LLM Confidence Evaluation Measures in Zero-Shot CSS Classification

David Farr [1,2], Iain Cruickshank [3], Nico Manzonelli [2], Nicholas Clark [1], Kate Starbird [1], Jevin West[1]

[1]University of Washington, [2]Army Cyber Technology and Innovation Center, [3]Carnegie Mellon University
Correspondence: dtfarr@uw.edu

Abstract

Assessing classification confidence is critical for leveraging large language models (LLMs) in automated labeling tasks, especially in the sensitive domains presented by Computational Social Science (CSS) tasks. In this paper, we make three key contributions: (1) we propose an uncertainty quantification (UQ) performance measure tailored for data annotation tasks, (2) we compare, for the first time, five different UQ strategies across three distinct LLMs and CSS data annotation tasks, (3) we introduce a novel UQ aggregation strategy that effectively identifies low-confidence LLM annotations and disproportionately uncovers data incorrectly labeled by the LLMs. Our results demonstrate that our proposed UQ aggregation strategy improves upon existing methods and can be used to significantly improve human-in-the-loop data annotation processes.

1 Introduction

Large Language Models (LLMs) have transformed the way artificial intelligence is integrated into professional workflows, with applications that span healthcare (Ray, 2024), academia (Meyer et al., 2023), cybersecurity (Zhang et al., 2024), software development (Rasnayaka et al., 2024), and many others. However, research shows that users struggle to identify incorrect LLM responses which poses a problem because LLMs are less likely to refrain from answering questions they do not know as they scale with size and complexity (Zhou et al., 2024). Despite these challenges, LLMs have proven effective in synthesizing vast amounts of data and applying contextual understanding, making them a popular choice for integration into natural language processing tasks, particularly in zero-shot classification settings where prior training data is unavailable (Yang et al., 2024).

With broad applications in critical industries, LLM-generated responses that are assumed to be correct can lead to drastic second- and third-order consequences when answered incorrectly and integrated into decision-making processes. Although some LLMs incorporate expressions of uncertainty (Tian et al., 2023), developers often restrict the output of the model to a predetermined set of responses to manage nondeterministic behavior or reduce token generation cost (Liu et al., 2024b). However, these constraints can cause LLMs to provide confident answers even when they lack the correct knowledge. While LLMs are useful for large-scale data annotation tasks, there remains uncertainty as to which labels are correct or how to best quantify label confidence in LLM-generated annotations, especially in multi-modal systems.

This paper evaluates various Uncertainty Quantification (UQ) methods to assess LLM confidence in data annotation tasks applied to Computational Social Science (CSS) problems. Based on these results, we present a simple UQ aggregation strategy to help identify misclassified LLM-labeled data. We constrain our settings to realistic industry scenarios where previously labeled data is unavailable to simulate common, real-world problems. Additionally, we propose a new evaluation metric that assesses the recall of misclassified LLM-labeled data at low-confidences and compare UQ techniques using the Area Under Curve (AUC) analysis by applying thresholds based on percentiles of confidence scores. Our methodology has significant implications for systems that use human-machine teaming for data annotation tasks by better identifying data on which humans should spend finite resources.

2 Related Works

Zhou et al., 2024 show that as LLMs scale, they become more confident and less avoidant in answering questions. However, this increased confidence comes at a cost: they answer questions

incorrectly more frequently compared to smaller LLMs, which were more likely to avoid answering altogether. In a related study, Liu et al., 2024b demonstrate the importance of constraining LLM outputs in software development workflows to ensure predictability. Together, these works highlight both the internal challenge of larger LLMs being more prone to incorrect answers instead of avoidance, as well as the common practice of imposing constraints on LLM outputs to improve workflow predictability.

In the field of LLM UQ techniques, Liu et al., 2024a demonstrates an effective method of UQ via supervised calibration from utilizing hidden activation layers. Wang et al., 2024 integrate a human annotated training set to train an external BERT-based verifier to select data that the LLM was likely to mislabel for later external human annotation. However, these methods require a labeled dataset for training an external supervised ML model which is not available in many contexts.

