Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

ALDANER: Active Learning based Data Augmentation for Named Entity Recognition

Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì *, Andrea Vignali

Department of Electrical Engineering and Information Technology (DIETI), University of Naples “Federico II”, Via Claudio 21, Naples, Italy

ARTICLE INFO

Keywords:
Data augmentation
Named Entity Recognition
Active Learning

ABSTRACT

Training Named Entity Recognition (NER) models typically necessitates the use of extensively annotated datasets. This requirement presents a significant challenge due to the labor-intensive and costly nature of manual annotation, especially in specialized domains such as medicine and finance. To address data scarcity, two strategies have emerged as effective: (1) Active Learning (AL), which autonomously identifies samples that would most enhance model performance if annotated, and (2) data augmentation, which automatically generates new samples. However, while AL reduces human effort, it does not eliminate it entirely, and data augmentation often leads to incomplete and noisy annotations, presenting new hurdles in NER model training. In this study, we integrate AL principles into a data augmentation framework, named Active Learning-based Data Augmentation for NER (ALDANER), to prioritize the selection of informative samples from an augmented pool and mitigate the impact of noisy annotations. Our experiments across various benchmark datasets and few-shot scenarios demonstrate that our approach surpasses several data augmentation baselines, offering insights into promising avenues for future research.

1. Introduction

Named Entity Recognition (NER), which involves the identification of entity mentions such as person names, organizations, and locations within unstructured text data [1], holds pivotal significance across diverse domains including scientific discovery [2], machine translation [3], and question answering [4]. The recent strides in pre-trained language models [5,6] have ushered in notable enhancements in the performance of NER systems. Nonetheless, the efficacy of these techniques often hinges upon access to substantial volumes of manually annotated text data, necessitating a labor-intensive and expensive process, particularly in domains necessitating specialized expertise for accurate annotations from domain experts (e.g. healthcare, legal).

To mitigate the annotation burden, Active Learning (AL) methods [7–9] have been leveraged. These methods operate by interactively prompting users to annotate new instances from an unannotated pool, thereby minimizing the number of annotations required through the selection of the most informative samples. Such samples are identified based on their potential to significantly enhance model performance. For example, one common strategy — known as uncertainty-based AL — involves selecting samples that yield the most uncertain model predictions, exemplified by scenarios such as $P(y = 0|\mathbf{x}) = 0.51$ and $P(y = 1|\mathbf{x}) = 0.49$ in a binary classification problem.

While human effort is indeed reduced through Active Learning AL, manual annotation remains a necessity, often constrained by a limited budget. In scenarios characterized by a scarcity of annotated data, a.k.a. *few-shot learning*, a common strategy to address such situations is data augmentation. This technique expands the available training set by manipulating training instances without altering their labels [10, 11]. In prior studies focusing on sentence- and token-level classification [12,13], automatic heuristic rules (e.g., mention replacement, word swapping) have been employed to augment the original few-shot training data. However, these methods can introduce significant noise into the dataset, leading to the generation of grammatically and semantically incorrect samples. Such noise may misguide the model, ultimately proving detrimental to its performance.

To tackle the challenges outlined above, this study introduces a novel data augmentation technique named *Active Learning Data Augmentation for NER (ALDANER)*. As illustrated in Fig. 1, we employ the principles of active learning in a slightly different context. After the data augmentation step, which generates new samples based on the original few-shot training set, we use active learning to select the most informative samples from this augmented pool. While it is true that the generated data can often be very similar, our method ensures that the selected samples are the most diverse and informative as possible,

* Corresponding author.
E-mail addresses: vincenzo.moscato@unina.it (V. Moscato), macro.postiglione@unina.it (M. Postiglione), giancarlo.sperli@unina.it (G. Sperlì), andrea.vignali@unina.it (A. Vignali).

https://doi.org/10.1016/j.knosys.2024.112682
Received 6 February 2024; Received in revised form 11 October 2024; Accepted 25 October 2024
Available online 4 November 2024

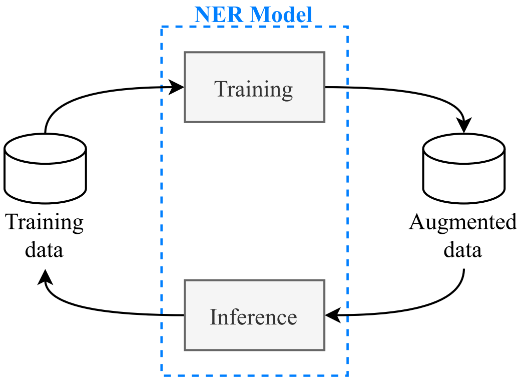


Fig. 1. Key idea in ALDANER. Active Learning is employed to select the most informative set of samples from an augmented pool to enlarge the training set and thus improve model performance.

given the data augmentation method chosen. This is because the active learning strategy we employ is designed to prioritize samples that are expected to bring the most benefit to the model's learning, based on their predicted uncertainty or informativeness. Furthermore, ALDANER can theoretically be applied on top of any data augmentation method, which could be designed to generate more diversified data.

Initially, the available low-resource training set is utilized to (1) train a NER model and (2) generate an augmented pool. Subsequently, AL cycles are executed, enabling the model to discern which augmented samples to incorporate, thus mitigating the impact of noisy annotations. We evaluate this approach using benchmark datasets from the biomedical domain, where the challenge of obtaining high-quality annotated corpora is particularly pronounced.

Our primary contributions are summarized as follows:

- We leverage an Active Learning (AL) pool-based methodology in the data augmentation domain, where the *pool* comprises augmented and labeled data, eliminating the need for manual intervention from a human oracle.
- We investigate the utilization of similarity metrics to retrieve as many new entities as possible while minimally altering the semantics of input samples.
- We conduct extensive experiments across three benchmark biomedical and three generic datasets, covering a range of few-shot scenarios. Our method demonstrates superiority over existing approaches and maintains original performance levels consistently.

The paper is structured as follows: Section 2 outlines the Related Work pertinent to the data augmentation task, while Section 3.1 formulates the challenged problem. In Section 3, we detail the proposed novel methodology for data augmentation. Section 4 presents all conducted experiments alongside the attained results. Finally, we discuss the conclusions and limitations of our approach in Section 6.

2. Related work

Natural Language Processing (NLP) has become a pivotal technology in the medical domain, enabling the extraction and analysis of valuable information from vast amounts of unstructured text data. This is crucial for enhancing patient care, streamlining clinical workflows, and advancing medical research. Recent studies have demonstrated the significant impact of NLP in various medical applications. For instance, NLP techniques have been successfully employed to analyze electronic health records (EHRs), aiding in the prediction of patient outcomes and the early detection of chronic diseases [14]. Additionally, NLP has been instrumental in automating the analysis of clinical notes, which helps in improving hospital triage systems and generating diagnostic models [14]. Moreover, the integration of NLP in healthcare has facilitated the extraction of insights from medical literature and patient feedback, thereby supporting personalized treatment plans and accelerating drug discovery [15]. These advancements underscore the importance of incorporating NLP methodologies in medical research and practice.

However, the annotation of datasets for NLP applications in healthcare is a time-consuming and costly process, posing significant challenges. The manual effort required for accurate and comprehensive annotation is substantial, often involving domain experts to ensure the quality and relevance of the annotated data. This process is not only labor-intensive but also financially burdensome, making it difficult to scale and sustain. To address these challenges, there is a pressing need for approaches that can effectively handle scenarios of data scarcity.

Few-shot Named Entity Recognition (NER) [16] pertains to the identification of entity mentions such as persons, locations, and organizations when the training dataset is severely constrained. Approaches to address few-shot scenarios can be categorized into two main strategies [17]: model-centric and data-centric. Model-centric methods primarily concentrate on refining model architectures to enhance performance under limited training data conditions. These approaches often involve architectural modifications, parameter tuning, or specialized training techniques tailored to mitigate the challenges posed by sparse datasets. On the other hand, data-centric strategies center on optimizing the data used for model training. These methods explore techniques for data augmentation, distant or weak supervision, aiming to enrich the training dataset and extract maximal utility from the available information. Both model-centric and data-centric approaches play vital roles in addressing the complexities of few-shot NER tasks, offering complementary avenues for improving model performance and adaptability in resource-constrained settings.

Among data-centric approaches, data augmentation has long been recognized as a pivotal technique in enhancing the robustness of machine learning models. While its application has been extensively explored for sentence-level classification [12,18], its potential in token-level classification tasks, such as Named Entity Recognition (NER), is only now being fully realized.

Foundational techniques, including Mention Replacement (MR), Label-wise Token Replacement (LwTR), and Synonym Replacement (SR) [13] leveraging resources like WordNet [19], laid the groundwork for contemporary advancements. Moreover, methods like Masked Entity Language Modeling (MELM) [20] and Cross-Domain Named Entity Recognition [21] have significantly enriched the augmentation landscape with more sophisticated strategies.

Recent studies have introduced innovative approaches tailored for low-resource NER settings. Yu et al. [22] proposed Embedded Prompt Tuning (EPT), embedding NER labels as prompts into pre-trained language models' hidden layers to address token-label misalignment and enhance data diversity. In a similar vein, Song et al. [23] introduced Robust Prompt-based Data Augmentation (RoPDA), leveraging continuous prompts and Self-Consistency Filtering to optimize the utilization of augmented samples, yielding substantial improvements over existing baselines. For Chinese NER, Liu et al. [24] presented LADA-Trans-NER, harnessing lexicons to construct directed graphs and fuse word information, coupled with advanced data augmentation techniques, showcasing notable performance enhancements across diverse datasets. Liang et al. [25] ventured into automatic speech recognition (ASR), introducing a novel data augmentation method utilizing a text-based speech editing model. Their approach yielded coherent and diversified augmented speech, demonstrating significant improvements in tasks like code-switching and NER. These advancements collectively underscore the evolving landscape of data augmentation techniques across various domains, propelling the robustness and adaptability of machine learning models.

