LLMs for Now, Fine-Tuning for Later: An Ensemble Approach to Data Drift in Domain-Specific Tasks

Yuxuan Lu
Northeastern University
lu.yuxuan@northeastern.edu

Bingsheng Yao
Northeastern University
b.yao@northeastern.edu

Shao Zhang*
Shanghai Jiao Tong University
shaozhang@sjtu.edu.cn

Yisi Sang
Amazon
yisisang@amazon.com

Yun Wang
Microsoft
wangyun@microsoft.com

Hansu Gu
Independent
gu.hansu@gmail.com

Peng Zhang
Fudan University
zhangpeng_@fudan.edu.cn

Tun Lu
Fudan University
lutun@fudan.edu.cn

Toby Jia-Jun Li
University of Notre Dame
toby.j.li@nd.edu

Dakuo Wang
Northeastern University
d.wang@northeastern.edu

Abstract

Deploying machine learning models in real-world domain-specific scenarios is challenged by the scarcity of expert annotations and by data drift, where the statistical properties of incoming data continuously evolve. Active Learning (AL) iteratively improves compact models with expert annotations but suffers from recurring cold-start degradation, while LLMs provide strong off-the-shelf performance yet cannot leverage newly accumulated labels, raising the question: how can we better leverage LLMs to assist the active learning process? Through an empirical study on five legal and biomedical datasets, we reveal a complementary temporal dynamic: LLMs excel during early and post-drift stages, while AL-assisted compact models eventually surpass them as annotations accumulate. Motivated by this finding, we propose an ensemble system that combines an LLM, an AL-assisted compact model, and an automatic switch module that routes predictions to the better-performing model in real time. Evaluated under simulated data drift on two mental health datasets, our system achieves 96–98% switch accuracy and consistently outperforms either model used alone.

1 Introduction

Deploying machine learning models in real-world domain-specific scenarios, such as clinical diagnosis, legal analysis, and education, is challenged not only by the scarcity of expert annotations (Wu et al., 2025; Yin et al., 2024; Wu et al., 2022; Chen et al., 2024), but more fundamentally by the non-stationary nature of real-world data. In practice, the statistical properties of incoming data continuously evolve over time, a phenomenon known as **data drift** (Žliobaitė et al., 2014). For instance, in the biomedical domain, the emergence of new diseases and treatments requires clinicians to constantly update their knowledge and decision-making criteria (Finlayson et al., 2021; Zhou et al., 2025). In legal practice, evolving regulations and precedents shift how contracts and terms of service should be interpreted. Due to such data drift, a model performing well today may become unreliable tomorrow, making static, one-time training insufficient for real-world deployment. A robust system must not only learn effectively from limited expert annotations, but also adapt quickly and continuously as the data distribution evolves.

Active Learning (AL) (Settles, 2009) is a natural fit for this setting: AL iteratively selects the most informative unlabeled examples for expert annotation, allowing a compact model to be fine-tuned continuously as new labeled data becomes available from domain experts' daily workflows (Shen et al., 2017; Yao et al., 2023). However, AL suffers from a well-known structural weakness: the **cold-start** problem (Jin et al., 2022; Bayer et al., 2026), where the model performs poorly in its early training stages due to insufficient labeled data. Crit-

*Work was done when Shao Zhang was visiting Northeastern University.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 861–876

ically, in the presence of data drift, this weakness is not a one-time obstacle but a recurring one. Each time the data distribution shifts, the model's previously learned patterns become misaligned, effectively resetting it to a cold-start state. This means AL-assisted models perform worst precisely when the user needs reliable predictions most: at initial deployment and immediately after each drift event.

Large Language Models (LLMs), through prompting strategies such as zero-shot and few-shot In-Context Learning (ICL) (Brown et al., 2020), offer a complementary strength. With zero or minimal annotations, LLMs can deliver reasonable performance on domain-specific tasks off-the-shelf (Wei et al., 2021; Zhang et al., 2024c; Jia et al., 2024; Zhang et al., 2025b; Wang et al., 2025a,b; Zhang et al., 2025a; Chen et al., 2025; Wang et al., 2026), making them well-suited for periods when labeled data is scarce. Yet unlike AL-assisted models that continuously improve with incoming annotations, LLMs cannot effectively leverage newly labeled data, and the model's capability remains largely fixed regardless of how much additional domain data becomes available. This raises a critical but underexplored question: rather than treating LLMs and AL-assisted compact models as competing alternatives, how can we better leverage LLMs to assist the active learning process itself?

To answer this question, we first conduct an empirical study comparing AL-assisted compact models against state-of-the-art LLMs (GPT-3.5 and GPT-4 (OpenAI, 2023a)) on five domain-specific datasets spanning the legal and biomedical domains. Our results reveal two clear findings. First, AL-assisted T5-base models, fine-tuned on only a few hundred expert annotations, can reliably surpass GPT-3.5 and reach performance comparable to or exceeding GPT-4 across all datasets. Second, all AL settings suffer from severe cold-start degradation in their early stages, during which they significantly underperform both LLMs. Together, these findings reveal a complementary temporal dynamic: LLMs provide strong and stable performance from the start, while AL-assisted compact models start weak but eventually overtake LLMs as domain-specific annotations accumulate.

Motivated by this complementary dynamic, we propose an ensemble system that combines an LLM and an AL-assisted compact model with an automatic switching mechanism. The system comprises three modules: (1) an LLM that provides zero-shot or few-shot ICL predictions to bootstrap performance during cold-start and post-drift periods, (2) an AL-assisted, locally fine-tunable compact model that continuously improves as expert annotations accumulate, and (3) a switch module that evaluates both models on a rolling validation set and routes predictions to the better-performing one at each time step. When data drift occurs and the compact model's performance degrades, the switch automatically falls back to the LLM; as the compact model retrains on newly annotated data and recovers, the switch hands control back to the AL-assisted model. We evaluate the system on two mental health datasets under simulated data drift, and results show that the switch achieves 96–98% accuracy in identifying the better model, and the ensemble consistently outperforms either model used alone.

