ALERTS: Active Learning and Ensemble LLM Real-Time Switch for Real-World Data Drift Challenges

Anonymous ACL submission

Abstract

In the rapidly changing real-world scenarios, data drift and "cold-start" issues present significant challenges for the development of machine learning models, along with the high cost and resource scarcity of domain experts. Traditional compact models fine-tuned on small number of domain-specific examples often outperform generic LLMs, despite the fine-tuned models struggling with rapid data changes. This study introduces ALERTS, an ensemble system designed to address these data challenges. The system comprises 1) an LLM to enhance early-stage performance and adapt to sudden data drifts, 2) an Active Learning (AL)-assisted compact model iteratively fine-tuned on annotations from daily human expert workflows, and 3) a switch mechanism that evaluates both models in real-time and selects the best performing ones. We conducted empirical studies to understand the performance between LLMs and AL-assisted compact models, then evaluated our system's effectiveness through AL simulations of real-world scenarios. Our work offers a novel framework for developing robust language model systems across various dynamic real-world scenarios.

1 Introduction

Developing and deploying machine learning models in real-world domain-specific scenarios (e.g., legal, clinical, and education (Xu et al., 2022; Pappas et al., 2020)) presents numerous challenges. For instance, such tasks generally require extensive domain-specific knowledge (Karabacak and Margetis, 2023) and high cost to recruit domain experts for high-quality and large-scale data annotation (Rasmussen et al., 2022; Wu et al., 2022), while facing domain experts' resource scarcity and unwillingness to be "data slaves". The difficulties in acquiring expert annotations motivate the rapid development of efficient low-resource training methods, such as Active Learning (AL, Settles (2009)), where a sampling strategy can efficiently select the most helpful data, query for annotation, and iteratively fine-tune the model.

Critically, real-world tasks and scenarios are constantly changing, leading to a common but very tough challenge called **data drift** (Žliobaitė et al., 2014)–when the statistical properties of the data drift from time to time. For example, in the biomedical domain, new diseases are discovered quite rapidly, enfocing clinicians to constantly learn new knowledge about diseases and treatments. A similar phenomenon can be found at the very beginning of model training, namely "cold-start" issues, which is also a long-lasting challenge for low-resource learning techniques.

Recently, Large Language Models (LLMs, Brown et al. (2020)) have shown great capability in a variety of generic tasks off-the-shelf. A variety of prompting strategies, without the need to fine-tune LLMs, were proposed to enhance LLMs' domain-adaptation capabilities, including few-shot In-Context Learning (ICL, Brown et al. (2020)), Retrieval-Augmented Generation (Guu et al., 2020), etc. However, recent work shows that traditional compact models fine-tuned on high-quality domain-specific datasets can reliably outperform much larger LLMs after being fine-tuned with a small number of examples (Xu et al., 2023).

Fine-tuned domain-specific models and LLMs are not without limitations in the rapidly changing real-world, nevertheless, we ask, can we benefit from both types of models to make up for each other's shortcomings? Precisely, can we develop a system that leverages LLMs' extraordinary task-solving capabilities to overcome the "cold-start" and data drift challenges for domain-specific fine-tuned compact models, while being able to switch back to the fine-tuned models once it can reliably outperform LLMs?

This work presents ALERTS: **A**ctive **L**earning and **E**nsemble LLM **R**eal-**T**ime **S**witch, an ensem-

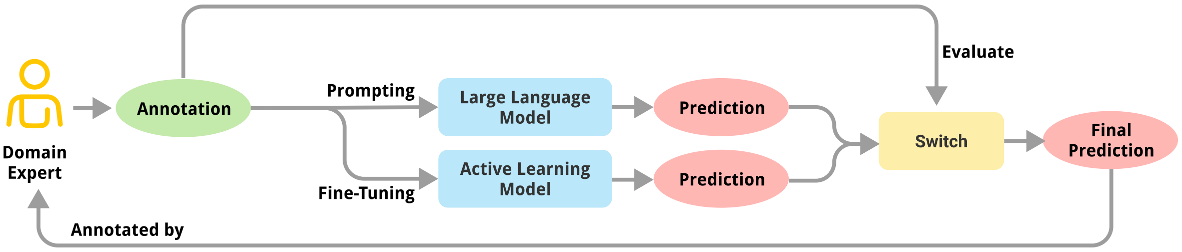


Figure 1: System Architecture of ALERTS

ble system we designed that aims to tackle the pervasive challenges of data drift in the real world with three primary modules: 1) An LLM supported by zero-shot and few-shot ICL prompting strategies to bootstrap the system's performance during early-stages and data drifts. 2) An AL-assisted, locally fine-tuneable compact model. We expect this model can be passively and iteratively trained with implicitly collected annotations during human experts' daily work. 3) A "Switch" mechanism that evaluates the predictions from different models in real-time using the collected data, determining which model's prediction should be used based on the current performance.

