# Reducing Annotation Burden in Prompt Injection Detection: An Empirical Study of Uncertainty Sampling in Active Learning

**Anonymous ACL submission**

## Abstract

Prompt injection attacks pose significant risks to the safe deployment of large language models (LLMs), yet detecting adversarial prompts typically requires costly human annotations. This work explores uncertainty-based active learning as a strategy to reduce annotation effort in prompt injection classification. Using sentence embeddings and a lightweight XGBoost classifier, we simulate a human-in-the-loop labeling process on a benchmark dataset. Our results demonstrate that entropy-based sampling consistently outperforms random selection, achieving higher accuracy and inter-annotator agreement with fewer labeled examples. Our approach avoids dependence on large LLMs during annotation, mitigating risks associated with prompt injection vulnerabilities. These findings suggest that uncertainty-driven active learning combined with classical classifiers provides an effective and practical solution for adversarial prompt detection under limited annotation budgets, with implications for safer and more scalable deployment.

## 1 Introduction

Large Language Models (LLMs) accept free-form natural language prompts, enabling a wide range of applications. However, this flexibility also creates challenges for responsible deployment. In particular, users can craft prompts that elicit responses misaligned with platform policies, a tactic known as prompt injection or jailbreaking.

To mitigate such behavior, LLM providers often deploy prompt classifiers that flag inputs as benign or adversarial. Training these classifiers typically requires human-annotated data, making annotation a costly bottleneck. This raises an important methodological question: how can we most effectively select which examples to label?

Active Learning (AL) provides a framework for selecting informative examples to label, potentially improving classifier performance with fewer annotations. In this work, we simulate a human-in-the-loop AL process using a fully labeled prompt injection dataset: labels are revealed incrementally based on either random or entropy-based acquisition strategies. We study whether active learning can improve classifier performance when labels are acquired incrementally under a limited annotation budget.

We pose two research questions:

- **RQ1:** Can uncertainty-based sampling improve the informativeness of data collected in a simulated human-in-the-loop setting?
- **RQ2:** Does it lead to improved classifier performance on held-out validation data?

We compare random sampling and entropy-based acquisition using sentence embeddings and an XGBoost classifier. Our results show that uncertainty sampling yields stronger test set performance, including gains in accuracy and inter-annotator agreement ($\kappa$). Because active learning performance can vary across acquisition steps and datasets, we assess statistical significance using bootstrap confidence intervals and a one-sided Wilcoxon signed-rank test.

## 2 Related Work

Uncertainty sampling remains a foundational strategy in active learning for NLP. Recent work has explored augmenting uncertainty-based acquisition with pseudo-labeling or self-training to improve sample efficiency. For instance, Schröder and Heyer (2024) show that integrating self-training with uncertainty sampling can yield gains in low-resource settings by leveraging model confidence to supplement annotated data. Classical entropy-based uncertainty also remains competitive when revisited in the context of fine-tuning large transformer models (Schröder et al., 2022), reinforcing

the robustness of this baseline. Hybrid approaches such as ALVIN (Korakakis et al., 2024) interpolate between uncertainty and diversity-based sampling to mitigate demographic shortcut learning. Active learning with complementary labels (ALCL) reduces annotation costs by replacing full labels with class exclusions, combining uncertainty sampling with efficient supervision (Liu et al., 2023). Our work focuses on pure entropy-based uncertainty sampling without additional augmentation, employing a lightweight classifier rather than a transformer model, and targeting a safety-critical classification setting.