To summarize, our contributions are as follows:

- We conduct an empirical study on five domain-specific datasets across the legal and biomedical domains, comparing AL-assisted compact models with state-of-the-art LLMs. Our results reveal a temporal dynamic: LLMs excel in early stages while AL-assisted models surpass them as human annotations accumulate.
- We propose an ensemble system that leverages this temporal complementarity through an LLM, an AL-assisted compact model, and an automatic switch module that selects the better-performing model in real time.
- We evaluate the ensemble system under simulated data drift scenarios on two mental health datasets, demonstrating that the system consistently outperforms either model used alone and accurately identifies the best model at each time step.

2 Related Works

2.1 Active Learning of Domain-Specific Models

Active Learning (AL) (Shen et al., 2017; Ash et al., 2019; Teso and Kersting, 2019; Kasai et al., 2019; Zhang et al., 2022; Yao et al., 2023) is an iterative process that involves the following steps at each iteration: 1) select a few examples from an unlabeled data repository with a specific AL sampling algorithm, 2) finetune the model after acquiring annotation for the selected examples, and 3) assess the model's performance. AL sampling strategies are commonly categorized into two high-level categories (Settles, 2009; Olsson, 2009; Fu et al., 2013):

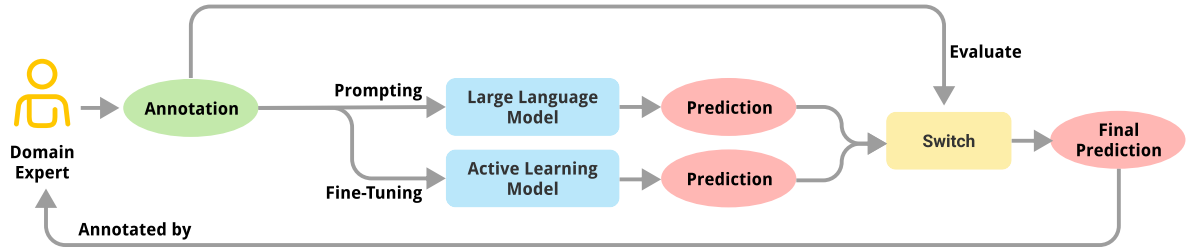


Figure 1: System architecture of the LLM+AL ensemble system. The system comprises three modules: (1) an LLM for zero-/few-shot predictions, (2) an AL-assisted compact model for continuous fine-tuning, and (3) a switch module that routes predictions to the better-performing model.

similarity-based and uncertainty-based strategies. The similarity-based strategies aim to identify the most representative examples based on similarities in data features, whereas the uncertainty-based approaches attempt to locate examples based on the model's confidence.

Although AL frameworks reduce human labor, they suffer from cold-start issues, with poor model performance during the early training stages. Many attempts have been made to assist AL models in the cold start stages, such as using the Masked Language Model (Yuan et al., 2020), representative sampling strategies (Jin et al., 2022), and Outlier-based Discriminative AL (ODAL) (Barata et al., 2021). Yet, these methods mostly focus on proposing new AL sampling strategies, which still fail to adapt to complicated data drift situations dynamically.

2.2 Domain Adaptation of Large Language Models

LLMs (Brown et al., 2020; OpenAI, 2023b; Touvron et al., 2023a,b) have shown promising task-solving abilities off-the-shelf without fine-tuning (Wei et al., 2021). However, generic LLMs do not perform reliably well in specialized domains, potentially due to the lack of domain-specific knowledge and adaptations. The research on domain adaptation of LLMs derives two directions: fine-tuning LLMs with domain-specific tasks, and effective prompting strategies that do not require updates to model's parameters. For instance, Xu et al. (2023) proposed Mental-LLM, which fine-tunes a generic LLM with multiple domain-specific tasks to achieve significantly enhanced and reliable performance in mental health detection scenarios. Moreover, innovative prompting methods, such as Chain-of-Thoughts (Wei et al., 2022) and

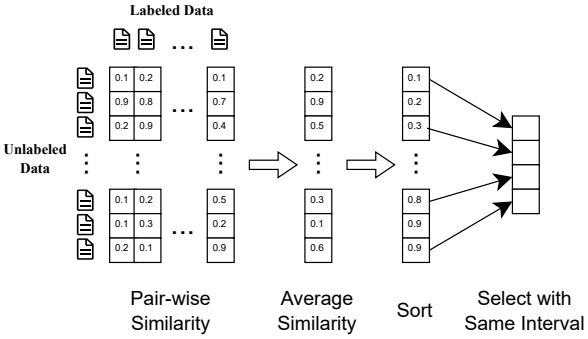


Figure 2: The sampling process of our data diversity-based strategy.

In-Context Learning (ICL) (Brown et al., 2020) were proposed to harness the potential of LLM for robust domain adaptations. For instance, Yao et al. (2024) proposed In-Context Sampling that augments multiple ICL prompts for reliable model performance in specialized domains. RAFT (Zhang et al., 2024b) utilized RAG to retrieve related documents to consistently improve the model's performance in domain-specific QA tasks.

3 Preliminary Empirical Study

We first conduct a preliminary empirical study to compare AL-assisted compact models against state-of-the-art LLMs, to formulate a better understanding of "When and how can AL-assisted small models outperform Large Language Models".

3.1 Active Learning-Assisted Models

We choose T5 (Raffel et al., 2020) as representatives for locally fine-tunable compact models based on existing works that demonstrate its strong performance for domain-specific fine-tuning (Yao et al., 2022; Mou et al., 2021). We initialize the T5 model with `T5-base`, a pre-trained weight that has been trained on many general-domain downstream tasks.

