Cost-aware LLM-based Online Dataset Annotation

Eray Can Elumar[1] Cem Tekin[2] Osman Yağan[1]

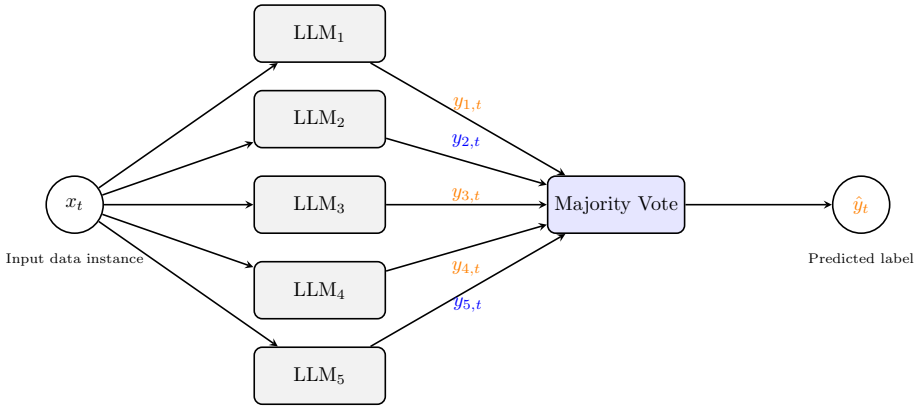
[1]Carnegie Mellon University [2]Bilkent University
{eelumar, oyagan}@andrew.cmu.edu, cemtekin@ee.bilkent.edu.tr

NeurIPS 2025

Motivation

- The rapid proliferation of data across different domains has created the need for accurate, large-scale annotation pipelines
- LLMs offer a promising remedy: automated dataset annotation with minimal human effort
- However, due to hallucinations and model bias, relying on a single LLM can have issues
- A common strategy to bolster label quality is ensembling: querying multiple LLMs, or multiple samples from the same model, and aggregating their outputs

Weighted Majority Voting

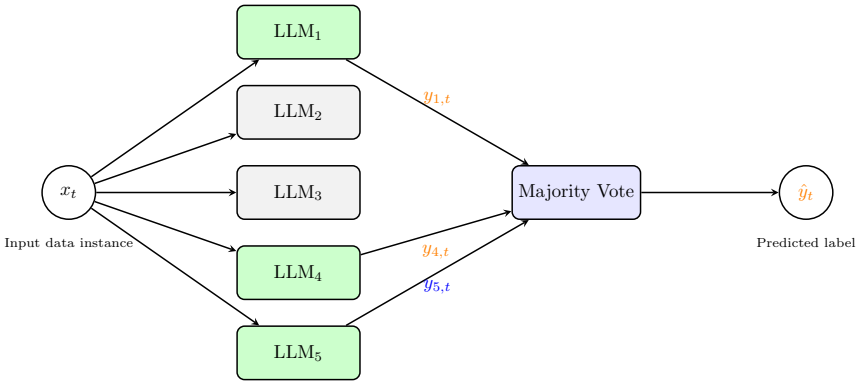


- Majority vote assigns the label that receives the highest weight from voters (LLMs)
- Weights: Accuracy, empirical accuracy, self-reported confidence, etc.

Motivation

- Majority voting is costly, queries all given LLMs
- Can reduce costs by performing majority voting on a smaller subset of LLMs
- Need to optimize when to select a smaller subset, and which LLMs to choose for the subset
- Some LLMs may perform better on specific types of data instances. For example, a model fine-tuned on geographic knowledge is likely to perform better on geography-related questions in a high school knowledge dataset.
- Some data instances can be easier to label than others

Motivation



- Based on the input data instance, we aim to **adaptively select and aggregate** LLMs to optimize labeling cost while adhering to a user-defined accuracy parameter.
- We assume there are no ground-truth labels and aim to learn the adaptive mechanism in an online manner

