INTUIT turbotax credit karma quickbooks mailchimp

NEURAL INFORMATION PROCESSING SYSTEMS

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

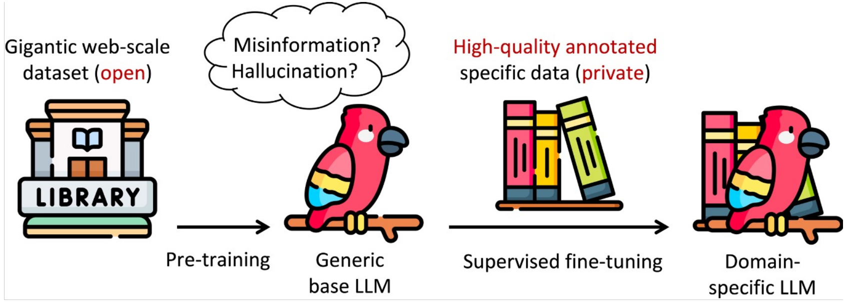
Jiaxin Zhang[1], Zhuohang Li[2], Kamalika Das[1], Sricharan Kumar[1]

[1]Intuit AI Research, [2]Vanderbilt University

12-13-2023

NeurIPS 2023, New Orleans, USA

Introduction



Challenges

- LLM suitability for domain-specific tasks, e.g., finance and healthcare, is limited due to their immense scale at deployment, susceptibility to misinformation.
- Tuning small LMs on target domain data requires extensive human effort and expert knowledge, making supervised fine-tuning very expensive

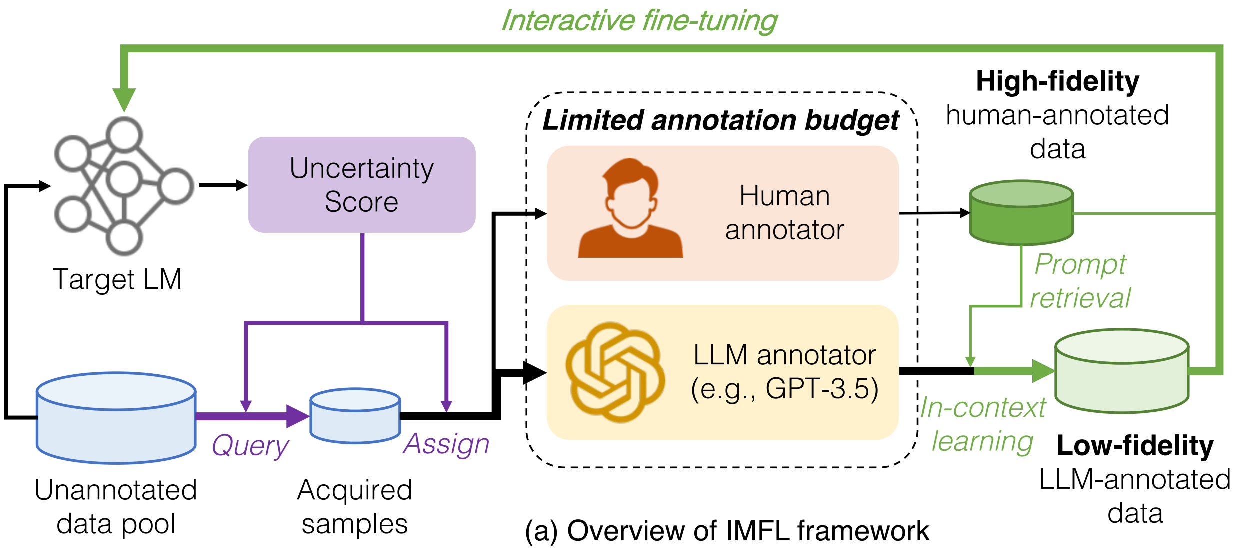
Intuitions

- High-fidelity human annotation + low-fidelity LLM annotation
- Interactive fine-tuning + knowledge distillation (prompt retrieval)
- Limited budget: less human effort with large LLM annotations

