Confidence Evaluation Measures for Zero Shot LLM Classification

Anonymous ACL submission

Abstract

Assessing classification confidence is critical for leveraging Large Language Models (LLMs) in automated annotation tasks, especially in the sensitive domains presented by Natural Language Processing (NLP). In this paper, we apply five different Uncertainty Quantification (UQ) strategies for three NLP tasks: stance detection, ideology identification and frame detection. We use three different LLMs to perform the classification tasks. To improve the UQ performance, we propose an ensemble-based UQ aggregation strategy. 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 (Barzyk 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 are 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.

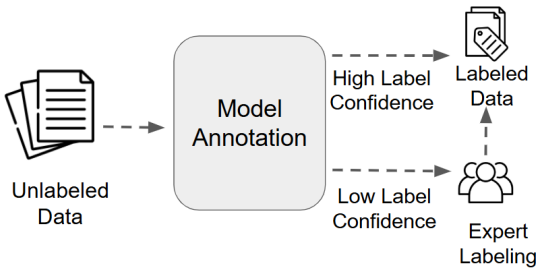

Figure 1: Graph depicts the human-in-the-loop proposed annotation methodology of routing based in confidence evaluation of individual data points.

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 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.

1.1 Key Contributions

In this paper, we make four key contributions:

1. **Comprehensive Evaluation of UQ Methods**: We benchmark multiple uncertainty quantification approaches, including logit-based methods and ensemble techniques, to assess their effectiveness in identifying incorrect LLM annotations.

2. **Proposed UQ Aggregation Strategy**: We define an ensemble based method that effectively identifies low-confidence LLM annotations and disproportionately uncovers data incorrectly labeled by the LLMs compared to other common techniques.

3. **New Evaluation Metric**: We define a recall-based metric that quantifies the ability of UQ methods to detect misclassified LLM-labeled data at low-confidence thresholds.

4. **Realistic Industry Simulation**: We constrain our experiments to zero-shot settings, simulating practical industry scenarios where labeled data is unavailable.

All code and data to produce these experiments can be found at (link removed for anonymity).

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 and Kaur et al., 2024 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. Berenbeim et al., 2023 propose intuitive scoring using logit results that can be implemented via bayesian or frequentist frameworks. 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

We evaluated five UQ techniques in our 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 probability for the token t^* is obtained by exponentiating the log probability.

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

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

where t^* is the common token predicted by the LLMs. This methodology is referred to as *C_ensemble*.

4 Tasks and Datasets

We evaluate our UQ technique on three CSS tasks: stance detection, ideology identification and frame detection. These tasks are chosen as they are often used in Computational Social Science studies (Ziems et al., 2024). To provide easy comparability and use, we leveraged common benchmark datasets for these three tasks.

4.0.1 Stance Detection

Stance detection is defined as "a classification problem where the stance of the author of the text is sought in the form of a category label from this set: Favor, Against, Neither" (Ng and Carley, 2022). For stance detection, we used the SemEval-16 dataset provided by (Mohammad et al., 2016). The SemEval-16 dataset consists of approximately 5000 tweets in relation to one of five targets: Hilary Clinton, Legalization of Abortion, Feminism, Climate Change, and Atheism.

In our experiments, we treated each target class as its own dataset. That is, the texts that are labeled Favor was one dataset, the texts labeled Against was one dataset, and the texts labeled Neither was one dataset. This allowed optimal prompt convergence in our LLM models.

Baseline For a baseline accuracy score, we reference the original work (Mohammad et al., 2016). This work constructs stance classification models ranging from Support Vector Machines to neural network models. The best performing model resulted in a 56.3% accuracy score.

4.0.2 Ideological Identification

Ideology can be defined as "the shared framework of mental models that groups of individuals possess

that provide both an interpretation of the environment and a prescription as to how that environment should be structured"(North and Denzau, 1994). We evaluated our methodology on the Ideological Books Corpus (IBC) from (Sim et al., 2013) with sub-sentential annotations (Iyyer et al., 2014). The IBC dataset consists of 4326 sentences of which 1701 are labeled conservative, 600 neutral, and 2,025 liberal.

Baseline The baseline accuracy score references the original work, which constructs a hidden markov model, and results in a 86.2% accuracy (Sim et al., 2013).

4.1 Frame detection

Erving Goffman defines frames as mental schema that people use to interpret the world (Goffman, 1974). Frame detection is used to explore media bias of news headlines (Verma and Jaidka, 2024), analyze arguments of proponents and critiques of global issues like climate change (Hirsbrunner, 2024), and visualize the episodic rise and fall of news trends (Mavridis et al., 2024).

