Human Experts Still Needed: Active Learning-Assisted Humans vs. Large Language Models (LLMs) on Domain-Specific Data Annotation Tasks

Anonymous ACL submission

Abstract

A considerable amount of effort has been delved into developing low-resource learning techniques, such as Active Learning (AL), to reduce human annotation costs. Despite Large Language Models (LLMs) demonstrating exceptional performance on benchmarking datasets, sometimes even exceeding human performance, **whether LLMs can substitute human experts in domain-specific data annotation tasks** has been a debatable question of significant importance. In this work, we conduct an empirical study on **five** expert-annotated datasets from **two** specialized domains (legal and bio-medical) to investigate the performance of generic LLMs versus a much smaller T5-base supported by different AL strategies. Although generic LLMs (i.e., GPT-3.5 and GPT-4) are hundreds of times larger, the AL-assisted T5-base can consistently outperform GPT-3.5 and is compatible with GPT-4 with only a few hundred expert annotations. Our study validates the irreplaceability of human annotations in real-world domain-specific scenarios, and we propose a future hybrid paradigm that leverages LLMs to "warm-up" AL-assisted models.

1 Introduction and Background

Human experts are usually very difficult to recruit for annotating large-scale and high-quality datasets, especially in many domains that require extensive domain expertise (e.g., legal, clinical, and education) (Xu et al., 2022; Pappas et al., 2020), due to the increasingly expensive annotation cost (e.g., expert resources, time, money, etc.) (Wu et al., 2022). To bridge the gap between the scarcity of high-quality data and the data demand for model training, the research communities have widely explored low-resource learning techniques, such as Active Learning (AL) (Settles, 2009).

AL-assisted Human in Data Annotation The expensive nature of expert-generated annotations has sparked interest in methods that can effectively learn from a constrained number of labeled samples. 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. AL has been shown to be effective in various low-resource scenarios (Sharma et al., 2015; Yao et al., 2023). 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 categories of selection concepts: data diversity-based strategies and model uncertainty-based strategies. We experiment with both types of AL strategies in our work.

LLM in Data Annotation Recently, LLMs (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b) have shown great capability in learning from the context and generating high-quality content 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), have come into the picture to harness the potential of LLMs. Chain-of-Thoughts encourages models to produce a sequence of rationales, while ICL teaches the LLMs based on a handful of insightful examples within the input.

In addition, several recent works claim that LLMs can outperform human annotators for text classification (Gilardi et al., 2023), task evaluations (Chiang and Lee, 2023; Liu et al., 2023), and even in specialized domains (Nori et al., 2023), etc (Törnberg, 2023). **Can LLMs substitute human experts in domain-specific tasks?**

We hypothesize that AL-assisted small language

Dataset	Domain	Task	# Test Data
BioMRC Pappas et al. (2020)	Biomed.	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.

models with a small number of expert annotations can outperform generic LLMs in domain-specific tasks. This work presents an empirical study with AL simulations on five datasets (CUAD, BioMRC, ContractNLI, Unfair_TOS, Casehold) from two real-world specialized domains (Biomedicine and Legal). We probe state-of-the-art (SOTA) LLMs (GPT-3.5 and GPT-4 (OpenAI, 2023)) with their best-performing prompting methodologies and compare them with an AL-assisted T5-based model (Raffel et al., 2020) leveraging two different AL strategies (i.e., data-based and model-based).

Our results show that AL-assisted T5-base with hundreds of human-annotated data can consistently outperform GPT-3.5 and perform compatible with GPT-4, justifying our hypothesis that human experts are irreplaceable in domain-specific data annotation tasks. We also propose a hybrid paradigm to leverage LLMs to "warm-up" AL models.

2 Empirical Study Design

2.1 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 of legal and biomedical domains and of different types of tasks, including Multiple Choice, Classification, Natural Language Inference (MacCartney and Manning, 2008), and Question Generation. We report dataset details in Table 1.

2.2 Experiment Setup

2.2.1 Models

We utilize two SOTA generic LLMs, specifically GPT-3.5 and GPT-4 (OpenAI, 2023)[1], in our study.