While transformers and large language models dominate many NLP tasks, they face challenges such as vulnerability to prompt injection attacks and restricted access in safety-critical applications. Consequently, lightweight and interpretable models like XGBoost remain valuable alternatives. Gradient-boosted tree ensembles such as XGBoost are competitive for various classification problems, especially when combined with uncertainty- or entropy-based sampling strategies. Their efficiency and interpretability suit scenarios with limited computational resources or labeled data. Prior work has demonstrated the effectiveness of entropy-aware XGBoost classifiers in domains including emotion recognition (Wang et al., 2018), malware detection (Prattipati et al., 2024), fraud identification in imbalanced datasets (Onur Erboy and Can Karaca, 2024), and human-in-the-loop loan default prediction (Khan et al., 2025). Notably, XGBoost has even outperformed large language models like GPT-4 in specific text classification tasks (Bohacek and Bravansky, 2024). We deliberately avoid transformers or large language models here due to concerns about prompt injection vulnerabilities and access restrictions, making XGBoost a preferable choice for safety-critical, human-in-the-loop classification.

Recent advances in active learning have addressed robustness and adversarial challenges, particularly in open-set settings where label spaces are partially unknown. For example, bidirectional uncertainty-based AL methods have been developed to handle such scenarios (Zong et al., 2024). Hybrid human-machine labeling frameworks have also been proposed to enhance robustness in neural machine translation (Azeemi et al., 2025). These approaches align with broader concerns around adversarial vulnerability in NLP, which have been extensively surveyed in recent literature (Goyal et al., 2023).

Despite advances in active learning algorithms, practical deployment faces persistent challenges such as annotation bottlenecks, batch size optimization, and label noise (Lowell et al., 2019). Recent work has highlighted the importance of annotator-centric approaches, especially for subjective or nuanced NLP tasks, by tailoring sampling strategies and interfaces to better accommodate human annotators (van der Meer et al., 2024). Efficiency improvements have been demonstrated through techniques like adapters on frozen transformer backbones, which reduce computational costs without sacrificing query quality (Galimzianova and Sanochkin, 2024). Additionally, studies specific to BERT confirm that uncertainty-based active learning remains effective for transformer fine-tuning, with cold-start problems addressed via self-supervised pretraining methods (Ein-Dor et al., 2020; Yuan et al., 2020).

Robust evaluation of active learning strategies necessitates rigorous statistical testing. The Wilcoxon signed-rank test is a widely adopted non-parametric method for comparing paired model performances, particularly when combined with cross-validation-based ranking approaches to ensure reliable and interpretable results (Dror et al., 2018; Sziklai et al., 2022).

## 3 Data And Experimental Setup

We use the QualiFire Prompt Injection Benchmark dataset (Qualifire, 2024), containing 5,000 labeled prompts with a near-even split between benign and injection examples. Its size, label quality, and benchmark status make it well-suited for reproducible supervised evaluation of classification models.

We reserve 20% of the data (1,000 prompts) as a held-out validation set. The remaining 80% (4,000 prompts) serve as the training and acquisition pool for active learning (AL). From this pool, an initial labeled set of 400 prompts (10%) is randomly and consistently selected to initialize both the AL and baseline models.

Two models are trained in parallel: one using active learning via uncertainty sampling (maximum entropy), and the other using random sampling as a baseline. At each of 1,600 acquisition steps, both models independently select one new prompt to label and retrain on all labeled data collected so far (including the original 400). This setup simu-

lates a human-in-the-loop annotation process with incremental model updates.

At each acquisition step, evaluation is performed on the subset of the 4,000-prompt pool not yet labeled by that model. This reflects the model's generalization to unseen examples during the active learning process and its decision-making prior to further annotation. While multiple metrics—including Accuracy, F1, AUC, and Cohen's $\kappa$—are computed, we focus on test-set Accuracy and Cohen's $\kappa$ to illustrate generalization trends: Accuracy reflects the model's performance as an automated annotator, and $\kappa$ measures agreement with human annotations.

Final model performance is assessed on the held-out validation set after acquiring 2,000 total labeled prompts (400 initial + 1,600 acquired). This allows standardized comparison between the uncertainty-based and random sampling models, evaluating generalization beyond the acquisition pool.

During active learning, we compute each model's validation accuracy at every acquisition step. To determine whether observed performance differences are statistically meaningful, we apply two tests: bootstrap confidence intervals (95%, 1,000 resamples) on paired validation accuracy differences, and a one-sided Wilcoxon signed-rank test across all 1,600 acquisition steps. These tests quantify whether performance gains from active learning are statistically reliable and generalize beyond the acquisition pool.

