RED-CT: A Systems Design Methodology for Using LLM-labeled Data to Train and Deploy Edge Linguistic Classifiers

David Farr [1], Nico Manzonelli [2], Iain Cruickshank [3], Jevin West[1]

[1]University of Washington, [2]Army Cyber Technology and Innovation Center, [3]Carnegie Mellon University

Correspondence: dtfarr@uw.edu

Abstract

Large language models (LLMs) have enhanced our ability to rapidly analyze and classify unstructured natural language data. However, concerns regarding cost, network limitations, and security constraints have posed challenges for their integration into industry processes. In this study, we adopt a systems design approach to employing LLMs as imperfect data annotators for downstream supervised learning tasks, introducing system intervention measures aimed at improving classification performance. Our methodology outperforms LLM-generated labels in six of eight tests and base classifiers in all tests, demonstrating an effective strategy for incorporating LLMs into the design and deployment of specialized, supervised learning models present in many industry use cases.

1 Introduction

Large Language Models (LLMs) have significantly improved the ability to rapidly evaluate large amounts of unstructured natural language data. Despite their promise, many organizations face internal obstacles integrating LLMs into production environments. Developing LLMs internally is resource, expertise, and time intensive. Likewise, relying on APIs to access external LLMs introduces other issues. For instance, many organizations often have cost constraints, data privacy concerns, air-gapped networks, or decision cycle times that make integrating commercially available APIs infeasible.

Prior work shows that LLMs can perform well across a variety of NLP tasks for computational social science (CSS) via zero-shot prompting (Ziems et al., 2024). Traditionally, these tasks, like emotion, stance, persuasion, and misinformation classification, are solved with classification via supervised learning techniques. Although using supervised models solves many issues associated with deploying LLMs in production environments, they are known to perform poorly on out-of-domain data and require a significant upfront investment in data labeling.

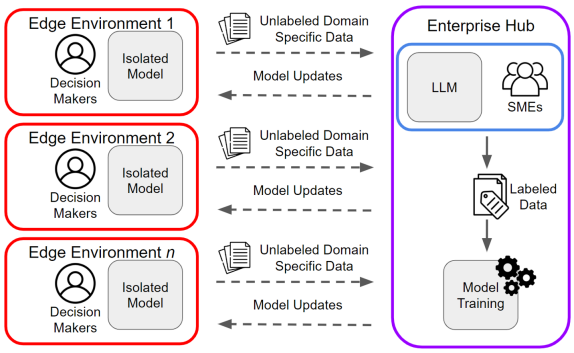

Figure 1: RED-CT design which allows LLM-like capabilities for NLP tasks deployed in edge environments.

To balance the flexibility associated with LLMs and advantages of supervised models, we propose Rapid Edge Deployment for CSS Tasks (RED-CT). RED-CT is a system that integrates traditional techniques from active learning such as confidence measurements and soft labels to pair LLM generated data labels with minimal selected human-annotated labels to deploy classifiers to edge environments fast. We define the *edge environment* as time- and / or resource-limited situations where users need to interface with NLP solutions. Additionally, the edge environment may be disconnected from the internet for security or privacy purposes or in crisis response settings where connection to the internet is either impracticable or unreliable.

In this paper, we introduce RED-CT and propose a confidence-informed sampling method to select LLM-labeled data for human annotation. In addition, we present a simple method to generate soft labels from LLM predictions to use during edge classifier training. We evaluate RED-CT with confidence-informed sampling and learning on soft labels across four CSS tasks: stance detection, misinformation identification, humor detection, and

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 58–67
January 19–24, 2025.

ideology detection. We further evaluate the proposed approach with two different common data-labeling prompting schemes and across three different sizes of distilled models. Our results show that it is possible to approximate or outperform LLMs on CSS tasks with minimal human data labeling (10% of dataset) in the distillation of edge models.

2 Related Works

One of the chief issues in creating ML solutions for CSS tasks is generalizing to out-of-domain data. CSS tasks, such as stance detection or sarcasm classification, often have very nuanced, context-dependent language (Ng and Carley, 2022; Ziems et al., 2024; Cruickshank and Ng, 2024). Due to high contextually-dependency, supervised approaches produce models that struggle to generalize between datasets. For example, previous research indicates that while model generalizability can be improved through the aggregation of datasets, cross-dataset stance detection models still generalize poorly (Ng and Carley, 2022).

