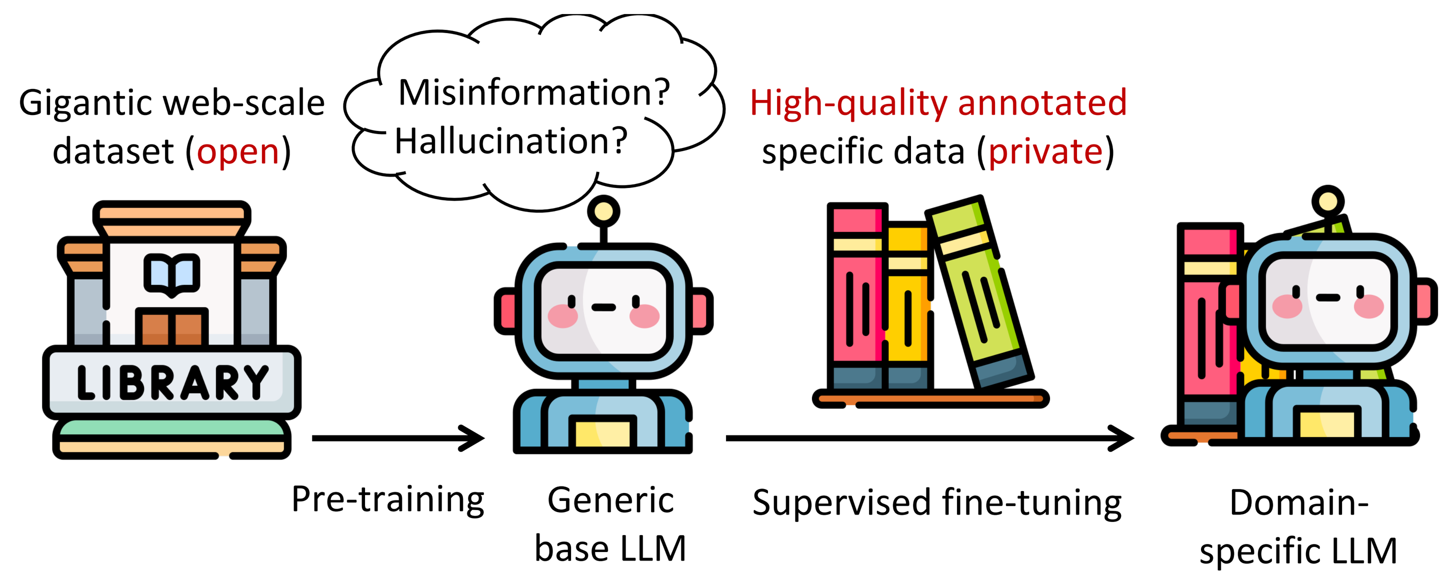
Interactive Multi-fidelity Learning for Cost-effective Adaptation of Language Model with Sparse Human Supervision

Jiaxin Zhang[1], Zhuohang Li[2], Kamalika Das[1], Sricharan Kumar[1]   [1]Intuit AI Research   [2]Vanderbilt University

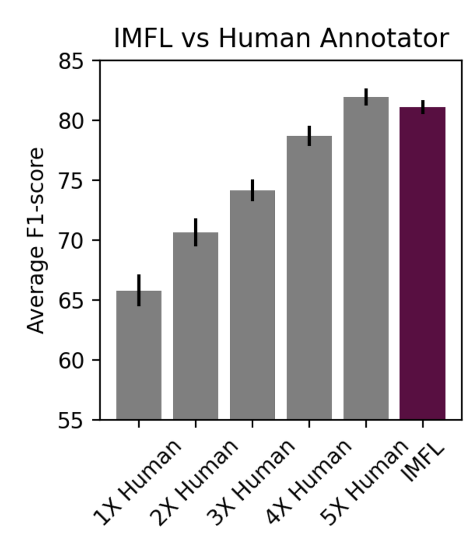
Introduction



- Generic LLMs for domain-specific tasks - immense scale at deployment, susceptibility to misinformation, e.g., healthcare and finance
- Fine-tuned small LMs for domain-specific tasks – faster development cycles, lower operating costs but need high data annotation costs