We use the Gun Violence Frames Corpus (GVFC) for our frame detection dataset (Liu et al., 2019). The GVFC consists of 1300 news headlines framing gun violence in one of nine target classes: 2nd Amendment, Gun control/regulation, Politics, Mental Health, School/Public space safety, Race/Ethnicity, Public opinion, Society/Culture, and Economic consequences.

For a baseline accuracy score, we reference the original work (Liu et al., 2019). This work classified texts into one of the nine frames using fine-tuned deep learning transformer-based models. The best performing model has an 84.2% accuracy.

5 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 annotation 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 identification, 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 (**?**).

5.1 LLM Parameters

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. We used a temperature of zero for all evaluated language models, allowing for more predictable outputs. For GPT-4o, we used a logit bias of 10 for each associated token in our constrained set of labels. We used a zero-shot prompting strategy for all experiments. All prompts are shown in the Appendix.

5.2 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 annotation, 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 misclassified 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. Figure 2 shows the curves.

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 2, 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 Figure 3. We also report the accuracy on each task for the LLM evaluated on the annotation tasks.

	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: Area Under Curve (AUC) across selected sampling strategy, dataset, and language model. The top performing sampling strategy for each task is **bolded**. Across all LLM types, the confidence ensemble method shows the most robustness. **Qual.**: Qualitative self-report metric, **Quant.**: Quantitative self-report metric, **Log Inverse**: Log Inverse metric, **C_Score**: Confidence Score metric, **C_Ensemble**: Confidence Ensemble metric.

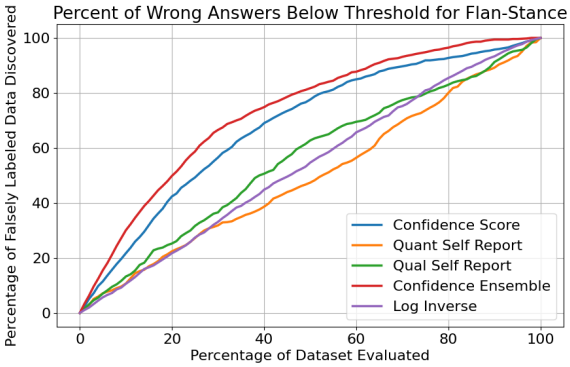

Figure 2: Percent of incorrect data annotations identified given the amount of data sampled for stance detection via Flan UL2. Half of all incorrect data annotations can be found by checking only the bottom 20% of data evaluated by our confidence ensemble method.

6 Results

The results of the Uncertainty Quantification metrics 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 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.

6.1 Area Under Curve threshold analysis

We calculated the AUC for each model and dataset. Figure 3 shows the graphs associated with the AUC threshold analysis, in correspondence to our UQ metrics. For all parameters, the AUC increases as the percentage of dataset increases. This indicates that larger dataset sizes results in better model performance.

6.2 LLM Annotation Accuracy

The three LLMs were used to perform stance, ideology and frame detection tasks. Table 2 shows the accuracy of LLM annotation for the three tasks. The LLMs do not necessarily perform better than baseline models. This observation is consistent with past studies that use LLMs for such tasks (Cruickshank and Ng, 2024; Zhao et al., 2024). These tasks are rather nuanced and require contextual, domain specific understanding.

	FLAN UL2	**GPT-4o**	**Mistral 8b**	**Baseline**
Stance	75.6	**77.4**	72.4	56.3
IBC	62.3	62.5	**65.2**	86.2
GVFC	58.7	**69.5**	58.3	84.2

Table 2: Accuracy of the LLM annotation on the tasks, in terms of percentage of correctly labeled data.

7 Discussion

In this work, we introduced UQ strategies to identify and correct erroneous LLM-generated annotations in CSS classification tasks. Our findings indicate that UQ techniques can significantly enhance the reliability of LLM-labeled data by enabling selective human intervention on low-confidence instances.

We find that our confidence ensemble is the most robust UQ strategy among the five UQ techniques evaluated, which consistently demonstrated superior performance across all LLMs and datasets, achieving the highest AUC scores. This suggests that leveraging multiple confidence scores from different models improves our ability to detect misclassified data. Unlike single-model approaches, ensemble-based uncertainty aggregation provides a more stable confidence estimate, reducing the risk of over-relying on a single LLM's confidence measure.

LLM self-assessment is effective but model-dependent. Our results indicate that RLHF-tuned models, such as GPT-4o, can provide reasonably effective self-reported confidence scores, particularly in the Quantitative Self-Report setting. However, open-source models, like Llama-3.1 8B Instruct and Flan UL2, performed worse when asked to assess their confidence, with qualitative self-reporting outperforming numeric responses. This aligns with prior research suggesting that models fine-tuned with human feedback better calibrate their internal uncertainty estimations.

