When Does Active Learning Actually Help? Empirical Insights with Transformer-based Automated Scoring

Justin O. Barber **Michael P. Hemenway** **Edward W. Wolfe**
Pearson Education
{justin.barber, michael.hemenway, ed.wolfe}@pearson.com

Abstract

Developing automated essay scoring (AES) systems typically demands extensive human annotation, incurring significant costs and requiring considerable time. Active learning (AL) methods aim to alleviate this challenge by strategically selecting the most informative essays for scoring, thereby potentially reducing annotation requirements without compromising model accuracy. This study systematically evaluates four prominent AL strategies—uncertainty sampling, BatchBALD, BADGE, and a novel GenAI-based uncertainty approach—against a random sampling baseline, using DeBERTa-based regression models across multiple assessment prompts exhibiting varying degrees of human scorer agreement. Contrary to initial expectations, we found that AL methods provided modest but meaningful improvements only for prompts characterized by poor scorer reliability (<60% agreement per score point). Notably, extensive hyperparameter optimization alone substantially reduced the annotation budget required to achieve near-optimal scoring performance, even with random sampling. Our findings underscore that while targeted AL methods can be beneficial in contexts of low scorer reliability, rigorous hyperparameter tuning remains a foundational and highly effective strategy for minimizing annotation costs in AES system development.

1 Introduction

Automated Essay Scoring (AES) systems have become integral to educational assessments by providing efficient, reliable, and scalable evaluation of student writing. State-of-the-art AES approaches typically utilize medium- to large-size pretrained transformer-based language models such as BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), and DeBERTa (He et al., 2021), fine-tuned on datasets of human-scored essays to produce scoring models aligned closely with human judgment. The development of robust AES models, however, usually requires extensive annotation efforts—often involving thousands of essays per prompt—posing significant practical limitations in terms of cost, time, and resources.

Active Learning (AL) mitigates these annotation burdens by scoring only the most informative essays. Although AL is well studied in NLP (Zhang et al., 2023; Li et al., 2024), few works test multiple strategies in AES or examine how scorer agreement moderates AL gains (Firoozi et al., 2023; Hellman et al., 2019). We compare four AL methods with a random sampling baseline across prompts of differing reliability.

Addressing this critical gap, our study evaluates four prominent AL strategies—uncertainty sampling, BatchBALD (Bayesian Active Learning by Disagreement), BADGE (Batch Active Learning by Diverse Gradient Embeddings), and a novel GenAI-based uncertainty sampling approach—across multiple writing and reading assessment prompts. These AL methods are benchmarked against random sampling as a baseline, examining their efficacy at annotation budgets ranging from 32 to 1,024 essays.

1.1 Research Questions

This study specifically investigates three research questions:

1. Which AL strategies yield the highest scoring agreement (measured via Quadratic Weighted Kappa [QWK]) with the minimal number of human-scored training examples, particularly across varying degrees of human inter-rater agreement?

2. Can a novel GenAI-guided AL approach effectively identify especially challenging-to-score essays, thereby enhancing the efficiency and quality of AES model training?

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 1–8
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

3. To what extent does comprehensive hyperparameter optimization alone (even with random sampling) significantly reduce the number of training essays required to achieve acceptable scoring accuracy across various prompts?

1.2 Contributions

Our contributions are:

- A four-way comparison of AL strategies versus a random baseline on prompts spanning different human scorer agreement;
- A novel GenAI sampler for small budgets;
- Evidence that hyperparameter tuning alone rivals AL when scorer reliability is moderate–high;
- Practical guidance for where AL is (and is not) worth the cost.

These findings hold significant implications for educational assessment organizations aiming to develop AES systems more efficiently. By understanding the nuanced contexts in which AL methods excel and the powerful impact of systematic hyperparameter tuning, stakeholders can better allocate annotation resources, enabling broader and more cost-effective application of automated scoring systems across diverse educational contexts.

2 Related Work

2.1 Active Learning in NLP

Active Learning (AL) reduces annotation costs by selecting the most informative unlabeled samples for labeling, enhancing model performance with fewer annotations (Settles, 2009; Zhang et al., 2023; Li et al., 2024). In NLP, prominent AL strategies include uncertainty-based, Bayesian, diversity-based, and hybrid approaches, which we adapt for Automated Essay Scoring (AES).