The classifier architecture is identical across conditions. Each prompt is embedded using the `all-MiniLM-L6-v2` sentence transformer (Hugging Face Model Hub, 2023), yielding a single vector representation per prompt. These embeddings are passed to an XGBoost classifier trained with binary logistic loss (log loss) and default hyperparameters.

## 4 Results

Figures 1 and 2 show the progression of model performance on the unlabeled portion of the training and acquisition pool as more prompts are labeled and added to the training set. The baseline random sampling method ("Random" in the figures) achieves a peak F1 score of approximately 0.80, precision of 0.80, recall of 0.81, accuracy of 0.84, AUC of 0.93, and Cohen's $\kappa$ of 0.67 around the 2,000 labeled samples mark.

In contrast, the active learning model using uncertainty sampling with maximum entropy acquisition ("Entropy" in the figures) demonstrates substantially improved performance, reaching a peak F1 score of 0.95, precision of 0.95, recall of 0.94, accuracy of 0.97, AUC of 0.98, and Cohen's $\kappa$ of 0.91 at a similar labeling budget. This difference occurs despite both methods drawing from the same pool of unlabeled data, differing only in the selection strategy for labeling.

Figure 3 displays the accuracy comparison on the held-out 20% validation set, which is entirely separate from the training and acquisition pool. This figure highlights the improved generalization ability of the entropy-based active learning model compared to random sampling as the number of labeled samples increases.

Table 1 reports the mean change in validation accuracy difference ($\Delta$Accuracy) between the two methods at various labeling budgets, averaged over samples within ±50 labeled prompts of the target budget. The positive mean differences and narrow 95% bootstrap confidence intervals indicate statistically meaningful improvements of the active learning approach at higher labeling budgets.

Overall, these results indicate that entropy-based uncertainty sampling can improve model performance and generalization in the prompt injection classification task, compared to random sampling, with relatively modest annotation budgets.

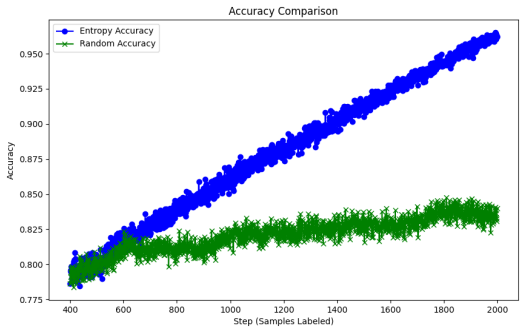


Figure 1: Accuracy on the unlabeled portion of the training and acquisition pool as additional prompts are labeled.

## 5 Discussion

Regarding RQ1, our results indicate that uncertainty-based active learning improves model generalization compared to random sampling on the QualiFire Prompt Injection Benchmark. The entropy acquisition strategy effectively prioritizes

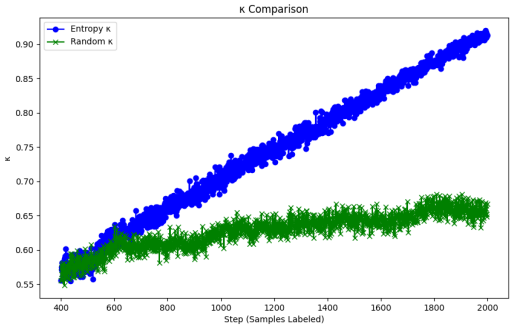


Figure 2: Test set Cohen's $\kappa$ over number of labeled samples, measured on the unlabeled portion of the training and acquisition pool.