Recent work has demonstrated that LLMs can perform well across various classification tasks within CSS (Cruickshank and Ng, 2024; Zhu et al., 2023). Ziems et al. (2024) provides best practices for prompting and benchmarks performance for a variety of CSS tasks across several LLMs. LLM-based classification methods work better with out-of-domain data due to the LLMs strong zero-shot classification capacity. However, these methods also require substantial resources and cannot scale, in terms of cost or compute time, to large CSS datasets. For example, just labeling the SemEval2016 dataset (Mohammad et al., 2016) (2,814 data points) with GPT-4 could cost over $30 USD. Additionally, ongoing research has found that LLMs still usually perform worse than in-domain supervised models at CSS tasks (Cruickshank and Ng, 2024; Ziems et al., 2024). (Tan et al., 2024) provide a survey paper of research using large language models for data annotation, including model distillation as a task. Some related works included show using synthetic generated data from larger LLMs to train smaller LLMs(Wang et al., 2023) and (Huang et al., 2022) which demonstrate LLMs can improve performance through self-annotation and subsequent fine-tuning based on self-annotated data. Further related work by (Wang et al., 2024) deploys an external verifier model to select samples LLMs are unlikely to classify correctly and routes them to human labelers for increased performance. Such previous work differs in data sampling methods, resource requirements, and distillation methodology.

In an effort to improve supervised model performance in other classification contexts, researchers have explored learning on soft labels. Soft labels employ a weighting mechanism to capture annotator uncertainty during labeling. Soft labels have been shown to enhance model generalization and better represent the confidence of the annotator (Alshahrani et al., 2021; Wu et al., 2023).

Researchers study LLM distillation techniques (Xu et al., 2024) to reduce model size and cost. These methods vary considerably in their use of LLMs. Some studies have focused on generating artificial data with LLMs useful for distilling small classification models (Ye et al., 2022b,a; Gao et al., 2022; Meng et al., 2023). Other works have explored few-shot prompting and active learning mechanisms, combined with LLMs for data labeling (Wang et al., 2021; Zhang et al., 2023; Hsieh et al., 2023). Many of these methods often require human intervention to filter low-quality data or LLM-generated rationales for labels which can be unreliable (Huang et al., 2023). Other works focus on reducing bias (Egami et al., 2023) without focusing on downstream classification performance (Wang et al., 2021). Pangakis and Wolken (2024) assess supervised classifiers performance on LLM generated labels, but do not offer a systems approach or intervention measures to improve downstream classification. None of the prior works attempt to integrate additional uncertainty information from LLMs into human intervention and model distillation.

3 Methodology

In this section, we outline our proposed methodology that contributes to the literature by presenting a systems approach that incorporates model uncertainty estimates. These estimates guide human intervention and improve model training for classifiers using LLM-labeled data.

3.1 RED-CT System Methodology

Rapid Edge Deployment for CSS Tasks (RED-CT) is designed with three tasks in mind: reducing latency for classification tasks, reducing the amount of data exposed to external API's, and decreasing the energy and monetary cost associated with

LLMs. By reducing LLM dependency, we can decrease energy expenditure, cost, and network dependency for CSS classification tasks. This also allows us to obfuscate batched data being sent to an LLM, opposed to needing to secure all data in a production environment. RED-CT is a system that enables users in edge environments to utilize ML tools for complex societal computing tasks. Figure 1 provides a high-level overview of our system.

RED-CT follows a framework in which classification and data collection are performed at the edge, model development is performed at a central point, and then model updates are pushed back to the edge. We refer to this framework of different, related contexts and devices as a data resupply framework. Transport mechanisms for data resupply include internet (when available) or physical devices transferred by personnel moving in and out of the edge environment.

Data delivered to the enterprise hub goes through a pipeline for labeling and model training. Unlike the edge environment, compute resources and connectivity are not restricted at the enterprise hub. This allows analysts at the hub to label the data via zero-shot LLM prediction for maximum expediency. Data label quality can be increased by integrating subject matter experts (SMEs) for prompt engineering, quality control, or expert labeling of small sample sizes. Edge classifiers are then trained or fine-tuned on the newly labeled data and deployed back to the edge environment.