We first conduct an empirical study to analyze "When and how can AL-assisted small models outperform LLMs?" Our results show that the AL-assisted T5 model, with hundreds of human annotations, can consistently outperform GPT-3.5 and perform on par with GPT-4 in existing domain-specific datasets. Based on the findings, we designed and developed our ensemble system. We conduct an AL experiment to evaluate the effectiveness of our system in a simulated real-world data annotation scenario with simulated data drifts. Results show that our ensemble system can successfully identify the best candidate model and consistently yield accurate predictions during cold starts and data drifts. More importantly, our work paves a broader avenue for the future design and development of AI systems, with significant practical implications, in different real-world scenarios.

2 Related Works

2.1 Active Learning

Active Learning (AL) (Sharma et al., 2015; 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 a cyclical process that involves: 1) selecting examples from an unlabeled data repository utilizing AL selection strategies to be labeled by human annotators, 2) training the model with the newly labeled data, and 3) assessing the tuned model's performance.

A few AL surveys (Settles, 2009; Olsson, 2009; Fu et al., 2013; Schröder and Niekler, 2020; Ren et al., 2021) of sampling strategies provide two high-level selection concepts: data diversity-based strategies and model uncertainty-based strategies. The diversity-based approach aims to identify the most representative examples from the unlabeled data space while maximizing the diversity, while the uncertainty-based approach attempts to locate examples that the model is least confident about.

Many attempts have been made to assist active learning models in the cold start stages, such as using the Masked Language Model (Yuan et al., 2020), representative sampling strategies (Jin et al., 2022), Outlier-based Discriminative AL (ODAL) (Barata et al., 2021), etc.) Yet, these methods mostly focus on the active learning sampling strategies, and thus are tied to specific task/dataset or sampling strategy.

2.2 Large Language Models and Domain Adaptation

Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b) have shown great capability in a variety of tasks (Wei et al., 2021; Chung et al., 2022). Moreover, innovative prompting methods, such as Chain-of-Thoughts (Wei et al., 2023; Chung et al., 2022), and In-Context Learning (ICL) (Brown et al., 2020) were proposed to harness the potential of LLMs.

In addition to the prompting strategies, domain adaption of LLMs has also been a recent focus. Xu et al. (2023) proposed Mental-LLM which fine-tunes the Alpaca and FLAN-T5 model on mental corpus. RAFT (Zhang et al., 2024) utilized RAG

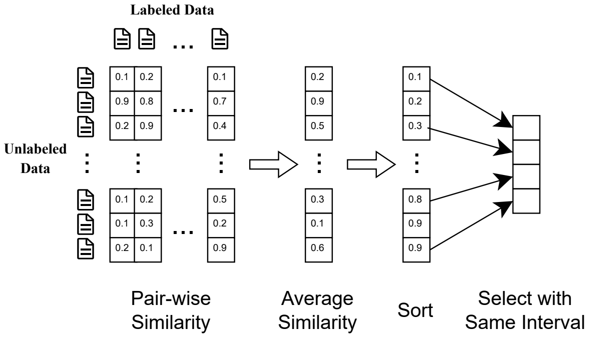


Figure 2: The sampling process of our data diversity-based strategy.

