LARGE LANGUAGE MODELS FOR ARGUMENT MINING IN BIOMEDICAL LITERATURE

BY

MENGFEI LAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in School of Information Sciences
with a concentration in Biomedical Literature Processing
in the Graduate College of the
University of Illinois Urbana-Champaign, 2025

Urbana, Illinois

Doctoral Committee:

Associate Professor Halil Kilicoglu, Chair
Professor Catherine Blake
Associate Professor Vetle Torvik
Assistant Professor Haohan Wang

Table of contents

Abstract

Biomedical literature plays a crucial role for scientific knowledge acquisition. However, the vast volume of biomedical literature presents significant challenges in efficiently locating and extracting the needed information. Biomedical natural language processing (BioNLP) techniques, which facilitate quick processing of large-scale biomedical texts, becomes increasingly important to address this challenge. Most existing BioNLP works focus on extracting biomedical knowledge in the form of named entities or entity relationships from literature, but overlooking the argument roles—such as hypothesis and novel findings—played by the entities and their relationships in the paper. These argument roles could influence the extracted knowledge reliability (e.g. hypothesis) and interpretability. Therefore, biomedical argument mining, as the process of automatic identification and analysis of arguments within biomedical literature, should receive more attention to advance knowledge interpretation and extraction. Recent advancements in large language models (LLMs) have opened new opportunities for biomedical argument mining with their strong abilities in understanding complex semantic dependencies from text. This thesis explores how to leverage LLMs to advance biomedical argument mining. Three specific challenges in the field are analyzed: extracting rhetorical roles for sentences in biomedical abstracts, contextualizing claims through understanding of study limitations, and claim extraction within a rich context.

Chapter 1

INTRODUCTION

In the current era of biomedical research, scientific literature serves as a cornerstone for knowledge acquisition [1]. The term "biomedical literature" broadly covers various publications, including clinical trial reports, case reports, and systematic reviews. These publications play an important role in advancing scientific discoveries, guiding clinical practice, and supporting evidence-based decision-making. For example, drug discovery can be supported by biomedical research on the underlying mechanisms of diseases [2]; the integration of clinical trial results with related case reports can enhance the efficacy of clinical treatments in practice [3]; systematic reviews help systematically identify, assess, and summarize the current state of research topics [4].

Biomedical literature is an important resource for scientific tasks, but it is challenging to extract task-relevant information from the large-scale data. The PubMed database, a central repository for biomedical and life sciences publications, has indexed over 38 million articles and continues to grow by approximately 1 million articles every year [5]. Quickly pinpointing the relevant publication from this vast and continually expanding collection is challenging. Additionally, the average length of the body text in a biomedical article is 2,378 words [6], with much of it unrelated to specific tasks, making it even more difficult to efficiently locate the needed information.

Given the challenges of pinpointing knowledge from the vast volume of biomedical literature, natural language processing (NLP) techniques have become increasingly important to automatically extract knowledge [7]. First, NLP methods streamline information retrieval, enabling stakeholders to quickly access relevant studies and reducing the time and effort required for decision-making [8]. Additionally, automated methods enable the quick overview of biomedical publications at large scale, uncovering the hidden trends of a research topic (e.g. identify trends in methodology reporting [9]).

Existing studies have extensively analyzed challenges in biomedical literature processing that can be addressed using NLP methods. We refer to the application of NLP methods in biomedical text processing as Biomedical Natural Language Processing (BioNLP) in this thesis. In the paradigm of classical machine learning, lexical features (e.g., word counting, n-grams), syntactic structures (e.g., part-of-speech tagging), and task-specific metadata (e.g., MeSH terms, ontologies) were used as features for models such as SVMs and logistic regression (e.g., drug discovery [10] and disease prediction [11]). However, these explicitly created features could not capture complex sequential patterns in the data [12]. Next, the development of deep neural networks, such as RNNs and LSTMs, has enabled the direct processing of raw sequential data without manual selection of features (e.g., [13]–[18]). However, these methods struggle with long-distance

implicit semantic information extraction due to the vanishing gradient problem [19], [20]. Then the development of transformers with the self-attention mechanism has enabled capturing of long-range dependencies, assigning weights according to the importance of every word in a sequence, regardless of distance [20]. Early transformer-based models, such as BERT [21] and its biomedical adaptation PubMedBERT [22], brought significant advancements in biomedical NLP tasks, such as text classification and named entity recognition [23]–[25]. However, these models had relatively few parameters and limited scalability, and their performance still highly depends on the quality and availability of human-annotated training data [26], [27]. Moving forward, methods that enable deeper and more implicit semantic understanding beyond annotated labels are crucial for further progress in BioNLP.

Named Entity Recognition (NER) and Relation Extraction (RE) are the main areas of interest for BioNLP researchers, given their fundamental role in structuring biomedical knowledge and supporting downstream applications [24], [28], [29]. However, existing automated NER and RE methods primarily extract explicitly stated factual knowledge while overlooking argument structure and contextual factors that influence knowledge interpretation. Therefore, entities and their relationships are identified without considering their roles within the context, such as background details, experimental context, or novel findings. The extracted entities and relationships might be insufficient for downstream applications. For example, protein-disease relationships derived from hypothesis sentences in biomedical literature should not be treated as reliable knowledge, as relying on such information in decision-making may pose risks [30].

Given the importance of understanding argument structure in knowledge interpretation, argument mining - defined as the automatic identification and analysis of arguments in text, such as claim extraction, claim verification, and understanding of reasoning steps - becomes a crucial area for advancing knowledge extraction [31], [32]. Several studies have explored biomedical argument mining from different aspects. For claim extraction, existing research has focused on automatically identifying explicit claims by leveraging lexico-syntactic features (e.g. dependency grammar labels and symbols indicating upward or downward trends) to enhance the scientific communication [33]–[35]. In claim verification, studies have explored how to retrieve claim-related evidence from reliable sources to assess factuality of claims [36], [37]. Regarding argument structure identification, researchers have analyzed extracting the rhetorical roles of sentences within a sequence, using frameworks like BOMRC (Background, Objective, Methods, Results, and Conclusions) [38]–[41].

Though explorations have been made, argument mining in biomedical literature is still an underexplored area with several gaps. From the perspective of specific argument mining tasks, firstly, claim extraction could expand beyond explicit sentence-level claims to include implicit ones that may span multiple sentences [42]. Moreover, claim verification tasks could be improved by extracting both the evidence sentences from reliable sources and the contextual information surrounding the evidence to improve evidence interpretation [43]. Additionally, argument structure detection should consider the cases where a single textual pattern serves multiple argument roles [31]. From the perspective of general argument mining problems, establishing gold standards for argument mining datasets is always challenging due to the complexity and diversity of language in biomedical literature [44]. The use of specialized terminology, implicit relationships between concepts, and varied writing styles, makes it difficult to decide the dependencies and separate the argument patterns. Furthermore, annotating argument components in lengthy texts requires annotators to thoroughly read and fully comprehend entire documents, which is time- and resource-consuming [45].

In recent years, the development of large language models (LLMs) brought new opportunities for argument mining [46]. LLMs possess strong generalization and in-context learning abilities to adapt to a task with no

or few task-specific training examples. With the appropriate prompt design, the LLMs could be guided to understand and effectively perform the specific argument extraction requirements [47]. Additionally, LLMs' natural language understanding and reasoning ability enable them to effectively identify the argument flow and implicit relationships over long sequential textual patterns, which could be applied to improve argument mining tasks [48]–[50]. Given these properties of LLMs, it is worth to further explore the application of LLMs on biomedical argument mining.

In this thesis, I aim to explore the potential advancements brought with LLMs to address the challenges related to argument mining in biomedical literature.

1.1 Statement of the Problems and Study Purposes

This thesis explores the potential advancements introduced by LLMs to biomedical argument mining through analyzing three challenges in the field. The first challenge arises from the complexity of understanding of argument structures within a given context, as the dependencies between different textual components are intricate. The second challenge relates to the study limitations under which the claims are made, as the certainty and generalizability of the claims could be influenced by the study limitations mentioned in the literature. The third challenge regards accurately interpreting claims within a rich context, as surrounding contextual information can influence the claim's interpretation. Detailed explanations of each challenge and how the studies could address these challenges are provided in the following subsections.

1.1.1 Intricate Biomedical Argument Structure

Argument structure refers to the way claims, premises, and reasoning components are organized within the long text to form a coherent argument [51]. For biomedical argument structure analysis, the BOMRC argument framework (Background, Objective, Methods, Results, and Conclusions) is the main focus of researchers due to its reasonable flow and logical order [38]. With the growing number of biomedical publications each year, structuring complex paper content based on argument roles of textual patterns can be helpful to quickly locate the needed information within the vast body of literature, and benefit the downstream tasks such as fine-grained information retrieval [52] and extractive summarization [53]. For example, since main findings are typically summarized in the conclusion sentences of scientific papers, focusing on these sentences can enhance precise information retrieval [54].

Existing studies have used NLP techniques to extract the BOMRC argument structure from biomedical abstracts. Recurrent neural networks (RNNs)-based methods followed a hierarchical structure: an encoding layer to represent word tokens and embed sentences, followed by context interaction layers to enhance sentence representation considering surrounding context, and finally the label optimization layer to generate label output [13], [14], [16]–[18], [55]–[57]. Other works utilized the masked token objective of BERT [21], introducing special token to encode contextual information, based on which to predict the argument labels [39]. Despite the progress, gaps remained in these studies. ANN-based methods struggle with information loss problem when projecting the long context in a fixed-length vector, while BERT-based approaches are constrained by the 512-token input limit. Additionally, the existing methods rely on supervised learning, making their performance dependent on the size and quality of annotated training data. Moreover, previous works proposed single-label classification models, while no exploration has been made in multi-label classification though that is essential since a sentence can serve multiple rhetorical roles within context [58].

The goal of this study is to utilize LLMs to bridge the gaps with the current biomedical argument structure extraction tasks, including information loss, context length constraint, reliance on human-annotated data, and the need for multi-label classification paradigm. It explores the question: *How to leverage LLMs to address the current gaps in biomedical argument structure extraction*?

1.1.2 Contextualizing Biomedical Claims through Understanding of Study Limitations

Claims made in biomedical literature are conditioned on the factors such as the study background, experimental methods, underlying assumptions, and other elements that can impact the results. The limitations within these factors would weaken the strength of the claims made based on them. For example, if a paper that acknowledges the limitation of a small sample size in experiments, as in "our study had a small sample size", the strength of the claims made in this paper would be constrained [59]. Therefore, before accepting a claim, it is important to consider the limitations of the context based on which the claim was made, many of which may be self-identified by the authors in their publications [60].

Existing works have explored self-acknowledged limitations (SALs) from quantitative and qualitative aspects in randomized controlled trials (RCTs) publications, an important subcategory of biomedical literature that reports experiment results comparing the effects of different treatments or interventions. A quantitative analysis showed that approximately 73% of RCTs articles report SALs [61], indicating that direct extraction of SALs from RCTs articles is feasible. The qualitative analysis categorized SALs mentioned in RCTs articles into hierarchical levels, with top-level categories such as "Sample Size", which further includes second-level categories like "Recruitment Less Than Expected" and "Convenience Sampling" under the Sample Size category [62]. This structure provides a detailed analysis of the content and nature of the SALs. Automated tools have been built to capture the existence of SALs from the RCT articles [60]. However, no exploration has been made to build the automated SALs categorization tool. To gain a deeper understanding of the specific aspects that constrain the strength of claims from RCT articles, it is essential to create SALs type datasets, based on which to build automated tools and capture the types and detailed contents in SALs.

The goal of this study is to develop supervised models tuned with LLM-augmented datasets for classifying SALs, addressing the gap in facilitating a deeper understanding of specific weaknesses in study claims. It explores the question: *How can we effectively classify study limitations in RCTs?*

1.1.3 Context-dependent Biomedical Claim Interpretation

In knowledge-driven natural language understanding tasks, such as biomedical fact checking where evidence sentences are extracted from reliable sources to confirm or refute a claim, the extracted sentences are originally embedded in nuanced contexts that might influence their interpretation [43]. For instance, consider the claim: "Toxic algae can affect the nervous system, liver, and kidney in humans and animals." Without contextual information, the term "toxic algae" is overly broad, implying that any algae classified as "toxic" would support the claim. However, examining the surrounding context-such as "blue-green algae is poisonous"-makes it clear that "blue-green algae" definitely satisfies the claim. Therefore, it is essential to extract sentences along with their relevant context to make the extracted knowledge interpretable.

Explorations in enhancing knowledge representation by incorporating relevant contextual information into extracted sentences (defined as "decontextualization" by previous work [43]) have advanced in the general domain, but several gaps remain. Supervised methods relied on the manually curated dataset to fine-tune

seq2seq models [43]. However, the dataset was shaped by annotator's subjective judgments, limiting models' ability to address complex contextual variations. Unsupervised methods leveraging LLM inference [63]–[65] have been explored under different decontextualization settings but either failing to preserve the original sentence meaning or introducing external knowledge beyond the surrounding context of the target sentence. Moreover, these existing works have focused on the general domain, while decontextualization in the biomedical domain is unexplored.

The goal of this study is to develop automated tools for incorporating relevant contextual information using LLMs to bridge existing gaps in the field, with a specific focus on textual data from the biomedical domain. It seeks to answer the question: *How can LLMs be leveraged to enhance decontextualization for extracted sentences from biomedical literature?*

1.2 Structure of the Proposal

In this thesis, Chapter 2 discusses the work on LLM-supported biomedical argument structure extraction, Chapter 3 presents classification of SALs from biomedical literature using automated tools trained on LLM-augmented dataset, and Chapter 4 explores how to improve the interpretation of extracted biomedical sentences within context using LLMs. Finally, Chapter 5 outlines the timeline for completing the remaining tasks in each of the three topics.

Chapter 2

Multi-label Sequential Sentence Classification via Large Language Models

This chapter has been adapted from: Lan, M., Zheng, L., Ming, S., & Kilicoglu, H. (2024, November). Multi-label Sequential Sentence Classification via Large Language Model. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 16086-16104). My contribution to the study are (in the format of CRediT taxonomy): conceptualization, formal analysis, data curation, investigation, methodology, software, validation, visualization, writing-original draft, writing-review & editing.

2.1 Introduction

With the increasing number of published biomedical scientific papers today, researchers face significant challenges in quickly pinpointing needed information. To address this problem, organizing complex paper content according to the argument roles of each sentence in a structured format has garnered interest [59], [66], [67]. Since the roles of each sentence are often informed by the context from neighboring sentences, this task is referred to as sequential sentence classification (SSC) [39]. SSC can enable the fine-grained information retrieval [68], enhance extractive summarization [53], and improve other downstream tasks. For example, labeling objective sentences in biomedical abstracts can support information retrieval based solely on the papers' objectives.

Existing studies have explored SSC using contextualized language representations. Artificial neural network (ANN)-based SSC methods typically follow a hierarchical structure: an encoding layer to represent word tokens and embed sentences, a context interaction layer to enhance sentence embedding with surrounding context, and a labeling optimization layer to produce optimized sequential labels [13]–[18], [56], [57]. Other research utilizes the masked token objective or transformers, introducing special tokens to encode contextual information and using these tokens to predict sequential labels [39].

Despite promising progress in the SSC task in biomedical domain, several gaps remain, including pretrained model size, input sequence length, multi-label annotation, and dataset creation. Specifically, current ANN- and transformer-based methods have only employed moderately sized pretrained models (e.g.,

word2vec [14], SciBERT [39], [56]), while the application of large language models (LLMs) in SSC is underexplored. Furthermore, the existing transformer-based methods rely on BERT, which is constrained by 512-token sequence length limit [39]. Additionally, SSC has not expanded from single-label to multi-label, which is essential since a sentence can serve multiple rhetorical roles within a context [58]. Moreover, the widely used SSC dataset in the biomedical field, PubMed 200K RCT, is automatically generated from structured abstracts in PubMed [38]. However, this dataset does not include unstructured abstracts with free-form writing styles, which may deviate from the common patterns found in structured abstracts [16], [39].

To bridge these gaps, this chapter explores the application of LLMs in multi-label SSC using manually created datasets from biomedical domain. We propose LLM-SSC, a novel unified framework for in-context learning and parameter-efficient finetuning (PEFT) using Gemma-2b [69] for this task. Unlike existing approaches that create contextual representations of sequential sentences, LLM-SSC leverages LLMs to generate SSC labeling results based on designed prompts, which include a demonstration part to showcase the task and a query part to introduce the prediction target. To address the challenge of multi-label annotation, we design an auto-weighting multi-label contrastive learning loss that relaxes the constraint of formation of positive and negative pairs in the contrastive learning and reweights the importance of positive and negative pairs based on their label information.