As such, recent research has investigated zero-shot UQ techniques for LLMs. Kadavath et al., 2022 and Tian et al., 2023 show that an effective technique to assess confidence in LLMs tuned with reinforcement learning human feedback (RLHF) is prompting the model to evaluate its confidence in its own answer. Kumar et al., 2023 find that the uncertainty estimates from conformal prediction are closely correlated with the accuracy of the prediction.

Instead of relying on the LLM to self-report confidence, other approaches analyze model output. For example, Ling et al., 2024 show that the approximation of entropy using measurements on a restricted set of returned tokens is a valid mechanism to assess confidence in multiple-choice questions. Additionally, Farr et al., 2024 present an effective mechanism for identifying mislabeled data is using the absolute difference between the two highest log probability values returned. Finally, Kuhn et al., 2023 look for semantic differences in responses can inform uncertainty. Our work builds on the existing literature by comparing a sample of the aforementioned UQ mechanisms for zero-shot classification while proposing a new methodology that takes advantage of existing UQ techniques through an ensemble method.

3 Uncertainty Quantification Techniques

In this section we describe the five UQ techniques used in the study.

3.1 Quantitative and Qualitative Self-report

Our first and second UQ techniques are driven by the work of Tian et al., 2023, who show that RLHF tuned LLMs can self-assess answer confidence. We accomplish this by prompting the model to give a quantitative assessment of its confidence on a scale between 0 and 100. We also assess the ability of language models to map its uncertainty in qualitative terms. Our hypothesis being that open-source LLMs may perform better using normal language as opposed to probabilistic quantitative values. We accomplish this by asking models to report either *no*, *low*, *medium*, *high*, or *absolute* confidence in their responses. Then we map those responses to quantitative values of 0, 0.25, 0.50, 0.75, and 1 to allow comparability to other confidence measures. Access to all prompt examples and datasets can be found in the availability section.

3.2 Confidence Score

We use the confidence score method from Farr et al., 2024, where the authors define the confidence score as the absolute value of the difference between the highest token label log probability and the second-highest token label log probability within a constrained set of tokens. Let $\mathcal{T}$ represent the set of given tokens, and $P(t)$ denote the distribution of log probabilities across each token $t \in \mathcal{T}$. The log probability is then computed using the formula

$$C = \left| \max_{t \in \mathcal{T}} P(t) - \max_{t \in \mathcal{T} \setminus \{t^*\}} P(t) \right|, \qquad (1)$$

where t^* is the token corresponding to the highest probability $\max_{t \in \mathcal{T}} P(t)$. We refer to this metric as *C_score* in our results section.

3.3 Log Inverse

We additionally test a commonly used method to convert the logarithmic probability of the highest returned token into a probability. This allows us to investigate whether the difference between the confidence score (based on the top two token probabilities) and the direct probability of the highest token leads to significant differences in sampling outcomes. Specifically, let t^* represent the token with the highest probability, and let $\log P(t^*)$ denote the log probability of this token. The probabil-

ity for the token t^* is obtained by exponentiating the log probability.

For our results, we refer to this methodology as the *log inverse*.

3.4 Confidence Ensemble

Finally, we introduce the following UQ aggregation strategy, which is more resource intensive than the previous three, requiring the aforementioned confidence score from multiple LLMs, but is meant to reward LLMs for converging on a single label, while not penalizing a divergence of LLM-responses. This is especially important in classification tasks with multiple target classes.

Let $\mathcal{L}$ represent the set of LLMs, and for each LLM L_i, the token with the highest probability is denoted by $t^*_{L_i}$, and the corresponding confidence score C_{L_i} is given in Equation 1. To aggregate confidence scores when multiple LLMs provide the same answer t^*, the overall confidence score C_{agg} is calculated as

$$C_{\text{agg}} = \sum_{\{L_i \in \mathcal{L} \mid t^*_{L_i} = t^*\}} C_{L_i}, \tag{2}$$

where t^* is the common token predicted by the LLMs. For our results section, this methodology is referred to as *C_ensemble*.