Table 1
Related Work overview.

Authors	Year	Ref.	Method	Key contributions
Dai & Adel	2020	[13]	Survey on simple data augmentation methods for NER: label-wise token replacement, synonym replacement, mention replacement, shuffle within segments, hybrid.	Empirical comparison. Results show the potential of simple augmentation methods that can improve strong baselines trained with large-scale pretrained transformers.
Chen et al.	2021	[21]	Projects data from a high-resource domain to a low-resource domain by leveraging the patterns (e.g. style, abbreviations) in the text.	Allows the transfer of knowledge about style patterns from high-resource datasets.
Zhou et al.	2022	[20]	Creates augmented data by generating diverse entities via masked entity prediction	Sentence context and entity labels are jointly leveraged; handles token-label misalignment with a labeled sequence linearization strategy.
Yu et al.	2023	[22]	Embed NER labels in prompts to the hidden layer of pre-trained language models, which are thus predicted during the fine-tuning phase.	Addresses the problem of token-label misalignment while increasing entity diversity with an implicit label embedding and a novel position embedding.
Song et al.	2023	[23]	Linearizes NER annotations to prompts and uses a variety of augmentation operations (e.g. augmenting the entity-related span, changing the entity type, adding an entity) to generate augmented samples.	Augments data w.r.t. entities and contexts, thus improving the diversity of the augmented set; uses bidirectional masking to improve the quality of augmented samples.
Liu et al.	2023	[24]	Incorporates lexicon in a character representation layer and uses a GAN-based approach for data augmentation	Adds vocabulary information to the representation layer; alleviates the problem of entity-labeled data imbalance.

Despite advancements, challenges persist in few-shot scenarios where the effectiveness of data augmentation may diminish due to the potential generation of syntactically and semantically flawed samples [21]. Li et al. [26] have proposed a framework where data augmentation is used to enhance the active learning process, improving model performance in early rounds of active learning. Similarly, Le et al. [27] explore active learning in NER tasks, emphasizing the importance of aligning train and test data distributions for better performance, particularly in clinical domains. Differently from both approaches, our work introduces a paradigm shift: we leverage active learning to improve the results of data augmentation. By applying uncertainty-based heuristics to select samples from an augmented pool, we mitigate the noisy annotations typically introduced by traditional data augmentation techniques, thereby fully automating the process in few-shot scenarios.

Traditional AL methods [7,8,28] typically involve querying users to annotate new instances from an unannotated pool. In contrast, our work introduces a novel re-interpretation of the AL pool-based methodology. Specifically, in our approach, the "pool" consists of augmented and labeled data, thereby eliminating the need for manual annotations by a human oracle. This paradigm shift not only streamlines the annotation process but also enhances the efficiency and scalability of the NER model in few-shot scenarios. In Table 1 we show a summary of the related work.

3. Methodology

In this section, we show how ALDANER works by first augmenting the original few-shot training set with a similarity-based mention replacement method and then employing an active learning strategy to select the most informative samples from the augmented pool. An overview of the proposed framework is shown in Fig. 2.

3.1. Problem formulation

In this work, we propose a data augmentation technique to improve the performance of *NER* systems. We start with a small[1] corpus of annotated sentences $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}\}$, where:

- $i \in \{1, \dots, N\}$, N being the length of the dataset
- $\mathcal{X}$ is the set of sentences
- $\mathbf{x}_i$ is a sentence, which is defined as a sequence of tokens $x_j \in \mathbf{x}_i$, $j \in \{1, \dots, H_i\}$, where H_i is the sequence length
- $\mathcal{Y}$ is the set of labels. In our work, we will refer to the IOB2 annotation scheme [29], thus $\mathcal{Y} = \{B, I, O\}$, where B indicates the *beginning*, I the *inside* and O the *outside* of an entity mention
- $\mathbf{y}_i$ is a sequence of labels $y_j \in \mathbf{y}_i$, $j \in \{1, \dots, H_i\}$, where H_i is the sequence length. Each label is associated to a token in $\mathbf{x}_i$.

Based on this corpus, the objective of a NER model is to assign the correct label in $\mathcal{Y}$ to each token in an input sentence.

3.2. Similarity-based mentions replacement

The core idea in our method is to augment input samples without relying on external data sources and generating the most plausible examples given the available training data. In light of this, we first collect all the entity mentions occurring in the training dataset $\mathcal{D}_{train}$ in a concepts vocabulary $\mathbb{C}$, and then augment samples by replacing mentions with the most similar concepts in $\mathbb{C}$. In the following, we detail the steps required for the augmentation to be performed.

Concepts vocabulary collection Our data augmentation procedure relies on a comprehensive representation of entity mentions through a structured *concepts vocabulary* $\mathbb{C}$. This vocabulary acts as a centralized repository containing all unique entity mentions extracted from the training set $\mathcal{D}_{train}$. Each mention serves as a key entry within the vocabulary, ensuring a standardized representation of entities across the dataset.

The construction of the concepts vocabulary involves a meticulous process of aggregating and organizing entity mentions encountered in the training data. Duplicates are intentionally omitted from the lexicon, except in cases where they denote nested entities, such as *cancer* and *breast cancer*. This exclusion helps maintain the clarity and efficiency of the vocabulary while accommodating diverse linguistic expressions and hierarchical relationships among entities.

In specialized domains, such as biomedicine, where the diversity of entities is vast and constantly evolving, the concepts vocabulary may be further enriched using external knowledge resources. For instance, leveraging domain-specific ontologies like the Unified Medical Language System (UMLS) or Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) enables the inclusion of novel entities and enhances the vocabulary's coverage and relevance to specific application domains. While our current methodology primarily relies on internal dataset analysis for vocabulary construction, we acknowledge

[1] Literature does not provide a standard definition of few-shot learning. We consider scenarios where the ratio between training and test data is small.

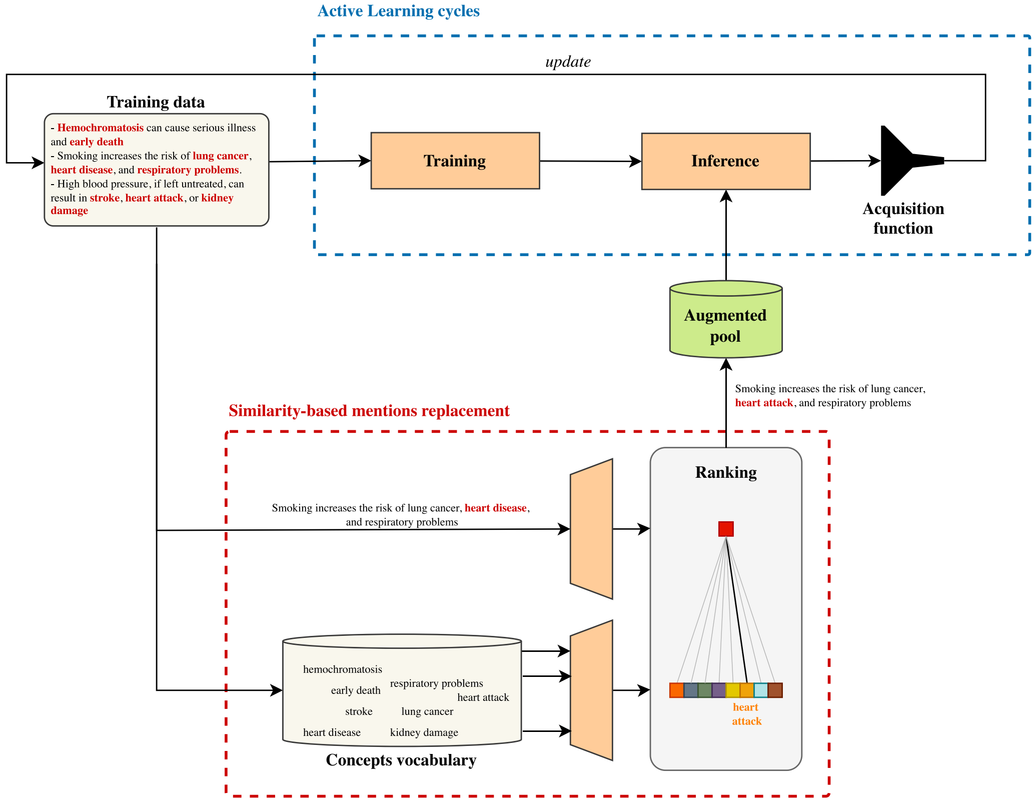

Active Learning cycles
update
Training data
- Hemochromatosis can cause serious illness and early death
- Smoking increases the risk of lung cancer, heart disease, and respiratory problems.
- High blood pressure, if left untreated, can result in stroke, heart attack, or kidney damage
Training
Inference
Acquisition function
Augmented pool
Smoking increases the risk of lung cancer, heart attack, and respiratory problems
Similarity-based mentions replacement
Smoking increases the risk of lung cancer, heart disease, and respiratory problems
Ranking
hemochromatosis
early death
respiratory problems
heart attack
stroke
lung cancer
heart disease
kidney damage
Concepts vocabulary
heart attack

Fig. 2. The ALDANER framework consists of two main components: Similarity-based mentions replacement and Active Learning cycles. In the Similarity-based mentions replacement section, entity mentions in the training dataset are replaced with similar concepts from a concept vocabulary. This process involves steps such as concepts vocabulary collection, entity mention replacement, and similarity-based ranking. The Active Learning cycles section outlines the process of iteratively selecting informative samples from the augmented dataset to enhance the Named Entity Recognition (NER) model. The cycle includes model training, model inference, and acquisition function steps.

the potential benefits of integrating external ontologies and plan to explore such enhancements in future iterations.

