On the Limitations of *Simulating* Active Learning

Katerina Margatina **Nikolaos Aletras**
University of Sheffield
`{k.margatina, n.aletras}@sheffield.ac.uk`

Abstract

Active learning (AL) is a *human-and-model-in-the-loop* paradigm that iteratively selects informative unlabeled data for human annotation, aiming to improve over random sampling. However, performing AL experiments with human annotations on-the-fly is a laborious and expensive process, thus unrealistic for academic research. An easy fix to this impediment is to *simulate* AL, by treating an *already* labeled and publicly available dataset as the pool of *unlabeled* data. In this position paper, we first survey recent literature and highlight the challenges across all different steps within the AL loop. We further unveil neglected caveats in the experimental setup that can significantly affect the quality of AL research. We continue with an exploration of how the *simulation* setting can govern empirical findings, arguing that it might be one of the answers behind the ever posed question *"why do active learning algorithms sometimes fail to outperform random sampling?"*. We argue that evaluating AL algorithms on available labeled datasets might provide a *lower bound* as to their effectiveness in real data. We believe it is essential to collectively shape the best practices for AL research, particularly as engineering advancements in LLMs push the research focus towards data-driven approaches (e.g., data efficiency, alignment, fairness). In light of this, we have developed guidelines for future work. Our aim is to draw attention to these limitations within the community, in the hope of finding ways to address them.

1 Introduction

Based on the assumption that "*not all data is equal*", active learning (AL) (Cohn et al., 1996; Settles, 2009) aims to identify the most informative data for annotation from a pool (or a stream) of unlabeled data (i.e., data acquisition). With multiple rounds of model training, data acquisition and human annotation (Figure 1), the goal is to achieve *data efficiency*. A data efficient AL algorithm entails that a model achieves satisfactory performance on a held-out test set, by being trained with only a fraction of the acquired data.

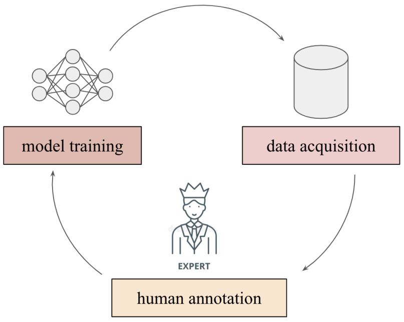

Figure 1: High-level overview of the *train-acquire-annotate* steps of the active learning loop.

AL has traditionally attracted wide attention in the Natural Language Processing (NLP) community. It has been explored for machine translation (Haffari et al., 2009; Dara et al., 2014; Miura et al., 2016; Zhao et al., 2020), text classification (Ein-Dor et al., 2020; Schröder and Niekler, 2020; Margatina et al., 2022; Schröder et al., 2023), part-of-speech tagging (Chaudhary et al., 2021), coreference (Yuan et al., 2022) and entity resolution (Qian et al., 2017; Kasai et al., 2019), named entity recognition (Erdmann et al., 2019; Shen et al., 2017; Wei et al., 2019; Shelmanov et al., 2021), and natural language inference (Snijders et al., 2023), *inter alia*. However, its potential value is still growing (Zhang et al., 2022d), driven by advancements in the state-of-the-art in language model pretraining (Tamkin et al., 2022). Given the initial assumption that "not all data is equal", it is reasonable to expect researchers to seek out the "most valuable" data for pretraining or adapting their language models.

The usual pool-based AL setting is to acquire data from an unlabeled pool, label it, and use it to

Findings of the Association for Computational Linguistics: ACL 2023, pages 4402–4419
July 9-14, 2023 ©2023 Association for Computational Linguistics

train a supervised model that, hopefully, obtains satisfactory performance on a test set for the task at hand. This is very similar to the general model-in-the-loop paradigm (Karmakharm et al., 2019; Bartolo et al., 2020, 2022; Kiela et al., 2021; Wallace et al., 2022), with the main difference being the AL-based data acquisition stage. The assumption is that, by iteratively selecting data for annotation according to an informativeness criterion, it will result into better model predictive performance compared to randomly sampling and annotate data of the same size.

However, this does not always seem to be the case. A body of work has shown that AL algorithms, that make use of uncertainty (Lewis and Gale, 1994; Cohn et al., 1996; Houlsby et al., 2011; Gal et al., 2017), diversity sampling (Brinker, 2003; Bodó et al., 2011; Sener and Savarese, 2018) or even more complex acquisition strategies (Ducoffe and Precioso, 2018; Ash et al., 2020; Yuan et al., 2020; Margatina et al., 2021), often fail to improve over a simple random sampling baseline (Baldridge and Palmer, 2009; Ducoffe and Precioso, 2018; Lowell et al., 2019; Kees et al., 2021; Karamcheti et al., 2021; Snijders et al., 2023). Such findings pose a serious question on the practical usefulness of AL, as they do not corroborate its initial core hypothesis that *not all data is equally useful for training a model*. In other words, if we cannot show that one subset of the data is "better"[1] than another, why do AL in the first place?

Only a small body of work has attempted to explore the pain points of AL. For instance, Karamcheti et al. (2021), leveraging visualisations from *data maps* (Swayamdipta et al., 2020), show that AL algorithms tend to acquire *collective outliers* (i.e. groups of examples that deviate from the rest of the examples but cluster together), thus explaining the utter failure of eight AL algorithms to outperform random sampling in visual question answering. Building on this work, more recently Snijders et al. (2023) corroborate these findings for the task of natural language inference and further show that uncertainty based AL methods recover and even surpass random selection when hard-to-learn data points are removed from the pool. Lowell et al. (2019) show that the benefits of AL with certain models and domains do not generalize reliably across models and tasks. This could be problematic since, in practice, one might not have the means to explore and compare alternative AL strategies. They also show that an actively acquired dataset using a certain model-in-the-loop, may be disadvantageous for training models of a different family, raising the issue of whether the downsides inherent to AL are worth the modest and inconsistent performance gains it tends to afford.