4 Experimental Design

We evaluated our five UQ techniques across three different LLMs and three distinct CSS tasks. For each LLM and task, we rank all annotated data from least confident to most confident, allowing us to sample low-confidence data for human-in-the-loop labeling or high-confidence data for downstream classifiers. Each CSS task is pulled from common benchmark datasets for stance, ideology, and frame detection. For stance detection, we use the SemEval-2016 dataset (Mohammad et al., 2016). For ideology detection, we use the ideological books corpus (IBC) from (Sim et al., 2013) with sub-sentential annotations (Iyyer et al., 2014). For frame detection, we use the Gun Violence Frames Corpus (GVFC) from (Liu et al., 2019). The LLMs chosen were Llama-3.1 8B Instruct, Flan UL2, and GPT-4o. This selection was intentional to show a variety of parameter sizes and the integration of a RLHF-tuned model to show utility in sampling strategy mechanisms across different LLMs.

4.1 Evaluation Metric

In order to demonstrate the efficacy of each UQ strategy, we devise a metric that measures a confidence scoring techniques' ability to recall misclassified LLM-labeled data at low-confidences. When used to inform sampling for human-in-the-loop labeling, we would like to send a small sample of LLM labeled data to human data annotators for evaluation. Ideally, human evaluation is only applied to the data that the LLM is likely to misclassify, which boosts overall classification accuracy under the assumption that the human will correctly label data misclassified by the LLM. By selecting data based on the lowest percentile of confidence scores, we aim to select misclassfied examples for humans to evaluate. Therefore, we measure the percentage of falsely LLM-labeled data recalled as a function of the percentage of the total dataset evaluated based on the same the bottom percentile of confidence scores. This curve is depicted in Figure 1.

Our goal is to succinctly measure performance across UQ techniques, LLMs, and datasets. In order to accomplish this, we report the Area Under Curve (AUC) of the proportion of wrong examples to evaluated examples. As indicated in Figure 1, a higher AUC indicates a better UQ-informed, data sampling strategy. The AUC is calculated for the curve evaluating the dataset from none of the dataset to the full dataset. All graphs for evaluated models, sampling strategies, and datasets are shown in Appendix E. We also report the accuracy on each task for the LLM evaluated on the labeling tasks in Appendix A.

5 Results

Our results are shown in Table 1. Overall, the confidence ensemble uncertainty quantification measure is the most robust evaluated UQ strategy, proving to be effective across all model types. In the RLHF model evaluated, GPT-4o, quantitative self-reporting seemed also to be an effective strategy. Interestingly, for GPT-4o the log inverse performance did not closely resemble the confidence score or ensemble metrics. In the evaluated data, GPT appeared to return less deterministic responses, meaning that it was not as likely to achieve a high log inverse score when searching for a selected token, even when the model found it to be an easy task when evaluated using our other UQ techniques. On the contrary, the difficulty or ease of the tasks is

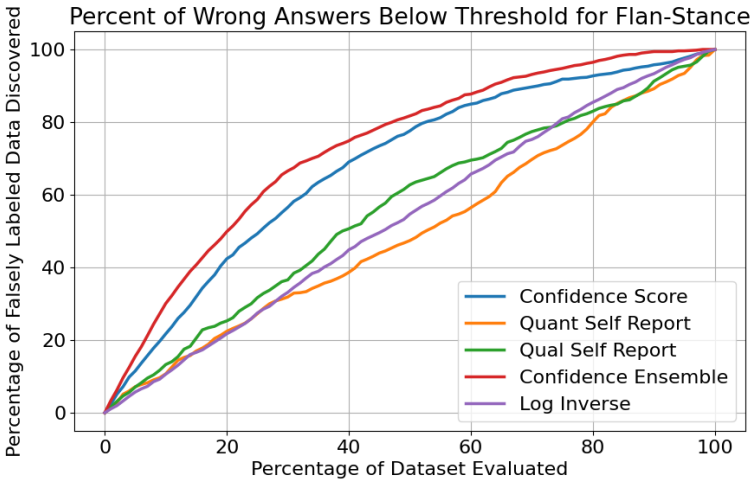


Figure 1: Graph depicts the percent of incorrect data annotations identified given the amount of data sampled for stance detection via Flan UL2. This shows we can find approximately half of all incorrect data annotations by checking only the bottom 20% of data evaluated by our confidence ensemble method. This graph also is meant to show a natural understanding of why AUC is a valuable measure for uncertainty quantification when measuring by percent of false labels detected.