2.1.1 Uncertainty Sampling

Uncertainty-based sampling selects samples where models exhibit the highest uncertainty. Lewis and Gale (1994) introduced entropy-based selection, while Gal et al. (2017) popularized Monte Carlo dropout to estimate uncertainty in deep learning models. Margin-based methods, recently highlighted by Doucet et al. (2024), select samples with minimal differences between top class probabilities and have frequently outperformed random sampling in NLP tasks.

2.1.2 Bayesian Active Learning

Bayesian Active Learning (BAL) focuses explicitly on maximizing information gain regarding model parameters (Siddhant and Lipton, 2018). Bayesian Active Learning by Disagreement (BALD) selects samples based on uncertainty across posterior predictions (Houlsby et al., 2011). BatchBALD (Kirsch et al., 2019) extends this to batch selection, reducing redundancy by jointly evaluating batch informativeness at increased computational cost.

2.1.3 Diversity-Based and Hybrid Sampling

Diversity-based methods select samples that represent diverse regions of input space, ensuring robust generalization. Hybrid strategies like BADGE (Ash et al., 2020) combine uncertainty and diversity by clustering gradient embeddings to identify diverse yet informative samples, demonstrating strong performance in various classification tasks.

2.1.4 LLM-Guided Active Learning

Emerging approaches integrate Large Language Models (LLMs) into AL for nuanced semantic evaluation of samples. Methods such as ActiveLLM (Bayer and Reuter, 2024), ActivePrune (Azeemi et al., 2024), SelectLLM (Parkar et al., 2024), and ranking-based approaches (Jeong et al., 2025) have shown promise in identifying linguistically complex or ambiguous samples relevant for AES.

2.2 Active Learning for Automated Scoring

Research explicitly addressing AL in automated scoring contexts remains sparse. Horbach and Palmer (2016) compared AL strategies on short-answer scoring, noting significant variability across prompts. Hellman et al. (2019) demonstrated batch-mode AL effectiveness in instructor-driven contexts. Firoozi et al. (2023) highlighted uncertainty sampling's efficiency in AES, although their work focused exclusively on shallow models without exploring transformer-based methods or comprehensive comparisons.

2.2.1 Our Study in Context

Existing AES-focused AL studies have not systematically evaluated how scorer reliability impacts AL strategy efficacy nor have they fully explored the independent impact of extensive hyperparameter optimization. Our study addresses these gaps by rigorously comparing multiple AL strategies, explicitly considering varying scorer reliability levels, and demonstrating the substantial efficiency gains achievable through hyperparameter optimization alone. These insights inform best practices for practical AES deployment.

3 Methods

3.1 Problem Formulation

Given an unlabeled pool $\mathcal{U}$ and budget B, we run four AL rounds. Each round (i) trains on current labels, (ii) selects $\lfloor B/4 \rfloor$ essays via an acquisition function, (iii) obtains scores, and (iv) updates the model. A final training pass with tuned hyperparameters follows. Performance is reported on a held-out validation set.

3.2 Model Architecture

We fine-tune DeBERTaV3-base (He et al., 2021) as a regression model by adding a single linear head on the [CLS] embedding and optimizing Mean-Squared-Error loss weighted by inverse score frequency. Essays are tokenized with the DeBERTaV2 tokenizer (512-token limit) and trained using AdamW with linear warm-up and gradient clipping.

3.3 Active Learning Strategies

We evaluate four AL strategies:

Uncertainty Sampling: Selects essays with highest predictive entropy based on Gaussian-derived probability distributions from regression outputs.

BatchBALD (Kirsch et al., 2019): Maximizes batch mutual information using Monte Carlo dropout, first filtering the unlabeled pool by predictive entropy to enhance computational efficiency.

BADGE (Ash et al., 2020): Combines uncertainty and diversity by clustering gradient embeddings derived from a temporary classification head on the model encoder.

GenAI-Uncertainty Sampling (novel approach): Uses large language models (LLMs) to identify challenging essays (rated 1–5 on scoring difficulty). Essays rated highly challenging (5) are prioritized, selecting diverse examples within difficulty strata using k-means clustering.