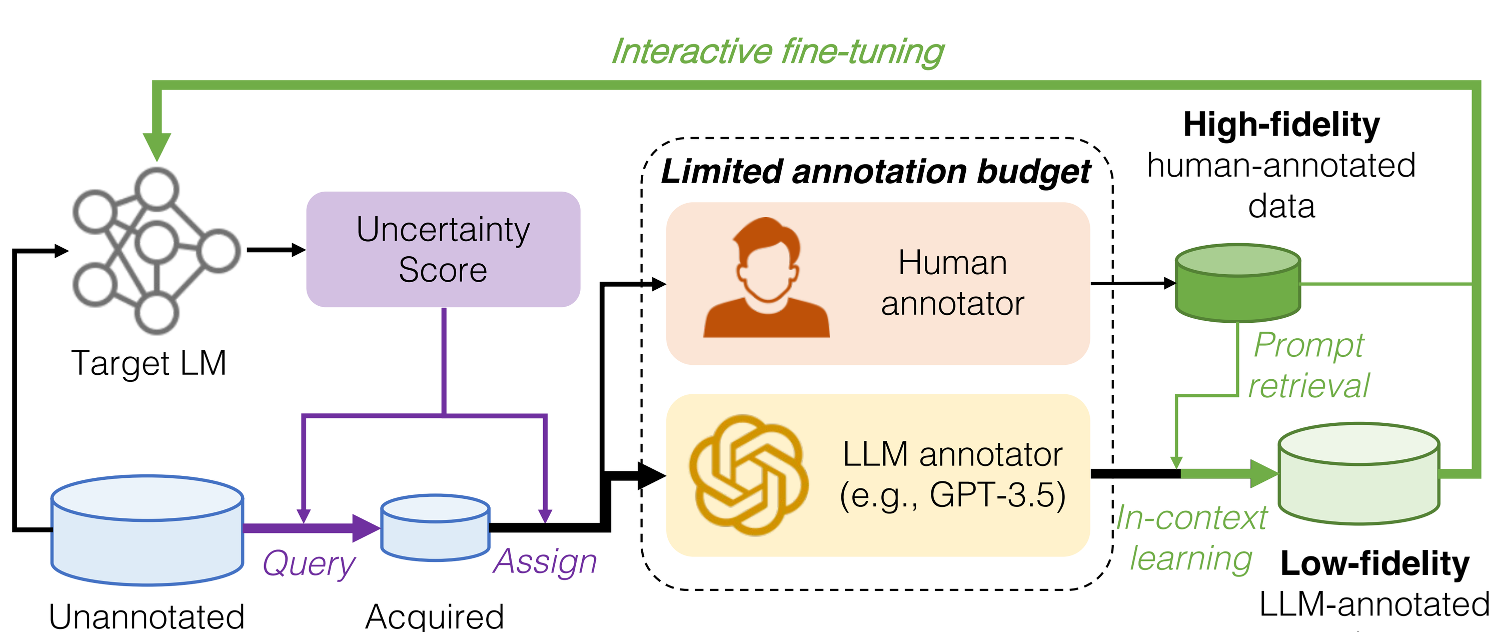
Table 1: A qualitative comparison of human annotation, LLM annotation, and IMFL .

	Human	LLM	IMFL
Cost Saving	Low	**Very High**	**High**
Quality	**Very High**	Low	**High**
Efficiency	Low	**Very High**	**High**
Performance	**Very High**	Low	**High/Very High**



Overview

IMFL proposes the best acquisition strategy that balances between low-fidelity automatic LLM annotations and high-fidelity human annotations to maximize model performance given limited annotation budgets.



(a) Overview of IMFL framework

The high human annotation cost in domain-specific tasks can be greatly reduced by employing IMFL, which utilizes fewer human annotations combined with cheaper LLM (e.g., GPT-3.5-turbo) annotations to achieve competitive performance.

Interactive Multi-fidelity Learning

Problem Formulation

Given a total annotation budget $\mathcal{B}$ and a computational cost $\mathcal{C}$, we aim to fine-tune a small LM $f(\boldsymbol{x};\theta^*): \mathcal{X} \rightarrow \mathcal{Y}$ on a downstream task by annotating samples from an unannotated data pool $\mathcal{U} = \{x_i\}_{i=1}^U$ to constitute the annotated sample set $\mathcal{A}$ ($|\mathcal{A}| \leq \mathcal{B}$ and initially $\mathcal{A} = \varnothing$) such that its performance is maximized.