	GPT			Flan			Llama			
UQ Metric	Stance	IBC	GVFC	Stance	IBC	GVFC	Stance	IBC	GVFC	**AVG**
Qual.	64.6	60.8	55.1	55.9	52.5	50.1	56.5	50.6	54.6	55.6
Quant.	66.3	64.6	59	49.8	50.6	46.25	51.6	51.7	55.3	55.0
Log Inverse	57.4	42.4	66.7	53.5	**60.8**	63.9	60.3	56.4	60.1	57.9
C_Score	67.1	63.3	**67.2**	68.1	60.3	62.7	66.5	**62.7**	59.6	64.2
C_Ensemble	**71.4**	**65.0**	66.6	**73.2**	54.3	**69.3**	**73.6**	58.4	**69.1**	**66.8**

Table 1: Depicts the Area Under Curve (AUC) metric across selected sampling strategy, dataset, and language model. The top performing sampling strategy for each task is in bold. We also report the average performance for each sampling strategy. Across all LLM types, the confidence ensemble method shows the most robustness.

highlighted in non-deterministic models with deterministic constraints by looking for the distribution between constrained tokens. For our non-RLHF models, Flan and Llama, our results indicate that self-assessment is a poor strategy; however, if underlying token log probabilities are not available, they seem to perform better when asked to qualitatively assess their confidence as opposed to answering with a numeric response. Like GPT-4o, the confidence ensemble appears to be the most robust metric, followed by the confidence score.

6 Conclusion

Through this work, we have evaluated several easy-to-implement UQ-based sampling strategies for finding erroneously LLM-labeled data in a zero-shot setting (i.e., common data annotation setting). We find that using confidence ensembles is the most effective mechanism for discovering erroneously labeled data. When only one LLM is being implemented, using the underlying distribution between the top two log probabilities is also an effective UQ mechanism. Using LLMs to label CSS data is a rapidly growing trend; however, it is important for humans to assess the quality of the labels generated. Our UQ strategies show that we can find a disproportionate amount of incorrectly annotated data, which should be evaluated by humans, by looking at small quantities driven by uncertainty quantification.

7 Availability and Resources

All code and data to produce these experiments can be found at https://anonymous.4open.science/r/UQMetrics-E69C. Two NVIDIA A6000 GPUs were used over the course of 18 hours for local LLMs. GPT was used to debug analysis graphs.

8 Limitations

We have only tested this methodology on three different datasets and three LLMs. Although it has seemingly extrapolated across the nine different testing combinations for the five separate sampling strategies, like all methodologies, it would benefit from testing across additional models and datasets for increased robustness. Furthermore, while we tested against different tasks, they were all broadly in the CSS space and against a constrained set of choices. For additional applications or labeling settings, more testing would need to be done. Finally, our most effective strategy required access to more than one LLM and underlying token log probability values, a combination that, while common, is not ubiquitous.

References

David Farr, Nico Manzonelli, Iain Cruickshank, and Jevin West. 2024. Red-ct: A systems design methodology for using llm-labeled data to train and deploy edge classifiers for computational social science. *Preprint*, arXiv:2408.08217.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 633–644.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *Preprint*, arXiv:2305.18404.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024a. Uncertainty estimation and quantification for llms: A simple supervised approach. *Preprint*, arXiv:2404.15993.

Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024b. " we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Sanka Rasnayaka, Guanlin Wang, Ridwan Shariffdeen, and Ganesh Neelakanta Iyer. 2024. An empirical study on usage and perceptions of llms in a software engineering project. In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 111–118.