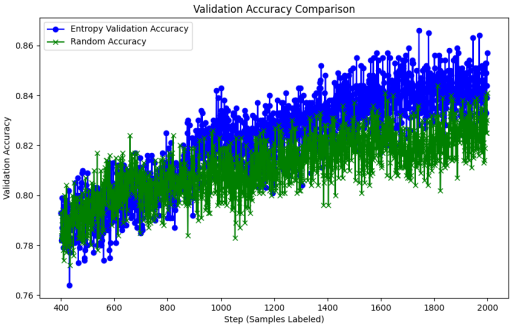


Figure 3: Validation Set Accuracy Comparison Across Number of Labeled Samples, measured on the held-out 20% validation set not used in active learning.

the most informative and uncertain prompts, enabling the model to learn more discriminative features with fewer labeled examples. This targeted selection leads to notably higher test-set accuracy, F1, AUC, and Cohen's $\kappa$, confirming the efficiency and effectiveness of uncertainty sampling in this context.

Our method leverages pre-trained sentence embeddings with a lightweight XGBoost classifier rather than relying on large language models (LLMs). This approach may avoid certain risks such as accidental prompt injection or reliance on costly API access, which could be advantageous in security-sensitive scenarios.

The test set performance reflects the model's ability to generalize from incremental annotations gathered from a single annotator, simulating a human-in-the-loop scenario where data is labeled bit by bit after an initial random start. In contrast, the held-out validation set evaluates broader generalization beyond this annotator-specific distribution, highlighting how well the model transfers to

Table 1: Change in accuracy ($\Delta$Accuracy) at different labeling budgets. Results are averaged over samples within $\pm 50$ of the target budget.

| Budget | Mean $\Delta$Accuracy | 95% CI Lower | 95% CI Upper |
|---|---|---|---|
| 600 | -0.0056 | -0.0077 | -0.0035 |
| 800 | 0.0003 | -0.0013 | 0.0021 |
| 1200 | 0.0116 | 0.0094 | 0.0136 |
| 1600 | 0.0165 | 0.0142 | 0.0188 |
| 1800 | 0.0207 | 0.0185 | 0.0228 |

truly unseen data. Regarding RQ2, the more modest validation gains compared to the test pool suggest that while active learning efficiently improves performance within the acquisition environment, challenges remain in ensuring robust generalization to diverse, out-of-distribution examples. This may be due to domain shift or dataset biases limiting transferability beyond the original data distribution.

The observed improvements are statistically significant, supported by bootstrap confidence intervals and a one-sided Wilcoxon signed-rank test. These results suggest uncertainty sampling can improve performance in iterative annotation settings.

Overall, our study suggests that active learning may improve prompt injection detection performance with limited labeled data, without requiring reliance on LLMs. These findings provide preliminary guidance for annotation strategies in adversarial NLP tasks and may contribute to developing more robust NLP systems.

## 6 Limitations

This study has several limitations that should be considered when interpreting the results. First, our experiments rely on a single primary benchmark dataset, the QualiFire Prompt Injection Benchmark. While this dataset is well-curated and widely used, its characteristics may not capture all the complexities of real-world prompt injection scenarios. To partially address this, we conducted supplementary experiments on a smaller, independent dataset (deepset, 2023), where similar performance trends were observed: entropy-based sampling consistently outperformed random sampling, achieving F1 scores of 0.94 and 0.78, respectively. However, this dataset's limited size and lack of cross-training restrict the strength of conclusions drawn from it.

Second, our study fixes the model architecture to an XGBoost classifier on top of sentence embeddings from the `all-MiniLM-L6-v2` transformer. While this choice balances performance and com-

putational efficiency, it does not explore the impact of alternative models, including large language models or fine-tuned transformers, which may further improve results or behave differently under active learning.

Third, our experimental setup assumes noise-free human annotations, not simulating errors or inconsistencies common in practical labeling workflows. Real-world human annotators may introduce label noise or disagreement, potentially affecting the robustness and reliability of active learning strategies.

Fourth, active learning acquisition was halted after labeling 50% of the training pool. While this cutoff aligns with resource constraints and simulates limited annotation budgets, extending the acquisition beyond this point could reveal longer-term learning dynamics or diminishing returns.