Our contributions are as follows:

- We present LLM-SSC, the first LLM-based framework supporting both single- and multi-label SSC that integrates complete contextual information within the prompt and consider neighboring context when making predictions.
- We propose a novel multi-label contrastive learning loss with auto-weighting scheme to reweight the importance of negative pairs.
- We introduce and release BIORC800, a manually annotated multi-label SSC dataset mainly using unstructured abstracts from the biomedical field using rhetorical labels (Background, Objective, Methods, Results, Conclusions, and Other).
- Extensive experiments demonstrate the strong capability of LLM-SSC in SSC tasks under both in-context learning and parameter-efficient finetuning settings.

2.2 Related Works

SSC datasets SSC datasets are from various domains. PUBMED 20K RCT [70] and NICTA-PIBOSO [71] are two datasets generally used in biomedical domain. CSABSTRUCT [39] and CS-ABSTRACTS [16] are datasets utilizing abstracts from computer science papers. EMERALD 100K [72] and MAZEA [73] contains samples from multiple domains. In addition to these abstract-based datasets, some others use the full paper, such as DR. INVENTOR [74] collecting samples from the computer graphics domain and ART-CORESC [75] from physics, chemistry, and biochemistry domains.

SSC methods Before the deep learning paradigm, traditional machine learning algorithms were applied to SSC [76], [77]. These methods rely on hand-selected features and the classification performance is limited to the annotation amount and quality [57]. Inspired by transfer learning and deep learning bringing pre-learned knowledge from external large datasets and simulating human-like thinking, recent SSC works

leverage neural networks [13], [14], [18], [56], [57]. The current SoTA methods commonly follow a hierarchical framework [57], including an encoding layer to represent word tokens (e.g. Word2Vec [78]) or embed sentences (e.g., CNN [79]), followed by a context interaction layer to enrich the embedding using the surrounding context (e.g. Bi-LSTM [18]), and a labeling optimization layer to output the optimized sequential labels (e.g. CRF [17]). In addition to the hierarchical framework, a BERT-based work leverages the BERT self-attention mechanism to handle the variable-length text by attending to features in context [39].

Supervised Contrastive Learning with LLMs Contrastive learning objectives could be widely applied in supervised LLM tasks. In text classification, these objectives enhance performance by providing a clearer understanding of class boundaries [80]–[82]. For named entity recognition, leveraging labeled entity types to create positive and negative pairs helps the model distinguish between different entities more effectively [83]–[85]. In semantic similarity evaluation, supervised contrastive learning improves the model's ability to recognize subtle semantic differences, thereby boosting task performance [86], [87].

2.3 Methods

In this section, we first introduce the notation and then present LLM-SSC, an LLM-based framework for sequential sentences in-context learning and parameter-efficient finetuning, integrating complete contextual information within the prompt and consider neighboring context when making predictions. To enable the multi-label classification, we propose auto-weighting multi-label contrastive learning loss. The overview of the proposed framework is shown in Figure 2.1.

2.3.1 Notation

We approach SSC as a task of conditional text generation. Specifically, for an SSC dataset with S text sequences, we denote S_i as the i^{th} text sequence, S_{ij} as the j^{th} sentence in S_i, C_i as the context where the sentence is located ($C_i = concat(S_{i1}, S_{i2}, ...S_{in})$), and Y_{ij} as the SSC label of S_{ij}. Our goal is to model the probability of generating the SSC label Y_{ij}.

2.3.2 In-context Learning

We utilize in-context learning to leverage the power of LLMs for this task. An overview of the ICL framework is provided in Figure 2.1. A prompt is created by combining a demonstration context with a query, which is then fed into the language model to generate the prediction. The demonstration samples are selected from the training set based on cosine similarity scores between the training samples and the prediction target sentence. These similarity scores are calculated using embeddings generated by the SimCSE pre-trained model [86][1]. Given a demonstration sample D_i, the label Y_i for the i^{th} sentence in D_i, and the set of rhetorical label candidates U, the demonstration part of the prompt D_{prompt} is constructed as:

<Start>The paragraph is [D_i]. Select from rhetorical labels including [U], the sentence [D_{i1}] plays a rhetorical role as <[Y_{i1}]>, the sentence [D_{i2}] plays a rhetorical role as <[Y_{i2}]>, ..., the sentence [D_{in}] plays a rhetorical role as <[Y_{in}]><End>.

Then we create the query part of prompt. Given the prediction target sentence S_{ij} and the context C_i where the target sentence is located, the query portion of the prompt Q_{prompt} is formatted as:

[1] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

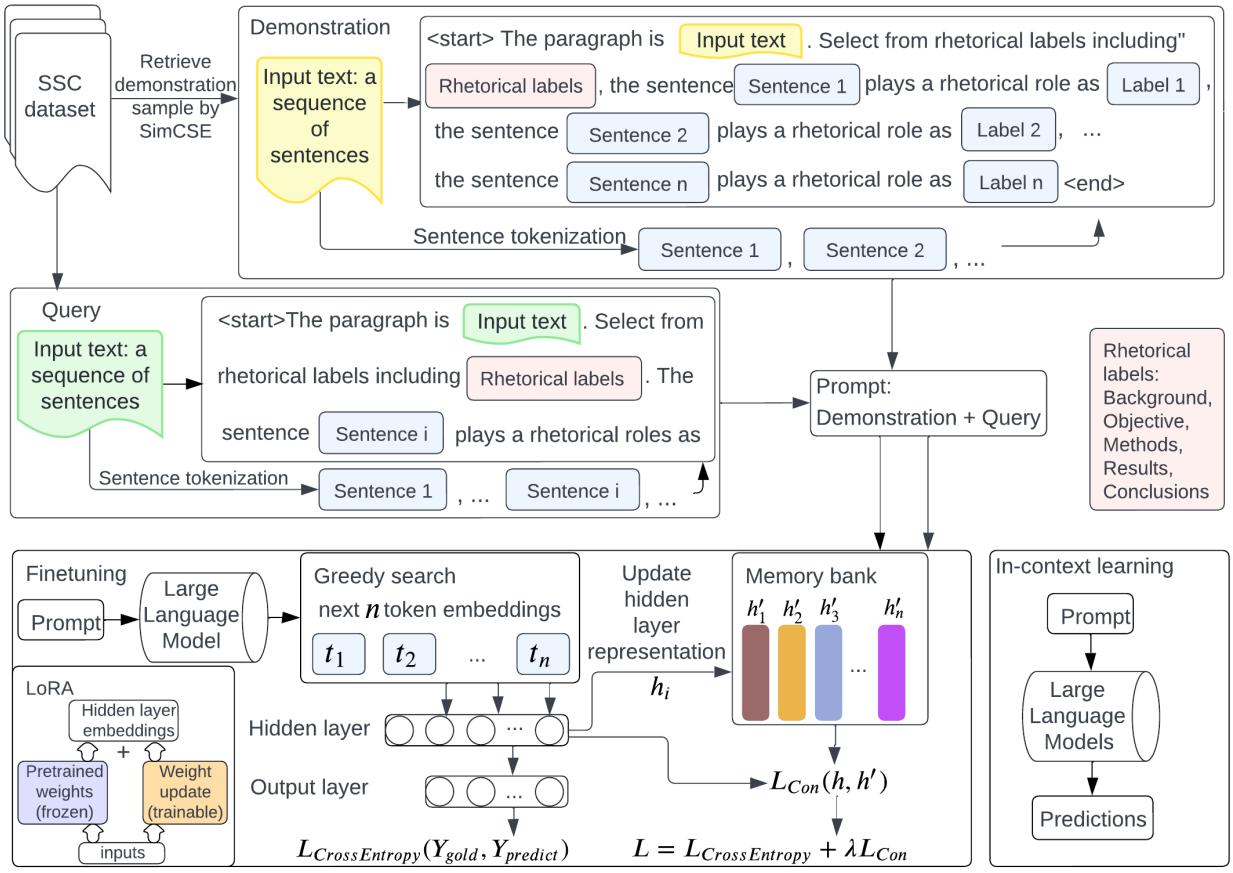

Figure 2.1: Structure of our LLM-based in-context learning and finetuning for SSC.

<Start>The paragraph is [C_i]. Select from rhetorical labels including [U], the sentence [S_{ij}] plays a rhetorical role as

The input prompt is constructed by combining the demonstration D_{prompt} and query Q_{prompt}:

$$X_{ICL} = D_{prompt} || Q_{prompt} \tag{2.1}$$

The goal for in-context learning is to generate the SSC label $Y_{predict}$ given the input prompt X_{ICL}:

$$Y_{predict} = \arg\max_{Y} P(Y|X_{ICL}) \tag{2.2}$$

where $Y_{predict}$ denotes the generated label that maximizes the conditional probability given the input prompt X_{ICL}. $P(Y|X_{ICL})$ denotes the conditional probability of generating outcome Y given the prompt X_{ICL}.

2.3.3 Task-specific Model Tuning

Although LLMs can recognize SSC labels using ICL due to their generalization ability without any parameter tuning, ICL underperforms the fine-tuning methods in text classification tasks [88]–[90]. To further explore the LLM application in SSC, we design a parameter-efficient fine-tuning framework of LLM. Figure 2.1

presents an overview of the fine-tuning framework.

Supervision with Demonstration. To bridge the gap between the pretrained model's original objective of predicting the next token and the goal of SSC to have the model generate the specific label for the classification target, we include one SSC demonstration within the input to guide the model's response. The format of the demonstrations and queries used in fine-tuning prompts $X_{finetune}$ is the same as that used in ICL (as described in Subsection 2.3.2). The tuning process modifies how the model adjusts the given demonstration and query within the prompt to generate appropriate token sequence $\hat{t}$:

$$\hat{t} = \arg\max_{t} P(t|X_{finetune}) \tag{2.3}$$

$$\hat{t} = \{t_1, t_2, ..., t_i, ...\} \tag{2.4}$$

where t_i denotes the hidden state of the ith token in the generated sequence.

Think Before Speak. Previous research found that giving space for the LLM model to produce additional tokens (delays) before generating the expected answer shows performance gains across various downstream tasks [91]. In our preliminary analysis, we observe a similar phenomenon. When employing the ICL approach outlined in Subsection 2.3.2, the model does not immediately generate the expected SSC label but first produces tokens not present in the label set. Motivated by this finding, we design the space-thinking mechanism [91] to provide some space for LLM to think before generating the expected answer.

The space-thinking mechanism requires LLM to generate the next n tokens after the prompt using greedy search. We assume that the predicted results are contained within one or more of these generated tokens. Therefore, we create a verbalizer to map the multiple generated tokens to the label space by concatenating the hidden states from the last layer of the generated tokens and feeding the combined results into a two-layer MLP. Specifically, given the prompt $X_{finetune}$, the next n token hidden states after the input context are generated as in Equation 2.3 and concatenated:

$$e_i = concat(t_1, t_2, ..., t_n) \tag{2.5}$$

Then a two-layer MLP is applied to map the concatenated representation to the label space:

$$h_i = ReLU(w_1 e_i + b_1) \tag{2.6}$$

$$p_{i,predict} = \sigma(w_2 h_i + b_2) \tag{2.7}$$

We use the cross entropy loss to compare the difference between the prediction probability $p_{i,predict}$ and the golden standard $y_{i,gold}$ for the i-th sequence, where N denotes the number of classes:

$$L_{CrossEntropy} = -\sum_{i=1}^{N} y_{i,gold} \log(p_{i,predict}) \tag{2.8}$$

Parameter-efficient Fine-tuning. Instead of fine-tuning all model parameters, we leverage low-rank adaptation (LoRA) method to tune the LLM in a parameter-efficient way [92]. LoRA keeps the pre-trained weights frozen and introduces trainable low-rank matrices in each layer of the transformer to approximate the weight updates needed for fine-tuning. It helps to reduce the computational cost and increase the memory efficiency during the tuning process.

2.3.4 Auto-weighting Multi-label Contrastive Learning

Supervised Contrastive Learning [93] has been widely employed in fine-tuning language models [94]–[96]. These methods typically construct positive and negative pairs based on the equivalence of label vectors in multi-class classification problems [97]. However, in the multi-label setting, treating two sentences with the same label vector as a positive pair is impractical due to the exponential growth in the number of unique label vectors with more labels (2^m unique label vectors for m binary labels), resulting in a scarcity of positive pairs for sentences with rare label vectors. In the worst-case scenario, it may be impossible to find two sentences with identical label vectors, thereby hindering the formation of positive pairs for contrastive learning. Additionally, minimizing the similarity of negative pairs in contrastive learning introduces the class collision issue [98], [99], where sentences with similar label vectors are erroneously pushed apart in the latent space, leading to sub-optimal solutions.

To address these issues, we propose an auto-weighting multi-label contrastive learning loss (WeighCon). Instead of requiring identical label vectors for positive pairs, we relax this constraint by forming positive pairs if two sentences share at least one common positive class. We introduce an auto-weighting scheme and propose the following multi-label contrastive learning loss:

$$L_{con} = -\sum_{c=1}^{m} E_i E_{j \in P_i(c)} \frac{\alpha_{ij} \text{sim}(h_i, h_j)}{\sum_k (1 - \alpha_{ik}) \text{sim}(h_i, h_k)} \tag{2.9}$$
$$\alpha_{ij} = \sigma(MLP(y_i, y_j))$$

Here, $\text{sim}(h_i, h_j) = \exp(\frac{h_i h_j^T}{|h_i||h_j|})$ is the exponential of the cosine similarity measurement, $\sigma(\cdot)$ is the sigmoid function, m is the number of labels, and $P_i(c) = \{j | y_i(c) = y_j(c) = 1\}$ represents the set of sentences with the same c^{th} label as the i^{th} sentence. The weighting function α_{ij}, parameterized by a one-layer MLP, takes two label vectors as input and outputs a scalar indicating the similarity between sentence representations h_i and h_j. Intuitively, α_{ij} is large when y_i and y_j are similar, peaking when they are identical. To mitigate the class collision issue, we use $1-\alpha_{ik}$ to reweight the importance of negative pairs in the denominator of Eq. 2.9. The weight (i.e., $1-\alpha_{ik}$) of a negative pair is large when label vectors differ significantly, decreasing as label vectors become more similar. By reducing the weights of negative pairs with similar label vector, we mitigate the negative impact of these negative pairs in minimizing the proposed contrastive loss. Furthermore, given the complexity and large number of parameters of the LLM, we incorporate supervised contrastive learning supported by a memory bank [100] into the training objective, thus reducing the memory requirement. The final loss function for task-specific model tuning contains two items and L_{Con} is weighted by a scaling factor λ (default 0.1):

$$L = L_{CrossEntropy} + \lambda L_{Con} \tag{2.10}$$

2.4 Experiments

In this section, we first present the a new dataset named BIORC800, a manually annotated multi-label SSC dataset mainly using unstructured abstracts from the biomedical field. Then, we evaluate the effectiveness of our proposed LLM-SSC by comparing it with state-of-the-art SSC methods and other contrastive learning based regularization. Additionally, we experiment on in-context learning setting and conduct an ablation

	Number of Structured Abstracts	Number of Unstructured Abstracts	Label Distribution		Average Sentences per Abstracts	Average Tokens per Sentence/Abstract
			Label	Distribution		
Train	60	420	BACKGROUND OBJECTIVE METHODS RESULTS CONCLUSIONS OTHER	805 (16.1%) 479 (9.6%) 1333 (26.6%) 1664 (33.2%) 654 (13.1%) 73 (1.5%)	9.86	22.94/225.19
Dev	20	140	BACKGROUND OBJECTIVE METHODS RESULTS CONCLUSIONS OTHER	228 (13.2%) 184 (10.7%) 521 (30.3%) 528 (30.7%) 243 (14.1%) 17 (1.0%)	9.98	22.97/229.26
Test	20	140	BACKGROUND OBJECTIVE METHODS RESULTS CONCLUSIONS OTHER	219 (13.3%) 164 (9.9%) 465 (28.1%) 565 (34.2%) 217 (13.1%) 22 (1.3%)	9.87	23.43/231.21
Total	100	700	BACKGROUND OBJECTIVE METHODS RESULTS CONCLUSIONS OTHER	1252 (14.9%) 827 (9.9%) 2319 (27.7%) 2757 (32.9%) 1114 (13.3%) 112 (1.3%)	9.89	23.03/227.80

Table 2.1: BIORC800 Detailed Statistics

study to further validate the assumptions outlined in the previous sections.

2.4.1 Datasets

Multi-label SSC Dataset: BIORC800 To enhance our multi-label sequential sentence classification (SSC) analysis and address the lack of manual SSC labels in unstructured biomedical texts, we manually annotated a corpus comprising 700 unstructured and 100 structured PubMed abstracts. Previous studies show that though sentences of the structured abstracts have author-assigned rhetorical categories, the categories might be erroneous [16], [39]. Therefore, we re-annotated those sentences from 100 structured abstracts to more accurately reflect their category. The collected biomedical abstracts were sampled from PubMed Central Open Access subset[2] using a modified version of Cochrane's sensitivity and precision-maximizing query. The annotation utilized the multi-label approach and followed the annotation schema of Background, Objective, Methods, Results, Conclusions, and Other.