3.4 Multi-Round Active Learning Framework

Our AL approach includes:

- Initial seed of 16 essays.
- Four AL rounds (one for GenAI), evenly dividing annotation budgets.
- Each round selects essays for scoring, expands the labeled set, and retrains the model.

3.5 Hyperparameter Optimization

Given its significant impact, we rigorously optimize hyperparameters using Optuna (Akiba et al., 2019):

Search Space:

- Learning rate: [1e-5 to 2e-5]
- Weight decay: [1e-3 to 1e-1]
- Batch size: [4, 8]

Optimization Approach:

1. **Discovery Phase**: Perform 40-trial hyperparameter optimization using random sampling at each annotation budget.
2. **Evaluation Phase**: Evaluate the top 16 discovered hyperparameter configurations across all AL strategies, limiting computationally intensive strategies (BatchBALD, BADGE) to budgets <= 384.

Final models train for up to 30 epochs with early stopping (patience=5) based on validation loss.

4 Experiments

4.1 Data Sources

Operational corpus. Our experiments utilize operational student response data from a large-scale summative K–12 assessment administered across multiple U.S. states. The dataset comprises both short constructed-response reading items and full-length essay prompts, capturing diverse aspects of student writing performance.

Prompt Selection Criteria. To establish a balanced and robust evaluation framework, prompts were selected based on sufficient availability of double-scored responses. This resulted in a set of eight suitable prompts: five reading items and three writing prompts.

Reading Tasks. The reading task subset consists of three Grade-8 items (R-8A, R-8B, R-8C) and two Grade-10 items (R-10A, R-10B). Reading responses were holistically scored on a three-point ordinal scale (0–2) or a five-point ordinal scale (0-4), each assessing a single construct.

Task	Grade	Genre / Trait	Scale	N
R-8A	8	Reading	0–2	5,305
R-8B	8	Reading	0–4	4,575
R-8C	8	Reading	0–2	3,911
R-10A	10	Reading	0–2	5,931
R-10B	10	Reading	0–4	4,987
W-5	5	Argumentative, Content	0–3	11,088
W-8	8	Informative, Content	0–3	10,754
W-11	11	Narrative, Content	0–3	10,416

Table 1: Descriptive statistics for the experimental corpus.

Writing Tasks. The writing tasks cover Grades 5 through 11, balanced across genres: W-5 (argumentative), W-8 (informative/explanatory), and W-11 (narrative). Although each essay includes multiple trait scores, we focus specifically on the *Content* trait, given its strong alignment with textual evidence and minimal confounding by surface-level mechanical features. Content scores range from 0 to 3.

Sample Sizes. Usable responses per task range from 3,911 to 11,088. Table 1 summarizes detailed counts.

Train–Validation Protocol. For each prompt–trait pair, we hold out a stratified sample of 500 responses as a validation set, preserving the marginal score distribution. This validation set is used exclusively for model checkpoint selection and hyperparameter optimization. Consequently, although it remains distinct from the training data used directly for gradient updates, it is not strictly unseen. This methodological choice may slightly overestimate absolute model performance but does not affect our comparative analysis of active learning strategies.

4.2 Evaluation Metrics

Quadratic Weighted Kappa (QWK): Our primary evaluation criterion measures the degree of agreement between model predictions and human raters and accounts explicitly for varying degrees of scoring discrepancy. We calculate QWK using Cohen's quadratic weighted kappa implementation from scikit-learn.

Metrics are calculated after rounding and clipping predictions: $\hat{y} = \text{clip}(\text{round}(f_\theta(x)), y_{\min}, y_{\max})$, where $y_{\min}$ and $y_{\max}$ represent score boundaries.

Model selection during training employs early stopping (patience=5) based on validation loss, with the best-performing model checkpoint saved according to QWK scores. Hyperparameter optimization also prioritizes QWK.

4.3 Implementation Details

Models are trained in PyTorch with Hugging Face Transformers on NVIDIA A10 GPUs. We use AdamW with 10% warm-up, gradient clipping (1.0), mixed precision, and smoothed inverse-frequency class weights (70% empirical frequency + 30% uniform distribution). Hyperparameter searches run in parallel round-robin across GPUs. For efficiency we drop BatchBALD and BADGE when budgets exceed 384 essays and subsample 500–2,560 essays for GenAI.