Dataset	Domain	Task	# Test Data
BioMRC (Pappas et al., 2020)	Biomedical	Multi-Choice	6,250
CUAD (Hendrycks et al., 2021)	Law	Classification	4,182
Unfair_TOS (Lippi et al., 2019)	Law	Classification	1,620
ContractNLI (Koreeda and Manning, 2021)	Law	NLI	1,991
Casehold (Zheng et al., 2021)	Law	Multi-Choice	3,600

Table 1: Datasets involved in our empirical study.

Algorithm 1 Active Learning Sampling Process

function SELECT($D_t, D_p, N, strategy$)
  D_t: unlabeled data in the training split
  D_p: previously selected data
  N: number of data needed
  $strategy$: Active Learning strategy
  if $strategy =$ `"similarity"` **then**
    $S \leftarrow \left(\frac{\sum_{d_p \in D_p} \cos(d_i, d_p)}{|D_p|}\right)_{1 \le i \le |D_t|}$
    $id \leftarrow \text{argsort}(S)$
    $step \leftarrow \frac{|D_t|}{N}$
    $result \leftarrow (id_i)_{i \equiv 0 (\text{mod } step), 1 \le i \le |D_t|}$
    return $result, id - result$
  end if
  if $strategy =$ `"uncertainty"` **then**
    $S \leftarrow (\text{Uncertainty}(d_i))_{1 \le i \le |D_t|}$
    $id \leftarrow \text{argsort}(S)$
    return $id_{<N}, id_{\ge N}$
  end if
end function

3.2 Active Learning Strategies

Following the established taxonomies of AL strategies (Schröder and Niekler, 2020), we designed and implemented one **data diversity-based** strategy and one **model uncertainty-based** strategy. We illustrate the details of each strategy below and in Algorithm 1.

Data Diversity-Based Strategy. During the data pre-processing stage, we utilize Sentence-BERT (Wang et al., 2020) to embed each data content as a vector to prepare for the diversity-based AL sampling. For each iteration of the diversity-based AL sampling strategy, we 1) calculate the average cosine similarity score between each unused training data and all previously used training data, 2) sort the unused data by the average similarity score, and 3) select representative examples with the same interval from the sorted list to ensure diversity. For instance, in order to select 4 examples from 10 unused data, we select the 1st, 4th, 7th and 10th data from the ranked list after Step 2. This strategy design allows us to ensure the diversity and representativeness of selected examples.

Model Uncertainty-based Strategy. The model Uncertainty-Based Strategy (Sener and Savarese, 2018) aspires to identify samples the model is least confident about. Within each iteration, the model operates on the training data, computing the logits and locating the samples holding the minimal average probability on the highest-ranked tokens.

In addition to the aforementioned two types of AL strategies, we also include a random AL sampling baseline. For each iteration in the AL simulations, we sample 16 data samples with a specified strategy and then evaluate the model on the test split. Each AL setting was executed 10 times, and we report the mean and standard errors.

3.3 Large Language Models

For the experiments with LLMs, we utilize two SOTA generic LLMs: GPT-3.5 and GPT-4 (OpenAI, 2023b). We probe the best-performing prompting strategy for each dataset with LLMs through extensive experiments on GPT-3.5 (reported in Table 4) and apply the same settings for GPT-4.

3.4 Datasets

We thoroughly examine existing expert-annotated datasets for specific real-world domains that require extensive expertise and choose BioMRC (Pappas et al., 2020), CUAD (Hendrycks et al., 2021), Unfair_TOS (Lippi et al., 2019), ContractNLI (Koreeda and Manning, 2021), and Casehold (Zheng et al., 2021) for our evaluation. The datasets are in legal and biomedical domains and are comprised of different types of tasks, including Multiple Choice, Classification, and Natural Language Inference (MacCartney and Manning, 2008). The dataset details are in Table 1.

Strategy	*Not-None* Ratio	*None* Ratio
Random	0.1247	0.8752
Diversity	0.1255	0.8744
Uncertainty	0.1458	0.8541
Complete dataset	0.1252	0.8747

Table 2: Label distributions of complete dataset and data sampled by different AL strategies in Unfair_TOS. The ratio is calculated by dividing the corresponding data type by all data counts.

3.5 Results

We plot the results on four legal domain datasets in Figure 3, and the results on BioMRC in Appendix A. The horizontal lines symbolize the best performance of GPT-3.5 and GPT-4, respectively. Unsurprisingly, all AL approaches suffer from the "cold-starting" problem. However, on all four datasets, the T5-base with AL can reliably **outperform GPT-3.5** and eventually reach a saturated performance that is **comparable with or even exceeds GPT-4**, leveraging a total of several hundred selected data. For BioMRC, as shown in Figure 5, the T5-base can also consistently beat GPT-3.5 but is saturated at a slightly lower performance compared to GPT-4. However, we believe GPT-4 might have seen or been trained on most of these datasets because they are publicly available text corpora. Regardless, our fine-tuned T5-base achieves comparable performance with GPT-4 despite having hundreds of times fewer parameters and requiring significantly less computational power.

Analysis of AL Strategies on Unfair_TOS. We observe the AL models in Unfair_TOS merely output "None" regardless of the input prior to the 20th iteration, but we can also observe clear advantage differences between AL strategies, where the uncertainty-based strategy can lead to better performance and saturate at higher results compared to the other settings.

The Unfair_TOS dataset consists of around 85% of data labeled *None*, and the rest of the data lies in eight other categories. We believe the AL model will be able to achieve a higher averaged F1 score if the AL strategy can select more *Not-None* data for the model to learn from. As a result, we calculate the label ratio for the original dataset and the data sampled by different AL strategies on the Unfair_TOS dataset, which can be found in Table 2. The ratio is calculated by dividing the corresponding data type by the count of all data. We sum the counts of all other eight data types and denote them as *Not-None*. We can observe the model uncertainty-based strategy selects significantly more *Not-None* labeled data than random ($t(14) = -2.46$, $p < 0.05$) and diversity ($t(14) = -2.51$, $p < 0.05$), which justifies the better performance of the uncertainty-based strategy.