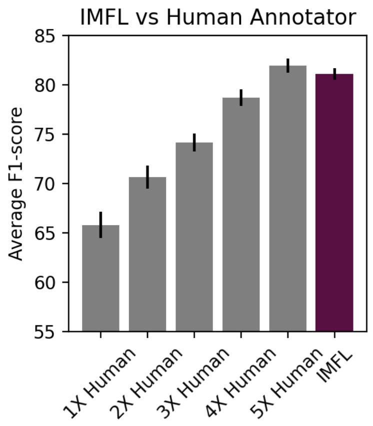
Table 1: A qualitative comparison of human annotation, LLM annotation, and `IMFL` .

	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



(a) Overview of IMFL framework



(b) Compare IMFL (1X Human + 4X GPT3.5) with human annotators

IMFL aims at solving 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. (b) IMFL significantly outperforms the 3X human annotation baselines in all four tasks and is very close to 5X upper bound in the Headline dataset (showed). This result indicates that the high human annotation cost in domain-specific tasks can be greatly reduced by employing IFML, which utilizes fewer human annotations combined with cheaper GPT-3.5 annotations to achieve competitive performance.

Interactive Multi-fidelity Learning (IMFL)

- 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

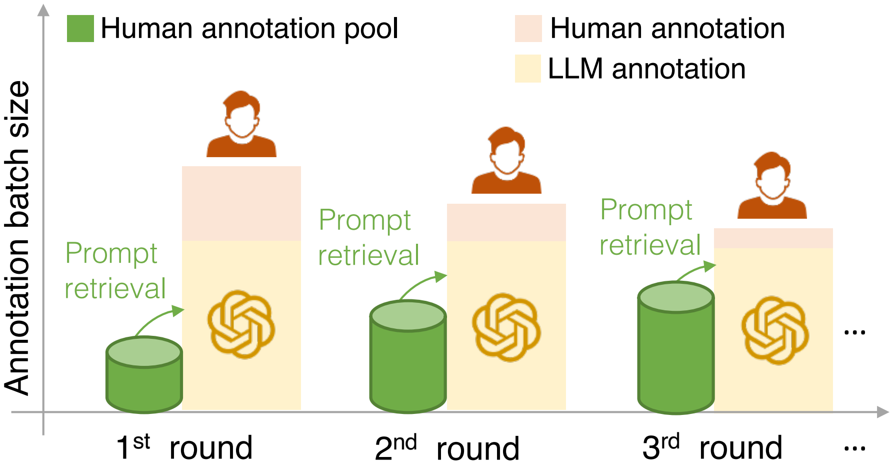
➢ Initialization $\theta^{(0)}=\underset{\theta^*}{\arg\min}\frac{1}{|\mathcal{A}_H^0|}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{A}_H^0}\mathcal{L}\left(f(\boldsymbol{x}_i;\theta^*),y_i\right),\quad i=1,\ldots,n_s$

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

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

Novel Designs in IMFL

- *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 $f(\boldsymbol{x}; \theta^{(r)}), \mathcal{A}$

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*

$$\boldsymbol{x}_i^* = \arg\max_{\boldsymbol{x}_i} \left[1 - p(f(\boldsymbol{x}_i;\theta^{(r)}) \mid \boldsymbol{x}_i;\theta^{(r)})\right]$$