Our framework is designed to work also on datasets comprising multiple entity types. To facilitate this, separate concepts vocabularies $\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_m$ are maintained, with each vocabulary dedicated to a distinct entity type. This strategy ensures that entity mentions are organized and cataloged according to their semantic context, facilitating targeted data augmentation strategies tailored to different entity categories.

Entity mention replacement A sample $\mathbf{x}$ in the original training set $\mathcal{D}_{train}$ is augmented by replacing the mention of an entity with one of the concepts stored in the concepts vocabulary $\mathbb{C}$. For each of the training samples, the user can select the number of new instances n to be generated. A critical consideration in the augmentation process is striking a balance between data expansion and noise mitigation. To address this, the parameter n is set to 0.2 times the minimum size of the concepts vocabularies for all entity types $(|\mathbb{C}_1|, |\mathbb{C}_2|, \dots, |\mathbb{C}_m|)$. This setting represents a carefully calibrated trade-off that ensures a substantial increase in dataset size while minimizing the risk of introducing excessive noise into the augmented samples.

Similarity-based ranking The overarching goal is to enrich the dataset with plausible examples while preserving the contextual integrity and semantic coherence of the input samples. Thus, the augmentation strategy leverages semantic similarity to identify suitable replacements for entity mentions. Specifically, we encode each concept c with a transformer-based backbone network [5] to arrive at a fixed-sized vector representation $\mathcal{V}_c$. Then, we perform the same operation with the entity mention e occurring in the input sample, obtaining $\mathcal{V}_e$. Note that when a mention consists of multiple tokens, their embeddings are averaged to get the final representation.

When an entity mention e from the input sample has to be augmented, we consider its embedded representation $\mathcal{V}_e$ and rank each possible candidate $c \in \mathbb{C}$ (with $\mathcal{V}_c$ as its embedding) based on its cosine similarity with $\mathcal{V}_e$:

$$\cos(\mathcal{V}_c, \mathcal{V}_e) = \frac{\mathcal{V}_c \cdot \mathcal{V}_e}{\|\mathcal{V}_c\| \cdot \|\mathcal{V}_e\|} \tag{1}$$

The candidates to be used for data augmentation are thus chosen from the generated ranked list. Algorithm 1 shows a pseudo-code for the entire mention replacement procedure.

3.3. *Active learning cycles*

The underlying principle of AL is that the model would achieve better results with a smaller set of labeled data if it could choose which data to use [30]. The methodology we apply does not distort this principle but projects it into the different context of data augmentation, that is our final goal. After an augmentation system generates a number of augmented (and thus already annotated) examples, we want our system to choose the best ones to learn from.

The unannotated pool will no longer be needed as it will be replaced by sentences resulting from the augmentation process. Considering that our procedure entails that the augmented sentences will already be labeled, it is possible to diminish the effort made by the human oracle through an automatic procedure that relies on a heuristic to determine whether a sentence should be included or not during the active learning cycle.

Starting from the initial corpus of annotated data, a lump sum augmentation process will be performed to generate a pool of augmented data from which we will draw the most significant examples during AL cycles, which consist of three steps:

1. *Model Training.* The NER model is trained by using the available annotated dataset $\mathcal{D}$. At the first iteration, $\mathcal{D}$ corresponds to the

Algorithm 1 Similarity-based Mentions Replacement

```
Concepts Vocabulary Collection
for each entity type $i$ from 1 to $m$ do
    Initialize an empty concepts vocabulary $\mathbb{C}_i$.
    for each sample $\mathbf{x}$ in the training set $\mathcal{D}_{train}$ do
        Extract entity mentions of type $i$ from $\mathbf{x}$.
        Add unique mentions to $\mathbb{C}_i$, except for nested entities.
    end for
end for
Combine all $\mathbb{C}_i$ to form the unified concepts vocabulary $\mathbb{C}$.
Similarity-based Entity Mention Replacement
Determine the number of new instances $n$: $n = 0.2 \times \min(|\mathbb{C}_1|, |\mathbb{C}_2|, \dots, |\mathbb{C}_m|)$.
for each sample $\mathbf{x}$ in the training set $\mathcal{D}_{train}$ do
    for each entity mention $e$ in $\mathbf{x}$ do
        Obtain the embedding $\mathcal{V}_e$ for $e$.
        for each concept $c$ in $\mathbb{C}$ do
            Obtain the embedding $\mathcal{V}_c$ for $c$.
            Calculate the cosine similarity between $\mathcal{V}_c$ and $\mathcal{V}_e$.
            Rank the concepts based on their cosine similarity with $\mathcal{V}_e$.
        end for
        Select the $n$ top-ranked concepts to augment the training set by replacing $e$.
    end for
end for
```

Table 2
Hyperparameters of ALDANER and their definition.

Hyperparameter	Definition
k	Number of augmented samples selected by an activation function $f(\cdot)$ during each AL cycle
n	Number of augmented samples to generate for each augmentable sentence
b	Split of the augmented set to use during the training process
m	Number of entity classes

original training set, which is then gradually extended cycle after cycle.

2. *Model Inference.* The NER model is used to make predictions on the augmented pool and returns probabilities of tokens being classified with labels $y \in \mathcal{Y}$.
3. *Acquisition function.* A function $f(\cdot)$ chooses the most informative k samples from the augmented pool to be added to the annotated dataset $\mathcal{D}$.

Usually, the AL cycle repeats until reaching a *budget* which corresponds to the maximum annotation effort affordable for the project. In our work, we use some heuristics to define a budget: it represents the number of AL cycles to be performed by the framework and we compute it based on the length of the augmented dataset with the following equation

$$number\ of\ AL\ cycles = \max\left\{\frac{b\cdot|C|}{k}, 1\right\} \tag{2}$$

where C is the augmented set already generated, k is the number of examples to take from the augmented set in each cycle, and $b \in [0,1]$ is a hyperparameter used to control the number of AL cycles. The tuning of k and b affects both the quality of the results and the execution times. Table 2 summarizes the most relevant hyperparameters of our framework and the respective symbols used in this paper.

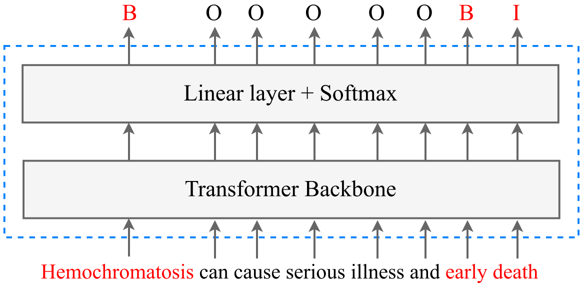


Fig. 3. Named Entity Recognition model. A Transformer Backbone network is used to compute embedded representations of input tokens, which are then classified with a linear layer followed by a softmax activation function. The input example here contains *disease* entity mentions.

3.4. Named entity recognition model

Following the recent advancements in self-supervised Pre-trained Language Models (PLMs) [5,31], we use a Transformer-based backbone network to extract the contextualized representation of each token x_j in an input sample $\mathbf{x}$, $\mathbf{z} = f_{\theta_{PLM}}(x_j)$, θ_{PLM} being the set of PLM parameters. Thereafter, a linear layer (a.k.a. *classification head*) with parameters $\theta_L = \{\mathbf{W}, \mathbf{b}\}$ project the Transformer-based representation $\mathbf{z}$ into the label space, $f_{\theta_L}(\mathbf{z}) = Softmax(\mathbf{Wz} + \mathbf{b})$. This pipeline is shown in Fig. 3. The model parameters are then optimized by minimizing cross-entropy:

$$\mathcal{L}_{CE} = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \sum_{i=1}^{H} KL\Big(y_i \Big| q(y_i|x_i)\Big), \tag{3}$$

where $KL(p|q)$ is the Kullback–Leibler divergence between the two distributions p and q, and q is the prediction probability vector for each token:

$$q(y|x) = Softmax(\mathbf{W} \cdot f_{\theta_{PLM}}(x) + \mathbf{b}) \tag{4}$$

3.5. Acquisition functions

We have chosen two simple acquisition functions available in current literature, both based on a measurement of *uncertainty*, i.e. we rank the augmented examples according to the uncertainty of the model in its predictions. Since model predictions are referred to each token in a sentence, we aggregate them to obtain a unique ranking value.

Least Confidence The *Least Confidence (LC)* criterion [32] ranks examples in descending order based on the probability of the model not predicting the most confident sequence of tags:

$$1 - \max_{y_1,y_2,\dots y_N} \mathbb{Q}(y_1, y_2, \dots, y_N|\mathbf{x}), \tag{5}$$

where $\mathbb{Q}(y_1, y_2, \dots, y_N|\mathbf{x})$ is the probability assigned by the model to the sequence of labels $y_1, y_2, \dots, y_N$, $y_i \in \{B, I, O\}$. Given the difficulty of computing this exact value, we rely on a greedy estimation [7]:

$$1 - \prod_{y_i\in\mathbf{y}} \Big(\max_{y_{pred}\in\{B,I,O\}} \mathbb{Q}(y_i = y_{pred}|\mathbf{x}) \Big), \tag{6}$$

where $\mathbb{Q}(y_i = y_{pred}|\mathbf{x})$ is the probability that the model assigns the label $y_i = y_{pred}$ to the token $x_i \in \mathbf{x}$.

Maximum Normalized Log-Probability Shen et al. [7] show that the LC method naturally favors longer sentences and propose an alternative approach, named *Maximum Normalized Log-Probability (MNLP)*, which consists in the following activation function:

$$\frac{1}{|\mathbf{y}|} \sum_{y_i\in\mathbf{y}} \log\Big(\max_{y_{pred}\in\{B,I,O\}} \mathbb{Q}(y_i = y_{pred}|\mathbf{x}) \Big), \tag{7}$$

where $|\mathbf{y}|$ is the number of elements in $\mathbf{y}$, thus corresponding to the sequence length.