Influence of Few-Shot Example Numbers. To establish a more solid evaluation, we conducted an additional experiment by evaluating GPT-4's performance when given different amounts of few-shot demonstrations. We used 1, 10, 50, and the maximum amount subject to the model limit. If GPT-4 can only handle less than 50 examples, we omit the results for the 50 shots and report the max-shot results instead. To ensure reproducibility and control cost, we randomly sample 200 examples from the original test split with fixed $seed = 42$.

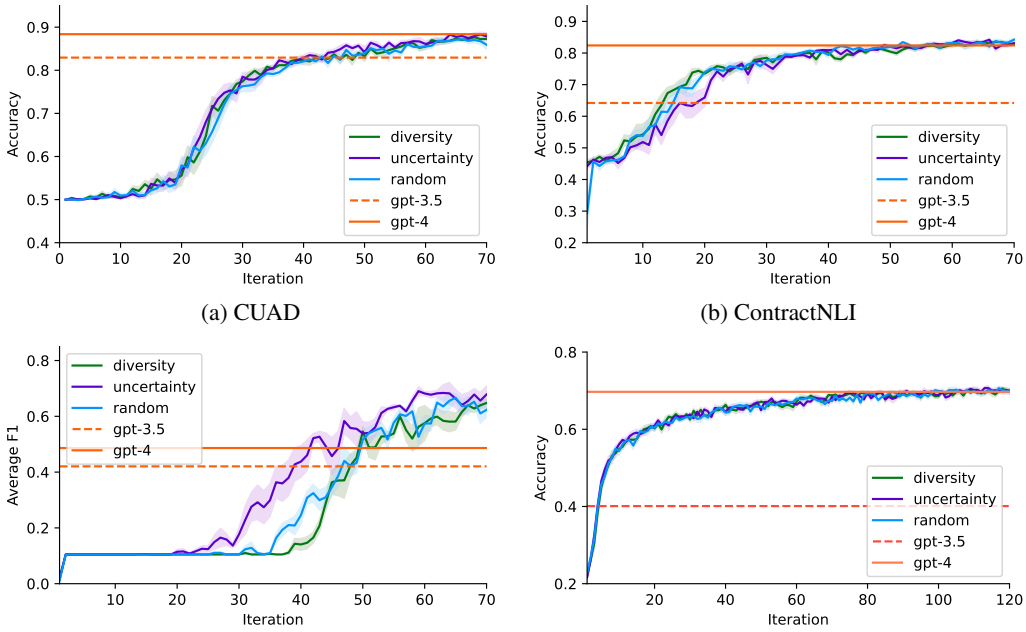
The result is reported in Table 3. We observe that generic LLM's (GPT-4) performance does not always increase when we add more and more data into the prompt, and with $\tilde{1}0$ shots can generally result in a saturated performance. Also, in three of the five datasets experimented, GPT-4 can only fit fewer than 20 few-shot examples in their context limit, justifying the need for small, fine-tuned models for domain-specific tasks.

4 LLM+AL Ensemble System

The architecture of our system is illustrated in Figure 1. We aim to develop a system capable of learning from a **stream of annotations**, delivering **high-quality predictions from the beginning**, and **continuously improving** as more data becomes available. When the data distribution shifts, i.e., **data drift**, the system should adapt to the new distribution and maintain high-quality predictions.

Our system consists of three modules:

1. An LLM: This component provides zero-shot or few-shot in-context learning predictions to quickly bootstrap the system's capabilities, thereby addressing the cold-start issue for the Active Learning (AL) model.
2. An AL-assisted, locally fine-tunable compact model: This model is trained with collected implicit annotations, enhancing performance as more data becomes available and adapting to the evolving data.
3. A "Switch" module: This module evaluates the predictions from different models in real-time using the collected data, determining



(a) CUAD (b) ContractNLI

(c) Unfair_TOS (d) Casehold

Figure 3: AL simulation results. The horizontal line represents two close-domain LLMs' best performance. We report the mean value (line) and standard error (colored shaded area) over 10 trials. Each AL iteration comprises 16 examples. We can observe the T5-base with AL can reliably **outperform GPT-3.5** and reach a saturated performance that is **comparable with or even exceeds GPT-4** on all four datasets.

Dataset	1-shot	10-shot	50-shot	max-shot (avg. # of shots)
BioMRC	0.835	0.810	-	0.760 (13 shots)
Unfair_tos	0.441	0.488	0.567	0.563 (137 shots)
ContractNLI	0.715	0.750	-	0.740 (47 shots)
CUAD	0.795	0.790	-	0.82 (18 shots)
CaseHOLD	0.660	0.790	-	0.735 (19 shots)

Table 3: GPT-4 result with different number of few-shot examples

which model's prediction should be used based on current performance.

4.1 Switch Module

The "switch" module is designed to evaluate the performance of the LLM and AL-assisted model in real-time, determining which model's prediction should be used based on current performance.

To address the data drift issue and maintain an up-to-date validation set, the switch will collect the first batch of data samples as the initial validation set. After obtaining this initial set, the switch will continuously update the validation data by replacing a random sample in the set with a newly annotated data sample at a configurable probability p, and the replaced sample will be used in model fine-tuning. This approach allows users to customize the switch's sensitivity to data drift by adjusting the value of p. For instance, if users are more concerned about data drift, they can set a higher p value to update the validation set more frequently. Conversely, if they prioritize the system's stability, they can set a lower p value to update the validation set less frequently. The candidate models will be evaluated on the selected validation set, and the best performing model's prediction will be used as the output of the ensemble system.

4.2 Evaluation Metric

To evaluate the performance of our ensemble, particularly during the cold start and data drift phases, we utilize the Area Under the Curve (AUC) of the Number of Data versus Accuracy plot as our evaluation metric, which is also the average accuracy

Dataset	Metric	GPT-3.5				GPT-4
		0-shot	1-shot	3-shot	10-shot	
CUAD	Accuracy	0.6404	0.8048	**0.8293**	0.8178	0.8837
BioMRC	Accuracy	0.4067	**0.5169**	0.5040	0.4532	0.8259
Unfair_tos	F1	0.4201	0.3847	0.3758	**0.4206**	0.4863
ContractNLI	Accuracy	0.4580	0.5990	0.5750	**0.6420**	0.8240
Casehold	Accuracy	0.3040	0.3020	0.3330	**0.4010**	0.6970

Table 4: Hyper-parameter tuning experiment results for GPT-3.5 and GPT-4.

across different time steps.