Annotation set – a human-annotated subset $\mathcal{A}_H$ and an LLM-annotated subset $\mathcal{A}_G$
Total annotation budget – human annotation budget $\mathcal{B}_H$ and LLM annotation budget $\mathcal{B}_G$

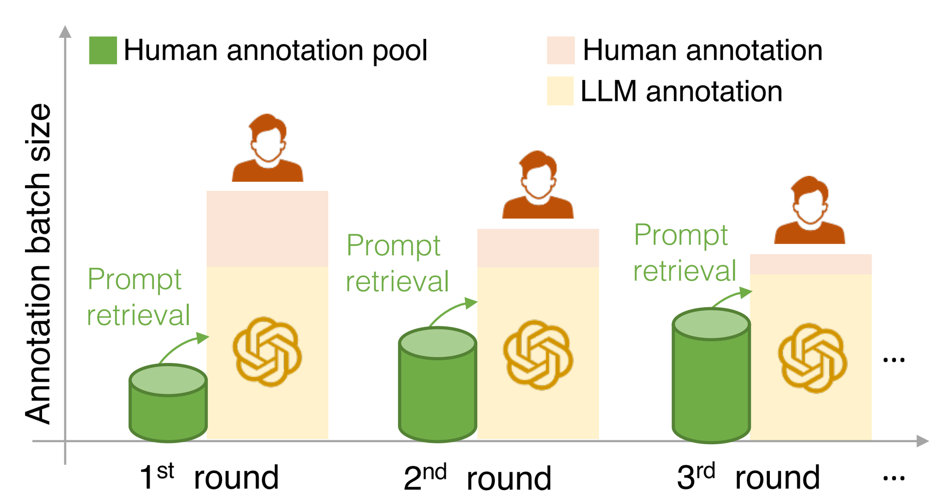
Multi-fidelity Learning Framework

➢ **Initialization** $\theta^{(0)} = \arg\min_{\theta^*} \frac{1}{|\mathcal{A}_H^0|} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{A}_H^0} \mathcal{L}\left(f(\boldsymbol{x}_i;\theta^*), y_i\right), \quad i = 1, \ldots, n_s$

➢ **Fine-tuning** $\mathcal{L}_{total} = \frac{1}{|\mathcal{A}_H^r|} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{A}_H^r} \mathcal{L}\left(f(\boldsymbol{x}_i;\theta^{(r)}), y_i\right) + \frac{1}{|\mathcal{A}_G^r|} \sum_{(\boldsymbol{x}_j, y_j) \in \mathcal{A}_G^r} \mathcal{L}\left(f(\boldsymbol{x}_j;\theta^{(r)}), y_j\right)$