Table 3
Statistics of the biomedical named entity recognition datasets in our experiments.

Dataset	Entity type	# Annotation	Train size	Dev size	Test size
NCBI-disease	Disease	6881	5425	924	941
BC5CDR	Chemical	15411	4561	4582	4798
BC2GM	Gene	20703	12575	2520	5039
Conll2003	Multiple	35089	14041	3250	3453
Ontonotes5	Multiple	113439	59924	13900	8262
Wikiann	Multiple	56035	20000	10000	10000

Table 4
Training set size in simulated few-shot scenarios. In brackets, the ratio between the training set and test set size.

Dataset	2%	5%	10%
NCBI-disease	108 (0.11)	271 (0.29)	542 (0.58)
BC5CDR	91 (0.02)	228 (0.05)	456 (0.09)
BC2GM	251 (0.05)	628 (0.12)	1257 (0.25)
Conll2003	281 (0.08)	702 (0.20)	1404 (0.41)
Ontonotes5	1198 (0.14)	2996 (0.36)	5992 (0.72)
Wikiann	200 (0.02)	500 (0.05)	1000 (0.10)

Bayesian Active Learning by Disagreement *The Bayesian Active Learning by Disagreement (BALD)* criterion [33] is computed as the mutual information between the model's parameters and the labels of the selected data point. It quantifies the expected reduction in uncertainty about the model's parameters when a particular data point is labeled with the following formula:

$$H[y|\boldsymbol{x}, \mathcal{D}] - \mathbb{E}_{\theta\sim(\theta|\mathcal{D})}[H[y|\boldsymbol{x}, \boldsymbol{\theta}]] \quad (8)$$

where the first term is the entropy of the label distribution in the current model, while the second term represents the expected entropy of the label distribution after observing the label of data point x, with θ representing the model parameters.

4. Experiments

To evaluate the effectiveness of our method, we conduct experiments on three popular benchmark datasets from the biomedical domain. The results show that our data augmentation method improves standard manipulation methods and achieves state-of-the-art results.

4.1. Datasets

We train and evaluate our model on six benchmark datasets. The first three are from the biomedical domain and contain a single entity type, while the second three contain multiple entity types and do not belong to a specific domain. Details are provided as follows:

- **NCBI-Disease** [34] consists of 793 PubMed abstracts, including 6881 Disease mentions.
- **BC5CDR** [35] consists of 1500 PubMed articles, including 15,935 Chemical and 12,852 Disease mentions. Since we test ALDANER with one entity type augmentation, we consider only chemical mentions in our experiments.
- **BC2GM** [36] contains 20,000 sentences from PubMed abstracts with 20,702 manually labeled gene entities.
- **Conll2003** [37] contains more than 14,000 sentences and 3 main entity types (PER, LOC, and ORG) which represent people, locations, and organizations. More generic entities of interest are identified with a miscellaneous label (MISC).
- **Ontonotes5** [38] is the fifth version of Ontonotes dataset which collects data from newswire, broadcast news, broadcast and telephone conversation, and web data in English, Chinese, and Arab. It contains 18 entity types that comprehensively identify people (PERSON), locations (LOC, FAC), organizations (ORG, GPE, NORP), objects (PRODUCT, WORK_OF_ART, MONEY), numbers (CARDINAL, ORDINAL, PERCENT, QUANTITY), time (TIME, DATE), and others (EVENT, LANGUAGE, LAW).
- **Wikiann** [39] is a multilingual dataset that contains a variable number of sentences for each one of the 176 languages covered. The English (en) dataset contains 40,000 sentences labeled with 3 different entity types (PER, ORG, and LOC) which represent people, organizations, and locations.

The statistics of the datasets are shown in Table 3.

4.2. Experimental setup

Computing infrastructure Our experiments have been performed on a local machine with a 13th Gen Intel(R) Core(TM) i7-13700K 3.40 GHz and an NVIDIA GeForce RTX 4080 GPU.

Few-shot scenarios To simulate few-shot scenarios, for each dataset, we consider randomly sampled subsets consisting of 2%, 5%, and 10% of the original corpus. Table 4 shows the size of the resulting few-shot training sets and the ratio with the test set. It is worth noting that the ratio between the training set and test set is variable since we use the whole test set across all the experiments and reduce only the size of the training set.

Metrics For the evaluation of the quality of NER models, we used the $F1$ scores computed with the seqeval[2] Python framework. In particular, scores are computed as in the following:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \quad (9)$$

where:

$$Precision = \frac{\text{\# correct entities}}{\text{\# predicted entities}} \quad (10)$$

$$Recall = \frac{\text{\# correct entities}}{\text{\# ground truth entities}} \quad (11)$$

Active Learning details For each training corpus and few-shot scenario, we run a number of iterations of our AL cycle as described by Eq. (2), where k is the number of examples used in each cycle and b is a control parameter needed to double the starting training set size. The model used to perform the active learning cycle is a BiLSTM with a CRF head that is considerably faster than a transformer. We trained the Active Learning model for 15 epochs for each cycle with a learning rate of 10^{-3}, an Adam optimizer and using the cross-entropy loss.

Training details Given the biomedical nature of training corpora for the single entity datasets, we use BioBERT [40] as the last classifying network for our methods, while for the multiple entities datasets we use BERT. Following previous work on few-shot learning [41], we assume the access to a small training set without a large development set for hyperparameters tuning. Hence, we choose hyperparameters based on previous work and practical considerations. Specifically, we train all our models for 5 epochs with a learning rate of $2 \cdot 10^{-5}$, an AdamW optimizer [42], a batch size of 8 and a maximum sequence length of 512. Each model is trained five times with different seeds and average results are reported with the relative confidence interval with an $\alpha = 0.95$.

[2] https://github.com/chakki-works/seqeval

4.3. Comparison with baselines

Experimental details. In this section, we compare ALDANER with several baselines described as follows:

- **Vanilla**: we train with the original corpus.
- **Random Oversampling (RO)**: the training set is doubled in size by randomly oversampling the few-shot dataset.
- **Mention replacement (MR)** [13]: we randomly select a mention from the original training set with the same entity type for each mention in the instance. For example, given the sentence "*Breast cancer* can occur in both women and men", the entity mention *Breast cancer* may be replaced by *Diabetes*.
- **Label-wise token replacement (LwTR)** [13]: for each word within a sentence we randomly chose whether or not to replace it with any other word within the dataset which has the same label. For example, given the sentence "*Breast cancer* can occur in both women and men", the word *occur* may be replaced by *complain*.
- **Synonym replacement (SR)** [13]: for each word within a sentence, a binomial distribution chooses whether or not to replace it with a synonym found using WordNet [19]. For example, given the sentence "*Breast cancer* can occur in both women and men", the word *cancer* may be replaced by *tumor*.
- **Masked Entity Language Modeling (MELM)** [20]: a XLM-RoBERTa-base pre-trained model is fine-tuned on randomly masked entities of the linearized training set, then the same model is used to generate the augmented dataset.
- **LLaMa-2** [43]: We perform Parameter Efficient Fine-Tuning (PEFT) on the 7 billion parameters model using Quantized Low-Rank Adaptation (QLORA) [44]. The training is performed on the original corpus, and in both the training and inference phases, we provide context within the prompt and include the list of labels the model must use. Additionally, we postprocess the outputs to address any hallucinations by removing extra labels the model generates or by integrating 'O' labels if the model provides insufficient labels.

For all the baselines, we generate as many new samples as necessary to double the original training set size assuming that the number of augmented sentences is a hyperparameter not tunable due to the lack of a development set. ALDANER has been used in its best configuration (i.e. BALD acquisition function and similarity-based ranking for data augmentation). We use the same backbone network as the basis for fine-tuning for each baseline.

Results. Results in Tables 5 and 6 demonstrate a comparative analysis of the different methods on single-entity and multi-entity Named Entity Recognition (NER), respectively.

The tables reveal that the performance of the methods tends to improve as the percentage of shots increases, which is expected as more data generally leads to better model performance. On single-entity NER datasets, ALDANER appears to outperform the other methods in the 10% shot scenario across all datasets, suggesting its potential robustness in few-shot learning scenarios. Different methods exhibit varying degrees of effectiveness across the datasets. For instance, in the 2% shot scenario, the SR method performs best on NCBI-disease and BC2GM while the LwTR method performs best on BC5CDR. On the contrary, RO worsens the results of the single-entity NER dataset since it does not introduce new contexts or different sentences, leading to overfitting. This suggests that the choice of method might need to be tailored based on the specific dataset or task at hand.

The results on multi-entity NER illustrate that the LLaMa-2 baseline generally outperforms other methods across all three datasets and in all few-shot scenarios, securing the highest F1 scores. ALDANER tends to perform better on the Wikiann dataset, especially in the 10% few-shot scenario. However, ALDANER's performance appears to lag on the Ontonotes5 dataset across all few-shot scenarios when compared to other methods like LLaMa-2, SR and LwTR. The superior performance of ALDANER on the Wikiann dataset compared to the other two datasets could be attributed to certain inherent characteristics of Wikiann. The multilingual nature of Wikiann might present a more complex or rich feature space that ALDANER could be leveraging better than other methods, since removing noisy augmented samples may be useful to filter out sentences augmented with mentions from other languages. Conversely, the relatively more focused or specialized nature of Conll2003 and Ontonotes5, in terms of language and genre of text, might favor methods like SR and LwTR.

In both single and multi-entity contexts, ALDANER consistently outperforms most of the baselines on average. However, its performance slightly diminishes in the 2% few-shot scenario due to the insufficient presence of concepts, which hampers the effectiveness of the vocabulary collection like in the NCBI-Disease case. This issue is particularly evident in the multi-entity case (Conll2003, and Ontonotes5), where the abundance of entity types results in a reduction in the overall length of concept vocabularies, exacerbating the challenge further.