4.3 Power Analysis of AL Models' Performance

The design goal of the switch is to identify the optimal candidate model with the fewest samples in the validation set. Therefore, we conduct a power analysis on the empirical study to determine the amount of test data for the switch module.

First, we calculate the effect size by measuring the difference between the performance of the best model and the second-best model in the AL simulation results. We compute Cohen's d effect size (Cohen, 1988) on Unfair_TOS, resulting in 0.34.

Next, using this effect size, we determine the required sample size for the comparison with a power of 0.8 and a significance level of 0.05. The resulting sample size is 137. This approach ensures our switch has an 80% chance of identifying the best model when there are differences between the two models' performance. Additionally, in our experiment, if 137 samples constitute less than 10% of the training samples used, we set the size of the validation set to 10% of the training set.

5 Evaluation and Results

We evaluate our LLM+AL Ensemble system by conducting a close simulation of a real-world scenario. We iteratively provide golden truth annotations to the system, which then selects a subset of these annotations as validation data and determines the backend model for output. We compare our system against the following baselines:

1. **LLM Few-Shot Prompting**: We use GPT-4's best performance as the baseline.
2. **Random Sampling AL**: We use an AL model (T5) with a random sampling strategy as the backend model.

Model	Accuracy	
	SDCNL	Dreaddit
Ensemble System	**83.49%**	**81.3%**
LLM Few-Shot	65.50%	<u>76.00%</u>
AL-Assisted T5	<u>71.55%</u>	71.67%
Switch Accuracy	96.00%	98.00%

Table 5: Results of the ensemble system, compared to the LLM few-shot in-context-learning and AL-assisted T5. The best performance is shown in **bold**, and the second best is <u>underlined</u>. Our ensemble system outperforms both models in all datasets, demonstrating the capability of our switch module. The accuracy of the switch module is shown in the bottom line.

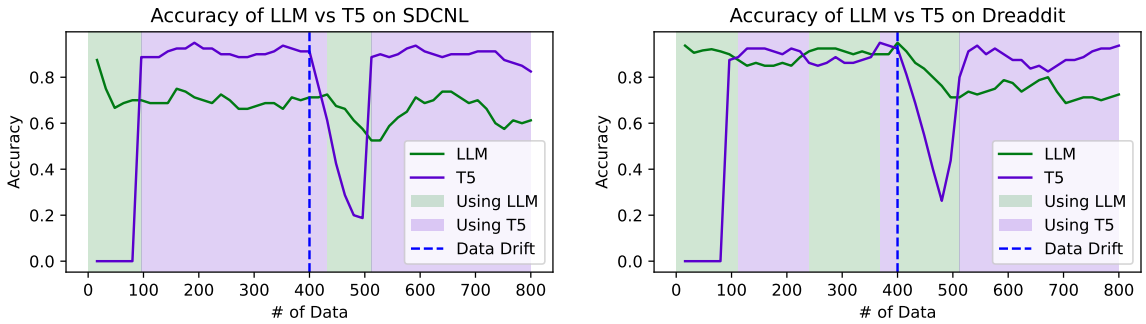
5.1 Datasets

To evaluate the system under controlled data drift, we select two binary classification datasets in the mental health domain: SDCNL (Haque et al., 2021) and Dreaddit (Turcan and McKeown, 2019). The binary nature of these tasks allows us to construct controlled data drift scenarios by flipping the label distribution ratio. Specifically, we construct two subsets for each dataset, one with a 9:1 label distribution ratio and the other with a 1:9 label distribution ratio. The system is first provided with the first subset. Then, in the middle of the experiment, we switch to the second subset to simulate a sudden data drift in the label distribution.

5.2 Results

The results are shown in Figure 4. The x-axis represents the number of data samples, and the y-axis represents the accuracy of two candidate models.

AL models in both datasets suffer from cold-start issues in the early stage and after the data drift. In both datasets, the system detects the change in model performance and uses the LLM's predictions as the output. Once the AL model surpasses its



(a) SDCNL (b) Dreaddit

Figure 4: System evaluation results. The accuracy of LLM versus T5 models is depicted for the SDCNL (a) and Dreaddit (b) datasets. The x-axis represents the number of data samples, while the y-axis denotes accuracy. The green and purple lines show the accuracy of the LLM and T5 models, respectively. The shaded regions indicate periods when either the LLM (green) or T5 (purple) models are in use. A data drift event (dashed blue line) occurs around step 400, leading to a temporary decline in the accuracy of the T5 model. The system dynamically switches between using LLM and T5 to maintain optimal performance.

starting stage and achieves better performance, the system switches to the AL model. In the experiment on the Dreaddit dataset, the AL model experiences a slight performance degradation due to overfitting before the data drift. The system adeptly switches back and forth between the AL model and the LLM to maintain the best performance.

The Accuracy AUC of our system is shown in Table 5. The results indicate that, despite the two models performing differently on the two datasets, our ensemble system consistently outperforms the two baseline models on all datasets. This demonstrates the effectiveness and generalization ability of our switch design.

The accuracy of the switch module, i.e., its ability to successfully identify the true performance difference between the two candidate models, is shown in the bottom line of Table 5. The results show that, even with minimal data, our switch design allows for accurate and sensitive measurement of the two candidate models' performances. This enables the system to achieve the best performance in addressing the data drift issue.

6 Discussion

In our empirical study, we observe that all AL strategies suffer from well-known "cold-start" issues (Chen et al., 2022; Jin et al., 2022), where the model performs poorly in the early iterations due to potential underfitting as a result of insufficient labeled data. On the other hand, LLMs, specifically GPT-4 in our case, yield reasonably good performance despite eventually being surpassed by AL models fine-tuned on domain-specific datasets.