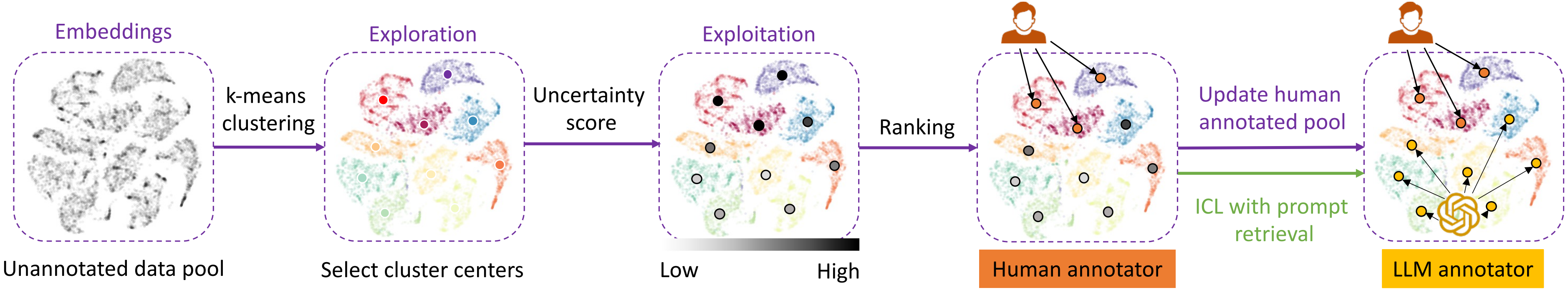


Illustration of exploration-exploitation query strategy with core components and steps

Experiments Setup

Fine-tuning. Dolly 2.0 as the target LM for fine-tuning on 8 NVIDIA V100 32G

Query and Annotation. GPT-3.5-turbo as the LLM annotator and limited our unannotated data pool to only contain 3000 data samples (sampled from the original training dataset)

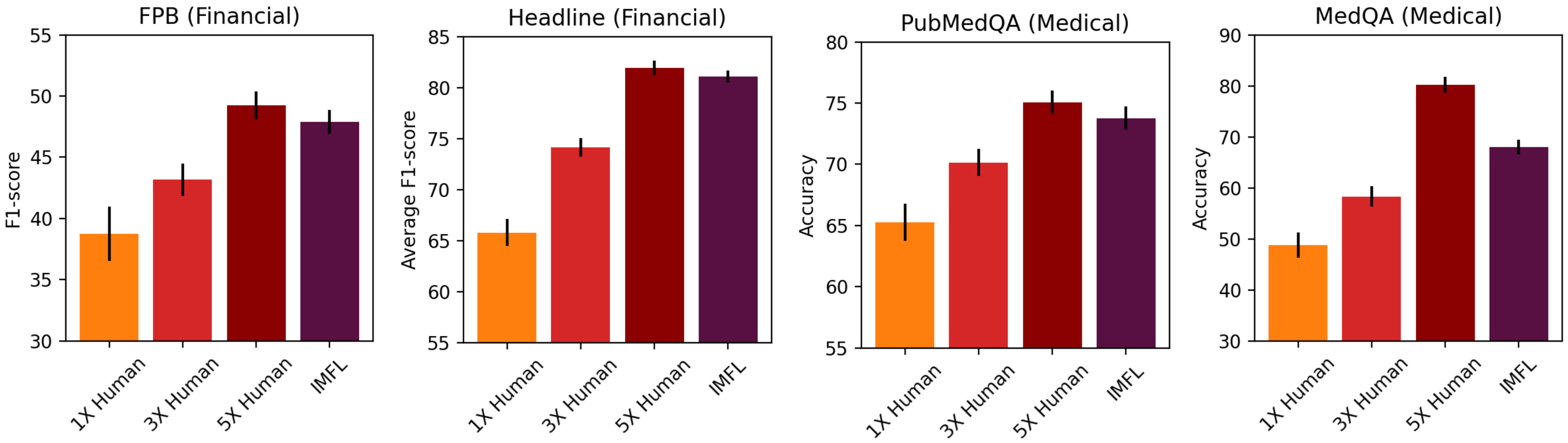
Annotation and Computational Budget. Annotation budget of 1000 for all datasets, human annotation is 200 and LLM annotation budget is 800. A total number of interaction rounds for fine-tuning is 5.

Table 2: Summary of the four domain-specific datasets used in our experiments.