RED-CT's modular design allows for increased performance as industry and academia continue to improve system components, such as LLMs, prompting techniques, and edge classifiers. Additionally, our method prevents model drift by enabling constant evaluation of data in a dynamic environment, with human-in-the-loop processes informing users.

3.2 Training Edge Classifiers on LLM-labeled Data

Due to the potential time-constrained setting in edge environments, RED-CT relies on fine-tuning BERT-based models on LLM-labeled data. BERT-based models require minimal text preprocessing, and their pretraining allows for fine-tuning on downstream tasks. BERT models exhibit strong performance when fine-tuned for a variety of classification tasks (Devlin et al., 2019).

Given that LLMs are prone to errors in the zero-shot prediction setting, we assume that our LLM labels will be imperfect. Naively fine-tuning BERT on the LLM-labeled data risks over fitting to noisy or incorrect labels. To improve edge model performance, we integrate several system interventions into the model fine-tuning process: including expert-labeled data into the training process, designing confidence scores to select samples for experts to label, and learning soft labels based on label weights.

3.2.1 Incorporating Confidence Informed Expert Labels

RED-CT helps streamline model deployment by reducing the number of personnel hours devoted to labeling data. Instead of using SMEs to label all available data, we only require them to label small samples of data. Integrating experts improves the quality of the LLM-labeled dataset and subsequently the edge classifier.

Randomly selecting samples for SME labeling within a bounded time or up to a certain percentage can improve edge model performance but may introduce inefficiencies where SME's analyze sample data in which the LLM is confident it has labeled correctly. To optimize sampling for SME analysis, we devise a confidence-based metric to identify examples where LLM labeling is less reliable.

The confidence score is defined as the absolute difference between the highest token label log probability and the second-highest token label log probability within this constrained set of expected tokens. Let $\mathcal{T}$ represent the set of given tokens, and $P(t)$ denote the distribution of probabilities across each token $t \in \mathcal{T}$. The confidence score, denoted as C, is then computed using the formula

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

where t^* is the token corresponding to the highest probability $\max_{t \in \mathcal{T}} P(t)$. To apply the confidence score, we stratify by each LLM-labeled class and sample the bottom p percentile.

To validate the proposed confidence estimate, we analyze the distribution of confidence scores for examples labeled correctly and incorrectly using the labels from zero-shot stance classification with gpt-3.5-turbo. Under the Kolmogorov-Smirnov test, we reject the null hypothesis that correctly and incorrectly labeled samples come from the same distribution of confidence scores (An, 1933). Highlighted in Figure 2, we are more likely to select

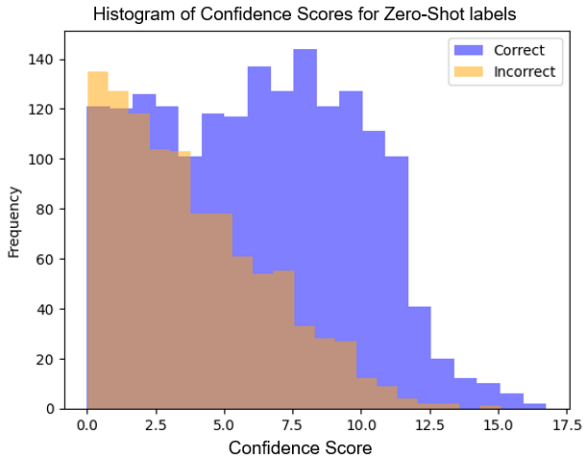

Figure 2: The distribution of confidence scores for examples labeled correctly and incorrectly using gpt-3.5-turbo zero-shot stance classification. The distributions are overlaid as opposed to stacked.

correctly labeled examples using random sampling, but less likely when sampling examples with very low confidence scores.

3.2.2 Learning on Soft Labels

Fine-tuning edge classifiers on the LLM-labeled data risks overfitting on incorrect labels. We help ease this problem by integrating SMEs into the labeling process; however, standard supervised training methods do not account for differences in label quality. To account for label confidence, we learn on *soft labels*.

To retrieve a soft label for model fine-tuning, we apply the expit function to the log probability of the token associated with each LLM label, resulting in a score between zero and one. For expert labeled examples, we assign a weight of 1 on the selected class and 0 for the others. Our experimental results show that learning with soft labels improves edge classification performance.