Strategy-specific details:

- **BatchBALD**: 10 Monte Carlo dropout passes with initial entropy-based filtering (top 10%, minimum 2,000 essays).
- **BADGE**: Temporary classification head derived from the regression model to compute gradient embeddings.

All experiments utilize fixed random seeds for reproducibility across NumPy, PyTorch, and strategy-specific operations.

Sample Size	Random	Uncertainty	BatchBALD	GenAI	BADGE
32	**0.79**	**0.79**	**0.79**	0.75	0.78
64	**0.81**	0.80	0.78	0.77	0.77
96	**0.81**	**0.81**	0.78	**0.81**	0.80
128	0.80	**0.81**	0.79	0.80	**0.81**
192	**0.82**	0.80	0.80	0.78	**0.82**
256	**0.83**	0.82	0.80	0.82	0.81
384	**0.83**	0.82	0.81	0.80	0.82
1024	**0.84**	–	–	–	–

Table 2: QWK results for prompts with good scorer agreement. Bold indicates the highest score(s) per row.

Sample Size	Random	Uncertainty	BatchBALD	GenAI	BADGE
32	**0.77**	0.70	0.74	0.75	0.73
64	**0.79**	0.73	0.77	0.71	0.77
96	**0.79**	0.76	0.75	0.76	0.75
128	**0.80**	0.76	0.78	0.70	0.78
192	**0.81**	0.75	0.77	0.78	0.76
256	**0.82**	0.78	0.79	0.79	0.78
384	**0.81**	0.76	0.80	0.79	0.80
1024	**0.82**	–	–	–	–

Table 3: QWK results for prompts with acceptable scorer agreement. Bold indicates the highest score(s) per row.

Sample Size	Random	Uncertainty	BatchBALD	GenAI	BADGE
32	0.69	0.67	0.66	**0.70**	0.65
64	0.67	0.71	0.69	**0.73**	0.69
96	0.71	0.66	0.70	**0.72**	0.71
128	0.71	0.70	0.71	0.72	**0.73**
192	**0.76**	0.71	0.75	0.73	0.73
256	0.73	0.71	**0.74**	0.73	**0.74**
384	**0.75**	0.74	**0.75**	0.74	**0.75**
1024	**0.77**	–	–	–	–

Table 4: QWK results for prompts with poor scorer agreement. Bold indicates the highest score(s) per row. Dashes indicate unavailable or omitted results.

5 Results

5.1 Performance of Active Learning Strategies by Scoring Quality

Tables 2, 3, and 4 present the Quadratic Weighted Kappa (QWK) performance of Active Learning (AL) strategies across three contexts of scorer reliability: good, acceptable, and poor. These tables explicitly compare random sampling against four AL methods (Uncertainty, BatchBALD, GenAI, and BADGE).

Table 2 highlights the scenario of good scorer agreement (approximately 80% agreement). Here, AL methods exhibit little advantage over random sampling. Even at small annotation budgets (e.g., n = 32 or 64), random sampling matches or surpasses AL approaches. For example, at $n = 256$, random sampling (QWK=0.83), uncertainty sampling (0.82), and GenAI (0.82) demonstrate similar effectiveness, but no AL method exceeds random sampling substantially.

Table 3 shows analogous results for acceptable scorer agreement contexts (about 60% agreement). Again, random sampling typically achieves a slightly higher or equal QWK compared to AL strategies across most sample sizes, though the GenAI method achieves competitive performance at several points. Notably, at $n = 256$ annotations, random sampling still yields the top performance (QWK=0.82), followed closely by BatchBALD, GenAI, and BADGE strategies, each achieving scores of at least 0.78.

In contrast, for prompts with poor scorer agreement (<60%), AL methods show clearer advantages over random sampling (Table 4). Particularly at lower annotation budgets, uncertainty-based strategies, including the GenAI and BADGE methods, consistently outperform random selection. For instance, at $n = 64$, the GenAI method (0.73) significantly surpasses random sampling (0.67).

Training Sample Size	QWK (Random Sampling)
32	0.77
64	0.78
96	0.78
128	0.80
192	0.81
256	0.82
384	0.82
1024	0.82

Table 5: Impact of 40-trial hyperparameter optimization on QWK using random sampling across sample sizes for all prompts.