We propose a promising future paradigm for real-world domain-specific tasks that incorporates LLMs and AL fine-tuned smaller models in parallel. Initially, the LLM's prediction will be presented to the human expert, and the collected annotations will be used to train the AL model. When the AL model begins to outperform the LLM, the system will "switch" to present the AL model's prediction. Thus, the LLM's prediction can help overcome the "cold-start" problem of AL, while the system can still benefit from AL's continually improving and up-to-date performance. Such an "human-in-the-loop" design is critical for the practical benefits of these models in real-world scenarios (Zhang et al., 2024a; Li et al., 2025).

In our system evaluation, we observe that our system consistently outperforms all baselines, and the accuracy of our switch design demonstrates the effectiveness of our system in real-world, domain-specific scenarios.

We also envision that LLMs' calibration ability (Zhu et al., 2023), where data samples that the LLM is least confident about tend to have lower accuracy, can also help cold-starting AL models. By utilizing a generic LLM as an assessor of the difficulty of the data samples, we can identify the hard-to-answer or incorrectly predicted examples during the sampling process for annotation, which may benefit the AL-assisted small models.

7 Conclusion

While LLMs such as GPT-4 have been endorsed for it's superior performance in many benchmarking datasets, whether they can substitute smaller

models, especially in real-world tasks and domains requiring extensive domain expertise, is critical but overlooked. In this work, we first present an empirical study evaluating the performance between SOTA generic LLMs (GPT-3.5 and GPT-4) and a much smaller language model (T5-base) fine-tuned with different Active Learning strategies on five specialized datasets representing real-world domain-specific tasks. Our evaluation demonstrates that AL-assisted models trained with expert annotation can consistently achieve or exceed best-performing LLMs with only a few hundred expert-annotated data, justifying that human experts remain indispensable in domain-specific tasks.

To better assist domain-experts' workflow without annotation burden and to facilitate real world's rapidly changing requirements, we propose an LLM+AL ensemble system. Results show that our system can identify the best performing model and consistently yield accurate prediction during cold starts and data drifts in real-world scenarios.

8 Limitations

Our empirical experiment of AL-assisted models solely utilizes a T5-base model, where the performance of other models, such as BART (Lewis et al., 2019) and even LLMs that can be efficiently fine-tuned with Parameter-Efficient Fine-Tuning techniques (Mangrulkar et al., 2022; Hu et al., 2021; Lester et al., 2021), remains to be explored. This work only benchmarks two generic LLMs (GPT-3.5 and GPT-4). Future work should benchmark newer and more capable LLMs such as GPT-5 and Claude Sonnet. We only implemented and evaluated two fundamental types (data diversity-based and uncertainty-based) of Active Learning strategies in our work. Future work should explore other families of AL strategies, e.g., hybrid or ensemble approaches (Krogh and Vedelsby, 1994; Qian et al., 2020), which may further improve annotation efficiency. Nevertheless, our empirical study with two fundamental Active Learning strategies justifies our primary statement that human experts are still needed in real-world domain-specific data annotation tasks.

Our system evaluation comprises two datasets from the mental health domain. While we acknowledge the existence of other domains and publicly available domain-specific datasets, we defer the analysis of the generalizability of our findings to other domains and tasks for future research. In our system evaluation, we only experiment with a random sampling strategy to closely mimic the daily work of domain experts. Designing and evaluating an AL sampling strategy that addresses real-world scenarios is also a future direction of research. Additionally, our system evaluation simulates data drift via an abrupt reversal of label distributions (from 9:1 to 1:9), representing an extreme case of sudden drift. Real-world data drift is often more gradual or involves shifts in feature distributions rather than label proportions alone. Evaluating the system under more realistic drift scenarios, such as gradual concept drift or covariate shift, is left for future work. The sample size for the switch module's statistical test is determined by a power analysis conducted on a single dataset (Unfair_TOS). The optimal sample size may vary across datasets with different characteristics, and an adaptive approach to determining the validation set size could improve robustness.

In addition, we primarily engage in model comparisons through automated metrics. However, these may not necessarily provide an accurate representation of a model's performance. Also, an error analysis on which type of questions LLMs may excel or fail is also meaningful for future work. Therefore, human evaluation including human agreement and error analysis, might be needed for a more comprehensive assessment.

References

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. **arXiv preprint arXiv:1906.03671**.

Ricardo Barata, Miguel Leite, Ricardo Pacheco, Marco O. P. Sampaio, João Tiago Ascensão, and Pedro Bizarro. 2021. Active learning for imbalanced data under cold start. In **Proceedings of the Second ACM International Conference on AI in Finance**, ICAIF '21, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Markus Bayer, Justin Lutz, and Christian Reuter. 2026. Activellm: Large language model-based active learning for textual few-shot scenarios. **Transactions of the Association for Computational Linguistics**, 14:1–22.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,

Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS'20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2024. StorySparkQA: Expert-annotated QA pairs with real-world knowledge for children's story-based learning. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pages 17351–17370, Miami, Florida, USA. Association for Computational Linguistics.

Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L. Yuille, and Zongwei Zhou. 2022. Making Your First Choice: To Address Cold Start Problem in Vision Active Learning. **Preprint**, arXiv:2210.02442.

Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. 2025. Unlearning isn't invisible: Detecting unlearning traces in llms from model outputs. **arXiv preprint arXiv:2506.14003**.

Jacob Cohen. 1988. **Statistical Power Analysis for the Behavioral Sciences**, 2 edition. Routledge, New York.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. 2021. The Clinician and Dataset Shift in Artificial Intelligence. **New England Journal of Medicine**, 385(3):283–286.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. **Knowledge and information systems**, 35:249–283.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction. In **Artificial Neural Networks and Machine Learning – ICANN 2021**, pages 436–447, Cham. Springer International Publishing.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. **Preprint**, arXiv:2103.06268.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. **Preprint**, arXiv:2106.09685.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pages 4276–4292.