4 System Implementation and Experiment Design

We replicate the available LLM labeling capabilities with two models: OpenAI's gpt-3.5-turbo, available closed-source from their API, and Mistral's Mistral-7B-Instruct-v0.2, available open-source on Huggingface.[1] We experiment with two different prompting styles for labeling: zero-shot and zero-shot chain of thought (CoT). We attempted to use the best prompting practices in literature for our classification tasks, integrating prompting techniques from (Ziems et al., 2024), (Cruickshank and Ng, 2024), and (Zhu et al., 2023). Examples of each prompt are provided in Appendix A.

For edge classifiers, we test three flavors of BERT: 'Distil-BERT', 'RoBERTa', and 'RoBERTa-Large' (Devlin et al., 2019; Liu et al., 2019). These models vary in size, allowing us to assess performance across model compute requirements. For each BERT model, we evaluate the effects of system intervention measures. The system settings we tested included a base classifier trained with no system interventions directly on the LLM labels, a classifier trained on soft labels (SL), a classifier trained on 10 percent randomly selected expert labeled data (RS 10%), a classifier trained on confidence-informed sampling (CI 10%), and a classifier trained with all system intervention measures (CI SL 10%). We train five classifiers on each LLM-labeled dataset and report the averages across each. For each edge model, we do full fine-tuning (i.e., unfreeze all model weights) from pre-trained models, but note that this process can be done with any type of fine-tuning or training a model with initialized weights.

4.1 CSS Tasks and Data Selection

For the purposes of testing our systems methodology, we selected four well known CSS tasks: stance detection, misinformation detection, ideology detection, and humor detection. We then selected a dataset for each task that had known benchmarks to compare our system design against.

4.1.1 Stance Detection

We define stance detection as an "automatic classification of the stance of the producer of a piece of text, towards a target, into one of these three classes: Favor, Against, Neither" (Küçük and Can, 2021). We use 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. There are three classification classes for each target: favor, against, and neutral.

4.1.2 Misinformation

We define misinformation as "false or inaccurate information that is deliberately created and is intentionally or unintentionally propagated" (Wu et al.,

[1]Model resources and information are contained in the Ethics and Availability section.

Task	Enterprise LLM	Edge Classifier - RoBERTa-L			
	GPT-3.5 Turbo	Random	Base	RS 10%	CI SL 10%
Stance	.667	.333	.626	.665	**.689**
Misinformation	.761	.500	.653	.703	.752
Ideology	.579	.333	.567	.597	**.626**
Humor	.565	.500	.534	.555	**.571**
	Mistral-7B-Instruct				
Stance	.529	.333	.439	.448	.486
Misinformation	.602	.500	.594	.629	**.665**
Ideology	.406	.333	.413	.441	**.451**
Humor	.492	.500	.384	.427	**.508**

Table 1: Zero-Shot LLM performance (weighted f1 score) compared to edge model performance. Random are dummy models predicting on a uniformed distribution, base edge models are trained without system interventions, RS 10% edge models are trained with 10% randomly sampled expert examples, and CI SW 10% is 10% confidence-informed sampling and learning with label weights. Results that out-performed the enterprise LLM are bolded.

2019). We evaluate misinformation detection on the Misinfo Reaction Frames corpus (Gabriel et al., 2022). The Misinfo Reaction Frames corpus consists of 25k news headlines consisting of topics such as COVID-19, climate change, or cancer. Each headline was fact checked and has an associated binary misinfo classification of misinformation or trustworthy.

4.1.3 Ideology

We define ideology 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 used the Ideology Books Corpus (IBC) dataset from (Sim et al., 2013) with sub-sentential annotations (Iyyer et al., 2014) to evaluate our system's utility in ideology detection. The IBC dataset contains 1,701 conservative sentences, 600 neutral sentences, and 2,025 liberal sentences.

4.1.4 Humor

For humor detection, we used a broad definition when prompting LLMs with the question, "Would most people find this funny?" This approach focused on binary humor classification. We evaluated our system using a curated collection of posts from Reddit's r/Jokes, where researchers labeled jokes as humorous or not based on the number of upvotes. The two classes were distinguished through binary cluster analysis (Weller and Seppi, 2019).