Domain	Name	Task	Size (train/test)	Metric
Financial	FPB [29]	Sentiment Analysis	3876/969	F-1 score
Financial	Headline [40]	News Classification	9130/2282	Average F-1 score
Medical	PubMedQA [18]	Biomedical QA	500/500	Accuracy
Medical	MedQA [17]	Medical knowledge QA	11450/1273	Accuracy

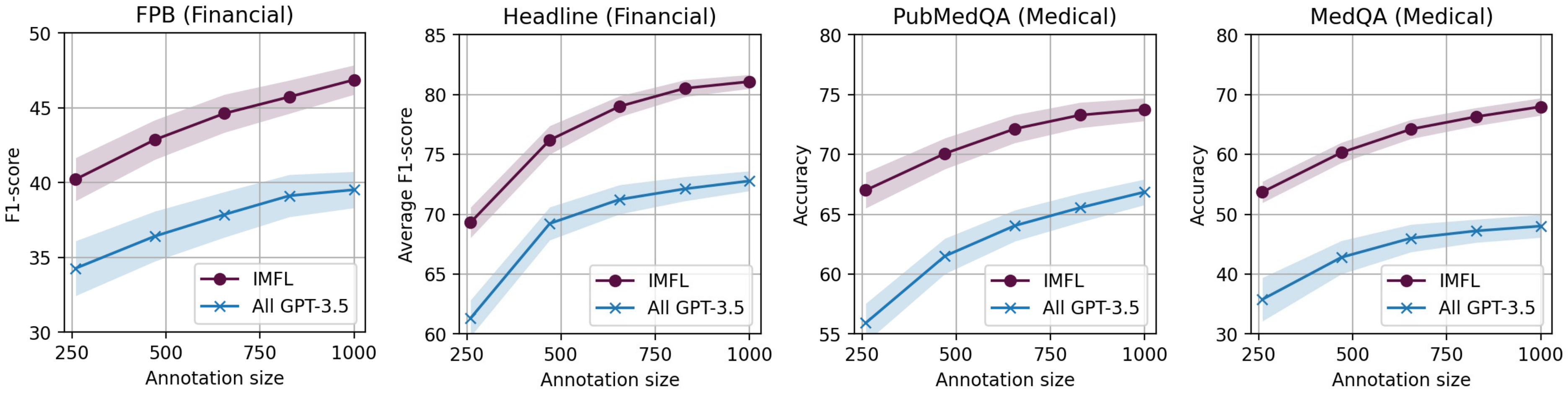
Main Results: IMFL vs Human

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



Main Results: IMFL vs GPT

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



Analysis: EEQ and Designs

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

Analysis: different LLMs and Human Ratio

A comparison of annotation accuracy by GPT-3 , GPT-3.5 and GPT-4 in zero/few-shot learning

	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

Ablation Study of Human Annotation Ratio

Method	Number of Annotations		Dataset			
	Human	GPT-3.5	FPB	Headline	PubMedQA	MedQA
IMFL	200 (1×)	800	**47.88 ± 0.98**	**81.09 ± 0.58**	**73.76 ± 0.95**	**67.98 ± 1.45**
IMFL	100 (0.5×)	900	43.66 ± 1.42	75.41 ± 1.01	70.88 ± 1.08	61.44 ± 1.83
IMFL	50 (0.25×)	950	40.76 ± 1.48	73.65 ± 1.09	68.18 ± 1.11	52.38 ± 1.93

Discussion and Limitation

Our achievement

- IMFL can significantly reduce the high cost of human annotation in domain-specific tasks.
- IMFL efficiently uses sparse human supervision to improve GPT-3.5/4 annotations through prompt retrieval and in-context learning, ultimately leading to enhanced performance.

Our future work

- IMFL framework assumes that the annotation budget is defined by the number of annotations, rather than reflecting the true cost which typically involves multiple complex factors.
- IMFL's performance is limited by the size of the unannotated dataset and the diversity of examples presented in the dataset as IMFL only seeks to improve performance through annotating existing samples rather than creating new samples.
- The performance of IMFL to continue to grow by incorporating stronger LLM annotators, such as GPT-4-turbo, to further improve annotation accuracy