In this paper, we aim to explore all possible limitations that researchers and practitioners currently face when doing research on AL (Zhang et al., 2022d). We first describe the process of pool-based AL (Figure 1) and identify challenges in every step of the iterative process (§2). Next, we unearth obscure details that are often left unstated and under-explored (§3). We then delve into a more philosophical discussion of the role of simulation and its connection to real practical applications (§4). Finally, we provide guidelines for future work (§5) and conclusions (§6), aspiring to promote neglected, but valuable, ideas to improve the direction of research in active learning.

2 Challenges in the Active Learning Loop

We first introduce the typical steps in the pool-based AL setting (Lewis and Gale, 1994) and identify several challenges that an AL practitioner has to deal with, across all steps (Figure 2).[2]

2.1 Problem Definition

Consider the experimental scenario where we want to model a specific NLP task for which we do not yet have any labeled data, but we have access to a large pool of unlabeled data $\mathcal{D}_{\text{pool}}$. We assume that it is unrealistic (e.g., laborious, expensive) to have humans annotating all of it. $\mathcal{D}_{\text{pool}}$ constitutes the textual corpus from which we want to sample a fraction of the most *useful* (e.g., informative, representative) data points for human annotation. In order to perform active learning, we need an initial labeled dataset $\mathcal{D}_{\text{lab}}$, often called "seed" dataset, to be used for training a task-specific model with supervised learning. To evaluate the model, we need a usually small validation set for model selection $\mathcal{D}_{\text{val}}$ and a held out test set $\mathcal{D}_{\text{test}}$ to evaluate the model's generalization. We use $\mathcal{D}_{\text{lab}}$ and $\mathcal{D}_{\text{val}}$ to train the first model and then test it on $\mathcal{D}_{\text{test}}$.

[1] We consider a labeled dataset $A \subset C$ to be "better" than a labeled dataset $B \subset C$, both sampled from a corpus C and $|A| = |B|$, if a model M_A trained on A yields higher performance on a test set compared to M_B, where both models are identical in terms of architecture, training procedure, etc.

[2] We point the reader to the comprehensive survey of Zhang et al. (2022d) for a more in-depth exploration of recent literature in AL.

In this stage, we start acquiring labeled data for model training. Data points are sampled from $\mathcal{D}_{\text{pool}}$ via an acquisition strategy and subsequently passed to human annotators for labeling. The acquisition function selects a batch of data $Q \subset \mathcal{D}_{\text{pool}}$ according to some informativeness criterion and can either use the model-in-the-loop or not. We employ crowdsourcing or expert annotators to label the selected batch Q which then is appended to the labeled dataset $\mathcal{D}_{\text{lab}}$.

Now that we have augmented the seed dataset with more data, we re-train the model on the new training dataset, $\mathcal{D}_{\text{lab}}$. We test the new model on $\mathcal{D}_{\text{test}}$ and we stop if we obtain satisfactory performance or if the budget for annotation has run out (or using any other stopping criterion). If we do not want to stop, we use the acquisition function to select more unlabeled data from $\mathcal{D}_{\text{pool}}$, which we annotate and append to $\mathcal{D}_{\text{lab}}$, etc. This is the AL loop shown in Figure 2.

2.2 Active Learning Design

Seed dataset We start the AL loop (§2.1) by defining an initial labeled "seed dataset" (Figure 2: $\boxed{1}$). The seed dataset plays an important role, as it will be used to train the the first model-in-the-loop (Tomanek et al., 2009; Horbach and Palmer, 2016). In AL research, we typically address the cold-start problem by sampling from $\mathcal{D}_{\text{pool}}$ with a uniform distribution for each class, either retaining the true label distribution or choosing data that form a balanced label distribution.[3] This is merely a convenient design choice, as it is simple and easy to implement. However, sampling the seed dataset this way, does not really reflect a real-world setting where the label distribution of the (unlabeled data of the) pool is actually unknown.

Prabhu et al. (2019) performed a study of such sampling bias in AL, showing no effect in different seed datasets across the considered methods. Ein-Dor et al. (2020) also experimented with different imbalanced seed datasets, showing that AL improves over random sampling in settings with highest imbalance.

Furthermore, the choice of the seed dataset has a direct effect on the entire AL design because the first model-in-the-loop marks the reference point of the performance in $\mathcal{D}_{\text{test}}$. In other words, the performance of the first model is essentially the baseline, according to which a practitioner will plan the AL loop based on the goal performance and the available budget. It is thus essential to revisit existing approaches on choosing the seed dataset (Kang et al., 2004; Vlachos, 2006; Hu et al., 2010; Yuan et al., 2020) and evaluate them towards a realistic simulation of an AL experiment.

Number of iterations & acquisition budget After choosing the seed dataset it is natural to decide the number of iterations, the acquisition size (the size of the acquired batch $\mathcal{Q}$) and the budget (the size of the actively collected $\mathcal{D}_{\text{lab}}$) of the AL experiment. This is another part where literature does not offer concrete explanations on the design choice. Papers that address the cold-start problem would naturally focus on the very few first AL iterations (Yuan et al., 2020), while others might simulate AL until a certain percentage of the pool has been annotated (Prabhu et al., 2019; Lowell et al., 2019; Zhao et al., 2020; Zhang and Plank, 2021; Margatina et al., 2022) or until a certain fixed and predefined number of examples has been annotated (Ein-Dor et al., 2020; Kirsch et al., 2021).

