguishing between surface-level memorization and true comprehension.

A.3 SYNTHETIC DATASET

We describe the algorithm for generating synthetic data in Algorithm 2. Further, the system prompt used for synthetic data is given in Figure 4. We use parameters $n = 2500, d = 2, c = 5$ for our experiments. We evaluate accuracy of the LLM as it's ability to correctly predict the class.

Find the correct option for the question, belonging to various domains.You are given the question followed by options 0, 1, 2 and 3....Your goal is to learn correlation between prompt and label.Format: Question: ⟨question⟩
Options:
0:⟨option1⟩
1:⟨option2⟩
2:⟨option3⟩
3:⟨option4⟩
. . .
Answer: ⟨type⟩
Just say the type label (0,1,2,3. . .). Do not say anything more. No explanations needed.

Don't respond by mentioning 'Answer'.

Figure 3: System Prompt for MMLU and MMLU-Pro

Algorithm 2 Algorithm to generate synthetic data points

Input: n (number of points), d (dimensionality), c (number of labeled prototypes)
Sample n points randomly in $[low, high]^d$ space and store in set S
Sample c points randomly in $[low, high]^d$ space
Assign each point in C a unique label from $\{0, 1, \ldots, c-1\}$
Let C be the set of labeled prototype points
for each point $x \in S$ **do**
    Find the closest point $y \in C$ based on a eucledian-distance metric
    Let y have label i
    Assign label i to x and store (x, i) in set D
end for
Return D (set of n labeled points)

You are given data described by 2 features X0, X1. Based on these features, the data belongs to one of five classes, CLASS = 0, 1, 2, 3 or 4. Given the data your goal is to output the class this point belongs to.
Format:
Question: X[0] = ⟨value0⟩; X[1] = ⟨value1⟩;
Answer: ⟨class⟩
Just say the class label (0,1,2,3,4). Do not say anything more. No explanations needed.

Don't respond by mentioning 'Answer'.

Figure 4: System Prompt for MMLU and MMLU-Pro