Cost-aware Majority Voting (CaMVo) Algorithm

NEURAL INFORMATION PROCESSING SYSTEMS

Input Parameters

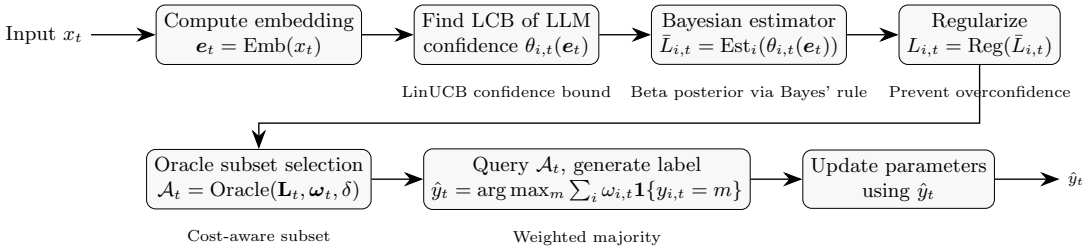
δ: the desired minimum relative confidence of the selected subset compared to the full majority vote using all LLMs

$k_{\min}$: minimum number of LLMs that needs to be queried for each data instance

Key Idea

CaMVo leverages the context of the data instance with a LinUCB-based confidence model to find the LCB of the probability that the LLM will produce a correct label. Based on this, CaMVo adaptively selects the minimum cost subset of LLMs that achieve the targeted relative confidence δ.

Cost-aware Majority Voting (CaMVo) Algorithm



Experiments: MMLU Dataset

- MMLU Dataset: It is a challenging multiple-choice benchmark spanning 57 diverse subjects including mathematics, U.S. history, law, and computer science. It demands broad world knowledge and strong reasoning capabilities—conditions under which majority voting is particularly effective.

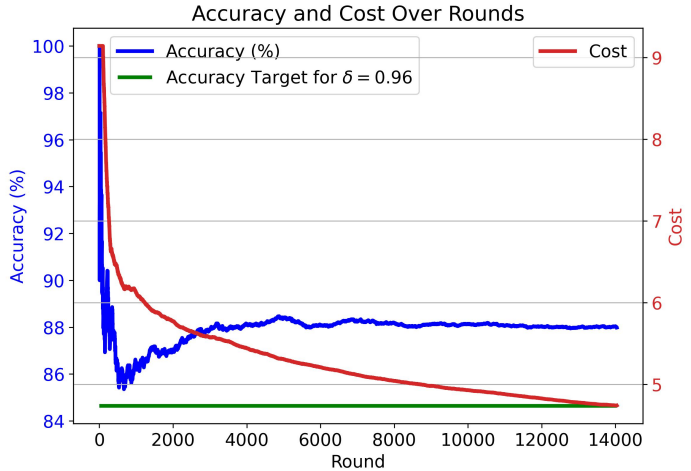
LLM / Method	Accuracy (%)	Cost
o3-mini	85.92	1.10
claude-3-7-sonnet	85.65	3.00
o1-mini	84.82	1.10
gpt-4o	83.58	2.50
llama-3.3-70b	81.70	0.59
llama-3.1-8b	68.01	0.05
claude-3-5-haiku	64.09	0.80
Majority Vote	88.18	9.14
Baseline Method	88.18	9.14

Experiments: MMLU Dataset

CaMVo δ	Target Acc. (%)	Acc. (%) $k_{\min} = 1$	Cost $k_{\min} = 1$	Acc. (%) $k_{\min} = 3$	Cost $k_{\min} = 3$
0.999	95.52	95.59	6.15	95.59	6.15
0.998	95.43	95.43	4.03	95.43	4.03
0.997	95.33	95.45	2.83	95.45	2.83
0.995	95.14	95.25	2.06	95.25	2.06
0.99	94.66	95.10	1.09	95.12	0.99
0.985	94.20	94.69	0.34	95.06	0.84
0.98	93.71	94.69	0.31	95.07	0.83
0.97	92.75	94.56	0.22	95.07	0.82
0.96	91.80	94.21	0.13	95.06	0.81
0.95	90.84	94.28	0.14	95.07	0.81
0.9	86.06	94.24	0.10	95.06	0.81

- CaMVo satisfies all target accuracy levels defined by the confidence parameter δ.
- As δ decreases, accuracy decreases, and the cost of labeling also declines in a controlled manner. This behavior highlights the flexibility of CaMVo in adapting to a wide range of practical scenarios with varying accuracy and budget constraints.

Experiments: MMLU Dataset



Conclusion

- Results show CaMVo matches or exceeds full-ensemble majority-vote accuracy while reducing labeling cost
- Hence, CaMVo is a practical solution for cost-efficient, automated annotation in dynamic labeling environments without any ground-truth labels or offline training.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!