Partha Pratim Ray. 2024. Timely need for navigating the potential and downsides of llms in healthcare and biomedicine. *Briefings in Bioinformatics*, 25(3):bbae214.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence

scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Jie Zhang, Haoyu Bu, Hui Wen, Yu Chen, Lun Li, and Hongsong Zhu. 2024. When llms meet cybersecurity: A systematic literature review. *arXiv preprint arXiv:2405.03644*.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, pages 1–8.

A LLM Annotation Accuracy

The table below shows the accuracy of LLM labeling for the three tasks given for each LLM evaluated.

	FLAN UL2	GPT-4o	Mistral 8b
Stance	75.6	77.4	72.4
IBC	62.3	62.5	65.2
GVFC	58.7	69.5	58.3

B LLM Parameters

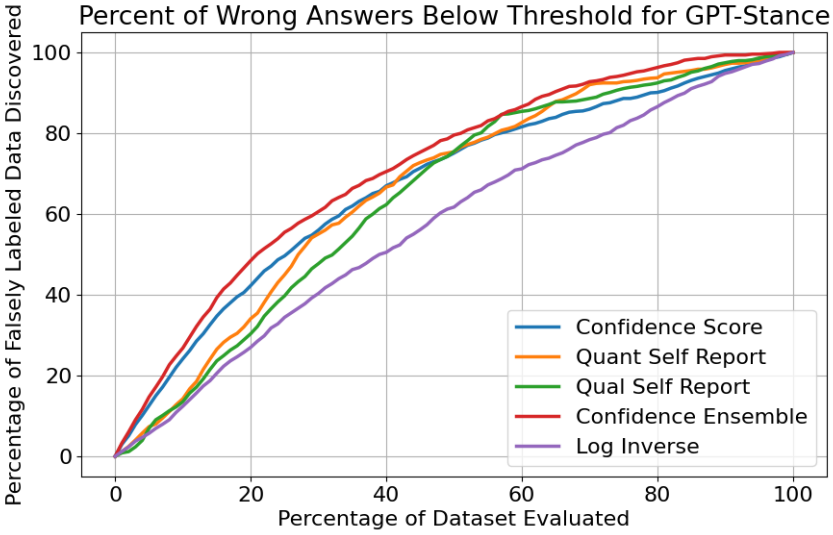
C Datasets Used

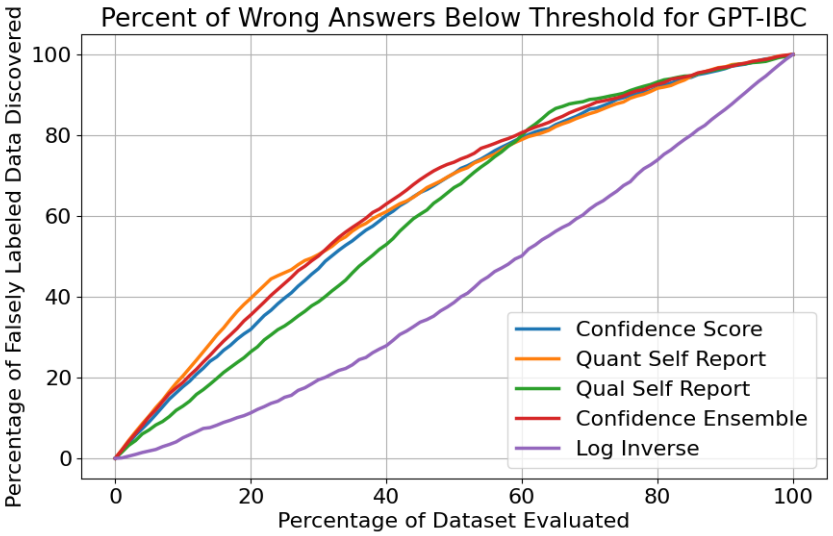
We used three main datasets for our experiments that are all available on our linked GitHUb in the availability section. Our first evaluated dataset was the SemEval-16 dataset