Similarly, uncertainty-based AL strategies continue to show small but consistent advantages at larger annotation sizes (e.g., $n = 128$ through $n = 384$), reflecting their capacity to effectively select informative and potentially challenging essays for model training.

Finally, in table 5 we explicitly examine how extensive hyperparameter optimization alone influences AES performance (Table 5). With careful tuning, random sampling swiftly achieves strong performance and approaches saturation quickly (QWK=0.81 at $n = 192$ annotations), demonstrating the significant impact of optimization without specialized AL. Indeed, this tuning reduces required annotation counts substantially, effectively narrowing the advantage that sophisticated AL methods could achieve in many practical scoring scenarios.

6 Discussion

6.1 Effectiveness of Active Learning for AES

Our findings indicate that active learning (AL) methods provide modest yet meaningful benefits specifically for prompts characterized by low scorer agreement (<60% agreement per score point). In these challenging scoring contexts, uncertainty-based methods, including BatchBALD and our novel GenAI-based approach, consistently yielded slight improvements over random sampling at smaller annotation budgets. This aligns with the intuition that uncertain, borderline scoring cases are particularly informative for model calibration, and extends prior findings by Firoozi et al. (2023), emphasizing AL's specific utility in challenging scoring scenarios.

However, contrary to initial expectations, AL methods provided no substantial advantage over random sampling in contexts with moderate to high scorer reliability (approximately 60–80% agreement). This lack of improvement can largely be attributed to our extensive hyperparameter optimization process, which significantly boosted the performance of random sampling, leaving limited room for AL methods to offer additional benefits.

Additionally, our GenAI-based approach demonstrated encouraging results in identifying challenging essays early in the annotation process, highlighting the potential of leveraging large language models to enhance targeted sampling. Although the overall improvement was modest, the interpretability and targeted nature of the GenAI sampling suggest potential future avenues for improving essay scoring models, especially in highly ambiguous scoring contexts.

6.2 Impact of Hyperparameter Optimization

A critical secondary finding of our study is the pronounced effectiveness of extensive hyperparameter optimization—even when employing random sampling. Our rigorous hyperparameter tuning approach (40 trials using Optuna) substantially reduced the annotation budget required to achieve robust model performance. This suggests that, in many practical AES contexts, careful model optimization can significantly improve annotation efficiency, often exceeding the marginal gains offered by more complex sampling strategies.

6.3 Practical Implications

The findings reported here offer important insights for the practical development and operational management of AES systems:

- **When to use active learning (AL):** Our findings suggest that AL methods demonstrate the strongest benefits in low-reliability scoring contexts. When scoring reliability is low and essays are challenging to rate, AL techniques—such as uncertainty-based sampling and GenAI methods—systematically identify the most informative instances, thus effectively improving model quality and calibration.

- **Tune first, apply AL second:** Extensive hyperparameter optimization alone produces highly competitive AES models, especially for scoring contexts with scorer reliability at or above 60%. Model builders should, therefore, devote significant attention initially to optimizing hyperparameters before turning to AL methods.

We thus propose the following operational framework for AES implementations based on these insights:

1. Begin with a modest-sized randomly sampled initial set (e.g., 16–32 essays), ensuring sufficient prompt coverage.
2. Immediately prioritize extensive hyperparameter optimization early in the model-development process.
3. After initial tuning, selectively apply uncertainty-based AL (particularly GenAI-driven sampling) as annotation proceeds, especially in cases of lower scoring reliability.
4. As more responses are collected, continuously revisit and adjust hyperparameters, since optimal settings may evolve with increasing data.

6.4 Limitations and Future Work

Our study offers valuable insights but has several limitations indicating promising directions for future research:

- **Prompt and Context Diversity**: Our analysis was limited to eight prompts from a single assessment context. Future work should explore broader prompt variability, scoring traits, and educational contexts.
- **Human-in-the-loop Validation**: Real-world AL implementations involve iterative human scoring. Future research should directly assess AL's practical implications within live annotation workflows.
- **Hyperparameter Exploration**: This work has highlighted the importance of hyperparameter optimization in model performance. Future experiments will consider an even wider hyperparameter space and optimization techniques that would be robust in operational contexts.
- **Fairness Considerations**: Further research could investigate how AL and targeted sampling methods, including GenAI, influence scoring fairness and demographic representation, potentially integrating fairness-aware constraints or regularizations.
- **Semi-Supervised Approaches**: Leveraging unlabeled data via semi-supervised or self-supervised learning methods (e.g., consistency regularization, pseudo-labeling, contrastive learning) may further enhance AES efficiency and warrants exploration.