Finally, the dataset itself is roughly balanced between benign and injection prompts, whereas real deployment environments are likely to exhibit significant class imbalance, with benign inputs being much more frequent. This imbalance could affect model calibration and active learning efficacy in practice.

Together, these limitations highlight avenues for future work, including broader dataset evaluations, alternative model architectures, incorporation of human annotation noise, extended active learning regimes, and deployment-focused class imbalance considerations.

## 7 Ethical Considerations

This work uses prompt injection datasets designed to evaluate and improve model robustness against adversarial inputs. While such benchmarks are essential for advancing security, it is important to recognize potential biases in labeling and dataset construction that could influence model behavior and generalization. Additionally, the responsible use of adversarial datasets requires careful handling to avoid misuse or overfitting to specific attack patterns. We emphasize transparency and reproducibility to foster trustworthy research and encourage ongoing efforts to develop fair and robust AI systems.

## References


Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2025. To label or not to label: Hybrid active learning for neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3071–3082, Abu Dhabi, UAE. Association for Computational Linguistics.

Matyas Bohacek and Michal Bravansky. 2024. When XGBoost outperforms GPT-4 on text classification: A case study. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 51–60, Mexico City, Mexico. Association for Computational Linguistics.

deepset. 2023. Prompt injections dataset. https://huggingface.co/datasets/deepset/prompt-injections.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Daria Galimzianova and Leonid Sanochkin. 2024. Efficient active learning with adapters. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14374–14383, Miami, Florida, USA. Association for Computational Linguistics.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in nlp. *ACM Comput. Surv.*, 55(14s).

Hugging Face Model Hub. 2023. all-MiniLM-L6-v2: Sentence embedding model. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2025-05-16.

Shamim Ahmad Khan, Rahama Salman, Al-Hussein Maysir Majid, M. Mythili, Malik Bader Alazzam, and Ponni Valavan M. 2025. Enhancing loan default prediction with human-in-the-loop and xgboost ensemble learning. In *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–6.

Michalis Korakakis, Andreas Vlachos, and Adrian Weller. 2024. ALVIN: Active learning via INterpolation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22715–22728, Miami, Florida, USA. Association for Computational Linguistics.

Shengyuan Liu, Ke Chen, Tianlei Hu, and Yunqing Mao. 2023. Uncertainty-aware complementary label

queries for active learning. *Frontiers of Information Technology & Electronic Engineering*, 24(10):1497–1503.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. 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 21–30, Hong Kong, China. Association for Computational Linguistics.

Mehmet Onur Erboy and Ali Can Karaca. 2024. Weighted xgboost based active learning framework for fraud detection with using small number of samples from imbalanced dataset. In *Recent Trends and Advances in Artificial Intelligence*, pages 674–686, Cham. Springer Nature Switzerland.

Anshitha Prattipati, S. Saravanan, and S. Veluchamy. 2024. Adaptive malware detection in android: An active learning and xgboost approach. In *2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1–6.

Qualifire. 2024. Qualifire prompt injection benchmark. https://huggingface.co/datasets/qualifire/Qualifire-prompt-injection-benchmark.

Christopher Schröder and Gerhard Heyer. 2024. Self-training for sample-efficient active learning for text classification with pre-trained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11987–12004, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. *Preprint*, arXiv:2107.05687.

Balázs R. Sziklai, Máté Baranyi, and Károly Héberger. 2022. Testing rankings with cross-validation. *Preprint*, arXiv:2105.11939.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Sheng-Hui Wang, Huai-Ting Li, En-Jui Chang, and An-Yeu (Andy) Wu. 2018. Entropy-assisted emotion recognition of valence and arousal using xgboost classifier. In *Artificial Intelligence Applications and Innovations*, pages 249–260, Cham. Springer International Publishing.

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

Chen-Chen Zong, Ye-Wen Wang, Kun-Peng Ning, Hai-Bo Ye, and Sheng-Jun Huang. 2024. *Bidirectional Uncertainty-Based Active Learning for Open-Set Annotation*, page 127–143. Springer Nature Switzerland.