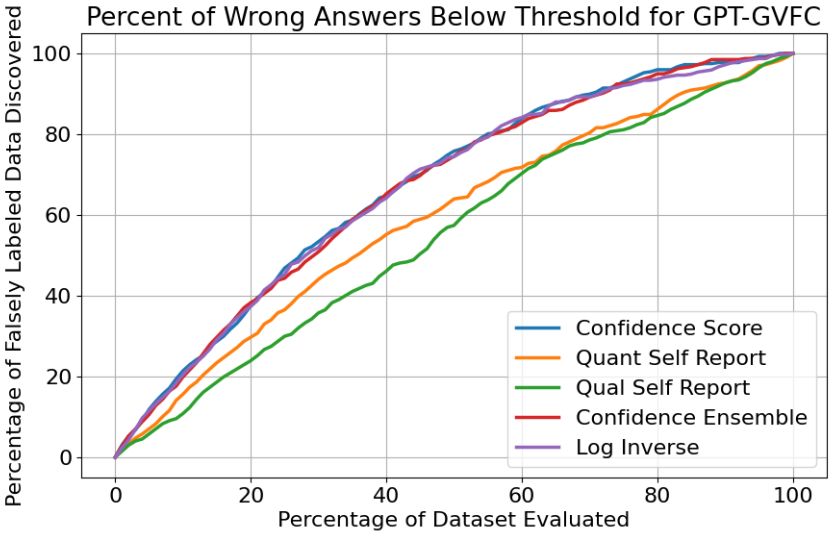
D Human in the Loop System

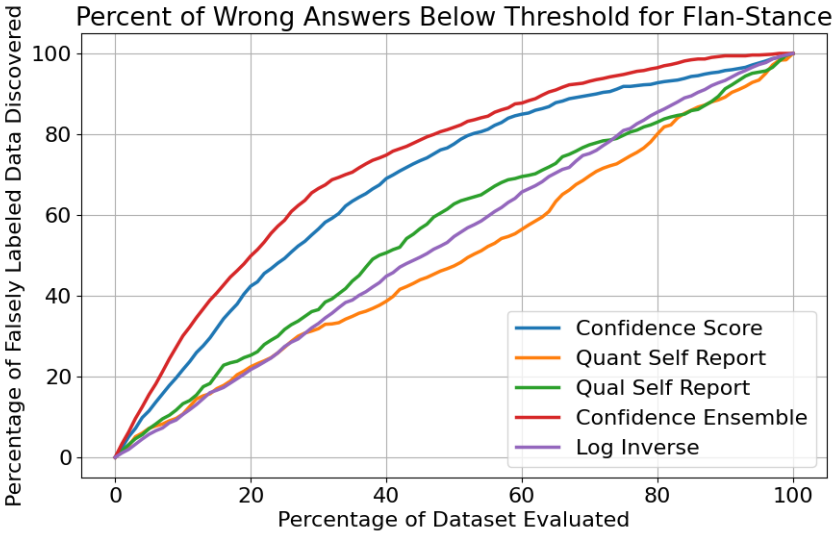
E Plots

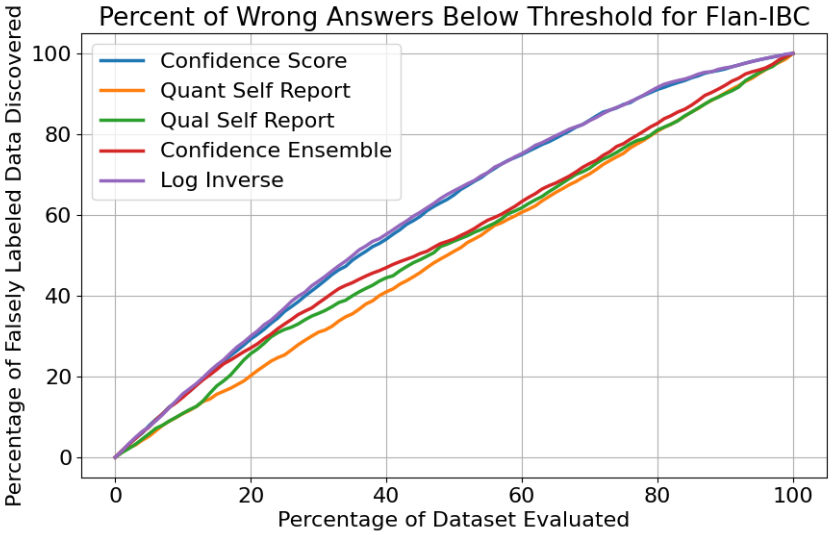
Below are all plots associated with reported AUC metrics.