In our comparative analysis of ALDANER and LLaMa-2 results, an intriguing finding emerges. LLaMa-2, benefiting from its extensive pre-training stage — where it has been exposed to approximately 2 trillion tokens and over 1 million human annotations — usually achieves high performance in zero- and few-shot scenarios. In our experiments, LLaMa-2 outperforms other methods on the datasets reported in Table 6, which are general-purpose and align closely with the model's training domain, even with a hallucination rate between 4% and 6%. However, the gap between LLaMa-2 and ALDANER narrows on the Wikiann dataset, which contains multilingual data. Remarkably, ALDANER surpasses LLaMa-2 on domain-specific datasets, as highlighted in Table 5 since the hallucination rate of the LLM increases up to 10% across all the single-entity datasets. This underscores the value of ALDANER's meticulous data selection process, particularly in domain-specific scenarios — the primary context for few-shot learning — where its significance is amplified.

It is important to acknowledge the potential for overfitting due to the small scale of the training data. Indeed, this is a challenge that many machine learning models face, especially when dealing with limited datasets [16]. However, our method is designed to mitigate this issue. While the concept vocabulary and augmented samples are selected from existing training data, the key point is that these selections are not arbitrary. They are carefully chosen based on their potential to enhance the model's understanding of the task at hand. Moreover, it is important to note that the models trained in few-shot scenarios after the selection of augmented samples with ALDANER are tested on the entire original test sets. This means that despite the training data being limited, the models are evaluated on a broad and diverse set of examples.

4.4. Effects of active learning-based data augmentation

Experimental details. While we have implemented our framework by using the mention replacement approach for data augmentation, it could be applied on top of any data augmentation technique. In this section, we measure the performance improvement brought by employing AL for data augmentation by comparing the usage of the three heuristics (LC, MNLP, BALD) with the standard *mention replacement* approach, i.e. we augment the original training set by randomly replacing a mention from the input sentence with a mention from the Entity Lexicon with the same entity type. We used the 5% scenario of the NCBI dataset to compare the 4 methods and to choose the best one to use for the comparison with the baselines across all the scenarios and datasets.

Table 5
Comparison with baseline approaches on single-entity Named Entity Recognition. Best results across few-shot scenarios are reported in bold, while second-best results are underscored.

Few-shotscenario	Method	NCBIDisease	BC5CDR	BC2GM	Avg.
2%	Vanilla	0.651 ±0.122	0.792 ±0.067	0.644 ±0.031	0.696
	RO	0.526 ±0.119	0.642 ±0.134	0.52 ±0.045	0.563
	MR	0.666 ±0.084	0.813 ±0.032	0.640 ±0.020	0.706
	LwTR	0.677 ±0.101	**0.828** ±0.019	0.640 ±0.042	0.715
	SR	**0.692** ±0.103	0.814 ±0.026	0.665 ±0.038	**0.724**
	MELM	0.597 ±0.045	0.768 ±0.042	0.592 ±0.005	0.652
	LLaMa-2	0.567 ±0.104	0.188 ±0.666	0.607 ±0.05	0.454
	ALDANER$_{MR}$	0.639 ±0.14	0.824 ±0.031	**0.676** ±0.01	0.713
	ALDANER$_{LWTR}$	0.609 ±0.12	0.812 ±0.032	0.653 ±0.047	0.691
	ALDANER$_{SR}$	0.638 ±0.118	0.810 ±0.036	0.670 ±0.031	0.706
5%	Vanilla	0.735 ±0.041	0.850 ±0.020	0.711 ±0.012	0.765
	RO	0.634 ±0.028	0.754 ±0.023	0.597 ±0.035	0.662
	MR	0.743 ±0.048	0.849 ±0.021	0.713 ±0.006	0.768
	LwTR	0.743 ±0.072	0.860 ±0.039	0.699 ±0.012	0.767
	SR	**0.758** ±0.044	0.858 ±0.030	0.710 ±0.011	0.775
	MELM	0.676 ±0.054	0.816 ±0.008	0.643 ±0.012	0.711
	LLaMa-2	0.71 ±0.051	0.565 ±0.536	0.687 ±0.027	0.654
	ALDANER$_{MR}$	0.751 ±0.055	**0.870** ±0.019	**0.781** ±0.003	**0.801**
	ALDANER$_{LWTR}$	0.696 ±0.056	0.863 ±0.021	0.75 ±0.012	0.770
	ALDANER$_{SR}$	0.737 ±0.049	0.866 ±0.02	0.774 ±0.006	0.792
10%	Vanilla	0.791 ±0.028	0.875 ±0.013	0.759 ±0.017	0.808
	RO	0.711 ±0.043	0.805 ±0.028	0.661 ±0.016	0.726
	MR	0.794 ±0.018	0.874 ±0.034	0.754 ±0.010	0.807
	LwTR	0.789 ±0.023	0.882 ±0.017	0.741 ±0.012	0.804
	SR	0.803 ±0.033	0.883 ±0.018	0.763 ±0.012	0.816
	MELM	0.749 ±0.024	0.852 ±0.013	0.681 ±0.006	0.761
	LLaMa-2	0.779 ±0.025	0.627 ±0.62	0.741 ±0.024	0.716
	ALDANER$_{MR}$	**0.811** ±0.010	**0.912** ±0.015	0.819 ±0.008	**0.847**
	ALDANER$_{LWTR}$	0.793 ±0.022	0.905 ±0.013	0.783 ±0.012	0.827
	ALDANER$_{SR}$	0.806 ±0.023	0.904 ±0.013	**0.825** ±0.013	0.845

Table 6
Comparison with baseline approaches on multi-entity Named Entity Recognition. Best results across few-shot scenarios are reported in bold, while second-best results are underscored.

Few-shotscenario	Method	Conll2003	Ontonotes5	Wikiann	Avg.
2%	Vanilla	0.693 ±0.047	0.746 ±0.009	0.507 ±0.079	0.649
	RO	0.745 ±0.403	0.751 ±0.045	0.58 ±0.16	0.692
	MR	0.666 ±0.091	0.739 ±0.039	0.402 ±0.121	0.602
	LwTR	0.725 ±0.044	0.753 ±0.009	0.579 ±0.032	0.686
	SR	0.762 ±0.02	0.764 ±0.016	0.596 ±0.024	0.707
	MELM	0.572 ±0.043	0.469 ±0.333	0.382 ±0.042	0.474
	LLaMa-2	**0.828** ±0.069	**0.832** ±0.042	**0.712** ±0.034	**0.791**
	ALDANER$_{MR}$	0.731 ±0.038	0.651 ±0.034	0.596 ±0.046	0.659
	ALDANER$_{LWTR}$	0.674 ±0.031	0.637 ±0.034	0.55 ±0.02	0.620
	ALDANER$_{SR}$	0.707 ±0.025	0.636 ±0.031	0.579 ±0.036	0.641
5%	Vanilla	0.811 ±0.017	0.797 ±0.042	0.665 ±0.019	0.758
	RO	0.815 ±0.325	0.809 ±0.063	0.672 ±0.189	0.765
	MR	0.773 ±0.034	0.791 ±0.039	0.608 ±0.076	0.724
	LwTR	0.802 ±0.023	0.806 ±0.006	0.669 ±0.034	0.759
	SR	0.827 ±0.02	0.818 ±0.01	0.686 ±0.044	0.777
	MELM	0.683 ±0.124	0.552 ±0.384	0.523 ±0.031	0.586
	LLaMa-2	**0.883** ±0.013	**0.882** ±0.013	**0.784** ±0.012	**0.850**
	ALDANER$_{MR}$	0.813 ±0.03	0.802 ±0.016	0.737 ±0.008	0.784
	ALDANER$_{LWTR}$	0.799 ±0.056	0.788 ±0.023	0.703 ±0.016	0.763
	ALDANER$_{SR}$	0.809 ±0.042	0.801 ±0.015	0.725 ±0.014	0.778
10%	Vanilla	0.851 ±0.013	0.828 ±0.054	0.716 ±0.008	0.798
	RO	0.847 ±0.227	0.838 ±0.055	0.718 ±0.165	0.801
	MR	0.825 ±0.053	0.814 ±0.04	0.666 ±0.051	0.768
	LwTR	0.840 ±0.015	0.834 ±0.011	0.716 ±0.011	0.797
	SR	0.858 ±0.006	0.845 ±0.004	0.735 ±0.009	0.813
	MELM	0.716 ±0.019	0.580 ±0.403	0.587 ±0.008	0.628
	LLaMa-2	**0.901** ±0.003	**0.905** ±0.014	**0.812** ±0.013	**0.873**
	ALDANER$_{MR}$	0.855 ±0.021	0.831 ±0.013	0.767 ±0.008	0.818
	ALDANER$_{LWTR}$	0.851 ±0.012	0.828 ±0.01	0.738 ±0.042	0.806
	ALDANER$_{SR}$	0.858 ±0.012	0.829 ±0.014	0.767 ±0.014	0.818

Results. Fig. 4 shows our experimental results in terms of $F1$ score. The standard mention replacement technique is the most stable, but the mean result represented by the red line inside the box is among the lowest approaches. The best results are given by the BALD acquisition function which is lightly skewed toward high results, contrary to MNLP which is skewed toward a worse score. LC is the most stable among the acquisition functions and its results are very similar to MNLP without surpassing it significantly. In the end, the score difference between the

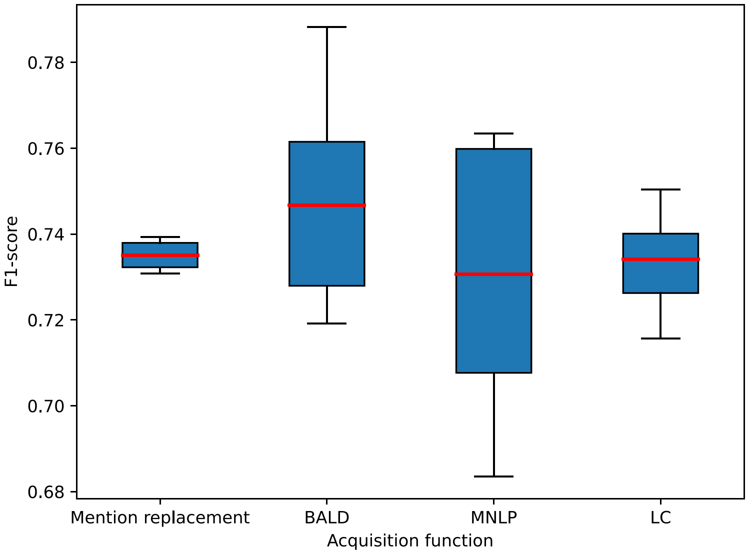


Fig. 4. Acquisition functions comparison. MR technique is used as a baseline to compare the three acquisition functions taken into account. The red line shows the mean of each box plot.