5 Results

Table 1 presents the high-level results across our four chosen CSS tasks using RoBERTa-L. Figure 3 is a more detailed analysis of the implementation of our system methodology in the stance detection task, including varying the type of BERT model in all combinations of system intervention strategies. A key takeaway is that through our methodology and associated system intervention measures, we were able to outperform LLM-labeled data in 6 of the 8 tested tasks, while approximating it in an additional task. Additionally, in GPT labeled data, we had an average improvement of 6.75% over the base classifier and we out performed the base and normal sampling techniques in 100 percent of tasks. In Annex B, we have included additional results analysis, including varying the percentage of expert labels in Figure 4 and a full table for each stance detection result in Table 2.

5.1 Discussion

Our results represent a significant improvement in system design for using LLMs as imperfect annotators for downstream classification tasks. Our system intervention measures were effective in both GPT 3.5 and Mistral-7B, but more consistent in GPT 3.5. We theorize that this is because the token log probabilities returned from GPT 3.5 provided more value to our confidence score and weighting interventions due to better associated log probability values with correct classification. Furthermore,

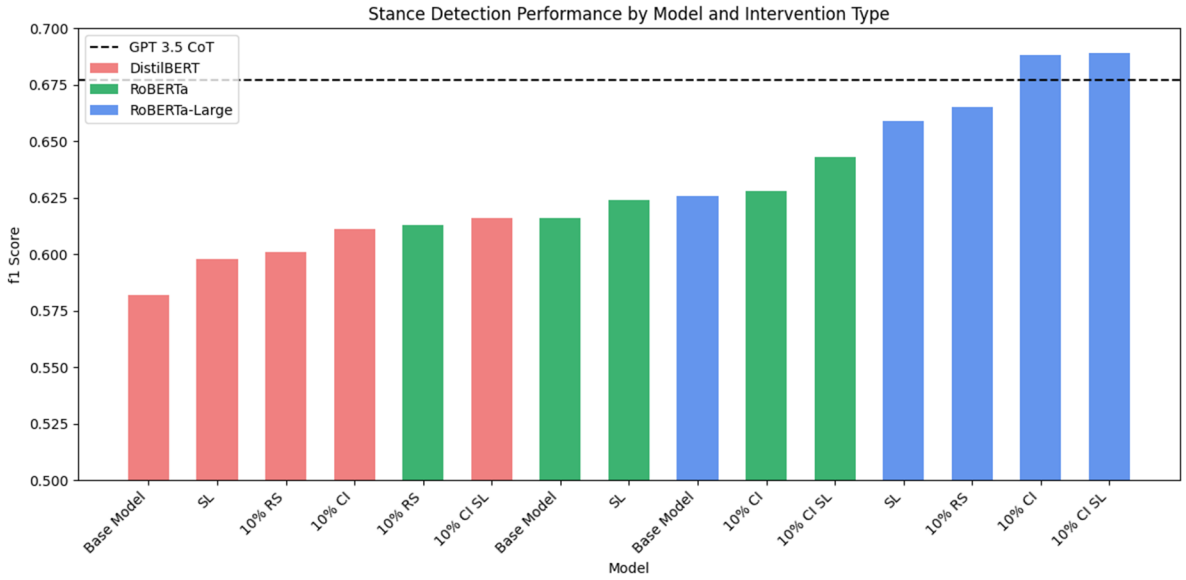

Figure 3: Comparing edge model F1 score as we change model and system interventions types for stance detection. We note steady improvements of edge model performance as we introduce more complex models and system intervention measures. The largest edge model with all system interventions out-performs gpt-3.5-turbo CoT.

we noticed some bias in LLM classification where the LLM was consistently incorrect in predicting a single class. This was represented in our confidence scores and caused our expert labels to focus on a single class, resulting in heavily weighted soft labels applied to a single class extrapolating existing error. To solve this problem, we stratified the expert sampling process, selecting the bottom 10 percent of confidence scores for each class instead of the bottom ten percent of the entire dataset. Doing so slightly decreased the accuracy on tasks where there was minimal bias, but greatly increased the accuracy where LLM bias was present such as ideology and stance classification. This difference in class performance for a given task has also been observed in other works. For example, LLMs consistently exhibit a discernible left and libertarian bias, as assessed by political orientation surveys, that likely arises due to the training data used for training LLMs (Motoki et al., 2023; Rozado, 2023; Rutinowski et al., 2023). This bias could affect performance on frequently politically charged tasks (which are also frequently important tasks for CSS), such as stance classification.