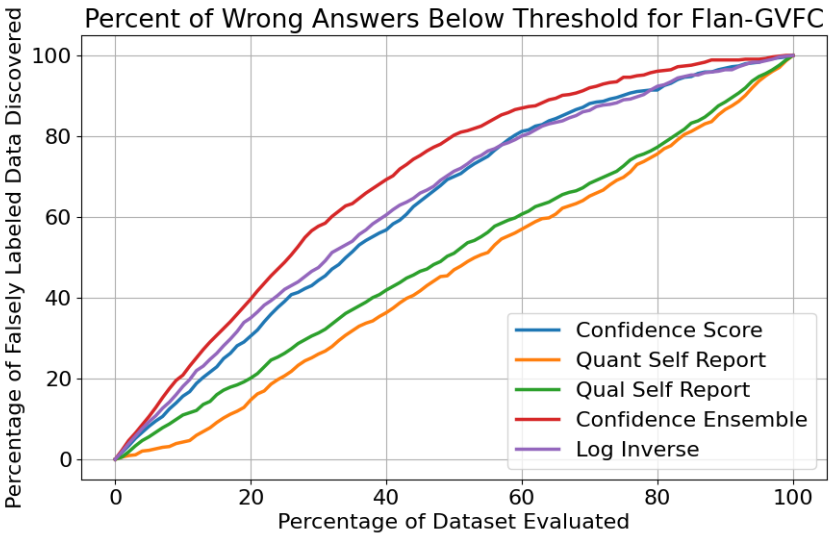

Percent of Wrong Answers Below Threshold for GPT-Stance
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
0
20
40
60
80
100
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse

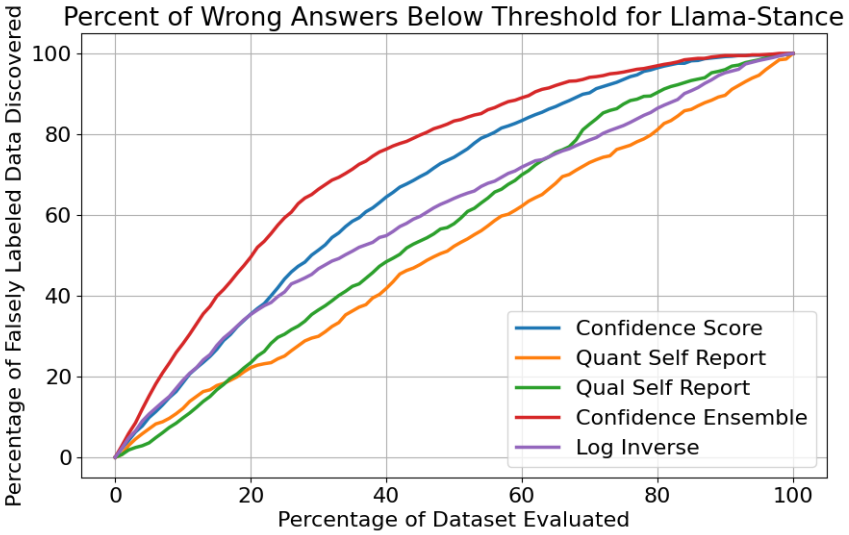

Percent of Wrong Answers Below Threshold for GPT-IBC
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
0
20
40
60
80
100
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse