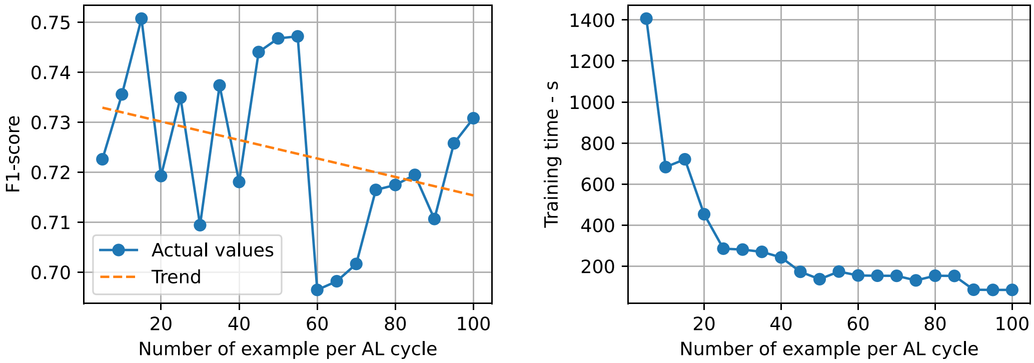


Fig. 5. The leftmost figure represents the performance of ALDANER (blue line) and its decreasing trend (orange dashed line) while increasing k. The rightmost figure shows the training times in seconds.

acquisition functions is between 1% and 2% with BALD being the best. As a result, we have chosen BALD as our acquisition function for further experimentation and analysis.

4.5. How many samples to be selected for each cycle

Experimental details. In this section, we explore the impact of the value k, i.e. the number of examples to be considered from the augmented set in each AL cycle with a fixed budget. This parameter affects both the quality of results and the execution times. The values of k used are $\{10, 15, 25, 50, 75\}$ to show the trend of the performance and the times. The dataset used, as in the previous section, is the 5% scenario of the NCBI.

Results. Fig. 5 shows our experimental results in terms of both quality and efficiency. In the leftmost figure, we can see how the performance becomes worse while increasing the number of samples for each AL cycle. Increasing k means selecting a wider range of uncertain samples during each active learning cycle. Thus, training the model on a coarse grain training set with a fixed budget lowers its precision and decreases the number of active learning cycles. Hence, in the rightmost figure, the training times decrease exponentially since the number of the AL cycles decreases with a bigger value of k and a fixed b.

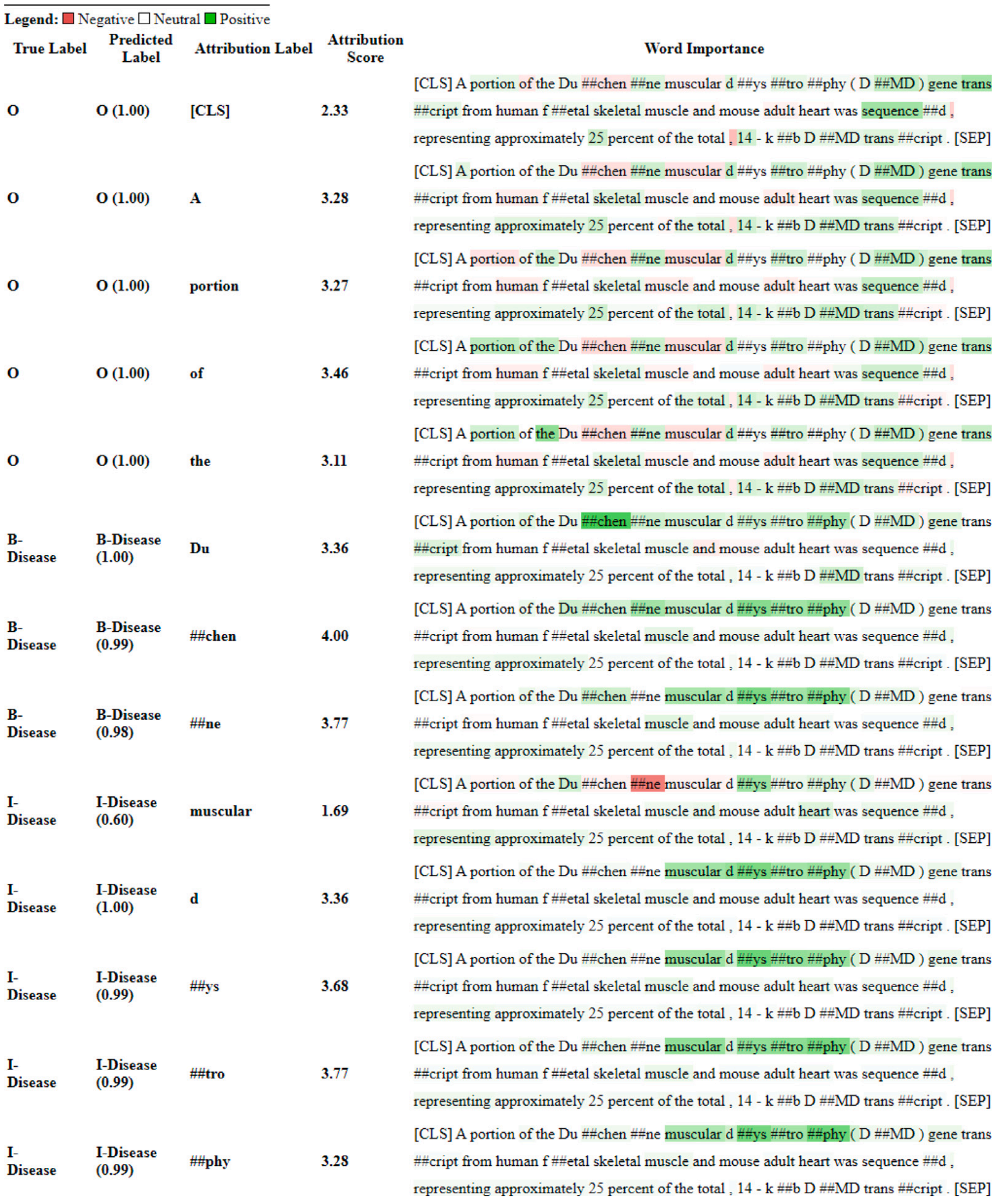
4.6. Effects of data augmentation on models interpretability

Experimental details. To enhance the comprehension of our methodology and demonstrate its capacity for elucidating underlying processes, we present interpretability results for sentences sourced from the NCBI-Disease and CoNLL datasets in Figs. 6, 7 and 8. This is achieved by employing the Integrated Gradients (IG) approach [45] to models trained under a 2% shots and 5% shots configuration, respectively.

Results. Based on an analysis of the sample from the NCBI corpus, the textual content in both the “Original” and “Augmented” sentences exhibits substantial similarity. In contrast, for the sample derived from the CoNLL corpus, there is a discernible alteration in mentions, leading to a pronounced difference between the sentences. However, in both instances, there is no significant disparity in the token-importance values between the “Original” and “Augmented” sentences. The visual representations for both consistently denote the same sequence and its pertinent role in the predictions.

5. Limitations

Our proposed system, Active Learning Data Augmentation for Named Entity Recognition (ALDANER), shows promise in few-shot NER tasks, but it has several limitations. Its performance is heavily influenced by the quality of the initial training data; biases or inconsistencies in this data can propagate during augmentation, leading



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
O	O (1.00)	[CLS]	2.33	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	A	3.28	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	portion	3.27	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	of	3.46	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	the	3.11	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
B-Disease	B-Disease (1.00)	Du	3.36	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
B-Disease	B-Disease (0.99)	##chen	4.00	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
B-Disease	B-Disease (0.98)	##ne	3.77	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.60)	muscular	1.69	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (1.00)	d	3.36	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.99)	##ys	3.68	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.99)	##tro	3.77	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.99)	##phy	3.28	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]

Fig. 6. Explainability analysis on original sample (NCBI-Disease 2% corpus). In the first column, the ground truth is presented. The subsequent column illustrates the predicted label corresponding to each token. The third column enumerates the tokens post-tokenization as executed by the model. The fourth column presents the attribution score. Lastly, the output from the Integrated Gradients (IG) algorithm pertaining to the specific token is delineated in the final column.

to degraded model performance. The effectiveness of ALDANER also depends on the selected data augmentation techniques, as some may introduce noise or semantic alterations that hinder the quality of augmented data. Additionally, incorporating active learning increases computational overhead, which can limit scalability, particularly in resource-constrained settings or with large datasets. Furthermore, while ALDANER aims to reduce overfitting, the risk remains, especially when augmented data is derived from a limited corpus, potentially leading to redundancy.