Qiuye Jin, Mingzhi Yuan, Shiman Li, Haoran Wang, Manning Wang, and Zhijian Song. 2022. Cold-start active learning for image classification. **Information Sciences**, 616:16–36.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource Deep Entity Resolution with Transfer and Active Learning. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. **Advances in neural information processing systems**, 7.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. **arXiv preprint arXiv:1910.13461**.

Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choube, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital insight: Assisting experts' context-driven sensemaking of multimodal personal tracking data using visualization and human-in-the-loop llm. **Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies**, 9(3):1–37.

Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service. **Artificial Intelligence and Law**, 27(2):117–139.

Bill MacCartney and Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In **Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)**, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative Question Answering with Cutting-Edge Open-Domain QA Techniques: A Comprehensive Study. **Transactions of the Association for Computational Linguistics**, 9:1032–1046.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

OpenAI. 2023a. Gpt-4 technical report. **ArXiv**, abs/2303.08774.

OpenAI. 2023b. GPT-4 Technical Report. **Preprint**, arXiv:2303.08774.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A Dataset for Biomedical Machine Reading Comprehension. In **Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing**, pages 140–149, Online. Association for Computational Linguistics.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Learning structured representations of entity names using Active Learning and weak supervision. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 6376–6383, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. **The Journal of Machine Learning Research**, 21(1):5485–5551.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. **arXiv preprint arXiv:2008.07267**.

Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In **International Conference on Learning Representations**.

Burr Settles. 2009. Active learning literature survey.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. In **Proceedings of the 2nd Workshop on Representation Learning for NLP**, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In **Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society**, pages 239–245.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. **Preprint**, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. **Preprint**, arXiv:2307.09288.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In **Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)**, pages 97–107, Hong Kong. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In **Advances in Neural Information Processing Systems**, volume 33, pages 5776–5788. Curran Associates, Inc.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, et al. 2025a. Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. **arXiv preprint arXiv:2506.05606**.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, Jiri Gesi, Xianfeng Tang, Chen Luo, Yisi Sang, Hanqing Lu, Manling Li, et al. 2026. Trajectory2task: Training robust tool-calling agents with synthesized yet verifiable data for complex user intents. **arXiv preprint arXiv:2601.20144**.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. 2025b. Customer-r1: Personalized simulation of human behaviors via rl-based llm agent in online shopping. **arXiv preprint arXiv:2510.07230**.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In **International Conference on Learning Representations**.

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.

Siyi Wu, Weidan Cao, Shihan Fu, Bingsheng Yao, Ziqi Yang, Changchang Yin, Varun Mishra, Daniel Addison, Ping Zhang, and Dakuo Wang. 2025. Cardioai: A multimodal ai-based system to support symptom monitoring and risk prediction of cancer treatment-induced cardiotoxicity. In **Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems**, pages 1–22.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. **Future Generation Computer Systems**.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. **Preprint**, arXiv:2307.14385.

Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.

Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. 2023. Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pages 11629–11643, Singapore. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Changchang Yin, Pin-Yu Chen, Bingsheng Yao, Dakuo Wang, Jeffrey Caterino, and Ping Zhang. 2024. Sepsislab: Early sepsis prediction with uncertainty quantification and active sensing. In **Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, pages 6158–6168.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start Active Learning through Self-supervised Language Modeling. **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 7935–7948.

Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. 2024a. Rethinking human-ai collaboration in complex medical decision making: a case study in sepsis diagnosis. In **Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems**, pages 1–18.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022. ALLSH: Active learning guided by local sensitivity and hardness. In **Findings of the Association for Computational Linguistics: NAACL 2022**, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024b. RAFT: Adapting Language Model to Domain Specific RAG. **Preprint**, arXiv:2403.10131.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. 2024c. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. **arXiv preprint arXiv:2402.11592**.

Yimeng Zhang, Jiri Gesi, Ran Xue, Tian Wang, Ziyi Wang, Yuxuan Lu, Sinong Zhan, Huimin Zeng, Qingjun Cui, Yufan Guo, et al. 2025a. See, think, act: Online shopper behavior simulation with vlm agents. **arXiv preprint arXiv:2510.19245**.

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, et al. 2025b. Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning. **arXiv preprint arXiv:2507.17842**.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In **Proceedings of the eighteenth international conference on artificial intelligence and law**, pages 159–168.

Yuyue Zhou, Yan Zhang, Yingxia Nie, Dalin Sun, Deyu Wu, Lin Ban, Heng Zhang, Song Yang, Jiansong Chen, Haishun Du, et al. 2025. Recent advances and perspectives in functional chitosan-based composites for environmental remediation, energy, and biomedical applications. **Progress in Materials Science**, 152:101460.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the Calibration of Large Language Models and Alignment. **Findings of the Association for Computational Linguistics: EMNLP 2023**, pages 9778–9795.

Indrė Žliobaitė, Albert Bifet, Bernhard Pfahringer, and Geoffrey Holmes. 2014. Active Learning With Drifting Streaming Data. **IEEE Transactions on Neural Networks and Learning Systems**, 25(1):27–39.

A Empirical Study Result on BioMRC

For BioMRC, as shown in Figure 5, the T5-base with AL can quickly **outperform GPT-3.5** and eventually reach a saturated performance that is slightly lower than GPT-4. We posit that GPT-4 may have performed exceptionally well due to its exposure or training on BioMRC, given its source's public accessibility. Nevertheless, our refined T5-base model demonstrates comparable performance to GPT-4. Remarkably, this is achieved despite the T5-base model's comparative parameter deficiency - in the hundreds of times less - and a significantly lower demand for computational resources.