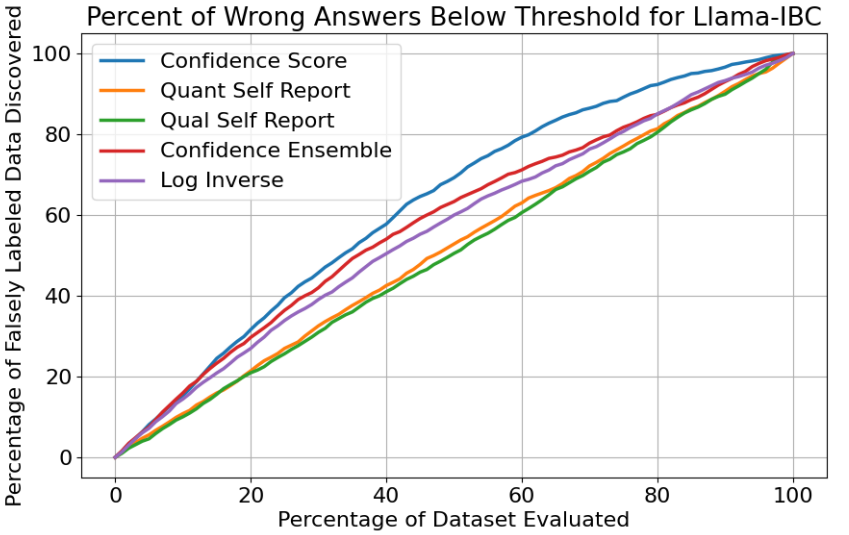

Percent of Wrong Answers Below Threshold for GPT-GVFC
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
0
20
40
60
80
100
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse

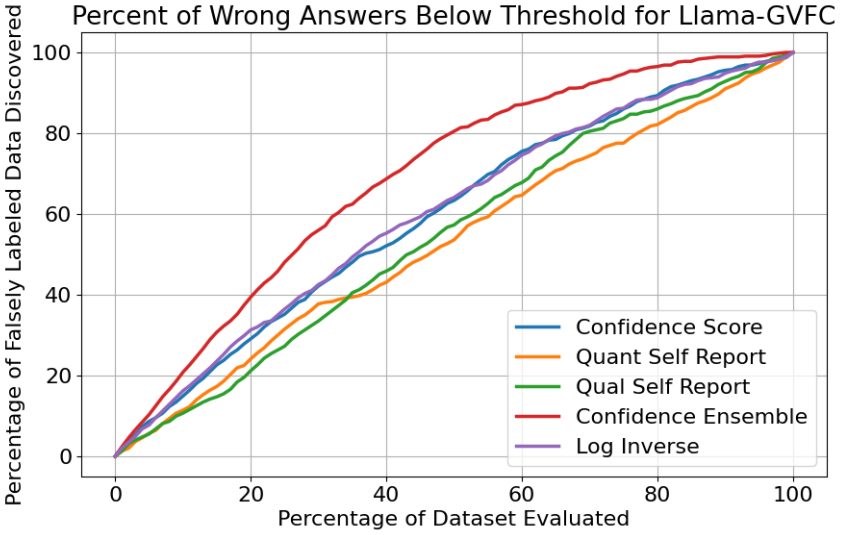

Percent of Wrong Answers Below Threshold for Flan-Stance
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse
0
20
40
60
80
100


Percent of Wrong Answers Below Threshold for Flan-IBC
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse
0
20
40
60
80
100


Percent of Wrong Answers Below Threshold for Flan-GVFC
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse
0
20
40
60
80
100


Percent of Wrong Answers Below Threshold for Llama-Stance
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse
0
20
40
60
80
100


Percent of Wrong Answers Below Threshold for Llama-IBC
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse
0
20
40
60
80
100


Percent of Wrong Answers Below Threshold for Llama-GVFC
Percentage of Falsely Labeled Data Discovered
Percentage of Dataset Evaluated
Confidence Score
Quant Self Report
Qual Self Report
Confidence Ensemble
Log Inverse
0
20
40
60
80
100