Token probability-based methods are effective for non-RLHF Models and RLHF models. The Confidence Score (C_Score), which measures the absolute difference between the two highest token log probabilities, emerged as a strong performer across all tasks. This metric outperformed simple Log Inverse probability estimates, particularly in deterministic model settings. These findings suggest that token probability-based confidence estimation remains a strong choice when limited to using a single model.

While LLMs such as GPT-4o achieved competitive performance in stance detection, their performance in ideology classification (IBC) and frame detection (GVFC) remained below the baseline accuracy of fine-tuned supervised models. This underscores the continued difficulty of zero-shot CSS classification, where context, domain-specific knowledge, and subtle language variations present unique challenges. The results further emphasize the need for hybrid human-machine approaches where LLM-generated labels are carefully validated and refined.

Our findings also suggest several practical implications for organizations seeking to integrate LLMs into data annotation workflows. By leveraging UQ-informed sampling strategies, human annotators can focus on correcting low-confidence, high-risk LLM-labeled examples, thereby reducing overall annotation workload. Improving Model Deployment in Zero-Shot Environments: Many real-world applications lack pre-existing labeled datasets, making LLMs an attractive option for initial data annotation. Our results indicate that confidence-aware selection of human-labeled data can improve overall label quality in such settings.

Organizations using proprietary RLHF-tuned models (e.g., OpenAI, Anthropic) may benefit from self-reported confidence scores, whereas those relying on open-source models (e.g., LLaMA, Mistral) should consider token probability-based UQ techniques.

Like all methodologies, our work would benefit from testing across additional models and datasets for increased robustness. Our study focuses on comparing uncertainty quantification methods in a zero-shot setting, and opens up future directions of post-hoc calibration techniques, an important consideration for real-world deployment to ensure that confidence scores align with empirical accuracies.

8 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.

References

Charles Barzyk, Joseph Hickson, Jerik Ochoa, Jasmine Talley, Mikal Willeke, Sean Coffey, John Pavlik, and Nathaniel D Bastian. 2024. A generative artificial intelligence methodology for automated zero-shot data tagging to support tactical zero trust architecture implementation.

Alexander M Berenbeim, Iain J Cruickshank, Susmit Jha, Robert H Thomson, and Nathaniel D Bastian. 2023. Measuring classification decision certainty and doubt. *arXiv preprint arXiv:2303.14568*.

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2024. Prompting and fine-tuning open-sourced large language models for stance classification. *Preprint*, arXiv:2309.13734.

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.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience*. Frame analysis: An essay on the organization of experience. Harvard University Press, Cambridge, MA, US. Pages: ix, 586.

Simon David Hirsbrunner. 2024. Computational methods for climate change frame analysis: Techniques, critiques, and cautious ways forward. *Wiley Interdisciplinary Reviews: Climate Change*, 15(5):e902.

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.

Ramneet Kaur, Colin Samplawski, Adam D Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander M Berenbeim, John A Pavlik, Nathaniel D Bastian, et al. 2024. Addressing uncertainty in llms to enhance reliability in generative ai. *arXiv preprint arXiv:2411.02381*.

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)*.

Panagiotis Mavridis, Oana Inel, Xander Wilcke, Mykola Makhortykh, Markus de Jong, Honorata Mazepus, Antoaneta Dimitrova, Jesse de Vos, Alessandro Bozzon, and Tobias Kuhn. 2024. Framing is mightier than the sword: Detection of episodic and thematic framing in news media. *Human Computation*, 11(1):1–28.

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.

Lynnette Hui Xian Ng and Kathleen M Carley. 2022. Is my stance the same as your stance? a cross validation study of stance detection datasets. *Information Processing & Management*, 59(6):103070.

Douglass North and Arthur Denzau. 1994. Shared mental models: Ideologies and institutions. *Kyklos*, 47:3–31.

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*.

Preetika Verma and Kokil Jaidka. 2024. Mediaspin: Exploring media bias through fine-grained analysis of news headlines. *arXiv preprint arXiv:2412.02271*.

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).

Jin Zhao, Jingxuan Tu, Han Du, and Nianwen Xue. 2024. Media attitude detection via framing analysis with events and their relations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17197–17210.

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.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Prompts for Large Language Models

A.1 Stance Detection Prompt

Statement: STATEMENT
Which of the following best describes the above social media statements' stance regarding the idea of TARGET?
A: for
B: against
C: neutral
Only respond with 'for', 'against', or 'neutral'

A.2 Ideology Detection Prompt

Statement: STATEMENT
Does the above statement lean 'Liberal', 'Neutral', or 'Conservative'? Answer with only one word.