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
        $S \leftarrow \left(\frac{\sum_{d_p \in D_p} \cos(d_i, d_p)}{|D_p|}\right)_{1 \le i \le |D_t|}$
        $id \leftarrow \mathrm{argsort}(S)$
        $step \leftarrow \frac{|D_t|}{N}$
        $result \leftarrow (id_i)_{i \equiv 0 (\mathrm{mod}\ step), 1 \le i \le |D_t|}$
        return $result, id - result$
    end if
    if $strategy$ = "uncertainty" then
        $S \leftarrow (\mathrm{Uncertainty}(d_i))_{1 \le i \le |D_t|}$
        $id \leftarrow \mathrm{argsort}(S)$
        return $id_{<N}, id_{\ge N}$
    end if
end function
```

to retrieve related documents, and by filtering out unrelated documents, RAFT can consistently improve the model's performance in domain-specific QA tasks. These methods have been proved effective for LLMs' domain adaption, still, in real-world domain-specific scenarios, it's often not feastable to use LLMs due to the privacy risks (Plant et al., 2022) and the high demand on computational power (Luccioni et al., 2023).

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 out-perform Large Language Models", a critical but overlooked research question.

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.

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 follow a common practice of sampling 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 LLM, we utilize two SOTA generic LLMs: GPT-3.5 and GPT-4 (Ope-

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.

nAI, 2023). 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.

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 data selected. 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 publically 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.

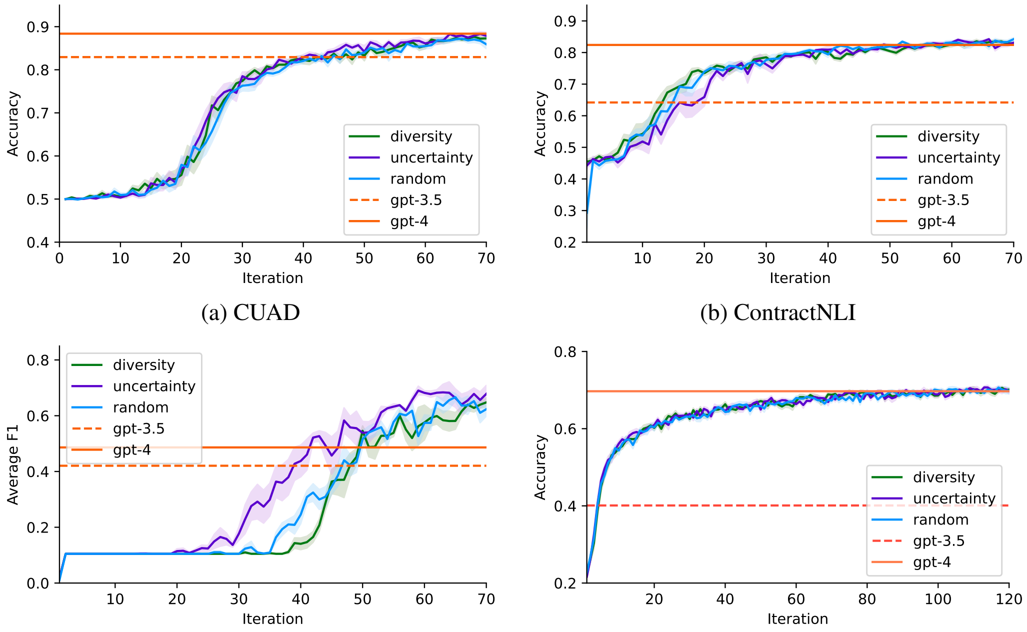
Strategy	*Not-None* Ratio	*None* Ratio
Random	0.1247	0.8752
Diversity	0.1255	0.8744
Uncertainty	0.1458	0.8541
Complete dataset	0.1252	0.8747

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.

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-



(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.

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 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

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

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 finetuning. 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 of different time stamp.

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.

5.1 Datasets

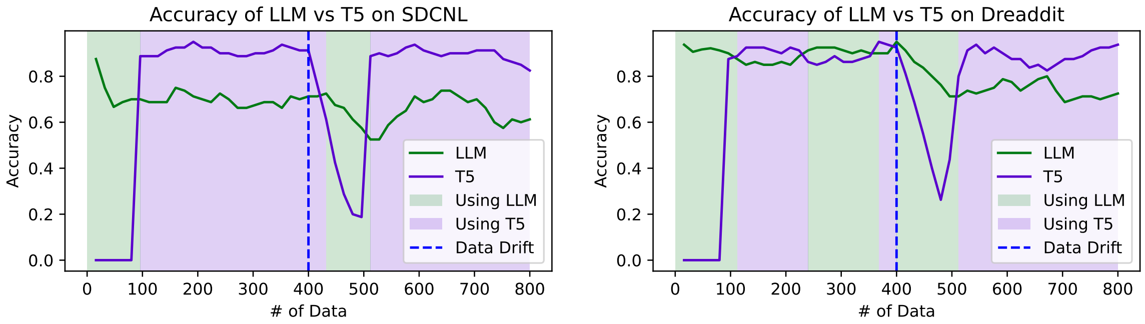
To mimic domain experts' daily work, we use two expert-annotated datasets in the mental health domain: SDCNL(Haque et al., 2021) and Dreaddit(Turcan and McKeown, 2019). Specifically, we construct two subsets for each dataset, one with a 90:10 label distribution ratio and the other with a 10:90 label distribution ratio, tosimulate data drift in real-world scenarios. The system is first provided with the first subset we constructed. Then, in the middle of the experiment, we switch to the second subset to simulate a data drift in the label distribution ratio. In this way, we can probe the system's performance when facing the real world's rapidly changing needs.

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.

Dataset	Metric	GPT-3.5				GPT-4
		0 shot	1 shots	3 shots	10 shots	
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.



(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.

Model	Accuracy	
	SDCNL	Dreaddit
Ensemble System	**83.49%**	**81.3%**
LLM Few-Shot	65.50%	76.00%
AL-Assisted T5	71.55%	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-assistede T5. The best performance is shown in **bold**, and the second best is shown in underline. Our ensemble system out-performs both models in all datasets, demonstrating the capability of our switch module. The accuracy of the switch module is shown in the bottom line.