We evaluated how consistently pairs of annotators agreed on sentence-level annotations using pairwise κ [101] over several stages. In the first stage, four annotators, who are also experts in biomedical text mining, used the first version of guideline to annotate the same 50 abstracts, with agreement scores between pairs ranging from 0.757 to 0.856. After discussing challenges and updating the guidelines, the second stage involved annotating a new set of 50 abstracts, improving the agreement scores to between 0.784 and 0.879. In the third stage, after further discussions and guideline updates, the remaining 700 abstracts were divided equally among the annotators. Finally, all 800 abstracts were combined, and one senior annotator reconciled the final set of labels. Compared to the author-assigned labels for the 100 structured abstracts, our re-annotation changed 4.1% sentence labels. We finally split the 800 abstracts into training (480 abstracts), development (160), and test sets (160), keeping the proportion of structured vs. unstructured abstracts the same in all three (12.5% - 87.5%). The descriptive statistics of BIORC800 are shown in Table 2.1.

[2]https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Single-label SSC Dataset In addition to the proposed **BIORC800** dataset, we test the models on the following three datasets in our experiments:

CS-ABSTRACT [16] contains 654 abstracts selected from computer science literature classified into Background, Objective, Methods, Results, and Conclusions sentences. It is the most recently published computer science RSC dataset annotated by crowdsourcing and collective intelligence[3].

PUBMED 20K RCT [38] contains 20k structured biomedical abstracts of randomized controlled trials with sentences automatically classified based on the author-assigned annotations as background, objective, method, result, or conclusion[4].

ART-CORESC [75] is a multi-domain dataset, containing sentence-level scientific discourse annotation for 265 full papers selected from physics, chemistry, and biochemistry fields. For the current SSC evaluation, we focus solely on the abstracts of these papers. In future work, we aim to extend the abstract-level SSC to the full-text level. The sentences in abstracts are annotated as background, hypothesis, motivation, objective, goal, methods, results, observation, experiment, or conclusion[5].

2.4.2 Baselines

The SSC methods that achieved state-of-the-art performance on SSC datasets and had publicly accessible code are selected as baselines for testing on our BIORC800 dataset. To adapt these methods, which were originally designed for single-label settings, to multi-label prediction, the code provided by the authors is modified by applying a threshold of 0.4 (chosen empirically to balance precision and recall for each label) to the output label probabilities.

Hierarchical Sequential Labeling Network (HSLN) [14] creates bi-RNN sentence representation, followed by attention-based pooling and a bi-LSTM layer to add contextual information from surrounding sentences. Finally, a CRF layer is concatenated to optimize the label sequence [6].

Sequential Sentence Classification (SSC) [39] employs BERT model [21] to encode both the semantics of the target sentence and the sequence's contextual information into a [SEP] token appended after the target sentence. This [SEP] token acts as the target sentence's representation, used to predict the rhetorical label[7].

Scientific Discourse Tagging (SDT) [56] uses token embeddings from SciBERT [102], an LSTM layer to encode sentences, and a bi-LSTM layer for sentence labeling, followed by a CRF layer with BIO tagging scheme to optimize the order of sequence labels[8].

SciBERT-HSLN [57] is built upon the HSLN model with SciBERT [102] as word embeddings[9].

The multi-label contrastive learning baseline is also used to compare to WeightCon:

[3] https://github.com/sergiog95/csabstracts
[4] https://github.com/Franck-Dernoncourt/pubmed-rct/tree/master/PubMed_20k_RCT
[5] https://live.european-language-grid.eu/catalogue/corpus/972/download/
[6] https://github.com/jind11/HSLN-Joint-Sentence-Classification
[7] https://github.com/allenai/sequential_sentence_classification
[8] https://github.com/jacklxc/ScientificDiscourseTagging
[9] https://github.com/arthurbra/sequential-sentence-classification

Dataset	0-shot		1-shot		5-shot		10-shot	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
BIORC800	0.476	0.322	0.642	0.507	**0.733**	**0.656**	0.159	0.068
CS-ABSTRACT	0.468	0.331	0.515	0.454	**0.581**	**0.562**	0.563	0.541
PUBMED 20K RCT	0.171	0.131	0.642	0.546	**0.712**	**0.659**	0.579	0.528
ART-CORESC	0.064	0.029	0.207	0.100	0.193	0.103	**0.217**	**0.102**

Table 2.2: In-context learning results with different number of demonstrations (shots).

HeroCon [98] is designed for multi-view and multi-label learning that applies weight to positive and negative label pairs by hamming distance of two label representations[10].

2.4.3 Experimental Setup

Implementation and Evaluation The Gemma-2b [69] model is selected as the backbone due to its lightweight design and advanced performance across various natural language tasks. This model supports an input sequence length of up to 8192 tokens, which is adopted as the maximum length for in-context learning. For fine-tuning, however, we limit the sequence length to 1200 tokens, a value chosen empirically to fit within the 40GB RAM of the GPU (experiments are performed on NVIDIA A100) and mitigate excessive computational demands and high memory usage. If the input sequence length exceeds this limit, the demonstration part of the input is removed and only the query part is used as input. To evaluate the proposed in-context learning method, the training set samples are used as demonstrations and test set samples as query. For validating the fine-tuning method, we tune the parameters on the training set, select the best model on the validation set, and finally test and report the performance of the selected model on the test set. PEFT [11] package is used to tune the model using LoRA. The default model is trained with the AdamW optimizer with zero weight decay.

Thresholding When evaluating the proposed model on the multi-label dataset (BIORC800), we apply dynamic thresholding, which utilizes different probability thresholds for each label. The optimal threshold for each label is determined by maximizing the label-specific F_1 score on the validation set. For single-label datasets, we apply softmax function to select the best label.

2.4.4 Results and Discussion

In-context Learning

In this subsection, we evaluate the model using 0-shot (no demonstrations in the prompt), 1-shot (one demonstration), 5-shot, and 10-shot settings, where the shots are chosen from the training set using SimCSE ranking, and the queries are from the test set. Table 2.2 presents the performance of our in-context learning approach across all datasets. Specifically, we have the following observations: (1) LLM-SSC with 5-shot setting achieves the highest micro F1 scores for BIORC800, CS-ABSTRACT, and PUBMED 20K RCT; (2) in the zero-shot setting, in-context learning on BIORC800 and CS-ABSTRACT datasets consisting of entire paragraphs of unstructured text achieves micro F1 scores as 0.476 and 0.468, and macro F1 scores as 0.322 and 0.331. This demonstrates the large language model's generative ability to recognize without seen any training data; (3) the 0-shot in-context learning performance on PUBMED 20K RCT is relatively poor, where

[10] https://github.com/Leo02016/HeroCon
[11] https://huggingface.co/docs/peft/en/index

		BIORC800		CS-ABSTRACT		PUBMED 20K RCT		ART-CORESC	
		Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
HSLN [14]		0.849	0.826	0.723	0.652	0.919	0.870	0.400	0.163
SSC [39]		0.905	0.892	**0.780**	0.714	0.924	0.859	0.470	0.253
SDT [56]		-	-	0.767	0.653	**0.940**	**0.903**	**0.534**	0.270
SciBERT-HSLN [57]		0.902	0.897	0.765	0.712	0.931	0.882	0.467	0.246
LLM-SSC	with HeroCon [98]	0.902	0.906	0.767	0.714	0.921	0.871	0.503	0.260
	with WeighCon (Ours)	**0.907**	**0.912**	0.768	**0.716**	0.925	0.879	0.524	**0.282**

Table 2.3: Task-specific model tuning results. In LLM-SSC, the next 2 tokens are generated. The SDT model performance on BIORC800 is not reported since it uses a BIO tagging mechanism to block different rhetorical sections within a paragraph, making it unsuitable for multi-label classification.

sentences are organized by the original authors to meet specific structural requirements at the expense of contextual dependence on each other; (4) the performance improvement from 0-shot to 1-shot emphasizes the importance of including samples in the prompt for LLMs to understand SSC tasks; (5) when provided with more samples (10-shots) on BIORC800, CS-ABSTRACT, and PUBMED 20K RCT, the performance are not as good as with fewer examples. We attribute this drop to the additional information introducing bias and confusing the large language model to capture the task-related features from the additional samples; (6) the poor in-context learning results yielded on CORESC can be attributed to the dataset's fine-grained rhetorical categories, which are challenging for large language models to recognize by simply relying on general common-sense reasoning or surface-level patterns without more detailed guidelines.

Task-specific Model Tuning

Table 2.3 presents the performance of the task-specific model-tuning methods. Under the multi-label setting, we observe: (1) our LLM-SSC with WeighCon achieves the highest micro and macro F1 scores (0.907 and 0.912, respectively) when tested on the BIORC800 dataset; (2) compared to HeroCon, the proposed WeighCon yields better performance, demonstrating its effectiveness with the auto-weighting design; (3) the SSC model delivers the second-best micro F1 score (0.905), together with the proposed LLM-SSC, highlighting the effectiveness of transformer-based methods in multi-label SSC; (4) although the SDT model achieves state-of-the-art (SOTA) micro-F1 performance (i.e., 0.940) on the PUBMED 20K RCT dataset, its BIO tagging, which"blocks" different rhetorical sections in a paragraph, is not applicable to the multi-label setting.

On the single-labeled datasets, we find: (1) the LLM-based method delivers promising macro F1 results (CS-ABSTRACT: 0.716, PUBMED 20K RCT: 0.879, ART-CORESC: 0.282), demonstrating its effectiveness in balancing performance across classes. Unlike previous baseline methods that predict rhetorical labels based on each sentence embedding, LLM-SSC leverages the contextual understanding ability of LLM to grasp the whole context before generating the SSC label, therefore treating each class more equally; (2) the micro F1 scores reveal that LLM-SSC's sample-specific performance is near SOTA (CS-ABSTRACT: 0.768, PUBMED 20K RCT: 0.925, ART-CORESC: 0.524) but does not outperform the SOTA, indicating a limitation in capturing the majority class compared to the fully fine-tuned baseline models.

Note that, different from the previous SOTA methods that tuned the parameters of the entire pre-trained

model, LLM-SSC is tuned using LoRA, keeping the original model parameters frozen while updates about 4% additional parameters relative to the size of the entire LLM. 10,100,736 parameters are trainable, and 2,516,273,152 parameters are frozen. This approach significantly reduces storage requirements, as only the task-specific additional parameters need to be stored.

Ablation Studies

Model	BIORC800		CS-ABSTRACT	
	Micro F1	Macro F1	Micro F1	Macro F1
LLM-SSC	0.907	0.912	0.768	0.716
w/o Demonstration	0.903	0.911	0.742	0.645
w/o WeighCon	0.896	0.901	0.746	0.682
w/o Space Thinking	0.892	0.899	0.749	0.685

Table 2.4: Ablation study.

We conduct ablation studies to assess the impact of various components of LLM-SSC when testing on BIORC800 and CS-ABSTRACT as shown in Table 2.4. Note that "w/o Space Thinking" refers to deleting space thinking mechanism by enabling the LLM to generate only one token directly after the prompt. For all four components, we observe the performance drops when each component is removed from LLM-SSC, indicating that all four components in LLM-SSC contribute to SSC performance on both single- and multi-label datasets. Note that the impact of each component is greater when the model is trained on CS-ABSTRACT compared to BIORC800. CS-ABSTRACT consists of 654 abstracts with an average of 7.23 sentences per abstract, while BIORC800 contains 800 abstracts with an average of 9.89 sentences. The small size of the CS-ABSTRACT dataset limits the model's performance, and adding three components mitigate the limitation. In contrast, this improvement is less evident on the larger BIORC800 dataset.

Think before Speak Analysis

Number of Generated Tokens	BIORC800		CS-ABSTRACT	
	Micro F1	Macro F1	Micro F1	Macro F1
1	0.897	0.904	0.749	0.685
2	0.907	0.912	0.768	0.716
3	0.895	0.904	0.739	0.686

Table 2.5: "Think before Speak" analysis results. Notice that generating the next one token equals to leaving no space for model to think.

We analyze the "Think before Speak" mechanism to determine whether generating more tokens introduces more bias or provides space for model to "think". Table 2.5 presents the performance of the model when

generating one, two, and three subsequent tokens. The results show that generating two tokens yields the best micro and macro F1 scores across both datasets. This suggests that generating two new tokens is sufficient to achieve optimal model performance for SSC task, whereas generating only one token restricts the model's ability to process information, and generating three tokens introduces bias into the SSC label prediction.

2.5 Conclusion

We introduce LLM-SSC, a unified framework for in-context learning and parameter-efficient LLM finetuning for multi-label sequential sentence classification problem. LLM-SSC integrates complete contextual information within the prompt and considers neighboring context when making predictions. Additionally, we present a multi-label contrastive learning loss with auto-weighting scheme to reweight the importance of negative pairs and address the multi-label sequential sentence classification problem. Furthermore, we release BIORC800, a manually annotated multi-label SSC dataset using unstructured abstracts from the biomedical field, contributing to the development of more robust methodologies for this task. Extensive experiments validate the remarkable capability of LLM-SSC in SSC tasks under both in-context learning and parameter-efficient finetuning settings.

2.6 Future Work

Currently, we have analyzed the sentence-level argument structure of biomedical abstracts, which provide a condensed summary of the full-text. However, this condensation process may omit detailed reasoning, supporting evidence, or essential context from the full-text. By analyzing the entire document, we can extract a more comprehensive and detailed argument structure across paragraphs. This broader perspective has the potential to improve downstream tasks such as fine-grained information retrieval and the identification of issues in methodological designs. As a next step, we plan to explore paragraph-level argument structure extraction within full-text biomedical literature.

Section headers in biomedical literature always provide the signals for the rhetorical purpose or topical focus of the information presented within each section of the literature. Inspired by this feature, our approach to full-text argument structure extraction will primarily rely on section header information. Specifically, we will explore how to assign a predefined argument role to a section content based on the rhetorical signals conveyed by its header, following the steps as:

(a) Collect a representative sample of approximately 50 full-text articles from the PubMed Central Open Access repository, ensuring diversity in biomedical domains, publication venues, and article types.

(b) Extract section headers and their corresponding content from the selected articles. Use the mapping standards defined in the Label List and NLM Category Mappings File[12] to align each section header with a standardized argument role. Assign the identified argument role to the associated section content.

(c) Evaluate the reliability of the rule-based argument structure identification results from step (b), and refine the mapping approach as needed—for example, by expanding from single-label to multi-label classification.

[12]https://wayback.archive-it.org/7867/20241213200409/https://lhncbc.nlm.nih.gov/ii/areas/structured-abstracts/downloads/Structured-Abstracts-Labels-111314.txt

Chapter 3

Automatic categorization of self-acknowledged limitations in randomized controlled trial publications

This chapter has been adapted from: Lan, M., Cheng, M., Hoang, L., Ter Riet, G., & Kilicoglu, H. (2024). Automatic categorization of self-acknowledged limitations in randomized controlled trial publications. Journal of biomedical informatics, 152, 104628. My contribution to the study are (in the format of CRediT taxonomy): conceptualization, formal analysis, data curation, investigation, methodology, software, validation, visualization, writing-original draft, writing-review & editing.

3.1 Introduction

Biomedical publishing has been transformed in recent years, spurred in part by the COVID-19 pandemic [103]. There has been a sharp rise in non-peer-reviewed preprint publications [104], and the speed with which these studies have been conducted and published has raised concerns in scientific and lay press about the methodological and reporting quality of COVID-19 research [105]–[108]; some publications in prominent journals were eventually discredited or retracted [103]. Such issues, of course, are not new or limited to COVID-19 research, or even to biomedical science. Problems in study design, execution, data analysis, and reporting affect the validity and applicability of the findings in any field. It is essential for researchers to acknowledge potential weaknesses and biases of their studies (i.e., its limitations) and discuss their magnitude when publishing their findings [109], [110]. Recognizing and discussing limitations is essential for scientific progress, as they help the reader contextualize the study, understand its findings, and assess the credibility of these findings [109]. Limitation discussions could also reveal future research directions and the caveats that need to be considered when incorporating the new findings into scientific evidence [109], [111], [112].

Study protocols are increasingly accessible, making it possible for experts to decide on the important study limitations themselves if authors are open about all discrepancies between study plan and execution; however, publications should also be optimally informative for readers who are not seasoned experts in a particular research field and frank limitations sections will probably remain important [113].