[1]Version: GPT-3.5-0613 and GPT-4-0613

Algorithm 1 Active Learning Sampling Process

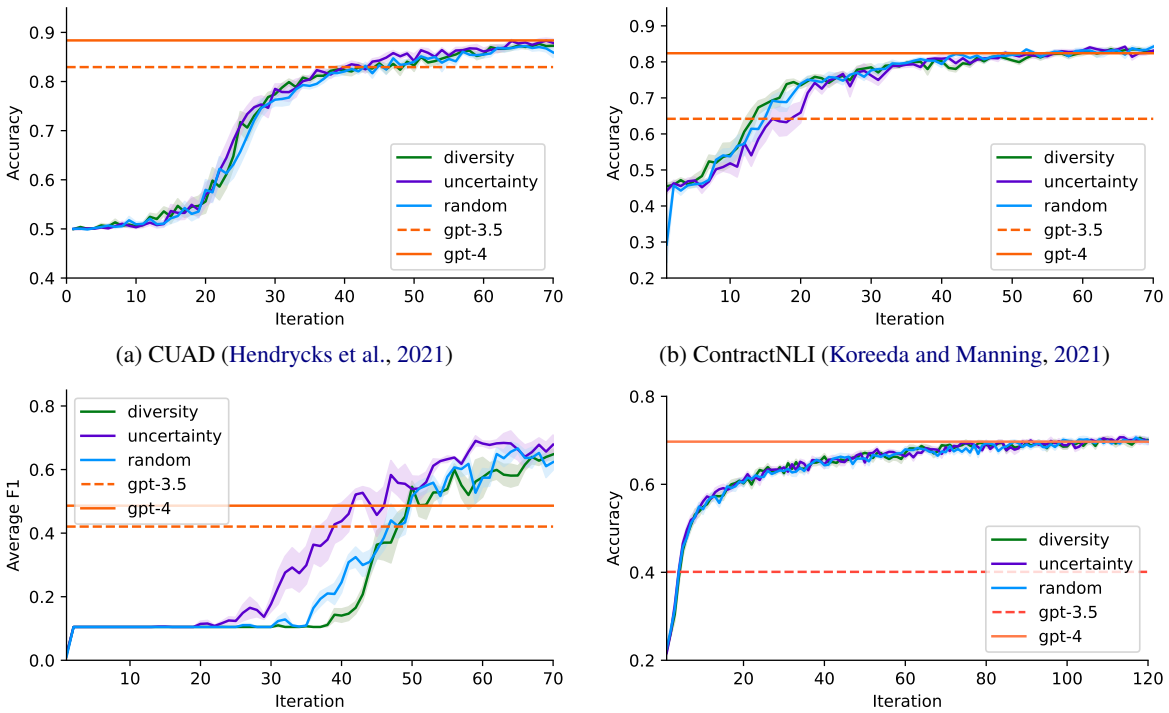
function SELECT($D_t, D_p, N, type$)
  D_t: unlabeled data in the training split
  D_p: previously selected data
  N: number of data needed
  $strategy$: Active Learning strategy
  if $strategy =$ `"similarity"` **then**
    $S \leftarrow \left(\frac{\sum_{d_p \in D_p} \cos(d_i, d_p)}{|D_p|} \right)_{1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

We probe the best-performing prompting strategy for each dataset with LLMs through extensive experiments on GPT-3.5 (reported in Appendix B) and apply the same settings for GPT-4. We choose T5-base (Raffel et al., 2020) as the backbone for AL because existing works demonstrate that T5 has strong performance for domain-specific fine-tuning (Yao et al., 2022; Mou et al., 2021).

2.2.2 AL-assisted Humans

Following the established taxonomies of AL strategies (Schröder and Niekler, 2020), we designed and implemented one **data similarity-based** and one **model uncertainty-based** strategy. The similarity-based approach aims to identify the most representative examples from the unlabeled data space while maximizing the diversity, regardless of the model, whereas the uncertainty-based approach attempts to locate examples that the model is least confident about. We illustrate the details of each strategy below and in Algorithm 1.

Data Diversity-based Strategy The objective of data diversity-based strategy (Schröder et al., 2022) is to identify the most representative and diverse data. In each iteration, we embed each unlabeled instance using SentenceTransformer (Wang et al., 2020) and then rank all the data in terms of the averaged cosine similarity with all previously selected examples. We then uniformly select (precisely, with the same step) a small batch of data from the



(a) CUAD (Hendrycks et al., 2021)

(b) ContractNLI (Koreeda and Manning, 2021)

(c) Unfair_TOS (Lippi et al., 2019)

(d) Casehold (Zheng et al., 2021)

Figure 1: Experiment results. The horizontal line represents two close-domain LLM's best performance. Plots report the mean value (line) and standard error (colored shaded area) over 10 trials. Each iteration comprises 16 examples. In the worst scenario (Casehold), humans only need 1,920 (16 examples $*$ 120 iterations) annotations to match or beat GPT-4 performance, which is critical for unseen tasks that have high domain specificity and data confidentiality constraints.

ranked list. This balanced approach ensures the model benefits from familiar and novel samples.

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 a randomly sampled subset of the training data, computing the model's 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. It is worth mentioning that the T5-base comprises only 220 million parameters, whereas GPT-4 is rumored to surpass trillion-level parameters. 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.

2.2.3 Evaluation Methods

We utilized the averaged F1 score for each label to evaluate Unfair_TOS to avoid the influence of unbalanced label distribution, which will also be discussed in Section 3.2 We evaluate the other datasets with average prediction accuracy. Detailed task instructions and experiment hyperparameters are shown in Appendix D and C.