Confidence-informed sampling allowed us to greatly improve our edge classifier and should be integrated into any knowledge distillation process where small batch labeling is incorporated. Our confidence score distributions were the most discernible when ensembling different prompting techniques or in zero-shot settings. Chain-of-thought prompting resulted in less clean distributions, but further testing is required to fully understand the causation of prompting mechanisms on returned log probability distributions.

6 Conclusions

In this work, we successfully replicated LLM performance in an edge environment on computational social science tasks using a systems methodology. Our approach, which integrates expert-in-the-loop data labeling for a small portion of the data (10% or less), enables the deployment of highly performant small models in environments where LLM access is restricted due to cost, security, or latency concerns.

Our results demonstrate generalizability across various labeling prompts and distilled models, providing a flexible and scalable solution. This methodology offers a practical mechanism to reduce labeling costs and dependence on large LLMs while improving performance and data annotation throughput, even in resource-constrained settings with minimal human intervention.

References

Ali Alshahrani, Meysam Ghaffari, Kobra Amirizirtol, and Xiuwen Liu. 2021. Optimism/pessimism prediction of twitter messages and users using bert with soft label assignment. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Kolmogorov An. 1933. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601. Curran Associates, Inc.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. *Preprint*, arXiv:2104.08790.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2022. Self-guided noise-free data generation for efficient zero-shot learning. *arXiv preprint arXiv:2205.12679*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *Preprint*, arXiv:2210.11610.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

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.

Dilek Küçük and Fazli Can. 2021. Stance detection: Concepts, approaches, resources, and outstanding issues. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2673–2676.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR.

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.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring chatgpt political bias. *Public Choice*, pages 1–21.

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.

Nicholas Pangakis and Sam Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 113–131, Mexico City, Mexico. Association for Computational Linguistics.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, Markus Pauly, et al. 2023. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024.

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.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *Preprint*, arXiv:2402.13446.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. 2023. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. *Preprint*, arXiv:2310.13671.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Don't waste a single annotation: improving single-label classifiers through soft labels. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5347–5355, Singapore. Association for Computational Linguistics.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.

Jiacheng Ye, Jiahui Gao, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. Progen: Progressive zero-shot dataset generation via in-context feedback. *arXiv preprint arXiv:2210.12329*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022b. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *Preprint*, arXiv:2304.10145.

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 Sample Prompts

This appendix contains example prompts using the same document in the SemEval-2016 dataset. The newlines in the example prompts correspond to actual newline characters ("\n") in the prompt string.

A.1 Stance Detection Zero-Shot

THIS IS AN EXAMPLE STATEMENT
Which of the following best describes the above social media statements' stance regarding TOPIC?
A) For
B) Against
C) Neutral
Only respond with 'For', 'Against', or 'Neutral'.

A.2 Stance Detection Zero-Shot CoT

Stance classification is the task of determining the expressed or implied opinion, or stance, of a statement toward a specific target. Think step-by-step and explain the stance (For, Against, or Neutral) of the following social media statement towards TOPIC.
target: TOPIC
statement: THIS IS AN EXAMPLE STATEMENT
explanation:
LLM Replies!
Therefore, based on your explanation, what is the stance of the following social media statement toward the target?
target: TOPIC
statement: THIS IS AN EXAMPLE STATEMENT
A) For
B) Against
C) Neutral
Only respond with 'For', 'Against', or 'Neutral'. If the statement is not relevant to TOPIC, select Neutral.

A.3 Misinformation

"This is an example headline"
Which of the following describes the above news headline?

A) Misinformation
B) Trustworthy
Only respond with 'Misinformation' or 'Trustworthy'

A.4 Ideology

statement: "This is an example ideological statement."
Which of the following leanings would a political scientist say that the above statement has?
A: Conservative
B: Neutral
C: Liberal
Only respond with 'Conservative', 'Neutral', or 'Liberal'

A.5 Humor

Joke: This is an example joke.
Would most people find the above joke humorous?
You must pick between 'True' or 'False'.
You cannot use any words other than 'True' or 'False'.

B Stance Classification Additional Results

The appendix contains additional experimental results for the stance detection task. Figure 3 shows the effect of different system intervention strategies across the three main flavors of BERT evaluated. Figure 4 shows the effect of varying the percentage of expert labels with various system interventions, and Table 2 contains a table of F1 scores across all interventions.