❖ *Design 1: In-context learning with similarity-based prompt retrieval*

❖ *Design 2: Variable batch-size query*



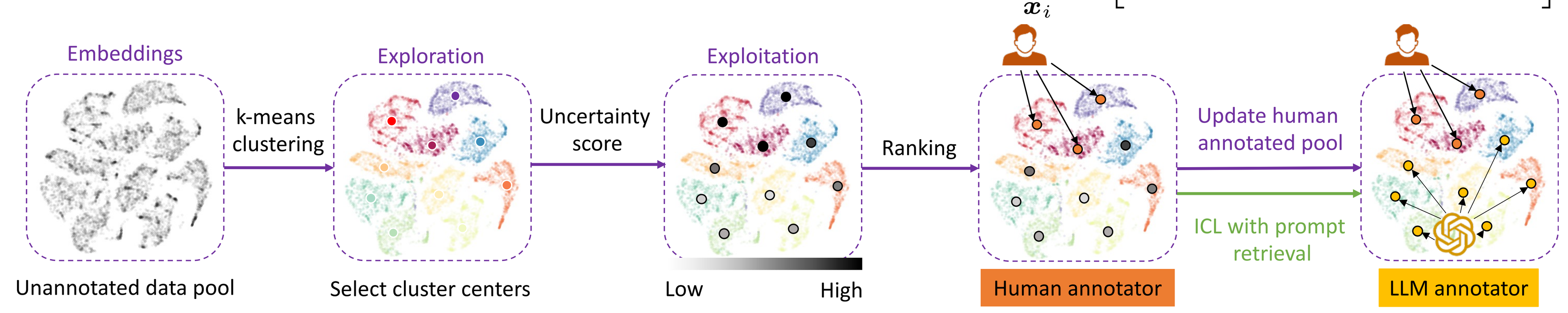
Algorithm 1 `IMFL` framework

Require: unannotated data pool $\mathcal{U}$, target LM model f, query strategy $\mathcal{S}$, annotation budget $\mathcal{B}$
Initialization: $\mathcal{A} = \varnothing$, $\theta = \theta^{(0)}$ on $\mathcal{A}_H^0$
for rounds $r = 1, \ldots, R$ **do**
  $\mathcal{U}_s^r \leftarrow$ Extract from $\mathcal{U}$ by random sub-sampling
  $[\mathcal{Q}_H^r, \mathcal{Q}_G^r] \leftarrow$ Acquire $[\mathcal{B}_H^r, \mathcal{B}_G^r]$ samples by query function $\mathcal{S}$ on model f, data $\mathcal{U}_s^r$
  $\mathcal{A}_H^r \leftarrow$ Annotate acquired samples $\mathcal{Q}_H^r$ by human
  $\mathcal{A}_H = \mathcal{A}_H \cup \mathcal{A}_H^r$
  Execute prompt retrieval from $\mathcal{A}_H$
  $\mathcal{A}_G^r \leftarrow$ Annotate acquired samples $\mathcal{Q}_G^r$ by LLMs
  $\mathcal{A}^r = \mathcal{A}_H^r \cup \mathcal{A}_G^r$
  $\mathcal{U} = \mathcal{U} \setminus \mathcal{A}^r$
  $f(\boldsymbol{x}_i;\theta^{(r)}) \leftarrow$ Fine-tune $f(\boldsymbol{x}_i;\theta^{(r)})$ on $\mathcal{A}^r$
return $f(\boldsymbol{x};\theta^{(r)}), \mathcal{A}$

➢ **Termination** two stopping criteria: (1) annotation budget and (2) computational cost

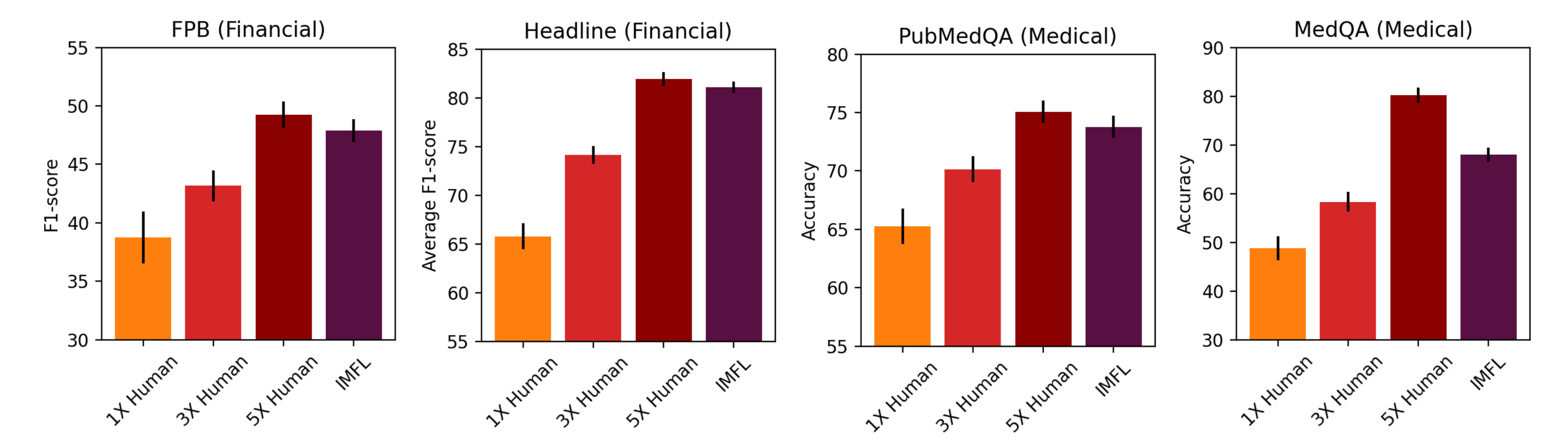
Exploration-Exploitation Query Strategy

EEQ harnesses human annotation for ***exploitation*** by maximizing informativeness through uncertainty sampling, and LLM annotation for ***exploration*** by enhancing representativeness through diversity sampling --- *two-stage selection* $\boldsymbol{x}_i^* = \arg\max_{\boldsymbol{x}_i} \left[1 - p(f(\boldsymbol{x}_i;\theta^{(r)}) \mid \boldsymbol{x}_i;\theta^{(r)})\right]$