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 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.

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 to outperform smaller models 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 ALERTS, a LLM+AL ensemble system. Results show that our ensemble system can identify the best performing model and consistently yield accurate prediction during cold starts and data drifts.

8 Limitation

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 SOTA generic LLMs (GPT-3.5 and GPT-4). We are aware other LLMs exist that we do not include in this work, such as Mistral-7B (Jiang et al., 2023), Llama-2 and 3 (Touvron et al., 2023b), etc. We only implemented and evaluated two fundamental types (data diversity-based and uncertainty-based) of Active Learning strategies in our work, and we are aware there exist other families of AL strategies that could extend our study, e.g., hybrid or ensemble approaches (Krogh and Vedelsby, 1994; Qian et al., 2020). 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.

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.

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.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. **Preprint**, arxiv:2210.11416.

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

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. **ArXiv**.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. **arXiv preprint arXiv:2310.06825**.

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.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. **Cureus**.

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**.

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.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. **Journal of Machine Learning Research**, 24(253):1–15.

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. 2023. 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.

Richard Plant, Valerio Giuffrida, and Dimitra Gkatzia. 2022. You Are What You Write: Preserving Privacy in the Era of Large Language Models. **Preprint**, arxiv:2204.09391.

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.

Christoffer Bøgelund Rasmussen, Kristian Kirk, and Thomas B. Moeslund. 2022. The Challenge of Data Annotation in Deep Learning—A Case Study on Whole Plant Corn Silage. **Sensors**, 22(4):1596.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. **ACM computing surveys (CSUR)**, 54(9):1–40.

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.

Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active Learning with Rationales for Text Classification. In **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pages 441–451, Denver, Colorado. Association for Computational Linguistics.

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.

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, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. **Preprint**, arxiv:2201.11903.

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.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension. **Preprint**, arxiv:2203.13947.

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. **Preprint**, arxiv:2305.12710.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby 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.

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

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. 2024. RAFT: Adapting Language Model to Domain Specific RAG. **Preprint**, arxiv:2403.10131.

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**, ICAIL '21, pages 159–168, New York, NY, USA. Association for Computing Machinery.

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.

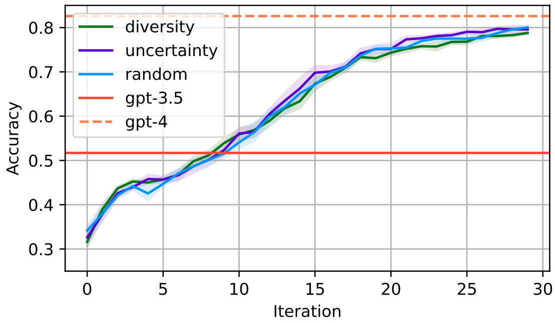


Figure 5: Result on BioMRC

B Hyperparameters and Settings

Dataset	Learning Rate	Training Epoch
BioMRC	1e-4	20
Unfair_TOS	1.5e-4	12
ContactNLI	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>:
```