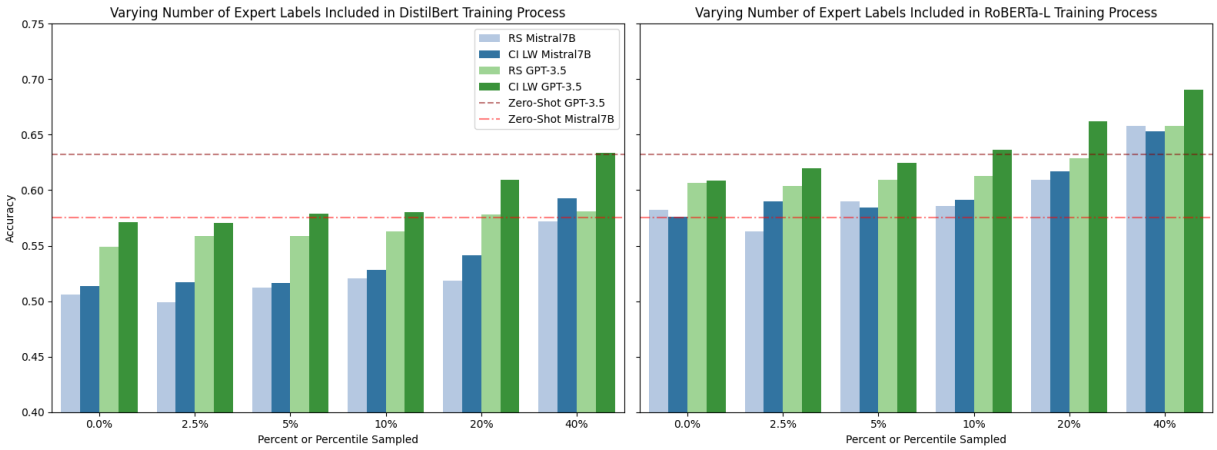

Figure 4: Varying the number of expert labels included amongst the LLM labels in the training process for DistilBERT and RoBERTa-L. RS implies randomly sampled expert labels for the training process and CI SL implies confidence informed sampling with label weighted training. Blue corresponds to the Mistral-7B-Instruct-2.0 LLM labeler and green corresponds to the GPT-3.5 LLM labeler. The horizontal dashed lines represent the zero-shot accuracy of each LLM.

Prompt Technique	Enterprise LLM	Edge Classifier - DistilBERT				
	GTP-3.5	Base	SL	RS 10%	CI 10%	CI SL 10%
Zero-Shot	.629	.549	.562	.559	.570	.582
Zero-Shot CoT	.677	.582	.598	.601	.611	.616
	Mistral-7B-Instruct					
Zero-Shot	.599	.485	.536	.534	.536	.552
Zero-Shot CoT	.589	.493	.452	.519	.496	.505
Prompt Technique	Enterprise LLM	Edge Classifier - RoBERTa				
	GTP-3.5	Base	SL	RS 10%	CI 10%	CI SL 10%
Zero-Shot	.629	.575	.587	.580	.594	.615
Zero-Shot CoT	.677	.616	.624	.613	.628	.643
	Mistral-7B-Instruct					
Zero-Shot	.599	.549	.539	.589	.588	.565
Zero-Shot CoT	.589	.530	.476	.561	.554	.532
Prompt Technique	Enterprise LLM	Edge Classifier - RoBERTa-L				
	GTP-3.5	Base	SL	RS 10%	CI 10%	CI SL 10%
Zero-Shot	.629	.603	.612	.617	.618	**.637**
Zero-Shot CoT	.677	.626	.659	.665	**.688**	**.689**
	Mistral-7B-Instruct					
Zero-Shot	.599	.578	.596	**.608**	**.613**	**.610**
Zero-Shot CoT	.589	**.597**	.560	**.603**	.559	**.597**

Table 2: F1 scores on SemEval2016. Edge classifier variants: 'Base' trained on LLM labels directly, SL trained with label weighting, RS 10% trained with 10% randomly sampled expert labels, CI 10% trained with 10% confidence informed expert labels, and CI SL 10% trained with 10% confidence informed expert labels and labeling weighting.