2.3 Model Training

We now train the model-in-the-loop with the available labeled dataset $\mathcal{D}_{\text{lab}}$ (Figure 2: $\boxed{2}$). Interestingly, there are not many studies that explore how we should properly train the model in the low data resource setting of AL. Existing approaches include semi-supervised learning (McCallum and Nigam, 1998; Tomanek and Hahn, 2009; Dasgupta and Ng, 2009; Yu et al., 2022), weak supervision (Ni et al., 2019; Qian et al., 2020; Brantley et al., 2020; Zhang et al., 2022a) and data augmentation (Zhang et al., 2020; Zhao et al., 2020; Hu and Neubig, 2021), with the most prevalent approach currently to be transfer learning from pretrained language models (Ein-Dor et al., 2020; Margatina et al., 2021; Tamkin et al., 2022). Recently, Margatina et al. (2022) showed large performance gains by adapting the pretrained language model to the task using the unlabeled data of the pool (i.e., task adaptive pretraining by Gururangan et al. (2020)). The authors also proposed an adaptive fine-tuning technique to account for the varying size of $\mathcal{D}_{\text{lab}}$ showing extra increase in $\mathcal{D}_{\text{test}}$ performance.

Still, there is room for improvement in this rather

[3] In AL research, a fully labeled dataset is typically *treated* as an *unlabeled* $\mathcal{D}_{\text{pool}}$ by entirely ignoring its labels, while in reality we *do* have access to them. Hence, the labels implicitly play a role in the design of the AL experiment. We analyze our criticism to this seemingly "random sampling" approach to form the seed dataset in §4.2.

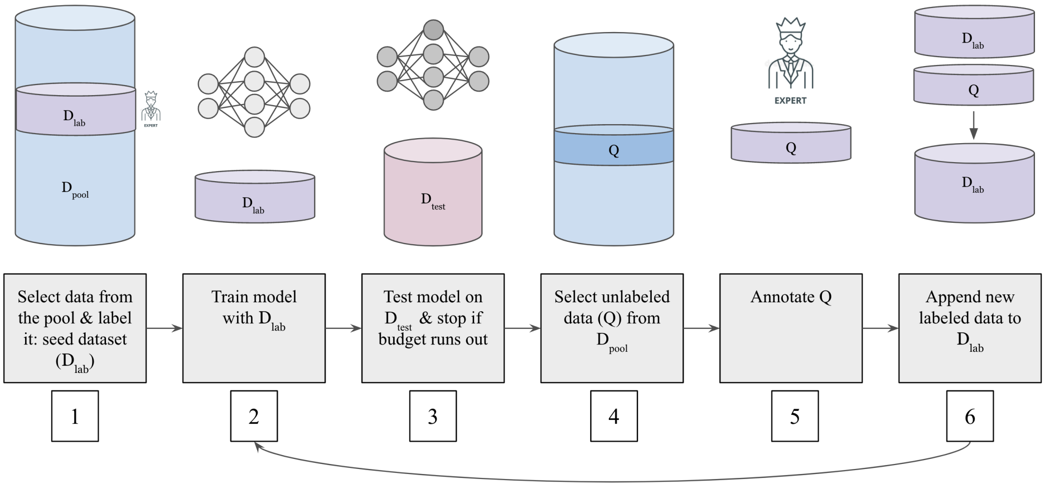
D lab
D pool
EXPERT
D lab
D test
Q
EXPERT
Q
D lab
Q
D lab
Select data from the pool & label it: seed dataset (D lab)
Train model with D lab
Test model on D test & stop if budget runs out
Select unlabeled data (Q) from D pool
Annotate Q
Append new labeled data to D lab
1
2
3
4
5
6

Figure 2: Distinct steps of the active learning loop (1–6). We use blue for the unlabeled data, purple for the labeled data and red for the (labeled) test data.

under-explored area. Especially now, state-of-the-art NLP pretrained language models consist of many millions or even billions of parameters. In AL we often deal with a small $\mathcal{D}_{\text{lab}}$ of a few hundred examples, thus adapting the training strategy is not trivial.

2.4 Data Acquisition

The data acquisition step (Figure 2: $\boxed{4}$) is probably the core of the AL process and can be performed in various ways.[4]

Zhang et al. (2022d) provide a thorough literature review of query strategies, dividing them into two broad families. The first is based on *informativeness*, and methods in this family treat each candidate instance individually, assign a score and select the top (or bottom) instances based on the ranking of the scores. Major sub-categories of methods that belong in the informativeness family are uncertainty sampling (Lewis and Gale, 1994; Culotta and Mccallum, 2005; Zhang and Plank, 2021; Schröder et al., 2022), divergence-based algorithms (Ducoffe and Precioso, 2018; Margatina et al., 2021; Zhang et al., 2022b), disagreement-based (Seung et al., 1992; Houlsby et al., 2011; Gal et al., 2017; Siddhant and Lipton, 2018; Kirsch et al., 2019; Zeng and Zubiaga, 2023), gradient-based (Settles et al., 2007; Settles and Craven, 2008) and performance prediction (Roy and Mccallum, 2001; Konyushkova et al., 2017; Bachman et al., 2017; Liu et al., 2018).