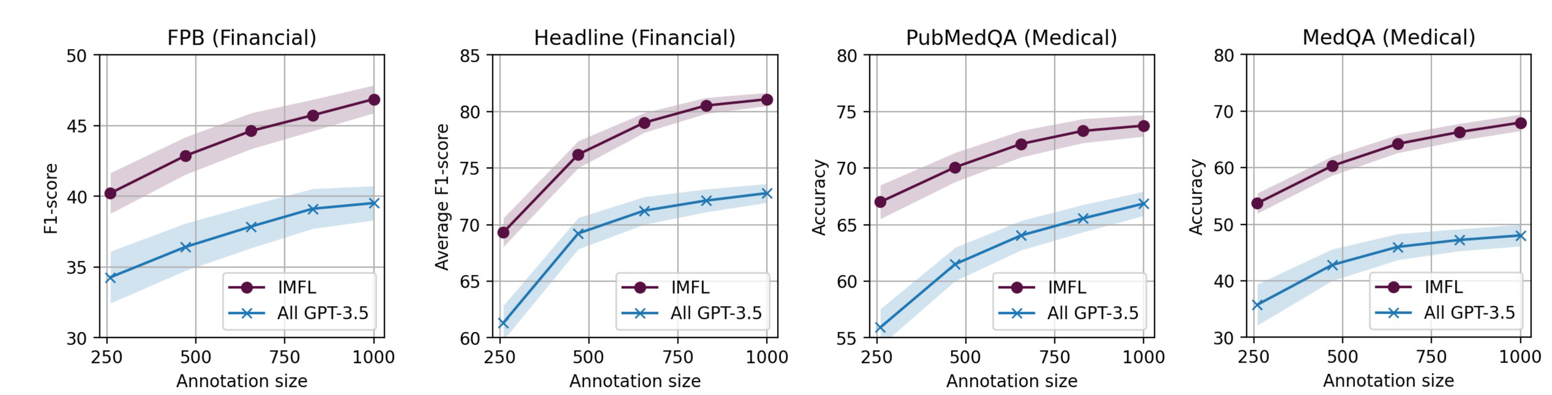


Main Results

Comparisons between our multi-fidelity learning (200 human + 800 GPT-3.5 annotations) and various sizes of human annotations.



Comparisons between our IMFL and single low-fidelity (all GPT-3.5) annotation on four domain-specific tasks given 1000 annotation budget.



Analysis

Exploitation-Exploration Query vs Random Query Strategy

Method	Budget		Query Strategy	Dataset			
Multi/Single	Human	GPT-3.5	EEQ/Random	FPB	Headline	PubMedQA	MedQA
Human + GPT-3.5	200	800	EEQ	**47.88**	**81.09**	**73.76**	67.98
Human + GPT-3.5	200	800	Random	41.94	74.32	66.03	63.77
Only Human	1000	0	Random	43.81	75.46	68.87	**70.17**
Only GPT-3.5	0	1000	Random	38.56	71.04	65.89	47.13

Effects of prompt retrieval, variable batch size, and batch orders

Method				Dataset			
Budget	Batch	Batch size	Retrieval	FPB	Headline	PubMedQA	MedQA
1000	5 Mini-Batch	Variable	Similar	**47.88**	**81.09**	**73.76**	**67.98**
1000	5 Mini-Batch	Equal	Similar	46.34	80.28	72.05	66.11
1000	5 Mini-Batch	Variable	Random	42.09	73.98	67.44	63.56
1000	5 Mini-Batch	Equal	Random	42.34	73.77	68.10	63.42
1000	1 Full-Batch	NA	Similar	43.72	75.48	68.90	63.79
1000	1 Full-Batch	NA	Random	39.80	72.11	65.94	57.23

Effects of prompt retrieval, variable batch size, and batch orders

	GPT-3 Annotation			GPT-3.5 Annotation			GPT-4 Annotation		
	retrieval	5-shot	0-shot	retrieval	5-shot	0-shot	retrieval	5-shot	0-shot
Headline	75.59	72.51	70.25	79.40	76.15	73.31	80.13	78.34	77.20
MedQA	51.42	44.89	42.03	59.45	53.57	50.82	82.67	81.38	78.87