Randomized controlled trials (RCTs) are a cornerstone of clinical medicine and provide the most robust

methodology to generate evidence on the effectiveness of therapeutic interventions [114], [115]. To fully realize their potential for informing patient care and health policies, they need to observe high methodological and reporting quality standards [114]–[116]. However, this is often not the case, leading to low-quality evidence and significant research waste [116]. One common reporting problem in RCT publications is that authors often do not properly acknowledge the limitations of their study [109], [113], [117], [118], making it difficult for stakeholders (peer reviewers, journal editors, systematic reviewers, clinicians, policymakers) to contextualize the findings and assess their trustworthiness [109]. CONSORT reporting guidelines for RCT publications [114], [119] recommends reporting of limitations in RCT publications and states that "trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses" [114] should be discussed in Discussion sections. However, reporting of limitations has been found to be inadequate [113], [120], similar to other checklist items [121]–[123]. Automated screening tools can reduce the time and effort for manual checks in journals, provide rapid reporting quality assessments of preprints, and raise awareness of poor reporting practices among researchers [124]–[126], leading to gradual improvements in RCT reporting. Natural language processing (NLP) techniques can underpin such screening tools [125]–[128].

Previous work has developed NLP models for recognizing sentences discussing self-acknowledged limitations (SALs) in clinical publications [60]. The work compared several models, including a rule-based approach, a logistic regression classifier, and a SVM classifier, and experimented with self-training to address the data scarcity problem. The rule-based method performed reasonably well (0.80 F_1 score and 0.915 accuracy). So the previous work used this method to study the impact of peer review on discussion of study limitations [129], finding support for the idea that editorial processes lead to more self-acknowledgment of study limitations. This method was later incorporated into a pipeline of tools for screening COVID-19 preprints for transparency and reproducibility [125], [126]. While automatically identifying limitation sentences are useful as a simple reporting check (i.e., have limitations been acknowledged?), extracting the types of limitations reported could help in: a) more precisely contextualizing a study's findings and assessing their credibility; and b) better understanding the common types of problems in large sets of research studies, including RCTs, to answer meta-research questions. In this study, we extend the previous work [60] by extracting self-acknowledged limitation types from RCT publications. We make the following contributions:

1. We create a SAL sentence recognition model by fine-tuning the PubMedBERT model [130], which outperforms the model created in the previous work [60].
2. We develop a fine-grained data model of limitation categories, taking related work [62] as our starting point.
3. We manually annotate a corpus of 200 RCT publications with SAL categories at fine granularity.
4. We fine-tune the PubMedBERT model for multi-label sentence level classification of SAL types.
5. We experiment with prompt-based data augmentation with the help of a large language model to address the data imbalance and scarcity problem.
6. We analyze the model output on a large corpus of RCT publications to describe the commonly reported limitations of RCTs.

Our results show that annotating SAL types and achieving good inter-annotator agreement (IAA) is challenging. Our NLP model shows that it is possible to recognize SAL types at coarse granularity, while

experiments with finer granularity yield more modest results. Prompt-based data augmentation improves NLP model performance. Sample size and population-related limitations are the most commonly discussed.

3.2 Related Work

On the other hand, there is a significant body of literature on natural language processing for RCT publications. Most of this literature has focused on extracting PICO classes (Population, Intervention, Comparator, Outcome) to assist systematic reviews and facilitate evidence-based medicine, often using sentence classification [56], [70], [131]–[135] and named entity recognition [135]–[139]. Some research focused on classifying RCT abstract sentences along IMRaD (Introduction-Methods-Results-Discussion) categories [14], [56], [70]. Other recent studies investigated extraction of information relevant to methodological assessment of RCT studies, such as risk of bias information [140], study characteristics (e.g., sample size) [141], [142], and CONSORT checklist items [9], [124], [143]. Recent state-of-the-art models have relied on domain-specific Transformer-based, pretrained language models [22].

To enable NLP model development and validation, a large amount of high-quality labeled data is often needed. Labeling such data, especially in biomedical domain, is challenging, because annotation is time-consuming and requires significant domain expertise. Methods to address small sample sizes have been proposed, often studied under the umbrella of weak supervision and data augmentation. Weak supervision attempts to use domain knowledge and subject matter expertise to assign somewhat noisy labels to unlabeled data [144]. Data augmentation generates realistic examples from a limited number of existing examples [145]. For example, simple transformations of individual sentences (e.g., synonym replacement, random deletion/insertion) have been used to improve modeling accuracy with small datasets [146], [147]. Given that such modifications may distort the original meaning of the text, recent research studies have used large language models to synthesize more meaningful sentences [148]–[150].

3.3 Material and methods

3.3.1 Improving SAL sentence classification

We fine-tuned the PubMedBERT model [130] from HuggingFace's model repository (microsoft/BiomedNLP-PubMedBERTbase-uncased-abstract-fulltext) on the limitation sentence dataset reported in prior work [60]. This dataset consists of 2,257 sentences, 467 (20.7%) of which include SALs. As the input representation for the model, we concatenated the target sentence, the top section header (e.g., Discussion), and the innermost section header (e.g., Study Limitations) and fed the [CLS] token representation generated by PubMedBERT to a linear layer for binary classification. Binary cross-entropy loss is used as the loss function. We set the maximum sentence length to 512 tokens and tuned all the hyperparameters on the development set. The following hyperparameters were used: AdamW optimizer, batch size: 4, learning rate: 3e-5, and epochs: 10.

We split the manually annotated dataset (200 articles) to train/dev/test sets as 120/40/40. To tune the PubMedBERT-based models SAL classification, we also set the maximum sentence length to 512 tokens and tuned the hyperparameters on the development set. The following hyperparameters are used for the reported models: AdamW optimizer, batch size: 4, learning rate: 1e-5, epochs: 20. All experiments were conducted on a Tesla T4 GPU.

3.3.2 Data collection for SAL type annotation

We collected a dataset of 200 RCT articles for annotation. A subset of the articles (52) came from our previous studies on limitation sentence classification [60] and CONSORT sentence classification [124]. Only those articles containing sentences manually labeled as limitation sentences were included. The rest of the RCT articles were sampled from PubMed Central Open Access Subset (PMC-OA) using a modified version of Cochrane's sensitivity and precision-maximizing query for RCTs as the search strategy. We split the downloaded RCT articles into sentences using the NLTK sentence tokenizer and identified the section to which a sentence belongs using a section recognizer. Then, the SAL sentence classifier based on PubMedBERT fine-tuning (described above) was used to identify SAL sentences from the abstract and discussion- and limitation-related sections, indicated by the keywords *discussion*, *limitation*, *weakness*, *conclusion*, *caveat*, *shortcoming*, and *drawback* in the header, consistent with previous work [60]. One of the authors (HK) manually assessed the accuracy of the sentence classifier predictions to verify that only RCT articles with at least one SAL sentence were included in the corpus.

3.3.3 Data model for SAL types

To create the data model for annotation, we started with the limitation categorization presented in a recent investigation of manual therapy trials [62]. Their categorization consists of 12 top-level categories (e.g., Blinding, Sample Size, Inadequate Control, Compromised Generalization) and 38 sub-categories (e.g., Underpowered Study, Convenience Sampling, and Recruitment Less Than Expected for the Sample Size category). After a test annotation of 10 articles, we recognized the need to simplify and adapt the categorization, as some categories related to specific characteristics of manual therapy trials (e.g., Therapist Profile) and some categories seemed difficult to reliably differentiate (e.g., Intervention vs. Compromised Generalization due to Intervention). As a result, two authors (HK and GtR) redesigned the data model in several iterations. The final data model consists of 15 top-level categories and 24 sub-categories. The data model and the definitions of the categories are provided in Table 3.1 and annotation examples in Table 3.2.

Top-level	Fine-grained categories	Description
StudyDesign		Limitations that have to do with the specific trial design used (e.g., crossover, factorial, cluster, etc).
Population		Limitations that have to do with the selection of subjects who participated in the trial.
	DiagnosticCriteria	Lack of standardized diagnostic criteria for including participants.
	VerySpecificPopulation	Inclusion criteria considered too restricted (e.g., single gender, athletes only, education level, or race).
	ConvenienceSampling	The sampling method was linked to specific study needs. Subjects were selected because they were convenient sources of data for the study.
Setting		Limitations related to where the study takes place.
	Unicentric	Study was conducted recruiting participants from a single center.
Intervention		Limitations that have to do with the active intervention treatment used.
	CompositeIntervention	It was not possible to know the net effect of every component in multimodal treatments.
	NonStandardTreatmentCharacteristics	The specific parameters for the intervention were not standardized (e.g., dosage, mode of administration).
Control		Limitations that have to do with the control intervention placebo.
	NoPlaceboGroup	No control intervention is included.
	ActivePlacebo	An active intervention (non-inert) was selected as control.
	CareAsUsualControlGroup	Due to the non-standardization of CAU, it is uncertain what it is that the experimental group is being compared to. In these cases, the control group treatment will sometimes be mentioned as "care as usual".
Outcome Measures		Limitations related to the outcomes used and how they are measured.

Table 3.1: Types and descriptions of self-acknowledged limitation types.

Top-level	**Fine-grained categories**	**Description**
	RelevantOutcomeExcluded	Some relevant data that would potentially provide interesting findings were not collected during the study.
	PrecisionOfMeasurement	Lack of or low precision of outcome measures. This refers to a limitation due to random errors that might have been introduced in measurement.
	ValidityOfMeasurement	The selected assessment instrument was not originally validated for the specific population or problem studied. This indicates that the outcome measurement may not correctly measure the concept that is the target of the measurement (systematic error, as opposed to random error).
	ResponsivenessOfMeasurement	Outcome measures were not sensitive enough to detect subtle changes (e.g., use of ordinal scales). Responsiveness is defined as the ability of an instrument to accurately detect change when it has occurred.
MissingData		Some data were not collected for some study participants. This indicates that some planned follow-up measurements, whether outcomes or co-variables (confounders) were not collected, regardless of the reasons for that missingness.
	HighLossToFollowUp	Many participants stopped participating before the planned duration of follow-up.
	UnbalancedDropout	Characteristics of dropped out patients differed between groups ('informative drop-out'). For example, relatively healthy patients dropped out from the experimental group, whereas patients in relatively poor health dropped out from the control group.
Underpowered Study		Inability to detect differences between groups due to sample size or insufficient number of outcome events.
	SampleSize	Limitations related to the insufficient number of patients participating in the trial. This may be a result of recruitment difficulties.
Randomization		Limitations that have to do with the randomization of patients into different trial arms.

Table 3.1: Types and descriptions of self-acknowledged limitation types.

Top-level	**Fine-grained categories**	**Description**
	UnbalancedGroups	After randomization, there were large differences between the groups with respect to (mean values of) prognostically important factors (confounders). This is problematic because the response to interventions may be due to these confounders, rather than the interventions.
	PoorRandomizationMethods	Randomization methods used (e.g., for sequence generation, restriction stratification, concealment) were not optimal. This also includes a lack of such methods, e.g., that allocation was not concealed (i.e., once a patient is assigned, the next assignment is predictable).
Blinding		Limitations related to how the study participants and personnel were blinded to the study groups.
	Patient	Patients were not blinded with respect to the study groups.
	StudyTeam	Some people in the study team (investigators, care providers, outcome assessors, statisticians etc.) are not blinded.
StudyDuration		Limitations that have to do with the length of the study. It could be the experimental phase or the follow-up.
	ExperimentPhaseDuration	The intervention phase is too short. It could be due to early stopping.
	FollowUpDuration	Only short term effects were evaluated. Long term (adverse) effects of the interventions were not considered.
Statistical Analysis		Limitations regarding the methods used for statistical analysis, indicating that the techniques used may not have been appropriate or were suboptimal.
	MultipleTesting	Simultaneous testing of more than one hypothesis.
	ConfoundingFactors	Findings were not adjusted for covariates.
Funding		The limited or lack of funding affected the study progress or completion.

Table 3.1: Types and descriptions of self-acknowledged limitation types.

Top-level	**Fine-grained categories**	**Description**
Generalization		The study results were compromised and may not generalize due to type of setting, specific population, intervention, and measurement instruments.
Other		Catch-all category for all limitation types that do not neatly fit in any of these categories.

Top-level	**Fine-grained categories**	**Description**
StudyDesign (19)	(19)	*One study limitation is* ***the use of a non-randomized control group****.*
Population (192)	(59)	***The unequal distribution of gender in the study sample*** *may also have influenced the results.*
	DiagnosticCriteria (23)	*We defined* ***no exclusion criteria related to severity of background illness or sepsis****.*
	VerySpecificPopulation (112)	*While the sample was randomly selected,* ***selection bias is possible*** *due to the 64% response rate.*
	ConvenienceSampling (2)	***We did not recruit a representative sample****; thus study results cannot be used to draw conclusions about hypothesis testing or population-level dynamics.*
Setting (12)	(5)	*Educational classes were an important activity in this trial and* ***the uptake of information could have been better in larger schools****.*
	Unicentric(7)	*Finally the characteristic of being* ***a monocentric trial*** *can be considered both a limitation but also an advantage due to reduction of variability of care.*
Intervention (109)	(67)	*Finally, while participant retention and compliance with outcomes protocols was high (loss to follow-up was 2%),* ***adherence to TC training*** *was lower than expected.*
	CompositeIntervention (10)	***We chose not to have an intervention group receiving BMD feedback alone*** *without any other educational intervention.*
	NonStandardTreatmentCharacteristics (33)	*One possible reason for the lack of effect is that* ***families did not comply with the intervention****.*
Control (23)	(6)	***The absence of another control comparator*** *makes it difficult to attribute the beneficial effects solely to the intervention.*
	NoPlaceboGroup (6)	*To prevent this bias,* ***a placebo treatment group with patients receiving an equal amount of therapist attention would be required****.*

Table 3.2: Limitation type categories with example sentences in the annotated dataset. Numbers in "()"s denote the number of samples for each category in our dataset. Spans relevant to the categorization are highlighted in bold.

Top-level	Fine-grained categories	Description
	ActivePlacebo (9)	*First, it is known that* ***the controlled substance propofol also has protective characteristics****.*
	CareAsUsualControlGroup (2)	*Our choice of* ***a usual care control*** *followed the overarching practical goal of our study–to evaluate the potential benefits to osteopenic women of adding TC to usual care.*
Outcome Measures (190)	(28)	*We are also limited by* ***our exclusive focus on intra-individual factors as moderators****.*
	RelevantOurcomeExcluded (66)	***Genotyping, which was not practical in our study setting****, might have aided the interpretation of our findings.*
	PrecisionOfMeasurement (59)	*However,* ***self-reported outcome by patients is necessarily subjective*** *and affected by many things besides knowledge of treatment allocation.*
	ValidityOfMeasurement (28)	***The difficulty of capturing all dimensions of physical activity by questionnaire*** *is well-known[38,39], particularly non-leisure activities in women[40].*
	ResponsivenessOfMeasurement (9)	*There are several possible explanations for this, including the possibility that* ***our measures were not sensitive enough over this time period****.*
MissingData (86)	(63)	*Furthermore,* ***we do not have complete information about adherence to study medication.***
	HighLossToFollowUp (15)	***Our loss to follow-up for our functional secondary outcomes measured at ICU and hospital discharge was also significant*** *as patients were often unable to participate in the assessments.*
	UnbalancedDropout (10)	*This impression of practical difficulty is reinforced by* ***the significantly higher proportion of participants that withdrew from the trial in the intensive arm****.*
Underpowered Study (185)	(82)	***The subgroup analysis is therefore confronted with an even poorer lack of power.***

Table 3.2: Limitation type categories with example sentences in the annotated dataset. Numbers in "()"s denote the number of samples for each category in our dataset. Spans relevant to the categorization are highlighted in bold.

Top-level	Fine-grained categories	Description
	SampleSize (115)	***Even when the remaining girls complete the study, the number evaluable at final height will still only be 92.***
Randomization (24)	(6)	*Another limitation is that* ***we performed simple randomization*** *instead of block randomization.*
	UnbalancedGroups (11)	*This method was not used in this study but probably would have avoided* ***this unlucky uneven distribution of severity factors****.*
	PoorRandomizationMethods (8)	*Although* ***open allocation*** *was an unavoidable limitation of the monitoring randomisation and was not undertaken for the ART-strategy randomisation, the endpoint review committee adjudicated endpoints masked to randomization.*
Blinding (64)	(31)	*Therefore, it is not clear why bias (or* ***lack of blinding****) would not have similarly led to more crossovers from the RF to the conventional needle.*
	Patient (19)	*First,* ***subjects were not blinded to their intervention group****.*
	StudyTeam (16)	*One major limitation was that* ***the first author conducted all aspects of the trail including provision of care to all study participants****.*
StatisticalAnalysis (38)	(9)	*Furthermore, due to the exploratory nature of the feasibility study,* ***no multiple outcomes or sample size calculation were performed.***
	MultipleTesting (12)	*Finally, we conducted a number of statistical tests, raising* ***the potential concern of alpha inflation****.*
	ConfoundingFactors (17)	***Differences in age, stress, depression, and other factors can potentially influence success in managing weight, and may have confounded results****.*
StudyDuration (51)	(6)	***The short duration*** *precluded the use of change in HbA levels as an efficacy end-point.*
	ExperimentPhaseDuration (21)	***Study duration*** *is another potential study limitation given the long half-life of OKZ (31days).*

Table 3.2: Limitation type categories with example sentences in the annotated dataset. Numbers in "()"s denote the number of samples for each category in our dataset. Spans relevant to the categorization are highlighted in bold.