A.3 Frame Detection Prompt

Framing is selecting aspects of data to communicate a point in text. Here is a framed statement about gun violence: STATEMENT
Which of the following frames best fits:
A: gun rights-the story focuses on the constitution or second amendment
B: gun regulation-story is about issues related to regulating guns through legislation and other institutional measures
C: politics-story is mainly about the political issues around guns and shootings
D: mental health-story is about issues related to individuals' mental illnesses or emotional well-being
E: public safety-issues related to institutional and school safety
F: race-story is about gun issues related to certain ethnic group(s)
G: public opinion-story is about the public's, including a certain community's reactions to gun-related issues
H: culture-Store is about societal-wide factors that are related to gun violence
I: economic-story is about financial losses or gains, or the costs involved in gun-related issues.
Constrain your answer to the single response of 'A','B', 'C', 'D', 'E', 'F', 'G', 'H', or 'I' that best fits and nothing else.

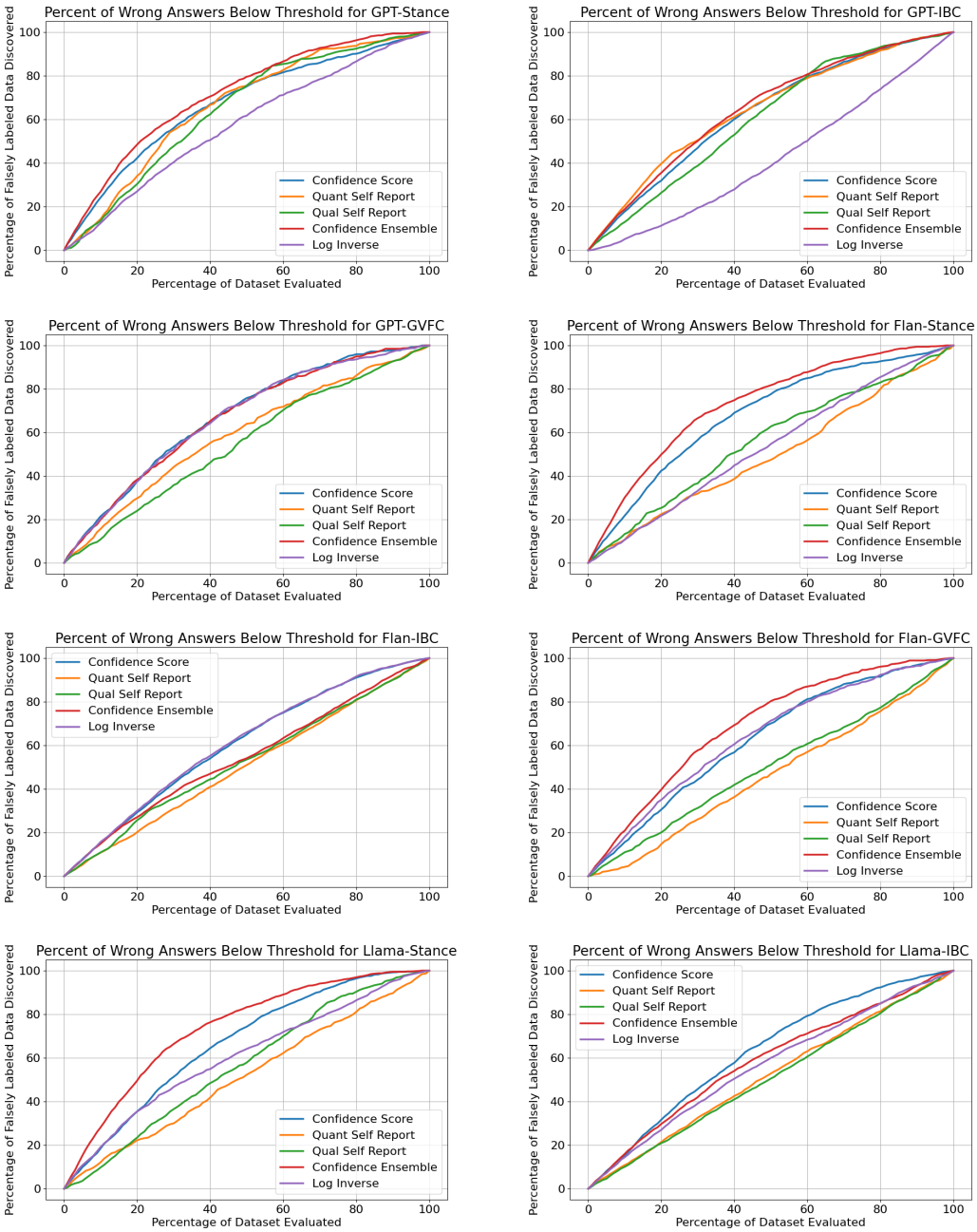

Figure 3: Area Under Curve threshold analysis