The second family is representativeness and takes into account how instances of the pool correlate with each other, in order to avoid sampling bias harms from treating each instance individually. Density-based methods choose the most representative instances of the unlabeled pool (Ambati et al., 2010; Zhao et al., 2020; Zhu et al., 2008), while others opt for discriminative data points that differ from the already labeled dataset (Gissin and Shalev-Shwartz, 2019; Erdmann et al., 2019). A commonly adopted category in this family is batch diversity, where algorithms select a batch of diverse data points from the pool at each iteration (Brinker, 2003; Bodó et al., 2011; Zhu et al., 2008; Geifman and El-Yaniv, 2017; Zhdanov, 2019; Yu et al., 2022), with core-set (Sener and Savarese, 2018) to be the most common approach.

Naturally, there are hybrid acquisition functions that combine informativeness and representativeness (Yuan et al., 2020; Ash et al., 2020; Shi et al., 2021). Still, among the aforementioned methods there is not a universally superior acquisition function that consistently outperforms all others. Thus, which data to acquire is an active area of research.

2.5 Data Annotation

Once an acquisition function is applied to $\mathcal{D}_{\text{pool}}$, a subset Q is chosen, and the obtained unlabeled data is subsequently forwarded to human annotators for annotation (Figure 2: $\boxed{5}$). In the context of simulation-based active learning, this aspect is not the primary focus since the labels for the actively acquired batch are *already* available. However, a

[4] In literature, the terms *data selection method*, *query strategy* and *acquisition function* are often used interchangeably.

question that naturally arises is: *Are all examples equally easy to annotate?* In simulation, all instances take equally long to label. This does not account for the fact that hard instances for the classifier are often hard for humans as well (Hachey et al., 2005; Baldridge and Osborne, 2004), therefore the current experimental setting is limiting and research for cost-aware selection strategies (Donmez and Carbonell, 2008; Tomanek and Hahn, 2010; Wei et al., 2019) is required. This would include explicit exploration of the synergies between random or actively acquired data and annotator expertise (Baldridge and Palmer, 2009).

2.6 Stopping Criterion

Finally, another active area of research is to develop effective methods for stopping AL (Figure 2: $\boxed{3}$). In simulation, we typically decide as a budget a number of examples or a percentage of $\mathcal{D}_{\text{pool}}$ up to which we "aford" to annotate. However, in both research and real world applications, it is not clear if the model performance has reached a plateau. The stopping criterion should not be pre-defined by a heuristic, but rather a product of a well-designed experimental setting (Vlachos, 2008; Tomanek and Hahn, 2010; Ishibashi and Hino, 2020; Pullar-Strecker et al., 2022; Hacohen et al., 2022; Kurlandski and Bloodgood, 2022).[5]

3 The Fine Print

Previously, we presented specific challenges across different steps in the AL loop that researchers and practitioners need to address. Still, these challenges have long been attracting the attention of the research community. Interestingly, there are more caveats, that someone with no AL experience might have never encountered or even imagined. Hence, in this section we aim to unveil several such small details that still remain unexplored.

3.1 Hyperparameter Tuning

A possibly major issue of the current academic status quo in AL, is that researchers often do not tune the models-in-the-loop. This is mostly due to limitations related to time and compute constrains. For instance, a paper that proposes a new acquisition function would be required to run experiments for multiple baselines, iterations, random seeds and datasets. For example, a modest experiment including $a = 5$ acquisition functions, $i = 10$ AL iterations, $n = 5$ random seeds and $d = 5$ datasets, would reach an outstanding number of minimum $a \times i \times n \times d = 1,250$ trained models in total. This makes it rather hard to perform hyperparameter tuning of all these models in every AL loop, so it is the norm to use the same model architecture and hyperparameters to train all models.

In reality, practitioners that want to use AL, apply it *once*. Therefore, they most likely afford to tune the one and only model-in-the-loop. The question that arises then, is "*do the findings of AL experiments that do not tune the models generalize to scenarios where all models-in-the-loop are tuned*"? In other words, if an AL algorithm A performs better than B according to an experimental finding, would this be the case if we applied hyperparameter tuning to the models of both algorithms? Wouldn't it be possible that, with another configuration of hyperparameters, B performed better in the end?

3.2 Model Stability

In parallel, another undisclosed detail is what researchers do when the models-in-the-loop are unstable (i.e., *crash*). This essentially means that for some reason the optimisation of the model might fail and the model never converges leading to extremely poor predictive performance. Perhaps before the deep learning era such a problem did not exist, but now it is a likely phenomenon.

Dodge et al. (2020) showed that many fine-tuning experiments diverged part of the way through training especially on small datasets. AL is by definition connected with low-data resource settings, as the gains of data efficiency are meaningful in the scenario when labeled data is scarce. In light of this challenge, there is no consensus as to what an AL researcher or practitioner should do to alleviate this problem. One can choose to re-train the model with a different random seed, or do nothing. Though, it is non-trivial under which condition one should choose to re-train the model, since it is common that not always test performance improves from one AL iteration to the next.

Furthermore, there is currently no study that explores how much AL algorithms, that use the model-in-the-loop for acquisition, suffer by this problem. For instance, consider an uncertainty-based AL algorithm that uses the predictive proba-

[5] Unless of course the actual budget is spent, where in real world settings this is effectively the stopping criterion.

bility distribution of the model to select the most uncertain data points from the pool. If the model crashes, then its uncertainty estimates are not meaningful, thus the data acquisition function does not work as expected. In effect, the sampling method turns to a uniform distribution (i.e., the random sampling baseline).