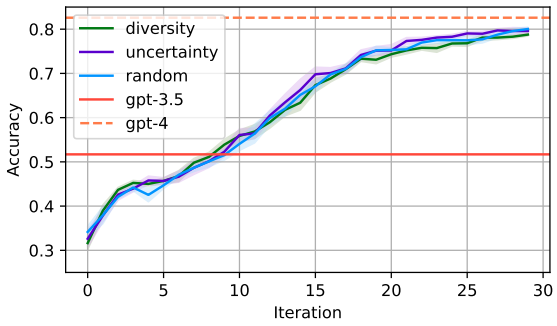


Figure 5: Result on BioMRC

B Hyperparameters and Settings

Dataset	Learning Rate	Training Epoch
BioMRC	1e-4	20
Unfair_TOS	1.5e-4	12
ContractNLI	1.5e-4	20
Casehold	4e-5	28
CUAD	6e-5	18
SDCNL	1e-5	20
Dreaddit	1e-5	20

Table 6: Hyperparameters for each dataset.

We report the experiment hyperparameters in Table 6. All our experiments are executed on one of two resources: 1) four NVIDIA V100 32G graphic cards and 2) eight NVIDIA V100 32G graphic cards. For GPT-3.5 and GPT-4, we used GPT-3.5-0613 and GPT-4-0613 respectively.

For model uncertainty-based strategies, we calculate the model probability on a randomly sampled subset of the training data to reduce the time complexity of the model uncertainty-based data sampling process. Compared to the naive approach's $O(n^2)$ time complexity, our implementation remains to have a time complexity of $O(n)$, which is the same as that of non-AL's (where n is the number of training data).

C Prompts Used for Each Dataset

Text in [[double brackets]] denotes input data.

C.1 BioMRC (Pappas et al., 2020)

```
I want you to act as an annotator for a
↪ question answering system. You will
↪ be given the title and abstract of a
↪ biomedical research paper, along
↪ with a list of biomedical entities
↪ mentioned in the abstract. Your task
↪ is to determine which entity should
↪ replace the placeholder (XXXX) in
↪ the title.

Here's how you should approach this
↪ task:

Carefully read the title and abstract of
↪ the paper.
Pay close attention to the context in
↪ which the placeholder (XXXX) appears
↪ in the title.
Review the list of biomedical entities
↪ mentioned in the abstract.
Determine which entity from the list
↪ best fits the context of the
↪ placeholder in the title.
Output only the identifier for the
↪ chosen entity (e.g., `@entity1`). Do
↪ not output anything else.

<INPUT>:
<title>:
[[TITLE]]
<abstract>:
[[ABSTRACT]]
<entities>:
[[ENTITY]]
<OUTPUT>:
```

C.2 UnfairTOS (Lippi et al., 2019)

```
I want you to act as an annotator for a
↪ Term of Service (ToS) review system.
↪ You will be given a piece of a Term
↪ of Service. Your job is to determine
↪ whether the ToS contains any of the
↪ following unfair terms:

Limitation of liability
Unilateral termination
Unilateral change
Content removal
Contract by using
Choice of law
Jurisdiction
Arbitration

If none of the above terms are present,
↪ you should output "None".

Here's how you should approach this
↪ task:

Carefully read the ToS.
Review the list of unfair terms.
For each unfair term, determine whether
↪ it is present in the ToS.
Output only the unfair terms that are
↪ present in the ToS. A ToS may have
↪ multiple unfair terms. \
You should output all of them, separated
↪ by a semicolon (;).
Do not output anything else.

<text>:
[[TEXT]]
<OUTPUT>:
```

C.3 ContractNLI (Koreeda and Manning, 2021)

```
I want you to act as an annotator for a
↪ question answering system. You will
↪ be given a contract and a hypothesis.
↪ Your task is to determine the
↪ hypothesis is contradictory,
↪ entailed or neutral to the contract.

Here's how you should approach this
↪ task:

Carefully read the contract.
Carefully read the hypothesis.
Determine whether the hypothesis is
↪ contradictory, entailed or neutral
↪ to the contract.
Output only the label (contradiction,
↪ entailment, neutral). Do not output
↪ anything else.

<INPUT>:
<premise>:
[[PREMISE]]
<hypothesis>:
[[HYPOTHESIS]]
<OUTPUT>:
```

C.4 CUAD (Hendrycks et al., 2021)

```
I want you to act as an annotator for a
↪ question answering system. You will
↪ be given the question and a piece of
↪ a contract. You will need to answer
↪ the question based on the contract.
↪ There are only two possible answers,
↪ "Yes" or "No".

Here's how you should approach this
↪ task:

Carefully read the question.
Carefully read the contract.
Determine the answer to the question is
↪ true or not.
Output only the exact answer (one of
↪ "Yes" or "No") of the questions. Do
↪ not output anything else.
<INPUT>:
<text>:
[[TEXT]]
<question>:
[[QUESTION]]
<OUTPUT>:
```

C.5 Casehold (Zheng et al., 2021)

```
I want you to act as an annotator for a
↪ Question Answering system. You will
↪ be given the question and several
↪ answers. Your job is to determine
↪ which answer best answers the
↪ question.

Here's how you should approach this
↪ task:

Carefully read the question.
```

```
Carefully read the answers.
Output the numeric index of the answers
↪  that best answers the question.
Do not output anything else.

<INPUT>:
<question>:
[[QUESTION]]
<answer>:
[[ANSWER]]
<OUTPUT>:
```

C.6 SDCNL (Haque et al., 2021)

```
This person wrote this paragraph on
↪  social media.
If you are a psychologist, consider the
↪  mental well-being condition
↪  expressed in this post and answer
↪  the question: does the person want
↪  to suicide?
Only return Yes or No.

<INPUT>:
<text>:
[[POST]]
<OUTPUT>:
```

C.7 Dreaddit (Turcan and McKeown, 2019)

```
This person wrote this paragraph on
↪  social media.
If you are a psychologist, consider the
↪  mental well-being condition
↪  expressed in this post and answer
↪  the question: is this person
↪  stressful?
Only return Yes or No

<INPUT>:
<text>:
[[POST]]
<OUTPUT>:
```