Top-level	**Fine-grained categories**	**Description**
	FollowUpDuration (24)	***Longer-term follow-up is needed*** *to confirm any lasting positive effects on BMD from ongoing calcium supplement use.*
Funding (4)	(4)	*The trial was stopped after achieving 86% of target recruitment owing to time and* ***financial limitations****.*
Generalization (66)	(66)	*As such,* ***our sample might not be generalisable to all those who experience grade 1 and 2 ankle injuries****.*
Other (2)	(2)	***This pilot trial involved co-funding and participation by the device manufacturer****.*

3.3.4 SAL type categorization

We formulated the task of identifying SAL types discussed in an RCT publication as multi-label sentence classification, despite associating these types with spans in annotation. This is due to a couple of reasons. First, we observed that spans indicating SAL types were very heterogeneous, ranging from short noun phrases which look like typical named entities to entire sentences. Secondly, IAA for span annotation was found to be low (results below), probably due to this heterogeneity. Because the goal of SAL type categorization is ultimately to understand which limitation types are described in an article, not finding the exact spans indicating them, a sentence multi-label classification approach was deemed appropriate. We converted and consolidated span-level annotations to sentence-level annotations. As our classification scheme, we experimented with both the top-level and the fine-grained categorization. Note that in fine-grained categorization, top-level categories are still considered, because it is possible that some sentences only have top-level labels. The dataset was split into training/development/test sets of 120, 40, and 40 articles, respectively.

As the baseline model for SAL type categorization, we also fine-tuned the same PubMedBERT model [22] using the target sentence prepended with the section headers as input. Similarly, in this case, the representation for the [CLS] token for the target sentence is fed into a multi-layer perceptron (MLP) and a softmax layer that calculates the probability distribution of each label for the target sentence. We then apply dynamic thresholding, which uses different probability thresholds for each label. The optimal threshold for each label is determined based on the label-specific F_1 score on the development set. Cross-entropy loss is used as the loss function.

To fine-tune the PubMedBERT model for limitation sentence classification, we set the maximum sentence length to 512 tokens and tuned all the hyperparameters on the development set. The following hyperparameters were used: AdamW optimizer, batch size: 4, learning rate: 3e-5, and epochs: 10.

We split the manually annotated dataset (200 articles) to train/dev/test sets as 120/40/40. To tune the PubMedBERT-based models SAL classification, we also set the maximum sentence length to 512 tokens and tuned the hyperparameters on the development set. The following hyperparameters are used for the reported models: AdamW optimizer, batch size: 4, learning rate: 1e-5, epochs: 20. All experiments were conducted on a Tesla T4 GPU.

Prompt-based Sentence Augmentation

Our annotation yielded an imbalanced dataset with few examples for some categories, including some top-level categories. To address these shortcomings, we used data augmentation to synthesize novel samples for the less frequent classes. As our primary data augmentation method, we adapted the PromDA method (Prompt-Based Data Augmentation) [150], which is built upon the *T5-Large* encoder-decoder model [151]. It keeps the entire pre-trained T5 model frozen, prepends additional soft prompts (i.e., a sequence of continuous and trainable vectors) in each layer of the model, and tunes the soft prompts only [152]. Soft prompts are pre-trained using a mechanism named *Task-agnostic Synonym Keyword to Sentence* pre-training. Next, a dual-view data augmentation approach is used to generate synthetic samples conditioned on the keywords in the input sample (Input View) and the input sample label (Output View), respectively. Keywords for the Input View are extracted using the unsupervised keyword extraction algorithm Rake [153]. Finally, a consistency filtering step is applied to only keep synthetic samples with consistent labeling.

In this study, we use the soft prompt parameters pretrained on the *realnewslike* dataset [151]. We fine-

tune the pretrained parameters using samples from the classes with fewer than 70 samples in our training set. Samples with multiple labels were excluded from data augmentation, since it was hard for the augmentation method to differentiate the features for each label. Funding label was excluded, because all sentences with this label were multi-label. In addition to original PromDA, we also experimented with augmentation with Input View and Output View only, because we observed that consistency filtering led to more examples for frequent labels and few examples for rare labels. We generated 10 synthetic examples for each original sentence. To augment the training set, we set the minimum number of samples for each class to 70. For classes with n samples in the training set ($n < 70$), we add 70 - n synthetic samples to the training set. The 70 – n samples are randomly selected from the synthesized sentences. Figure 3.1 depicts the PromDA process.

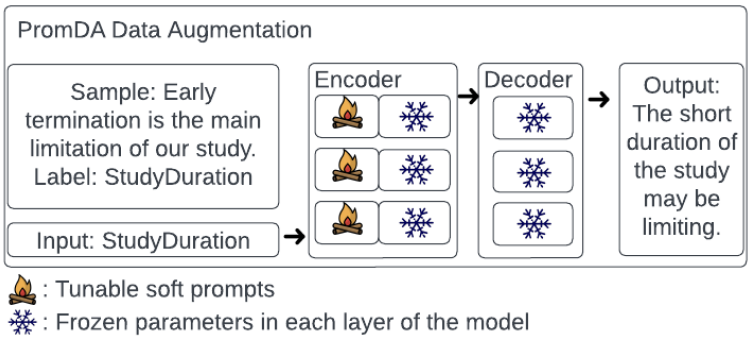

Figure 3.1: Overview of our soft-prompt based data augmentation (PromDA).

Oversampling

We also experimented with oversampling to address the issue of imbalanced data. Oversampling randomly duplicates samples from the minority classes to increase their representation in the training set. As in PromDA, we set the target size of each class to 70 and added $70 - n$ duplicate samples for classes with fewer than 70 samples, where n represents the number of original samples for the class.

Easy Data Augmentation (EDA)

Another method we used for data augmentation was EDA [146], a simpler, rule-based method that synthesizes samples via simple modifications to the original sentence, including word order shuffle, random deletion/insertion, and synonym replacement. As in other methods, we set the target size of each class to 70.

Rule-based identification

To address the Funding class, which was not augmented, we use a simple rule-based method, which labels a sentence as Funding, if stemmed tokens in the sentence contain the stems *"financi"* or *"fund"*.

3.3.5 Evaluation

We evaluated the SAL sentence classifier using precision, recall, and F_1 score for the positive class and accuracy, in line with previous work. For SAL type categorization, we trained a baseline model by fine-

tuning the PubMedBERT model on the manually annotated training set. We compared this model to those trained with augmented training sets (PromDA, PromDA – Input View, PromDA – OutputView, EDA). For comparison, we also trained the baseline model with the fine-grained labels. The model performances were measured using micro-precision, recall, and F_1 score. To obtain reliable estimates of model performance, the models were trained and tested using five randomly initialized runs. We report the performance averages of these runs and the standard deviations. We use McNemar's test [154] to determine whether the performance difference between the rule-based method and the PubMedBERT model for limitation sentence classification is statistically significant. To observe whether the data augmentation methods lead to statistically significant differences, we use Bhapkar test [155], a multi-class extension of McNemar's test, and treat multi-label cases as additional classes.

3.3.6 Large-scale analysis of SALs

To describe SAL reporting at large scale, we used a set of 11,988 RCT articles (not included in our manually labeled dataset). This unlabeled dataset, curated from the PMC-OA subset in prior work [143], includes articles published from 2011 to 2020. We first applied the SAL sentence classifier to this dataset and extracted SAL sentences from the abstract, discussion- and limitation-related sections. We then applied the best-performing SAL type classification model to these sentences to predict limitation types.

3.4 Results

3.4.1 Dataset Statistics

We annotated a total of 200 RCT articles, published between 2001 and 2022. 52 articles (26%) were published in general medical journals (e.g., BMJ), while the rest were published in specialty journals. 66 (33%) of the articles were published in journals with high-impact factors (defined as journal impact factor >= 10).

A total of 1090 limitation types in 952 limitation sentences were annotated (1.15 and 5.45 limitation instances per sentence and per article, respectively). The top-level distribution and the count and percentage of sub-categories under each top-level category at the sentence level are presented in Figure 3.2. Among the top-level categories, the most common limitation type was Population (192 out of 1090, 17.6%), closely followed by OutcomeMeasures (190) and UnderpoweredStudy (185). Limitations related to Funding (4) and Setting (12) were least reported. At fine-grained level, SampleSize (115), VerySpecificPopulation (112) and UnderpoweredStudy (82) were most discussed. The least mentioned fine-grained limitation types were ConvenienceSampling (2), CareAsUsualControlGroup (2) and Funding (4).

When we considered the unique limitation types reported in publications, we found that UnderpoweredStudy appears in 55% of the articles, closely followed by OutcomeMeasures (53.5%) and Population (52.5%).

Articles published in general medical journals had an average of 5.38 limitation sentences [95% CI: 5.26-5.51], and those published in specialty journals had an average of 4.56 sentences [95% CI: 4.50-4.62]. The average number of SAL types per general medical journal article was 3.75 [95% CI: 3.70-3.80] and that per specialty journal article was 3.45 [95% CI: 3.40-3.49].

Articles published in high-impact journals had an average of 5.23 limitation sentences [95% CI: 5.12-5.34], and those in lower-impact journals an average of 4.53 [95% CI: 4.46-4.60]. The average number of SAL types

per high-impact journal article was 3.86 [95% CI: 3.79-3.94] and that per lower-impact journal article was 3.34 [95% CI: 3.30-3.37].

3.4.2 IAA

Sentence-level IAA (Krippendorff's α with MASI) was 0.45 at the top level and 0.3 at the fine-grained level for 50 multiple-annotated articles over two annotation stages. In the second (last) stage of multiple annotation, they were 0.61 and 0.39, respectively, indicating some improvement in consistency over the first stage. Pairwise token-level agreement (pairwise κ [101]) showed a range of 0.2-0.46 for the top-level categories and 0.11-0.3 for the fine-grained categories. There are no standard guidelines for interpreting Krippendorff's α; however, the range of 0.6-0.8 is traditionally considered substantial agreement in the literature on agreement coefficients [156].

Given the low agreement at the token level and for fine-grained categories, we made the decision to focus on top-level categories and sentence-level classification for our NLP models. We note that all annotations were examined by at least two annotators, and verified for consistency by the annotator with the highest agreement with others (HK), which increases our confidence that the annotations can be used for training NLP models.

3.4.3 SAL Sentence Classification

The performance of the PubMedBERT-based SAL sentence classification model is reported in Table 3.3, along with the model performances from our previous work [60]. We obtained the best overall results with the PubMedBERT-based model proposed in this study (F1 score 0.821 vs. 0.806), mostly due to improvements in recall (0.907).

Method	**Precision**	**Recall**	**F_1**	**Accuracy**
Rule-based method*	0.758	0.848	0.800	0.915
SVM*	0.766	0.693	0.728	0.896
SVM + self-training*	0.778	0.835	0.806	0.919
PubMedBERT (this work)	0.751	0.907	0.821	0.929

Table 3.3: Performance comparison of limitation sentence classifiers. * denotes the results from previous work [60]. The performance difference between the PubMedBERT model and the rule-based method is statistically significant (McNemar's test: $p < .001$). We were unable to calculate the statistical significance of the performance difference with the SVM models, because their predictions were unavailable.

3.4.4 SAL Type Classification

Table 3.4 shows the performances of the SAL type classification models trained with and without data augmentation. For the baseline model (PubMedBERT fine-tuning), we present the model performances for both the top-level and fine-grained labels. Unsurprisingly, using a smaller set of labels (top-level) leads to better classification performance overall (about 18 absolute F_1 points better, 0.671 vs. 0.494). We consider the model using the top-level categories our primary model.

Oversampling led to an overall degradation in model performance. While EDA improved recall, it also led to a drop in precision, F_1 score remaining essentially the same. Although the original PromDA uses the

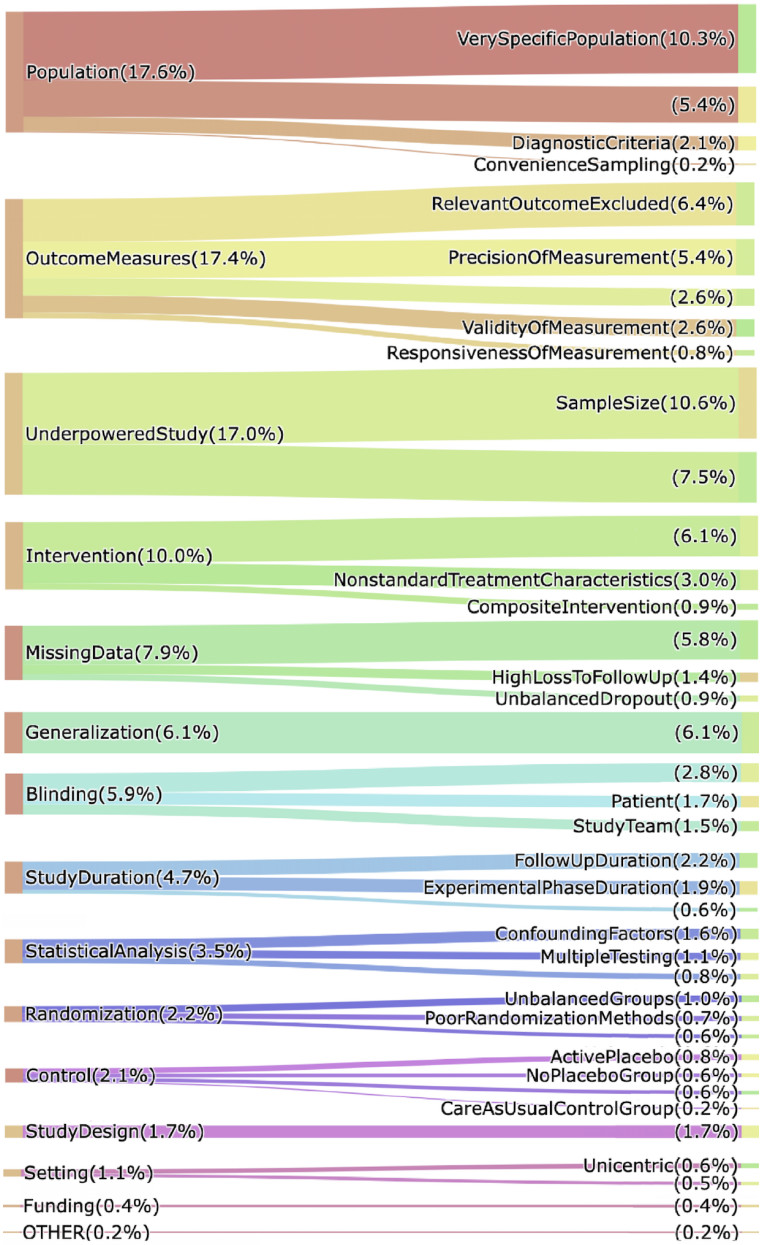

Figure 3.2: The sentence-level distribution of SAL types on the manually annotated dataset. Note that in some cases, the total number of fine-grained labels in a top-level category exceeds the total number for the top-level category, because the same sentence could be labeled with a top-level category as well as a fine-grained label belonging to the same top-level category (e.g., $10.6\% + 7.5\% > 17\%$ for the UnderpoweredStudy category).

Prediction Level	Method	Precision±SD	Recall±SD	F_1±SD
Top-level	PubMedBERT	0.680±0.024	0.666±0.011	0.673±0.010
	+ Oversampling	0.659±0.024	0.646±0.035	0.652±0.028
	+ EDA*	0.664±0.015	0.679±0.003	0.671±0.008
	+ PromDA Original*	0.631±0.043	0.619±0.039	0.625±0.039
	+ PromDA (Input View)**	0.643±0.020	0.633±0.027	0.638±0.022
	+ PromDA (Output View)**	0.690±0.011	0.711±0.015	0.700±0.007
Fine-grained	PubMedBERT	0.488±0.019	0.500±0.021	0.494±0.017

Table 3.4: Micro-precision, recall and F_1 scores for SAL type classification models. The average and the standard deviation over 5 randomly initialized runs are reported. The statistical significance of the performance difference between vanilla PubMedBERT model and models that use data augmentation are calculated using Bhapkar test and are shown with asterisks (*: $p < .05$, **: $p < .001$). The performance difference with the oversampling model is not statistically significant ($p = .055$).SD: standard deviation.

most advanced technique, it led to a reduced performance (0.625 micro F_1 score), which we attributed to the consistency filtering limiting the generation of examples for rare classes. When we only use the examples generated by PromDA (Output View) as additional training data, we obtain the best model performance (0.700 F_1 score). We note that using only PromDA (Input View) also yields poorer results compared to the baseline. The effect of including data augmentation methods on model performance is statistically significant, except for oversampling ($p < .05$ for EDA and PromDA, and $p < .001$ for PromDA (Input View) and PromDA (Output View).