3.3 Active Learning Evaluation

Another important challenge is the evaluation framework for AL. Evaluating the *actual* contribution of an AL method against its competitors would require to perform the same iterative *train-acquire-annotate* experiment (Figure 1) for all AL methods in the exact same data setting and with real human annotations. Certainly, such a laborious and expensive process is prohibitive for academic research, which is why we perform simulations by treating an *already* labeled and open-source dataset as a pool of unlabeled data.

Still, even if we were able to perform the experiments in real life, it is not trivial how to properly define when one method is better than another. This is because AL experiments include multiple rounds of annotation, thus multiple trained models and multiple scores in the test set(s). In cases with no clear difference between the algorithms compared, how should we do a fair comparison?

Previous work presents tables comparing the test set performance of the last model, often ignoring performance in previous loops (Prabhu et al., 2019; Mussmann et al., 2020). The vast majority of previous work though uses plots to visualize the performance over the AL iterations (Lowell et al., 2019; Ein-Dor et al., 2020) and in some cases offer a more detailed visualization with the variance due to the random seeds (Yuan et al., 2020; Kirsch et al., 2021; Margatina et al., 2021).

3.4 The Test of Time

Settles (2009) eloquently defines the "test of time" problem that AL faces: "*A training set built in cooperation with an active learner is inherently tied to the model that was used to generate it (i.e., the class of the model selecting the queries). Therefore, the labeled instances are a biased distribution, not drawn i.i.d. from the underlying natural density. If one were to change model classes—as we often do in machine learning when the state of the art advances—this training set may no longer be as useful to the new model class*".

Several years later, in the deep learning era, Lowell et al. (2019) indeed corroborates this concern. They demonstrate that a model from a certain family (e.g., convolution neural networks) might perform better when trained with a random subset of a pool, than an actively acquired dataset with a model of a different family (e.g., recurrent neural networks). Interestingly, Jelenić et al. (2023) recently showed that AL methods with similar acquisition sequences produce highly transferable datasets regardless of the model architecture. Related to the "test of time" challenge, it is rarely investigated whether the training data actively acquired with one model will confer benefits if used to train a second model (as compared to randomly sampled data from the same pool). Given that datasets often outlive learning algorithms, this is an important practical consideration (Baldridge and Osborne, 2004; Lowell et al., 2019; Shelmanov et al., 2021).

4 Active Learning in *Simulated* vs. Real World Settings

> *Is it truly logical to consider an already cleaned (preprocessed), typically published open-source labeled dataset as an unlabeled data pool for pool-based active learning simulation, with the expectation that any conclusions drawn will be applicable to real-world scenarios?*

The convenience and scalability of simulation make it an undoubtedly appealing approach for advancing machine learning research. In NLP, when tackling a specific task, for instance summarization, researchers often experiment with the limited availability of labeled summarization datasets, aiming to gain valuable insights and improve summarization models across various domains and languages. While this approach may not be ideal, it is a practical solution. *What makes the sub-field of active learning different?*

Admittedly, progress has, and will be made in AL research by leveraging simulation environments, similar to other areas within machine learning. Thus, there is no inherent requirement for a radically different approach in AL. We believe that simulating AL is indispensable for developing new methods and advancing the state-of-the-art.

Nonetheless, we argue that a slight distinction should be taken into account. AL is an iterative process that aims to obtain the smallest possible amount of labeled *data* given a substantially larger pool of unlabeled data for maximizing predictive performance on a given task. The difference between developing models and constructing datasets lies in the fact that if a model is poorly trained, it can simply be retrained. Conversely, in AL, there exists a finite budget for acquiring annotations, and once it is expended, *there is no going back*. Consequently, we must have confidence that the AL state-of-the-art established through research simulations will perform equally well in practical applications.

Given these considerations, we advocate for a more critical approach to conducting simulation AL experiments. We should be addressing all the challenges (§2) and the experimental limitations (§3) discussed previously, while acknowledging the disparities between the simulation environment and real-world applications (§4.1). Given that datasets tend to outlast models (Lowell et al., 2019), we firmly believe that it is crucial to ensure the trustworthiness of AL research findings and their generalizability to real-world active data collection. This will contribute to the generation of high-quality datasets that stand the test of time (§3.4).

4.1 Simulation as a *Lower* Bound of Active Learning

The distribution gap between benchmark datasets in common ML tasks and data encountered in a real world production setting is well known (Bengio et al., 2020; Koh et al., 2021; Wang and Deng, 2018; Yin et al., 2021).

High Quality Data It is common practice for researchers to carefully curate the data to be labeled properly, often collecting multiple human annotations per example and discarding instances with disagreeing labels. When datasets are introduced in papers published in prestigious conferences or journals, it is expected that they should be of the highest quality, with an in-depth analysis of its data collection procedure, label distribution and other statistics. Nonetheless, it is important to acknowledge that such datasets may not encompass the entire spectrum of language variations encountered in real-world environments (Yin et al., 2021). Consequently, it remains uncertain whether an AL algorithm would generalize effectively to unfiltered raw data. Specifically, we hypothesize that the filtered data would be largely *more homogeneous* than the initial "pool". Assuming that the simulation $\mathcal{D}_{\text{pool}}$ is a somewhat homogeneous dataset, we can expect that *any* subset of data points drawn from it would, consequently, be more or less identical.[6] Therefore, if we train a model in each such subset, we would expect to obtain similar performance on test data due to the similarity between the training sets. From this perspective, random (uniform) sampling from a homogeneous pool can be considered a rudimentary form of diversity sampling.