3 Study Result

3.1 LLM vs. AL-assisted Humans

We plot the results on four legal domain datasets in Figure 1, and the results on BioMRC in Appendix A. The horizontal lines symbolize the best performance of GPT-3.5 and GPT-4, respectively. On all four datasets, the T5-base with AL can quickly **outperform GPT-3.5** and eventually reach a saturated performance that is compatible with or even exceeds GPT-4, leveraging a total of several hundred data selected. For BioMRC, as shown in Figure 2, 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, which results in excep-

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.

tional performance. Regardless, our fine-tuned T5-base achieves compatible performance with GPT-4 despite having hundreds of times fewer parameters and requiring significantly less computational power. It is worth mentioning that even in the worst scenario (Casehold), humans only need to annotate a total of $16 * 120$ iterations $= 1,920$ examples to match or beat GPT-4 performance, which is critical for unseen tasks that have high domain specificity and data confidentiality constraints.

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. We hypothesize that the significant uneven data distribution of Unfair_TOS might be the primary reason for the aforementioned observations. Thus, we conduct an ablation study on Unfair_TOS to investigate the correlation between data distribution and AL strategies.

3.2 Ablation Study of AL strategies on Unfair_TOS

The Unfair_TOS dataset consists of around 85% of data labeled *None*, and the rest of the data comprises eight other types of data. 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. We sum the counts of all eight "non-None" data types and denote them as *Not-None*. The ratio is calculated by dividing the corresponding data type by all data counts, which can be found in Table 2. We can observe the model uncertainty-based strategy selects significantly more *Not-None* labeled data than random ($t(14) = -2.46$, p-value < 0.05) and diversity ($t(14) = -2.51$, p-value < 0.05), which justifies the better performance of the uncertainty-based strategy and our hypothesis.

3.3 Discussion

We can observe all AL strategies suffer from well-known "cold-start" issues (Chen et al., 2022; Jin et al., 2022), where the model performs poorly at the early iterations due to potentially under-fitting issues due to too few data to be fine-tuned with. 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 envision a promising future paradigm in real-world domain-specific tasks of incorporating LLMs and AL fine-tuned smaller models in parallel. Specifically, the LLMs are utilized to overcome the "cold-start" problem at early iterations of AL, and the paradigm will incorporate a switch mechanism to determine when to switch from LLM to smaller AL models. The design of such a switch mechanism to efficiently and reliably evaluate the performance between LLMs and smaller AL models will be a crucial component of such a paradigm, which will be investigated in our future work.

4 Conclusion

While LLMs such as GPT-4 have been endorsed to outperform humans in many benchmarking datasets, whether they can substitute human experts, especially in real-world tasks and domains requiring extensive domain expertise, is of significant importance and debatable. In this work, we 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 domains specific tasks.

Our evaluation demonstrates that AL-assisted expert annotation can consistently and rapidly achieve or exceed best-performing LLMs with only a few hundred expert-annotated data, justifying that human experts remain indispensable in domain-specific tasks. Derived from our results, we posit a future paradigm that utilizes LLMs to overcome the "cold-start" issue of AL models as a "warm-up" strategy and eventually switch back to small models fine-tuned on domains-specific data once the latter outperforms LLMs.

5 Limitations

This work primarily presents an empirical study of generic LLMs versus AL-assisted small language models fine-tuned on experts-annotated domain-specific data. Our 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 (Touvron et al., 2023c), 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 evaluation comprises five datasets from two specialized real-world domains (legal and biomedical). We identify there are other domains and publically available domain-specific datasets, and we leave the analysis of the generalizability of our observations from this work to other domains and tasks as future work. 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. Therefore, human evaluation of these datasets 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*.

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.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

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

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.

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.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.

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.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

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

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

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*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

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.

Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.

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.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023c. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

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*.

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.

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.

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.

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.

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.

A Empirical Study Result on BioMRC

For BioMRC, as shown in Figure 2, 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.

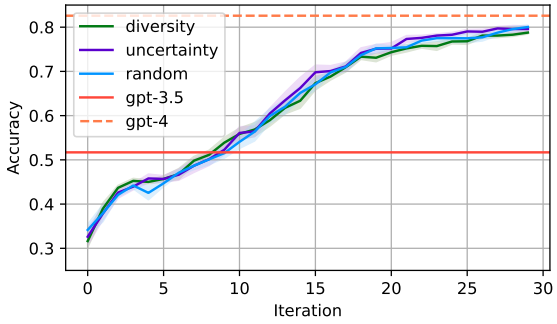


Figure 2: Result on BioMRC

B LLM Prompting Experiments

The LLM prompting experiments can be found in Table 3. To obtain the SOTA performance, we experiment with GPT-3.5 under zero-shot and few-shot (1,3, 10 shots) to find the best-performing setting (bolded) for each dataset and execute GPT-4 with the same settings.

C Hyperparameters and Settings

We report the experiment hyperparameters in Table 4. 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.

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

Table 4: Hyperparameters for each dataset.

D Prompts Used for Each Dataset

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

D.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>:
```

D.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
```

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

```
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>:
```

D.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>:
```

D.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>:
```

D.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>:
```