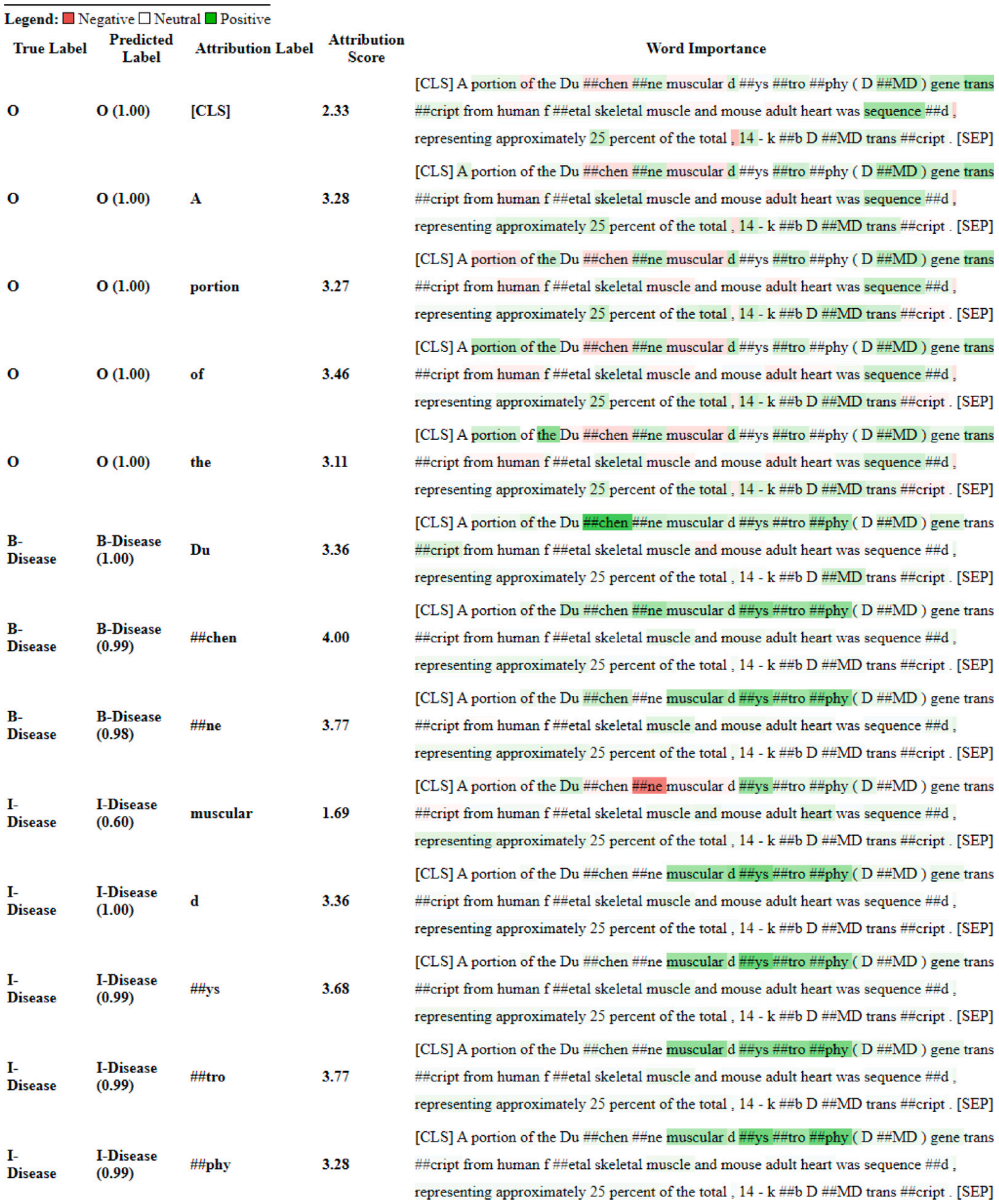
6. Conclusion & future work

In this paper, we introduced a novel data augmentation method, Active Learning Data Augmentation for NER (ALDANER), which seamlessly integrates the principles of Active Learning (AL) with data augmentation techniques to enhance the performance of Named Entity Recognition (NER) systems in few-shot scenarios. Our approach uniquely reinterprets the AL pool-based methodology, where the pool consists of augmented and labeled data, thereby eliminating the need for manual annotations by a human oracle.

Our experiments on various benchmark datasets, especially in the biomedical domain, demonstrated the superiority of ALDANER over traditional data augmentation methods. Notably, our method consistently outperformed other techniques, even in scenarios with extremely limited annotated data. This highlights the potential of our approach in real-world applications where obtaining large amounts of annotated data is challenging.

Furthermore, our method's ability to select the most informative samples from an augmented pool, based on uncertainty-based heuristics, proved to be a significant advantage. This not only reduced the effects of noisy annotations but also ensured that the model was trained on high-quality augmented data, leading to improved performance.

While ALDANER has shown promising results, there are several avenues for future work:



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
O	O (1.00)	[CLS]	2.33	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	A	3.28	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	portion	3.27	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	of	3.46	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
O	O (1.00)	the	3.11	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
B-Disease	B-Disease (1.00)	Du	3.36	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
B-Disease	B-Disease (0.99)	##chen	4.00	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
B-Disease	B-Disease (0.98)	##ne	3.77	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.60)	muscular	1.69	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (1.00)	d	3.36	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.99)	##ys	3.68	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.99)	##tro	3.77	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]
I-Disease	I-Disease (0.99)	##phy	3.28	[CLS] A portion of the Du ##chen ##ne muscular d ##ys ##tro ##phy (D ##MD) gene trans ##cript from human f ##etal skeletal muscle and mouse adult heart was sequence ##d , representing approximately 25 percent of the total , 14 - k ##b D ##MD trans ##cript . [SEP]

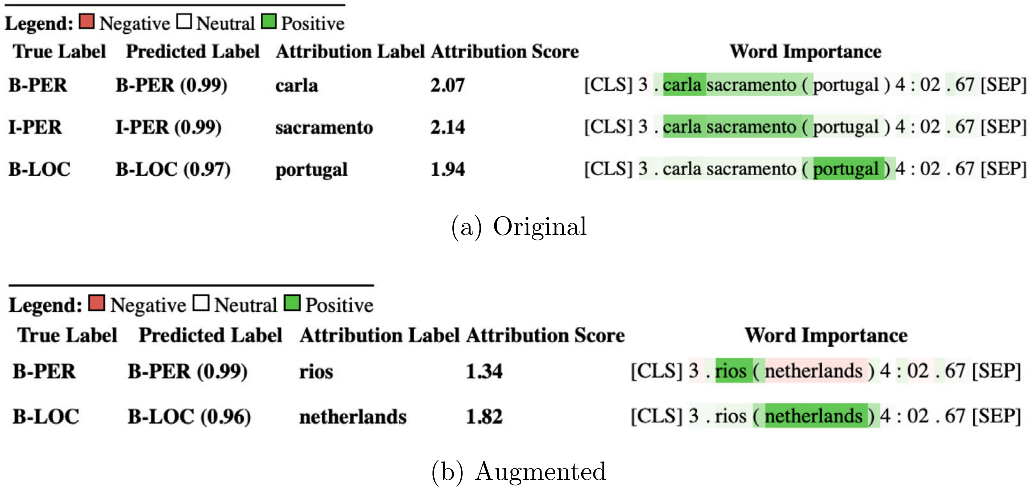
Fig. 7. Explainability analysis on augmented sample (NCBI-Disease 2% corpus). In the first column, the ground truth is presented. The subsequent column illustrates the predicted label corresponding to each token. The third column enumerates the tokens post-tokenization as executed by the model. The fourth column presents the attribution score. Lastly, the output from the Integrated Gradients (IG) algorithm pertaining to the specific token is delineated in the final column.

- *Integration with External Ontologies:* In domain-specific scenarios, such as the biomedical field, integrating external ontologies like UMLS or SNOMED-CT could further enhance the quality of the augmented data. This would allow the model to be exposed to a broader range of entities and relationships, potentially improving its generalization capabilities.
- *Exploration of Other Acquisition Functions:* While we employed uncertainty-based heuristics for sample selection, exploring other acquisition functions, especially those tailored for NER tasks, could lead to even better performance.
- *Scalability and Efficiency:* As datasets grow in size and complexity, optimizing the efficiency of the augmentation and active learning processes will be crucial. Future work could focus on developing more scalable algorithms and leveraging hardware accelerations.
- *Transfer Learning and Domain Adaptation:* Investigating how ALDANER can be combined with transfer learning techniques to adapt models trained on one domain to perform well on another could be a promising direction. This would be especially valuable in scenarios where data from the target domain is scarce.

In conclusion, ALDANER represents a significant step forward in the quest for high-performing NER systems in few-shot scenarios. By intelligently combining active learning principles with data augmentation, our method offers a robust solution to the challenges posed by limited annotated data. We believe that our work lays the foundation for future research in this area, with the potential to drive significant advancements in the field of NER and beyond.

CRediT authorship contribution statement

Vincenzo Moscato: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Marco Postiglione:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Giancarlo**



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-PER	B-PER (0.99)	carla	2.07	[CLS] 3 . carla sacramento (portugal) 4 : 02 . 67 [SEP]
I-PER	I-PER (0.99)	sacramento	2.14	[CLS] 3 . carla sacramento (portugal) 4 : 02 . 67 [SEP]
B-LOC	B-LOC (0.97)	portugal	1.94	[CLS] 3 . carla sacramento (portugal) 4 : 02 . 67 [SEP]

(a) Original

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-PER	B-PER (0.99)	rios	1.34	[CLS] 3 . rios (netherlands) 4 : 02 . 67 [SEP]
B-LOC	B-LOC (0.96)	netherlands	1.82	[CLS] 3 . rios (netherlands) 4 : 02 . 67 [SEP]

(b) Augmented

Fig. 8. Explainability analysis (CoNLL 5% corpus). In the first column, the ground truth is presented. The subsequent column illustrates the predicted label corresponding to each token. The third column enumerates the tokens post-tokenization as executed by the model. The fourth column presents the attribution score. Lastly, the output from the Integrated Gradients (IG) algorithm pertaining to the specific token is delineated in the final column.