Low Quality Data In contrast, it is possible that a publicly available dataset used for AL research may contain data of inferior quality, characterized by outliers such as repetitive instances, inadequate text filtering, incorrect labels, and implausible examples, among others. In such cases, an AL acquisition strategy, particularly one based on model uncertainty, may consistently select these instances for labeling due to their high level of data difficulty and uncertainty. Previous studies (Karamcheti et al., 2021; Snijders et al., 2023) have demonstrated the occurrence of this phenomenon, which poses a significant challenge as it undermines the potential value of AL. In a real-world AL scenario, it is plausible to have a dedicated team responsible for assessing the quality of acquired data and discarding instances of subpar quality. However, within the confines of a simulation, such data filtering is typically absent from the researcher's perspective, leading to potentially misleading experimental outcomes. Snijders et al. (2023) tried to address this issue in a multi-source setting for the task of natural language inference, and showed that while uncertainty-based strategies perform poorly due to the acquisition of collective outliers, when outliers are removed (from the pool), AL algorithms exhibited a noteworthy recovery and outperformed random baselines.

4.2 Simulation as an *Upper* Bound of Active Learning

However, one might argue for the exact opposite.

Favored Design Choices Previously, we mentioned that when selecting the seed dataset (§2.2) we typically randomly sample data from $\mathcal{D}_{\text{pool}}$, while keeping the label distribution of the true training set.[7] Hence, a balanced seed dataset is typically obtained, given that most classification datasets tend to exhibit a balanced label distribution. In

[6] Here we do not hint that all textual instances of a dataset are actually identical, but that they are more similar between them compared to the larger pool that they were created from.

[7] The "true training set" is the original one used as the pool ($\mathcal{D}_{\text{pool}}$) by removing the labels.

effect, the label distribution of $\mathcal{D}_{\text{pool}}$ would also be balanced, setting a strict constraint for AL simulation, as the actual label distribution of the unlabeled data should in reality be *unknown*. In other words, such subtle choices in the experimental design can introduce bias, making the simulated settings more trivial than more challenging real world AL settings where there is uncertainty as to the quality and the label distribution of data crawled online, that typically constitute the unlabeled pool.

Temporal Drift & Model Mismatch Datasets intended for research purposes are often constructed within a fixed timeframe, with minimal consideration for temporal concept drift issues (Röttger and Pierrehumbert, 2021; Lazaridou et al., 2021; Margatina et al., 2023b). However, it is important to recognize that this may not align with real-world applications, where the data distribution undergoes changes over time. The utilization of random and standard splits, commonly employed in AL research, can lead to overly optimistic performance estimates (Søgaard et al., 2021), which may not generalize to the challenges presented by real-world scenarios. Consequently, practitioners should consider this limitation when designing their active learning experiments. Lowell et al. (2019) also raises several practical obstacles neglected in AL research, such as that the acquired dataset may be disadvantageous for training subsequent models, and concludes that academic investigations of AL typically omit key real-world considerations that might overestimate its utility.

4.3 Main Takeaways

In summary, there exist compelling arguments that support both perspectives: simulation can serve as a lower bound by impeding the true advancement of AL methods, or it can implicitly favor AL experimental design, thus providing an upper bound for evaluation. The validity of these arguments likely varies across different cases. We can claim with certainty that this simulation setting, as described in this paper, is a far from perfect framework to evaluate AL algorithms among them and against random sampling. Nevertheless, we hypothesize that the lower bound argument (§4.1) might be more truthful. It is conceivable that AL data selection approaches may exhibit similar performance levels, either due to a lack of variation and diversity in the sampled pool of data or due to the presence of outliers that are not eliminated during the iterations. Hence, we contend that *simulation can be perceived as a lower bound for AL performance*, which helps explain why AL methods struggle to surpass the performance of random sampling. We undoubtedly believe that we can only obtain such answers by *exploring the AL simulation space in depth and by performing thorough analysis and extensive experiments to contrast the two theories.*

4.4 Active Learning in the LLMs Era

The field of active learning holds considerable importance in the current era of Large Language Models (LLMs). AL has recently been explore as a framework to identify the most useful demonstrations for in-context learning with LLMs (Zhang et al., 2022c; Diao et al., 2023; Margatina et al., 2023a). Additionally, AL is inherently intertwined with data-driven approaches that underpin recent advancements in artificial intelligence, such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2023; OpenAI, 2022, 2023; Bai et al., 2022a). AL and RLHF represent two distinct approaches that tackle diverse aspects of the overarching problem of AI alignment (Askell et al., 2021). AL primarily focuses on optimizing the data acquisition process by selectively choosing informative instances for labeling, primarily within supervised or semi-supervised learning paradigms. On the other hand, RLHF aims to train reinforcement learning agents by utilizing human feedback as a means to surmount challenges associated with traditional reward signals. Despite their disparate methodologies, both AL and RLHF emphasize the criticality of incorporating human involvement to enhance the performance of machine learning and AI systems. Through active engagement of humans in the training process, AL and RLHF contribute to the development of AI systems that exhibit greater alignment with human values and demonstrate enhanced accountability (Bai et al., 2022a,b; Ganguli et al., 2022; Glaese et al., 2022; Sun et al., 2023; Kim et al., 2023). Consequently, the synergistic relationship between these two approaches warrants further exploration, as it holds the potential to leverage AL techniques in order to augment the data efficiency and robustness of RLHF methods.

5 Guidelines for Future Work

Given the inherent limitations of simulated AL settings, we propose guidelines to improve trustworthiness and robustness in AL research.

Transparency Our first recommendation is a call for transparency, which essentially means to *report everything* (Dodge et al., 2019). Every detail of the experimental setup, the implementation and the results, would be extremely helpful to properly evaluate the soundness of the experiments. We urge AL researchers to make use of the Appendix (or other means such as more detailed technical reports) to communicate interesting (or not) findings and problems, so that all details (§3) are accessible.