Overall, our results highlight both the nuanced effectiveness of active learning methods under specific conditions and the crucial foundational role of rigorous hyperparameter optimization. These insights provide clear guidance for enhancing annotation efficiency and scoring reliability within AES deployments.

7 Conclusion

This study highlights two key findings for automated essay scoring (AES): First, active learning (AL) offers modest improvements over random sampling primarily in low-reliability scoring contexts. In prompts with higher scorer agreement, random sampling—when paired with wide hyperparameter sweeps—achieves near-optimal performance, often matching or exceeding AL strategies. Second, our novel GenAI-based sampling approach shows promise in identifying challenging essays early, but its benefits diminish as budgets increase.

These results suggest that rigorous hyperparameter optimization may be more impactful than AL in many AES scenarios. For practical deployment, AL may still provide value in identifying difficult examples and supporting scorer calibration in ambiguous contexts. Future research should explore how AL interacts with fairness, human-in-the-loop scoring, and hybrid semi-supervised learning strategies to further improve scoring efficiency and transparency.

Acknowledgments

We thank our colleagues at Pearson for their support and feedback on this research. We also acknowledge the computational resources provided by Pearson's technology team.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv preprint*. ArXiv:1907.10902 [cs].

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *arXiv preprint*. ArXiv:1906.03671 [cs].

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2024. Language Model-Driven Data Pruning Enables Efficient Active Learning. *arXiv preprint*. ArXiv:2410.04275 [cs].

Markus Bayer and Christian Reuter. 2024. ActiveLLM: Large Language Model-based Active Learning for Textual Few-Shot Scenarios. *arXiv preprint*. ArXiv:2405.10808 [cs] version: 1.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Doucet, Benjamin Estermann, Till Aczel, and Roger Wattenhofer. 2024. Bridging Diversity and Uncertainty in Active learning with Self-Supervised Pre-Training. *arXiv preprint*. ArXiv:2403.03728 [cs] version: 1.

Tahereh Firoozi, Hamid Mohammadi, and Mark J. Gierl. 2023. Using Active Learning Methods to Strategically Select Essays for Automated Scoring. *Educational Measurement*, 42(1):34–43. ArXiv:2301.00628 [cs].

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. *arXiv preprint*. ArXiv:1703.02910 [cs].

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Scott Hellman, Mark Rosenstein, Andrew Gorman, William Murray, Lee Becker, Alok Baikadi, Jill Budden, and Peter W. Foltz. 2019. Scaling Up Writing in the Curriculum: Batch Mode Active Learning for Automated Essay Scoring. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, pages 1–10, Chicago IL USA. ACM.

Andrea Horbach and Alexis Palmer. 2016. Investigating Active Learning for Short-Answer Scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 301–311, San Diego, CA. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian Active Learning for Classification and Preference Learning. *arXiv preprint*. ArXiv:1112.5745 [stat].

Daniel P. Jeong, Zachary C. Lipton, and Pradeep Ravikumar. 2025. LLM-Select: Feature Selection with Large Language Models. *arXiv preprint*. ArXiv:2407.02694 [cs].

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *arXiv preprint*. ArXiv:1906.08158 [cs].

David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. *arXiv preprint*. ArXiv:cmp-lg/9407020.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024. A Survey on Deep Active Learning: Recent Advances and New Frontiers. *arXiv preprint*. ArXiv:2405.00334 [cs].

Ritik Sachin Parkar, Jaehyung Kim, Jong Inn Park, and Dongyeop Kang. 2024. SelectLLM: Can LLMs Select Important Instructions to Annotate? *arXiv preprint*. ArXiv:2401.16553 [cs].

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. *arXiv preprint*. ArXiv:1808.05697 [cs].

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2023. A Survey of Active Learning for Natural Language Processing. *arXiv preprint*. ArXiv:2210.10109 [cs].