Sperlì: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Andrea Vignali:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been funded by the project NextGenerationEU via PNRR - DM 352 (CUP: E66G22000400009).

Data availability

The authors do not have permission to share data.

References

[1] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. Knowl. Data Eng. 34 (1) (2022) 50–70, http://dx.doi.org/10.1109/TKDE.2020.2981314.

[2] X. Wang, V. Hu, X. Song, S. Garg, J. Xiao, J. Han, ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5227–5240, http://dx.doi.org/10.18653/v1/2021.emnlp-main.424, URL https://aclanthology.org/2021.emnlp-main.424.

[3] Z. Gekhman, R. Aharoni, G. Beryozkin, M. Freitag, W. Macherey, KoBE: Knowledge-based machine translation evaluation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3200–3207, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.287, URL https://aclanthology.org/2020.findings-emnlp.287.

[4] B.Z. Li, S. Min, S. Iyer, Y. Mehdad, W.-t. Yih, Efficient one-pass end-to-end entity linking for questions, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 6433–6441, http://dx.doi.org/10.18653/v1/2020.emnlp-main.522, URL https://aclanthology.org/2020.emnlp-main.522.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

[6] A. Akbik, T. Bergmann, R. Vollgraf, Pooled contextualized embeddings for named entity recognition, 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, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 724–728, http://dx.doi.org/10.18653/v1/N19-1078, URL https://aclanthology.org/N19-1078.

[7] Y. Shen, H. Yun, Z. Lipton, Y. Kronrod, A. Anandkumar, Deep active learning for named entity recognition, in: Proceedings of the 2nd Workshop on Representation Learning for NLP, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 252–256, http://dx.doi.org/10.18653/v1/W17-2630, URL https://aclanthology.org/W17-2630.

[8] M. Liu, W. Buntine, G. Haffari, 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1874–1883, http://dx.doi.org/10.18653/v1/P18-1174, URL https://aclanthology.org/P18-1174.

[9] J. Yao, Z. Dou, J.-Y. Nie, J.-R. Wen, Looking back on the past: Active learning with historical evaluation results, IEEE Trans. Knowl. Data Eng. 34 (10) (2022) 4921–4932, http://dx.doi.org/10.1109/TKDE.2020.3045816.

[10] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, 2017, CoRR. arXiv:1712.04621, URL http://arxiv.org/abs/1712.04621.

[11] J. Li, B. Chiu, S. Feng, H. Wang, Few-shot named entity recognition via meta-learning, IEEE Trans. Knowl. Data Eng. 34 (9) (2022) 4245–4256, http://dx.doi.org/10.1109/TKDE.2020.3038670.

[12] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388, http://dx.doi.org/10.18653/v1/D19-1670, URL https://aclanthology.org/D19-1670.

[13] X. Dai, H. Adel, An analysis of simple data augmentation for named entity recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3861–3867, http://dx.doi.org/10.18653/v1/2020.coling-main.343, URL https://aclanthology.org/2020.coling-main.343.

[14] S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, G.B. Kitchen, Natural language processing in medicine: A review, Trends Anaesth. Crit. Care 38 (2021) 4–9.

[15] R. Garg, A. Gupta, A systematic review of NLP applications in clinical healthcare: Advancement and challenges, in: International Conference on Advances in Data-Driven Computing and Intelligent Systems, Springer, 2023, pp. 31–44.

[16] V. Moscato, M. Postiglione, G. Sperlí, Few-shot named entity recognition: Definition, taxonomy and research directions, ACM Trans. Intell. Syst. Technol. 14 (5) (2023) 94:1–94:46, http://dx.doi.org/10.1145/3609483.

[17] V. Moscato, M. Postiglione, G. Sperlí, Few-shot named entity recognition: Definition, taxonomy and research directions, ACM Trans. Intell. Syst. Technol. (ISSN: 2157-6904) 14 (5) (2023) http://dx.doi.org/10.1145/3609483, URL https://doi.org/10.1145/3609483.

[18] J. Min, R.T. McCoy, D. Das, E. Pitler, T. Linzen, Syntactic data augmentation increases robustness to inference heuristics, in: Proceedings of the 58th Annual

Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2339–2352, http://dx.doi.org/10.18653/v1/2020.acl-main.212, URL https://aclanthology.org/2020.acl-main.212.

[19] G.A. Miller, WordNet: A lexical database for english, Commun. ACM 38 (1992) 39–41.

[20] R. Zhou, X. Li, R. He, L. Bing, E. Cambria, L. Si, C. Miao, MELM: Data augmentation with masked entity language modeling for low-resource NER, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2251–2262, http://dx.doi.org/10.18653/v1/2022.acl-long.160.

[21] S. Chen, G. Aguilar, L. Neves, T. Solorio, Data augmentation for cross-domain named entity recognition, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5346–5356, http://dx.doi.org/10.18653/v1/2021.emnlp-main.434, URL https://aclanthology.org/2021.emnlp-main.434.

[22] H. Yu, K. Ni, R. Xu, W. Yu, Y. Huang, EPT: Data augmentation with embedded prompt tuning for low-resource named entity recognition, Wuhan University Journal of Natural Sciences 28 (4) (2023) 299–308, http://dx.doi.org/10.1051/wujns/2023284299.

[23] S. Song, F. Shen, J. Zhao, RoPDA: Robust prompt-based data augmentation for low-resource named entity recognition, 2023, arXiv preprint arXiv:2307.07417.

[24] J. Liu, C. Liu, N. Li, S. Gao, M. Liu, D. Zhu, LADA-trans-NER: Adaptive efficient transformer for Chinese named entity recognition using lexicon-attention and data-augmentation, in: AAAI, 2023, http://dx.doi.org/10.1609/aaai.v37i11.26554.

[25] Z. Liang, Z. Song, Z. Ma, C. Du, K. Yu, X. Chen, Improving code-switching and named entity recognition in ASR with speech editing based data augmentation, 2023, arXiv preprint arXiv:2306.08588.

[26] Q. Li, Z. Huang, Y. Dou, Z. Zhang, A framework of data augmentation while active learning for Chinese named entity recognition, in: Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part II 14, Springer, 2021, pp. 88–100.

[27] L. Le, G. Demartini, G. Zuccon, G. Zhao, X. Zhang, Active learning with feature matching for clinical named entity recognition, Natural Lang. Process. J. 4 (2023) 100015.

[28] Q. Wang, W. Wu, Y. Qi, Y. Zhao, Deep Bayesian active learning for learning to rank: A case study in answer selection, IEEE Trans. Knowl. Data Eng. 34 (11) (2022) 5251–5262, http://dx.doi.org/10.1109/TKDE.2021.3056894.

[29] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995, URL https://aclanthology.org/W95-0107.

[30] B. Settles, Active learning literature survey, Technical Report 1648, University of Wisconsin–Madison, 2009, URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, URL http://arxiv.org/abs/1907.11692, cite arXiv:1907.11692.

[32] A. Culotta, A. McCallum, Reducing labeling effort for structured prediction tasks, in: Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2, AAAI '05, AAAI Press, 2005, pp. 746–751.

[33] N. Houlsby, F. Huszár, Z. Ghahramani, M. Lengyel, Bayesian active learning for classification and preference learning, 2011, arXiv:1112.5745.

[34] R.I. Dogan, R. Leaman, Z. Lu, NCBI disease corpus: A resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.

[35] J. Li, Y. Sun, R. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A.P. Davis, C. Mattingly, T. Wiegers, Z. lu, BioCreative V CDR task corpus: A resource for chemical disease relation extraction, Database 2016 (2016) baw068, http://dx.doi.org/10.1093/database/baw068.

[36] L.L. Smith, L.K. Tanabe, R.J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C.A. Struble, R.J. Povinelli, A. Vlachos, W.A. Baumgartner, L.E. Hunter, B. Carpenter, R.T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P.W. Adriaans, C. Blaschke, R. Torres, M.L. Neves, P. Nakov, A. Divoli, M. Maña-López, J. Mata, W.J. Wilbur, Overview of BioCreative II gene mention recognition, Genome Biol. 9 (2008) S2.

[37] E.F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning At HLT-NAACL 2003, 2003, pp. 142–147, URL https://aclanthology.org/W03-0419.

[38] S. Pradhan, A. Moschitti, N. Xue, H.T. Ng, A. Björkelund, O. Uryupina, Y. Zhang, Z. Zhong, Towards robust linguistic analysis using OntoNotes, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 143–152, URL https://aclanthology.org/W13-3516.

[39] A. Rahimi, Y. Li, T. Cohn, Massively multilingual transfer for NER, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 151–164, URL https://www.aclweb.org/anthology/P19-1015.

[40] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[41] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 255–269, http://dx.doi.org/10.18653/v1/2021.eacl-main.20, URL https://aclanthology.org/2021.eacl-main.20.

[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019, URL https://openreview.net/forum?id=Bkg6RiCqY7.

[43] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P.S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E.M. Smith, R. Subramanian, X.E. Tan, B. Tang, R. Taylor, A. Williams, J.X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023, arXiv:2307.09288.

[44] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, 2023, arXiv preprint arXiv:2305.14314.

[45] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3319–3328, URL http://proceedings.mlr.press/v70/sundararajan17a.html.