Thorough Experimental Settings We aim to incentivize researchers to thoughtfully consider ethical and practical aspects in their experimental settings. It is crucial to compare a wide range of algorithms, striving for generalizable results and findings across datasets, tasks, and domains. Moreover, we endorse research endeavors that aim to simulate more realistic settings for AL, such as exploration of AL across multiple domains (Longpre et al., 2022; Snijders et al., 2023). Additionally, we advocate for investigations into active learning techniques for languages beyond English, as the prevailing body of research predominantly focuses on English datasets (Bender, 2011).

Evaluation Protocol We strongly encourage researchers to prioritize the establishment of fair comparisons among different methods and to provide extensive presentation of results, including the consideration of variance across random seeds, in order to ensure robustness and reliability of findings. Generally, we argue that there is room for improvement of the active learning evaluation framework and we should explore approaches from other fields that promote more rigorous experimental and evaluation frameworks (Artetxe et al., 2020).

Analysis We place additional emphasis on the requirement of conducting comprehensive analysis of AL results. It is imperative to delve into the nuances of how different AL algorithms diverge and the extent of similarity (or dissimilarity) among the actively acquired datasets. It is incumbent upon AL research papers to extend beyond the results section and include an extensive analysis component, which provides deeper insights and understanding, as in Ein-Dor et al. (2020); Yuan et al. (2020); Margatina et al. (2021); Zhou et al. (2021); Snijders et al. (2023), among others. If we aim to unveil why an AL algorithm fails to outperform another (or the random baseline), we need to understand which data it selected in the first place, and why.

Reproducibility Reproducing AL experiments can be challenging due to the complex nature of a typical AL experiment, involving multiple rounds of model training and evaluation, which can be computationally demanding. However, we strongly advocate for practitioners and researchers to prioritize the release of their code and provide comprehensive instructions for future researchers aiming to build upon their work. By making code and associated resources available, the research community can foster transparency, facilitate replication, and enable further advancements in AL methodologies.

Efficiency Finally, we propose the release of actively acquired datasets generated by different AL algorithms, which would greatly contribute to data-centric research and interpretability aspects of AL. Particularly when employing AL with large-scale models, it becomes crucial to establish the actively acquired data from other studies as baselines, rather than re-running the entire process from the beginning. Such an approach would not only enhance transparency, but also promote efficiency and eco-friendly practices within the research community.

6 Conclusion

In this position paper, we examine the numerous challenges encountered throughout the various stages of the active learning pipeline. Additionally, we provide a comprehensive overview of the often-overlooked limitations within the AL research community, with the intention of illuminating obscure experimental design choices. Furthermore, we delve into a thorough exploration of the limitations associated with simulation in AL, engaging in a critical discussion regarding its potential as either a lower or upper bound on AL performance. Lastly, we put forth guidelines for future research directions, aimed at enhancing the robustness and credibility of AL research for effective real-world applications. This perspective is particularly timely, particularly considering the notable advancements in modeling within the field NLP (e.g., ChatGPT[8], Claude[9]) . These advancements have resulted in a shift of emphasis towards a more data-centric approach in machine learning research, emphasizing the significance of carefully selecting relevant data to enhance models and ensure their alignment with human values.

[8] https://openai.com/blog/chatgpt

[9] https://www.anthropic.com/index/introducing-claude

Limitations

In this position paper, we have strived to provide a comprehensive overview, acknowledging that there may be relevant research papers that have inadvertently escaped our attention. While we have made efforts to include a diverse range of related work from various fields, such as machine learning and computer vision, it is important to note that our analysis predominantly focuses on AL papers presented at NLP conferences. Moreover, it is worth mentioning that the majority, if not all, of the AL papers examined and referenced in this survey are centered around the English language, thereby limiting the generalizability and applicability of our findings and critiques to other languages and contexts. We wish to emphasize that the speculations put forth in this position paper carry no substantial risks, as they are substantiated by peer-reviewed papers, and our hypotheses (§4) are explicitly stated as such, representing conjectures rather than definitive findings regarding the role of simulation in AL research. We sincerely hope that this paper stimulates robust discussions and undergoes thorough scrutiny by experts in the field, with the ultimate objective of serving as a valuable guideline for AL researchers, particularly graduate students, seeking to engage in active learning research. Above all, *we earnestly urge researchers equipped with the necessary resources to conduct experiments and analyses that evaluate our hypotheses, striving to bridge the gap between research and real-world settings in the context of active learning.*

Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback. Both authors are supported by an Amazon Alexa Fellowship.

References

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active semi-supervised learning for improving word alignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, ALNLP '10, pages 10–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Philip Bachman, Alessandro Sordoni, and Adam Trischler. 2017. Learning algorithms for active learning. In *Proceedings of the International Conference on Machine Learning*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Barcelona, Spain. Association for Computational Linguistics.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on*

Empirical Methods in Natural Language Processing, pages 296–305, Singapore. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2020. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Proceedings of the Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16, pages 127–139.

Kianté Brantley, Amr Sharaf, and Hal Daumé III. 2020. Active imitation learning with noisy guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2093–2105, Online. Association for Computational Linguistics.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

Aron Culotta and Andrew Mccallum. 2005. Reducing labeling effort for structured prediction tasks. In *Association for the Advancement of Artificial Intelligence*.

Aswarth Abhilash Dara, Josef van Genabith, Qun Liu, John Judge, and Antonio Toral. 2014. Active learning for Post-Editing based incrementally retrained MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 185–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sajib Dasgupta and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709, Suntec, Singapore. Association for Computational Linguistics.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 619–628, New York, NY, USA. Association for Computing Machinery.