3.4.5 Large-scale Characterization of SALs in the RCT literature

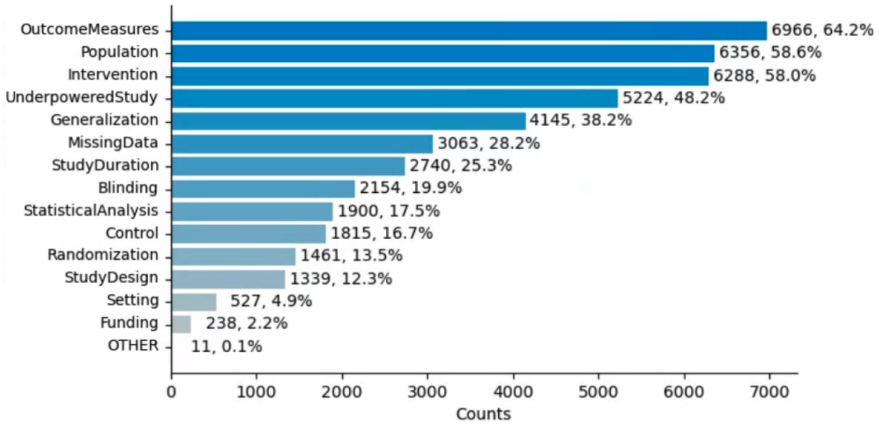

Figure 3.3: Document-level distribution of SAL types on the large-scale RCT dataset. x-axis shows the number of articles that contain a specific SAL type.

From 11,988 RCT articles, our SAL sentence classifier identified 74,670 sentences out of 2,198,534 sentences as limitation sentences (4.23%). 10,843 RCT articles out of 11,988 had SAL sentences (90.4%, 6.2 sentences per article), which is close to the finding by Alvarez et al. [62] that 9% of RCT articles did not report SALs, while being higher than earlier estimates [60], [109], [113]. In Figure 3.3, we present the SAL type distribution in this dataset, in terms of the number of RCT articles reporting the limitation, extracted by the PubMedBERT model trained with PromDA (Output View) data augmentation. OutcomeMeasures are most prevalent (64.2% of articles), followed by Population (58.6%), Intervention (58.0%) and UnderpoweredStudy (48.2%). These four types are also most common in the annotated dataset. The least common types in the large-scale RCT dataset were Setting (4.9%) and Funding (2.2%), also the least common in the annotated dataset.

3.5 Discussion

3.5.1 Limitation Reporting

In our annotated dataset, Population, UnderpoweredStudy, and OutcomesMeasures were most common limitation types with similar prevalence, while Setting and Funding were the least common. This is largely consistent with the findings of Alvarez et al. [62], who found that UnderpoweredStudy was the most common limitation type, and Setting and Funding related limitations were the least reported. In the annotated dataset, the RCTs published in general medical journals and high-impact factor journals show higher limitation reporting compared to those in specialty journals and lower-impact journals. The finding for the general vs. specialty journal is consistent with the finding from ter Riet et al. [113].

In our large-scale analysis, we found that limitation types OutcomeMeasures, Intervention, in addition to UnderpoweredStudy and Population, were highly reported, while Funding and Setting remained the least reported types. About two-thirds of limitations reported seem to relate to the top four categories (approximately 62% in the annotated dataset and 71% in the larger corpus).

3.5.2 Data Model and Annotation

We started with the limitation type categorization in Alvarez et al. [157]. Because they focused on manual therapy RCTs, they included categories specific to that domain. Therefore, we created a modified categorization, which we believe is more generalizable. Nonetheless, the fact that researchers in specific medical specialties often report domain-specific limitations highlights the potential need for further extensions to accommodate specialized areas.

Annotation of SAL types was challenging. Our initial plan to annotate spans indicating fine-grained types yielded modest IAA due to the large number of fine-grained categories and the heterogeneity of limitation type expressions. Therefore, we mainly focused on a top-level sentence-level characterization for NLP. IAA at this level was comparable to agreement of experts in similar work [132], [136], [138]. We attempted to ensure high-quality annotations by having at least two annotators evaluate each article. The resulting dataset is relatively small, due to limited resources, and imbalanced, due to the nature of limitation reporting. However, we believe that it represents a good first step towards understanding the limitations of RCT studies, both explicit and implicit. We make the dataset publicly available to enable further studies in this area.

3.5.3 NLP Models

Our results confirm PubMedBERT as a strong baseline system for supervised biomedical sentence classification and data augmentation as an effective strategy to address the data scarcity problem, common in biomedical NLP tasks. While the performance of the SAL sentence classifier seems reasonable for practical use, there is significant room for improvement for the SAL type classifier.

In this work, a prompt-based data augmentation method that uses a large language model (PromDA) helped the SAL type classifier achieve better performance. Interestingly, while Output View augmentation improved the performance, other mechanisms led to performance degradation. One potential explanation for the degradation due to Input View could be that the keyword extraction method, Rake, does not work well on biomedical text. Replacing this method with more recent methods, such as Yake [158] or KeyBERT [159], or using biomedical named entity recognition tools to identify important concepts could be a future direction. We also observed that consistency filtering diminished the possibility of generating examples of rare classes. The improved performance due to Output View augmentation suggests that label names are informative prompts for synthetic sentence generation using large language models. It seems plausible that SAL category definitions (Table 3.1) could additionally be leveraged for further performance improvements.

To better understand the generalizability of the best model, we assessed the accuracy of 250 predictions in the large-scale analysis set, which showed that 82% of sentences were correctly identified as the limitation sentences, and the SAL types in 77% of these limitation sentences were predicted correctly. These figures compare favorably to the precision of the sentence classifier and the SAL type classifier on the test sets (0.75 and 0.69, respectively) and suggest the generalizability of the model.

3.5.4 Error Analysis

Table 3.5 shows three errors made by the best-performing model (PubMedBERT + PromDA-Output View). We leverage the saliency map created by integrated gradient algorithm [160] to gain insights into how the model focuses on sentence features. Specifically, the gradient integrals of the model's output with respect to input features are calculated and presented by different colors; tokens assigned positive attention are highlighted in green and those assigned negative attention in red. Color intensity corresponds to the feature weight. In the first sentence, the features in the sentence, including "impossible", "some", and "between subjects", were positively weighted, while tokens that seem more relevant for the gold label Blinding ("prevent", "communication") are negatively weighted, resulting in the incorrect prediction Intervention. In the second sentence, the model attends to the token "device" and less to the seemingly relevant tokens "consistency" and "standardization" for the gold label Generalization. In the third case, the token "participant" seems to have high weight, leading to the correct prediction Population, while the tokens relevant for the label StudyDuration ("short intervention period") receive negative weight. This case also reveals the limited capability of our model in the multi-label setting.

3.5.5 Limitations of the Study

Our study has limitations. First, the annotated corpus is small due to limited resources. We attempted to address this issue by using data augmentation for NLP. The inter-annotator agreement was modest (although similar to IAA in similar work, such as PICO classification [132], [136], [138]). We also took additional steps to improve the data quality of the annotated corpus used for NLP. More specific RCT domain expertise, more detailed annotation instructions, and more extensive discussions of annotations could have further

Sentence	True	Predicted	Word Importance
It was impossible to prevent some communication between subjects in the FNI and SC groups.	Blinding	Intervention	[CLS] it was impossible to prevent some communication between subjects in the fn # # i and sc groups . [SEP]
We chose a single device and manufacturer to ensure consistency and standardization.	Generalization	Setting	[CLS] we chose a single device and manufacturer to ensure consistency and standardization . [SEP]
Limitations include the short intervention period and predominantly educated and white participant group.	Population, StudyDuration	Population	[CLS] limitations include the short intervention period and predominantly educated and white participant group . [SEP]

Table 3.5: Examples of errors made by the best-performing PubMedBERT PromDA-Output View model. Word Importance column indicates how the classifier focuses on the sentence features. If a feature is assigned positive attention, it is highlighted in green. Conversely, a feature is assigned negative attention and highlighted in red indicates bias might be introduced. The intensity of the color corresponds to the feature's weight.

improved data quality. The data model extended a data-driven characterization based on manual therapy RCTs [62]. While we believe our characterization is more broadly applicable, a more theoretically sound characterization based on causal inference literature [161] could have further improved generalizability. At the same time, our model may lack more fine-grained categories that might be more practically relevant to specialized domains. Our large-scale analysis is limited by the fact that we only considered a subset available from PMC-OA and the underlying model is imperfect. Lastly, we note that SALs may differ from the "real" limitations of a study as perceived by an RCT methodologist.

3.6 Conclusions

We presented the first NLP work on labeling and automatic identification of SAL types reported in RCT articles (and scientific literature, more broadly). We also improved a previously reported SAL sentence classification model. While the latter model performs well and has been incorporated into a COVID-19 preprint screening pipeline [125], [126], there is significant room for improving the performance of the SAL type model, which we will explore in future work. We also reported a large-scale analysis of RCT literature based on our model, which is the first of its kind.

3.7 Future Work

Given the importance of SALs in RCT publications for contextualizing the findings, it is essential to perform a quantitative analysis of how biomedical researchers have reported SALs over time. As a next step, we will

apply our proposed SAL type classification model to a collection of RCT publications spanning different time periods. This will allow us to assess whether the reporting of SALs has improved over time and to examine how the distribution of SAL types has changed over the years. Our exploration will follow the steps as:

(a) Data Collection: Gather a set of RCT articles for each year (200 articles per year) from 2001 to 2024 using the PubMed Central Open Access repository.

(b) SAL Identification Analysis: Apply our SAL identification model to quantitatively assess trends in SAL reporting over the years.

(c) SAL Type Classification Analysis: Use our SAL type classification model to evaluate changes in the distribution and prevalence of different SAL types across years.

Chapter 4

Sentence Decontextualization via LLM-driven Open Information Extraction

The chapter is adapted from a paper under review of ACL2025. My contributions to the study are (in the format of CRediT taxonomy): conceptualization, formal analysis, data curation, investigation, methodology, software, validation, visualization, writing-original draft, writing-review & editing.

4.1 Introduction

Claims are key components of argument structures, often containing the main findings and novel contributions of biomedical publications—making them the central focus of argument mining. However, the interpretation of a claim could be influenced by the rich context in which it appears, where the rich context refers to the surrounding textual and discourse elements such as the document's topic, discourse structure, causal links, and rhetorical cues. Effectively representing extracted knowledge from claims within such rich context where it originally appears is crucial for knowledge-driven natural language understanding tasks. For example, in fact verification tasks, evidence sentences to confirm or refute a claim are originally embedded in nuanced contexts that might influence the evidence interpretation [37], as illustrated in Figure 4.1; in claim extraction tasks, the claims are shaped by a claim sentence as well as its surrounding context in the document, as illutrated in Figure 4.4 [42]. We refer to the task of extracting a sentence together with the context to make the extracted knowledge interpretable while preserving the original sentence meaning as "decontextualization"[43].

The definition of decontextualization by previous work emphasized preserving the meaning of the original sentence while rewriting it by incorporating surrounding contextual information to make it interpretable [43]. However, subsequent works have introduced varying definitions of the term, such as the "decontextualization" for atomic textual patterns extracted from sentences without maintaining the original sentence meaning [64], [65], or "decontextualization" using both the context surrounding the sentence and external context resource [63]. In this work, we treat decontextualization as "preserving the original sentence while adding relevant contextual information to improve knowledge representation of the sentence". Our definition closely

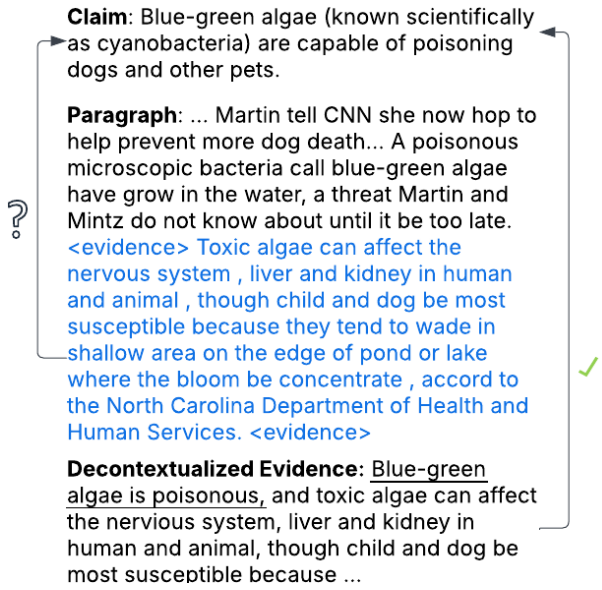

Figure 4.1: A sample for evidence decontextualization to support claim fact-checking [162]. The original evidence sentence does not mention "blue-green algae", while the context indicates "blue-green algae" is "poisonous". By integrating the contextual information and the evidence sentence (decontextualization), it becomes clear that the evidence supports the claim.

aligns with the first definition [43], while not emphasizing sentence rewriting due to rewriting risks distorting the original meaning (e.g. Figure 4.2).

There has been progress in explorations of decontextualization, but several gaps remain in this area. From a supervised learning perspective, a decontextualization dataset is annotated through crowdsourcing, which is then used to fine-tune sequence-to-sequence models [43] and support downstream tasks such as continuous learning for LLMs [163] and fact verification [42]. However, the dataset is created based on each annotator's subjective judgment with a focused collection on the straightforward context information [43], limiting the ability of the finetuned model to address complex contextual dependencies and variations. From the perspective of pretrained models, given the strong language understanding abilities of LLMs, existing frameworks mainly rely on inferencing LLMs and have been designed under different decontextualization settings. Some used context from cross-document [63], and some others decontextualized atomic textual patterns extracted from sentence [64], [65]. However, no generic framework has been developed that satisfies the requirement of retraining the original sentence meaning while adding surrounding contextual information only to provide a standardized and wide-ranged sentence-level decontextualization solution. In addition to the challenges of model development, automatic evaluation of sentence decontextualization remains an open problem, as existing studies use human evaluation [43], [63], or skip the sentence-level decontextualization evaluation [64], [65].

LLMs have advanced the task of extracting structured information from unstructured textual data with their implicit reasoning capabilities [49], [164], [165], which brings opportunities to understand better the connection between a single sentence and the contextual entities, relationships, and patterns. As a sub-domain of structured information extraction, open information extraction (OpenIE) breaks the reliance on predefined schemas [166]–[168]. It works on open-domain data, which has the potential to introduce flexibility for decontextualization to address diverse downstream tasks.

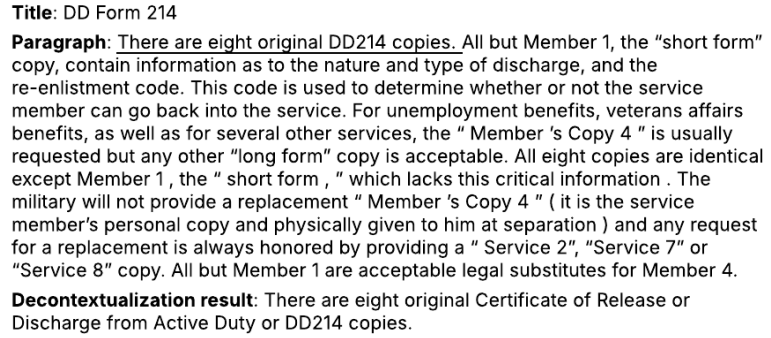

Figure 4.2: An illustration for rewriting risks distorting the original sentence meaning. In the sample, the original sentence refers to eight copies of the DD214. However, the decontextualized sentence is misrepresented as referring separately to DD214 copies, the "Certificate of Release", or "Discharge from Active Duty". Furthermore, a single decontextualization outcome cannot fully cover all possible interpretations, as mentioned in the previous work [43]. The decontextualization could retain more contextual details, such as specifying that the DD214 copies contain information on discharge, re-enlistment, unemployment benefits, and veterans affairs benefits.