Melanie Ducoffe and Frédéric Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *CoRR*, abs/1802.09841.

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.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice

Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2223–2234.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *CoRR*, abs/1711.00941.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active learning on a budget: Opposite strategies suit high and low budgets. *CoRR*, abs/2202.02794.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423.

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*.

Junjie Hu and Graham Neubig. 2021. Phrase-level active learning for neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1087–1099, Online. Association for Computational Linguistics.

Rong Hu, Brian Mac Namee, and Sarah Jane Delany. 2010. Off to a good start: Using clustering to select the initial training set in active learning.

Hideaki Ishibashi and Hideitsu Hino. 2020. Stopping criterion for active learning based on deterministic generalization bounds. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 386–397. PMLR.

Fran Jelenić, Josip Jukić, Nina Drobac, and Jan Šnajder. 2023. On dataset transferability in active learning for transformers.

Jaeho Kang, Kwang Ryel Ryu, and Hyuk chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online. Association for Computational Linguistics.

Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. 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): System Demonstrations*, pages 115–120, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Nataliia Kees, Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2021. Active learning for argument strength estimation. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 144–150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback.

Andreas Kirsch, Tom Rainforth, and Yarin Gal. 2021. Test distribution-aware active learning: A principled approach against distribution shift and outliers.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Neural Information Processing Systems*, pages 7026–7037.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*.

Luke Kurlandski and Michael Bloodgood. 2022. Impact of stop sets on stopping active learning for text classification. *CoRR*, abs/2201.05460.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1874–1883, Melbourne, Australia. Association for Computational Linguistics.

Shayne Longpre, Julia Reisler, Edward Greg Huang, Yi Lu, Andrew Frank, Nikhil Ramesh, and Chris DuBois. 2022. Active learning over multiple domains in natural language tasks.

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.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023a. Active learning principles for in-context learning with large language models.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023b. Dynamic benchmarking of masked language models on temporal concept drift with multiple views.

In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2881–2898, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrew McCallum and Kamal Nigam. 1998. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 350–358, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Akiva Miura, Graham Neubig, Michael Paul, and Satoshi Nakamura. 2016. Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Mussmann, Robin Jia, and Percy Liang. 2020. On the importance of adaptive data collection for extremely imbalanced pairwise tasks.

Ansong Ni, Pengcheng Yin, and Graham Neubig. 2019. Merging weak and active supervision for semantic parsing. In *AAAI Conference on Artificial Intelligence*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4056–4066.

Zac Pullar-Strecker, Katharina Dost, Eibe Frank, and Jörg Wicker. 2022. Hitting the target: stopping active learning at the cost-based optimum. *Machine Learning*, pages 1–19.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Learning structured representations of entity names using Active Learning and weak supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6376–6383, Online. Association for Computational Linguistics.

Kun Qian, Lucian Popa, and Prithviraj Sen. 2017. Active learning for Large-Scale entity resolution. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1379–1388. ACM.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Roy and Andrew Mccallum. 2001. Toward optimal active learning through monte carlo estimation of error reduction. In *in Proceedings of the International Conference on Machine Learning*.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *CoRR*, abs/2008.07267.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

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

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Burr Settles, Mark Craven, and Soumya Ray. 2007. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA. Association for Computing Machinery.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the Workshop on Representation Learning for NLP*, pages 252–256.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Ard Snijders, Douwe Kiela, and Katerina Margatina. 2023. Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2187–2209, Dubrovnik, Croatia. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alex Tamkin, Dat Pham Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. 2022. Active learning helps pretrained models learn the intended task. In *Advances in Neural Information Processing Systems*.

Katrin Tomanek and Udo Hahn. 2009. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047, Suntec, Singapore. Association for Computational Linguistics.

Katrin Tomanek and Udo Hahn. 2010. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. 2009. On proper unit selection in active learning: Co-selection effects for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17, Boulder, Colorado. Association for Computational Linguistics.

Andreas Vlachos. 2006. Active annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Andreas Vlachos. 2008. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, Dublin, Ireland. Association for Computational Linguistics.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, and Hua Xu. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Wenpeng Yin, Shelby Heinecke, Jia Li, Nitish Shirish Keskar, Michael Jones, Shouzhong Shi, Stanislav Georgiev, Kurt Milich, Joseph Esposito, and Caiming Xiong. 2021. Combining data-driven supervision with human-in-the-loop feedback for entity resolution.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. AcTune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436, Seattle, United States. Association for Computational Linguistics.

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.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Xia Zeng and Arkaitz Zubiaga. 2023. Active PETs: Active data annotation prioritisation for few-shot claim verification with pattern exploiting training. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 190–204, Dubrovnik, Croatia. Association for Computational Linguistics.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. 2022a. Prompt-based rule discovery and boosting for interactive weakly-supervised learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 745–758, Dublin, Ireland. Association for Computational Linguistics.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022b. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022c. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022d. A survey of active learning for natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Fedor Zhdanov. 2019. Diverse mini-batch active learning.

Yilun Zhou, Adithya Renduchintala, Xian Li, Sida Wang, Yashar Mehdad, and Asish Ghoshal. 2021. Towards understanding the behaviors of optimal deep active learning algorithms. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1486–1494. PMLR.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section. The section is not numbered. It is in the end of the paper after Conclusion.

☑ A2. Did you discuss any potential risks of your work?
Limitations Section. The section is not numbered. It is in the end of the paper after Conclusion.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Not applicable. Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.