In this paper, we investigate LLM-based structured knowledge extraction to support sentence decontextualization, with an emphasis on OpenIE. We propose LLM-DeCon - a unified framework that leverages LLMs to decontextualize sentences within rich contexts. To address the problem with the existing supervised decontextualization models that lack adaptability to the downstream tasks, the framework utilizes prompts following OpenIE paradigm to flexibly extract structured knowledge from the context as supporting information for decontextualizing the sentence. For automatic evaluation, we establish an evaluation benchmark for decontextualization outcomes using three downstream datasets across two tasks—scientific claim contradiction detection and fact verification—comparing baseline models trained on original sentences versus decontextualized sentences.

We summarize our contributions as follows:

- We introduce LLM-DeCon, the first unified framework leveraging LLMs to tackle sentence decontextualization by structured knowledge extraction through OpenIE, enhancing the flexibility and adaptability of the decontextualized sentences.
- We set the evaluation method for sentence decontextualization by comparing models trained on original sentences versus decontextualized sentences. The evaluation is conducted based on three datasets originally designed for different text analysis tasks: Cardiology [169] for scientific claim contradiction detection, and Rawfc [162] and SciFact [36] for fact verification.
- Our LLM-DeCon framework achieves state-of-the-art performance in decontextualization across all evaluation benchmark datasets.

4.2 Related Works

4.2.1 Decontextualization

The concept of "Decontextualization" was first introduced in the table-to-text dataset ToTTo [170], where a sentence within rich context is modified to include additional contextual information to be interpretable

with the aligned table. Choi, Palomaki, Lamm, *et al.* [43] formalize the concept, defining it as "taking a sentence together with its context and rewriting it to be interpretable out of context, while preserving its meaning". They also introduced a dataset designed for fine-tuning sequence-to-sequence models for decontextualization. The fine-tuned decontextualization model has been utilized in various downstream tasks to enhance sentence representations in rich contexts, including continuous knowledge learning for LLMs from new textual sources [171], trustworthiness revision of LLM-generated text using external knowledge [172], and fact verification through claim extraction from documents [42].

In addition to fine-tuned models, pretrained LLMs have also been employed for decontextualization in downstream tasks. For fact verification, LLMs have been used to decontextualize atomic evidences extracted by decomposing the evidence sentences [65] or paragraphs [64]. To improve the representation of user-facing snippet, such as the answers provided to users in a question-answering system, LLMs have been leveraged to decontextualize the snippet through cross-document references [63].

4.2.2 Open Information Extraction

OpenIE is the task of identifying and extracting all possible entity relationships from unstructured text without relying on predefined schemas [173], [174], enabling broader knowledge extraction compared to closed information extraction. Before the era of deep learning, researchers developed rule-based methods with statistical analysis to identify open named entities and relations, which heavily relied on syntax structure extraction [166], [175]–[180]. In the deep learning paradigm, deep neural networks with its ability to automatically learn complex linguistic patterns dominate the supervised OpenIE frameworks, including the RNN model that supports learning of semantic role labeling [181], the bi-LSTM model that enhances syntactic and semantic Learning from bi-directional context [55], [182], and the BERT models that enhances language understanding by attention machenism [183]–[185]. In recent years, LLMs with their ability of comprehending the text without substantial training, have been leveraged in few-shot inference for open information extraction and showing competitive results with supervise-finetuned models [186].

4.3 Methods

In this section, we first introduce the decontextualization notation, then present LLM-DeCon, a framework that supports sentence decontextualization by leveraging structured knowledge extraction with large language models (LLMs). The overview of the proposed framework is shown in Figure 4.3.

4.3.1 Notation

For a sentence S_i extracted from the context C ($\{S_1, S_2, ..., S_n\} \in C$) where the sentence originally appears, the decontextualization outcome S_i^* of S_i within C should satisfy the requirements: a) S_i^* should include enough contextual information to be understood unambiguously and independent from the content; b) S_i^* should preserve the original meaning of S_i, ensuring no information is lost or misrepresented in the sentence.

4.3.2 Structured Knowledge Extraction with LLMs

We leverage the in-context learning ability of LLMs to extract structured knowledge R from C, where the decontextualization target sentence S_i occurs. Specifically, we create a knowledge extraction prompt that combines a demonstration with a query to guide LLMs in generating structured knowledge. The

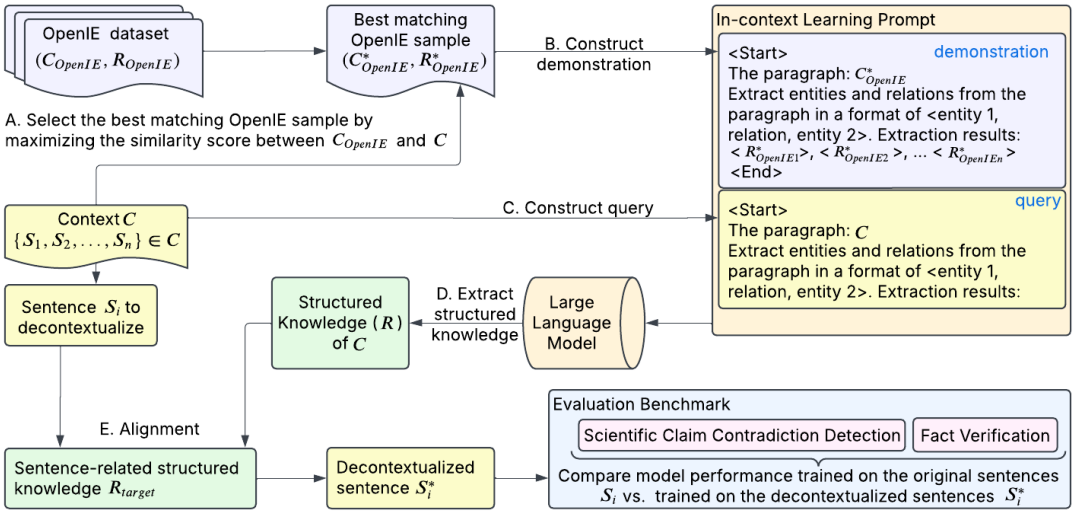

Figure 4.3: Overview of LLM-DeCon

demonstration samples are selected from a pre-annotated document-level OpenIE dataset, which contains context samples C_{OpenIE} and open relations between entities R_{OpenIE} within C_{OpenIE}. The open relations in R_{OpenIE} are represented as "<entity 1, relation, entity 2>". To select the most relevant demonstration, we calculated the cosine similarity scores between the embedding of context C, where the target sentence S_i originally appears, and embeddings of all contexts in the OpenIE dataset C_{OpenIE}. The embeddings are created using the e5-large-v2 model [187], chosen for its promising performance in information retrieval tasks [188]. The best-matching OpenIE context C^*_{OpenIE} used for demonstration is identified as the one with the highest similarity score to C:

$$C^*_{OpenIE} = \arg\max Sim(C_{OpenIE}|C) \tag{4.1}$$

Given the a list of open relations R^*_{OpenIE} corresponding to C^*_{OpenIE}, the demonstration part of the in-context learning prompt Q_{demo} is constructed as:

<Start>
The paragraph: C^*_{OpenIE}.
Extract entities and relations from the paragraph in a format of <entity 1, relation, entity 2>. Extraction results:$<R^*_{OpenIE1}>$, $<R^*_{OpenIE2}>$, ... $<R^*_{OpenIEn}>$
<End>

where $\{R^*_{OpenIE1}, R^*_{OpenIE2}, ..., R^*_{OpenIEn}\} \in R^*_{OpenIE}$, and $<R^*_{OpenIEn}>$ denotes the representation as "<entity 1, relation, entity 2>" using the entities and relation in $R^*_{OpenIEn}$.

Then we create the query prompt Q_{query}. Given the decontextualization target sentence S_i and the context C where the S_i originally occurs, the query portion of the prompt is formatted as:

<Start>
The paragraph: C.

Extract entities and relations from the paragraph in a format of <entity 1, relation, entity 2>. Extraction results:

The input prompt for in-context learning is constructed by combining the demonstration and query:

$$Q = Q_{demo}||Q_{query} \tag{4.2}$$

We then treat the OpenIE for the context C in query Q_{query} as a text generation task. Specifically, we guide the LLM with Q to generate a string $Y_{relations}$ of a list of relation triplets that similar to R^*_{OpenIE} from Q_{demo}:

$$Y_{relations} = \prod_{t=1}^{T} \arg\max_{y_t} P(y_t|Q, y_{0:t-1}) \tag{4.3}$$

The relation triples from the generated relation prediction $Y_{relations}$, which is formatted as a list of *"<entity_ 1, relation, entity_ 2>"* triples, collectively form the structured knowledge set R of context C.

4.3.3 Aligning Sentence to Structured Knowledge

We assume that if an entity in the sentence S_i also appears in some relation tuples R_{target} in R, R_{target} contains contextual information for S_i. Therefore, to decontextualize the sentence, we check the occurrence of entities from the extracted structured knowledge in the decontextualization target sentence, then combine the relation tuples R_{target} that contains those sentence-related entities to form a single context string $R_{target\_combine}$. Next, $R_{target\_combine}$ is combined with the original sentence to form the decontextualization outcome S^*_i:

$$R_{target\_combine} = R^S_{target1}||R^S_{target2}||...||R^S_{targetn} \tag{4.4}$$

$$S^*_i = R_{target\_combine}||S_i \tag{4.5}$$

where $R^S_{targeti}$ denotes the sentence-like representation as *"entity_ 1 relation entity_ 2"* using the entities and relation in $R_{targeti}$, and $\{R_{target1}, R_{target2}, ..., R_{targetn}\} \in R_{target}$.

4.4 Evaluation Benchmark

In this section, we present the evaluation benchmark for decontextualization quality, which covers three datasets across two downstream tasks: fact verification and scientific claim contradiction detection. In the fact verification tasks, the input consists of a claim sentence and related evidence, with the goal of determining whether the claim is supported or refuted by the evidence. In the scientific claim contradiction detection task, the input includes claim sentences from scientific documents, with the objective of identifying whether the claims from different documents contradict each other. The baseline models chosen for each dataset are originally designed to operate on single sentences, without incorporating the surrounding contextual information. The impact of decontextualization on these tasks is evaluated by comparing models trained on the original sentences versus decontextualized sentences.

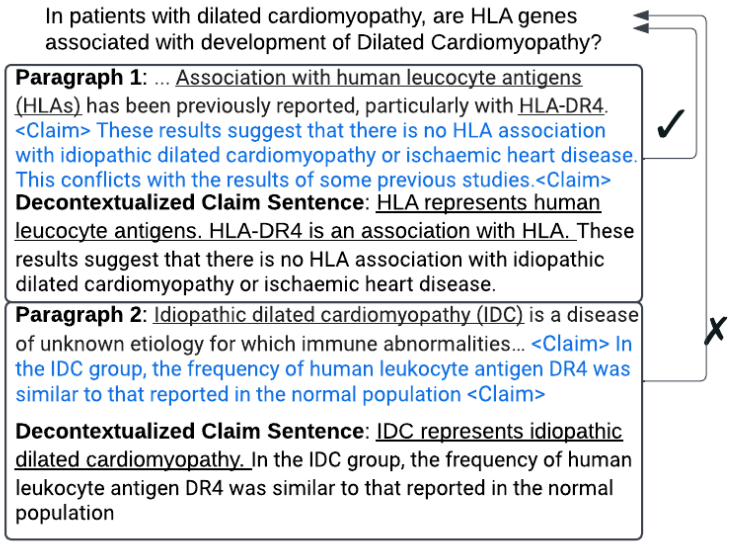

Figure 4.4: Example of claim decontextualization for cross-document contradiction detection [169]: The original claim sentences in paragraphs 1 (PMID: 8297701) and 2 (PMID: 9220309) do not provide sufficient information to determine if they agree with each other. However, after decontextualization (additional information marked by underline), it becomes more obvious that both statements agree that human leukocyte antigen DR4 is not associated with idiopathic dilated cardiomyopathy.

4.4.1 Contradiction Detection in Scientific Claims

Identifying contradictions between claims in scientific documents is crucial for assessing the validity of the source of knowledge [169], [189]–[191]. Given that the scientific claims might be embedded in a rich context, as illustrated in Figure 4.4, decontextualizing claim sentences—especially those whose meaning is influenced by surrounding information—is crucial.

CARDIOLOGY dataset consists of questions paired with scientific claim sentences that either support (labeled as "yes") or refute ("no") the claim presented in the question [169]. The cardiology disease-related questions are manually written by annotators, and the scientific claim sentences are extracted from the PubMed abstracts. We transform the corpus into claim-to-claim pairs, labeling the pair as "non-contradict" if both claims make the same assertion about a question and "contradict" if they differ, aligning with the setting of prior works [192], [193]. A sample from the dataset is presented in Figure 4.4.

We adopt the model achieving state-of-the-art performance on contradiction detection of scientific claims as the baseline [192], [193]. Specifically, the model forms the sentence pairs as "[CLS] sentence 1 [SEP] sentence 2" and predicts the label based on the "[CLS]" token with BERT model as backbone [194].

4.4.2 Fact Verification

Fact verification of statements depends on evidence sentences to support or refute a claim [195], [196]. However, because these sentences are often situated within nuanced contexts, their meaning can be influenced by surrounding information, as illustrated in Figure 4.1, highlighting the importance of decontextualization in ensuring that evidence sentences are accurately represented.

RAWFC [162] and SCIFACT [36] are fact verification datasets that include annotator-written claims, factual

document sources, and sentence-level evidence from documents to support or refute the claims. We use these two datasets to construct claim-evidence pairs, where each claim could be associated with multiple pieces of evidence and labeled as ["true", "false", "half"] (RAWFC), or ["support", "contradict", "not_enough_info"] (SCIFACT).

We use the model that predicts labels based on claim – single-evidence sentence pairs [36] as our baseline for fact verification. Unlike other approaches that incorporate contextual information for better performance [37], [41], [197], this model focuses on single sentences, making it ideal for comparing original and decontextualized sentences. To train the model, we break down the claim–evidence pairs into separate pairs, each consisting of one claim and one evidence sentence, assuming the label for the separated pair aligns with the original claim–evidence pair. We decontextualize the evidence sentences that originate from rich context. To evaluate the performance of the models, we aggregated the predictions from all separate pairs to label each claim through a majority vote.

4.5 Experiments

4.5.1 OpenIE Datasets

We utilize OpenIE datasets from diverse domains to support structured knowledge extraction with in-context learning:

CaRB is a sentence-level dataset for OpenIE [198] that builds upon the OIE2016 dataset [199], which consists of sentences from news articles and encyclopedias. While sharing the same underlying data, CaRB improves the label quality of OIE2016 by crowd-sourcing annotation.

DocOIE is a document-level dataset for OpenIE, containing data from health-care and transportation domains [185]. It addresses the gap in document-level context-aware extraction of relational tuples within the OpenIE domain.

BioRED is a document-level dataset for biomedical information extraction [200]. This biomedical-domain-specific dataset, which broadly encompasses potential relation types among biomedical entities—including 'positive correlation,' 'negative correlation,' 'association,' 'bind,' 'comparison,' 'cotreatment,' 'conversion,' and 'drug interaction'—is included to analyze the impact of domain discrepancy between the OpenIE demonstration in LLM-DeCon and the downstream data on the decontextualization outcomes.

We also created **One-shot** document-level OpenIE annotations for a randomly selected sample from each benchmark dataset. Due to the limited availability of document-level OpenIE datasets, manually annotating a sample from the same domain as the downstream task helps to bridge the domain gap between the OpenIE demonstrations and the queries formulated for the current task.

4.5.2 Decontextualization Baselines

T5 model is employed by previous work as a sequence-to-sequence backbone, fine-tuned on a manually annotated decontextualization dataset [43]. We adopt this model as the decontextualization baseline in our study.

Full-context is included as an additional decontextualization baseline in our study to compare model performance trained on the decontextualized sentences versus the entire context. Specifically, given the original sentence S_i, the decontextualized sentence is presented as "Full text: <full text>. Target sentence: S_i".

4.5.3 Experimental Setup

We choose Llama-3.1-8B [201] as our base LLMs for extracting structural knowledge. Additionally, we employ Ministral-8B-Instruct [202] to compare different LLMs' effectiveness in supporting the LLM-DeCon framework for structured knowledge extraction. All experiments are conducted on a NVIDIA V100 GPU with 32GB of RAM.

4.6 Results and Discussion

4.6.1 Decontextualization

Tasks & Datasets		Contradiction Detection		Fact Verification			
		Cardiology		Rawfc		SciFact	
		Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Baseline models (trained on original sentences)		0.857	0.835	0.455	0.416	0.854	0.829
+ trained on T5 decontextualized sentences		0.873	0.861	0.455	0.419	0.854	0.839
+ trained on full context		0.829	0.824	0.410	0.356	0.642	0.391
+ trained on LLM-Decon decontextualized sentences	CaRB	0.863	0.858	**0.483**	**0.454**	0.837	0.817
	DocOIE	**0.911**	**0.905**	0.455	0.425	0.854	0.833
	BioRED	0.888	0.879	0.410	0.378	**0.870**	0.852
	One-shot	0.875	0.869	0.477	0.439	**0.870**	**0.857**

Table 4.1: Overall performance. The LLM-DeCon presented in this table uses Llama-3.1-8B as the backbone.

Table 4.1 presents the performance of the models trained on original versus decontextualized sentences. From the table, we observe:

(1) Models trained on decontextualized sentences generated by the LLM-DeCon framework with OpenIE datasets from domains similar to the downstream tasks (e.g. CaRB for Rawfc) achieve the highest micro and macro F1 performance across all evaluation benchmarks. This demonstrates the effectiveness of our proposed LLM-DeCon framework in decontextualization. The performance gains by decontextualization compared to baseline models (e.g. Cardiology baseline micro F1: 0.857, LLM-Decon micro F1: 0.911) highlights the importance of decontextualizing sentence extracted from rich context before making further predictions.

(2) Models trained on the decontextualized sentences generated by a T5 model finetuned on the human-annotated decontextualization dataset [43] show no or slight improvement over the baselines (e.g. micro and macro F1 on SciFact as 0.837 and 0.817 respectively, lower than the baseline as 0.854 and 0.829). The results indicate that the straightforward decontextualization in the dataset limits the ability of the fine-tuned model to determine the appropriate contextual extent for a sentence from rich context.

Model	Cardiology		SciFact	
	Micro F1	Macro F1	Micro F1	Macro F1
LLM-DeCon with BioRED demonstration	0.888	0.879	0.870	0.852
w/o Embedding-based Demonstration Selection	0.882	0.875	0.865	0.851
w/o Demonstration	0.859	0.847	0.854	0.839
w/o Sentence Alignment to structured knowledge	0.779	0.773	0.809	0.829

Table 4.2: Ablation Study

(3) The selection of OpenIE datasets in the LLM-DeCon framework impacts the decontextualization quality. Using OpenIE datasets from similar domains as the decontextualization target improves the quality of generated decontextualized sentences. For example, DocOIE, which contains textual data from the healthcare domain, helps the contradiction detection task based on Cardiology dataset (containing cardiology-related scientific abstracts, a sub-domain of the biomedical domain) to achieve the highest micro and macro F1 scores of 0.911 and 0.905, respectively.

In contrast, the decontextualization using OpenIE datasets from the different domains as demonstrations in LLM-DeCon leads to declines in downstream task performance. For example, CaRB, which primarily consists of data from news, aligns well with the dataset Rawfc, which contains data of news or social media (achieving best micro and macro F1 scores 0.483 and 0.454 respectively) but not with scientific text-focused tasks such as SciFact (achieving decreased micro and macro F1 0.837 and 0.817 respectively compared to baseline).

(4) Models trained on decontextualization outcomes generated by LLM-DeCon with the manually annotated OpenIE demonstrations—created by a sample random-selected from each dataset and annotated with human expertise—achieve first or second-best performance on fact verification datasets, and the third-ranked performance on claim contradiction task. These results show the effectiveness of manual OpenIE annotations in bridging the domain gap between demonstrations and queries in the in-context learning prompt for improved decontextualization.

(5) Incorporating full context leads to a decline in model performance, as indicated by that models trained on sentences with full context perform the worst across all settings. This highlights the importance of finding a balance—providing enough information for a sentence to stand alone from context while minimizing contextual bias that could distort original sentence representations.

4.6.2 Ablation Study

We conduct ablation studies on Cardiology and SciFact to evaluate the impact of different LLM-DeCon components using the OpenIE dataset generated from BioRED. Table 4.2 presents the results. "w/o Embedding-based Demonstration Selection" replaces similarity-based demonstration selection with random selection; "w/o Demonstration" removes demonstrations from the in-context learning prompt: "w/o Sentence Alignment to Structured Knowledge" excludes sentence alignment procedure, incorporating all extracted structured knowledge from the context for a single sentence decontextualization.

The performance drop in the "w/o Demonstration" setting underscores the importance of including demonstrations in the prompt to guide the LLM toward generating task-specific results. Similarly, the drop in the "w/o Sentence Alignment" setting highlights that incorporating all extracted knowledge from context

can introduce additional bias into the decontextualized results. However, the model trained under the "w/o Embedding-based Demonstration Selection" setting performs comparably to the full model, indicating that embedding-based demonstration selection plays a relatively minor role in LLM-DeCon.

4.6.3 LLM Comparison

LLM-Decon	Ministral-8B-Instruct			
	CARDIOLOGY		SCIFACT	
	MicF1	MacF1	MicF1	MacF1
CaRB	0.863	0.855	0.820	0.820
DocOIE	0.898	0.891	0.878	0.850
BioRED	0.901	0.896	0.859	0.846
Manual	0.865	0.857	0.870	0.854

Table 4.3: LLM Comparison

Table 4.3 compares the performance of Llama-3.1-8B and Ministral-8B-Instruct on CARDIOLOGY and SCIFACT. The results show that the Ministral model performs comparably with the Llama model in supporting structured knowledge extraction, which demonstrates the generalizability of our proposed LLM-DeCon framework across LLMs. Additionally, the consistently lowest performance observed with CaRB for both models further confirms that selecting an OpenIE dataset from a domain different from the downstream task negatively impacts decontextualization quality.

4.6.4 Error Analysis

We present a sentence decontextualization sample in Figure 4.5. The T5 model rewrites the original sentence by removing "therefore" but fails to incorporate additional context that could enrich the knowledge and support the claim. In contrast, our LLM-DeCon model successfully integrates contextual information into the decontextualized output, providing sufficient support for the given claim.

4.7 Conclusions

This work introduced LLM-DeCon, a framework leveraging LLM-supported structured knowledge extraction via OpenIE to enhance sentence decontextualization. Our approach improves adaptability and handles complex contextual dependencies for decontextualization. We also established an automatic evaluation benchmark, showing that LLM-DeCon achieves strong performance in claim and evidence sentence decontextualization. The results highlight the potential of structured knowledge extraction in improving decontextualization. Future research can refine evaluation metrics and explore the framework's adaptability in broader domains.

Claim: Vaccinating the gastrointestinal tract induces protection of rectal and vaginal mucosa.
Evidence Sentence: ... Therefore, we designed a large intestine-targeted oral delivery with ph-dependent microparticles containing vaccine nanoparticles, which induced colorectal immunity in mice comparably to colorectal vaccination and protected against rectal and vaginal viral challenge...
Evidence Context: Abstract of PMC3475749
LLM-DeCon (one-shot): rectal and vaginal mucosal surfaces are protected by vaccination, therefore, we designed a large intestine-targeted oral delivery with pH-dependent microparticles containing vaccine nanoparticles, which induced colorectal immunity in mice comparably to colorectal vaccination and protected against rectal and vaginal viral challenge.
T5: We designed a large intestine-targeted oral delivery with pH-dependent microparticles containing vaccine nanoparticles, which induced colorectal immunity in mice comparably to colorectal vaccination and protected against rectal and vaginal viral challenge.

Figure 4.5: Decontextaulization sample by T5 and LLM-DeCon on SciFact.

4.8 Future Work

In the current work, we have explored how to use LLMs to decontextualize sentences. In the next step, we hope to to apply this method to a downstream task: identifying and analyzing conflicting claims in biomedical publications. Scientific claims are often shaped by their surrounding context—such as experimental design or research objectives—which can make cross-study comparisons challenging. By enriching claim sentences with essential contextual information, decontextualization can improve their interpretability in isolation. This, in turn, may allow for more accurate and meaningful comparisons of claims across different studies. Motivated by this potential, we will focus on leveraging decontextualized claim sentence to support the detection of contradictions between claims reported across multiple articles.

We plan to develop a multi-step methodology to achieve the goal of detecting claim contradictions across biomedical articles:

(a) Select a representative subset of articles from the PubMed Central Open Access dataset (around 200 publications), covering a predefined list of research topics (e.g. breast cancer and Alzheimer's disease)

(b) The argument structure detection tool developed in Chapter 2 will be utilized to detect the claim sentences from a given article.

(c) Utilize a topic clustering model to group articles that present claims on similar topics.

(d) The decontextualization method developed in this chapter (Chapter 4) will be utilized to enrich the claim sentences with enough context to make them interpretable out of context.

(e) The sentence pair classification model will be developed using existing, widely adopted biomedical natural language inference datasets (e.g., MedNLI [203]) to enable effective detection of contradictions between sentence pairs.

(d) The sentence-pair classification model developed for contradiction detection in the previous step is then utilized to detect the contradictions among the decontextualized claims from step (b).

(e) The quality of the automatically detected contradictions across articles will be checked manually.

Chapter 5

Timeline for the Remaining Work

Figure 5.1 shows the proposed timeline for the future works described at the end of Chapters 2-4. I plan to complete the works by May 2026, focusing on the remaining tasks in the next 10 months, while leaving dedicated time for the dissertation writing for the whole year. The final dissertation is expected to be completed in May 2026.

	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May
Chapter 2												
Full-text Collection & Rule-based Argument Labeling	■	■	■	■								
Checking and Improving Quality of Rule-based Annotation Results			■	■	■	■	■	■				
Chapter 3												
Data Collection (from Different Years)	■	■	■	■								
SAL Identification Analysis Over Years			■	■	■	■						
SAL Type Classification Analysis Over Years			■	■	■	■						
Chapter 4												
Article Collection and Clustering	■	■	■	■								
Claim Detection and Decontextualization			■	■	■	■						
Contradiction Detection among Claims					■	■	■					
Contradiction Detection Quality Checking by Human						■	■	■				
Dissertation Writeup	■	■	■	■	■	■	■	■	■	■	■	■
Dissertation Defense											■	■

Table 5.1: Timeline for the remaining work.

As for publications, the current contents in the Chapters 2-4 are either published or under review. I also aim to write follow-up papers as described in the future work section in each chapter.

References

[1] N. R. Smalheiser, G. Hahn-Powell, D. Hristovski, and Y. Sebastian, "From knowledge discovery to knowledge creation: How can literature-based discovery accelerate progress in science?" In *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, 2023.

[2] R. Bhatnagar, S. Sardar, M. Beheshti, and J. T. Podichetty, "How can natural language processing help model informed drug development?: A review," *JAMIA open*, vol. 5, no. 2, ooac043, 2022.

[3] D. C. Angus, A. J. Huang, R. J. Lewis, *et al.*, "The integration of clinical trials with the practice of medicine: Repairing a house divided," *Jama*, vol. 332, no. 2, pp. 153–162, 2024.

[4] A. Johnston, S. E. Kelly, S.-C. Hsieh, B. Skidmore, and G. A. Wells, "Systematic reviews of clinical practice guidelines: A methodological guide," *Journal of clinical epidemiology*, vol. 108, pp. 64–76, 2019.

[5] Q. Jin, R. Leaman, and Z. Lu, "Pubmed and beyond: Biomedical literature search in the age of artificial intelligence," *EBioMedicine*, vol. 100, 2024.

[6] W. Kim, L. Yeganova, D. C. Comeau, W. J. Wilbur, and Z. Lu, "Towards a unified search: Improving pubmed retrieval with full text," *Journal of biomedical informatics*, vol. 134, p. 104 211, 2022.

[7] S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, and G. B. Kitchen, "Natural language processing in medicine: A review," *Trends in Anaesthesia and Critical Care*, vol. 38, pp. 4–9, 2021.

[8] Z. Lu, Y. Peng, T. Cohen, M. Ghassemi, C. Weng, and S. Tian, "Large language models in biomedicine and health: Current research landscape and future directions," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1801–1811, 2024.

[9] H. Kilicoglu, L. Jiang, L. Hoang, E. Mayo-Wilson, C. H. Vinkers, and W. M. Otte, "Methodology reporting improved over time in 176,469 randomized controlled trials," *Journal of Clinical Epidemiology*, vol. 162, pp. 19–28, 2023.

[10] R. Rodrıguez-Pérez and J. Bajorath, "Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery," *Journal of Computer-Aided Molecular Design*, vol. 36, no. 5, pp. 355–362, 2022.

[11] R. J. Carroll, A. E. Eyler, and J. C. Denny, "Naıve electronic health record phenotype identification for rheumatoid arthritis," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2011, 2011, p. 189.

[12] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Machine learning techniques for biomedical natural language processing: A comprehensive review," *IEEE Access*, vol. 9, pp. 140 628–140 653, 2021.

[13] A. Agibetov, K. Blagec, H. Xu, and M. Samwald, "Fast and scalable neural embedding models for biomedical sentence classification," *BMC bioinformatics*, vol. 19, pp. 1–9, 2018.

[14] D. Jin and P. Szolovits, "Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3100–3109.

[15] X. Jiang, B. Zhang, Y. Ye, and Z. Liu, "A hierarchical model with recurrent convolutional neural networks for sequential sentence classification," in *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, Springer, 2019, pp. 78–89.

[16] S. Gonçalves, P. Cortez, and S. Moro, "A deep learning classifier for sentence classification in biomedical and computer science abstracts," *Neural Computing and Applications*, vol. 32, pp. 6793–6807, 2020.

[17] K. Yamada, T. Hirao, R. Sasano, K. Takeda, and M. Nagata, "Sequential span classification with neural semi-markov crfs for biomedical abstracts," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 871–877.

[18] X. Shang, Q. Ma, Z. Lin, J. Yan, and Z. Chen, "A span-based dynamic local attention model for sequential sentence classification," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 198–203.

[19] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[20] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[21] J. Devlin, M.-W. Chang, K. Lee, and 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)*, 2019, pp. 4171–4186.

[22] Y. Gu, R. Tinn, H. Cheng, *et al.*, *Domain-specific language model pretraining for biomedical natural language processing*, 2020. eprint: arXiv:2007.15779.

[23] S. Zhao, C. Su, Z. Lu, and F. Wang, "Recent advances in biomedical literature mining," *Briefings Bioinform.*, vol. 22, no. 3, 2021. DOI: 10.1093/BIB/BBAA057. [Online]. Available: https://doi.org/10.1093/bib/bbaa057.

[24] B. Song, F. Li, Y. Liu, and X. Zeng, "Deep learning methods for biomedical named entity recognition: A survey and qualitative comparison," *Briefings in Bioinformatics*, vol. 22, no. 6, bbab282, 2021.

[25] Q. Li, H. Peng, J. Li, *et al.*, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.

[26] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artificial intelligence review*, vol. 52, no. 1, pp. 273–292, 2019.

[27] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," *Artificial intelligence review*, vol. 56, no. 9, pp. 9401–9469, 2023.

[28] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[29] B. Wang, Q. Xie, J. Pei, *et al.*, "Pre-trained language models in biomedical domain: A systematic survey," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–52, 2023.

[30] M. Shardlow, R. Batista-Navarro, P. Thompson, R. Nawaz, J. McNaught, and S. Ananiadou, "Identification of research hypotheses and new knowledge from scientific literature," *BMC medical informatics and decision making*, vol. 18, pp. 1–13, 2018.

[31] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 45, no. 4, pp. 765–818, 2020.

[32] A. Vassiliades, N. Bassiliades, and T. Patkos, "Argumentation and explainable artificial intelligence: A survey," *The Knowledge Engineering Review*, vol. 36, e5, 2021.

[33] C. Blake, "Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles," *Journal of biomedical informatics*, vol. 43, no. 2, pp. 173–189, 2010.

[34] D. H. Park and C. Blake, "Identifying comparative claim sentences in full-text scientific articles," in *Proceedings of the workshop on detecting structure in scholarly discourse*, 2012, pp. 1–9.

[35] A. A. Prabhakar, S. Mohtaj, and S. Möller, "Claim extraction from text using transfer learning.," in *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 2020, pp. 297–302.

[36] D. Wadden, S. Lin, K. Lo, *et al.*, "Fact or fiction: Verifying scientific claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609. [Online]. Available: https://aclanthology.org/2020.emnlp-main.609/.

[37] D. Wadden, K. Lo, L. L. Wang, A. Cohan, I. Beltagy, and H. Hajishirzi, "MultiVerS: Improving scientific claim verification with weak supervision and full-document context," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 61–76. DOI: 10.18653/v1/2022.findings-naacl.6. [Online]. Available: https://aclanthology.org/2022.findings-naacl.6/.

[38] F. Dernoncourt and J. Y. Lee, "Pubmed 200k rct: A dataset for sequential sentence classification in medical abstracts," *arXiv preprint arXiv:1710.06071*, 2017.

[39] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. Weld, "Pretrained language models for sequential sentence classification," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3693–3699. DOI: 10.18653/v1/D19-1383. [Online]. Available: https://aclanthology.org/D19-1383.

[40] D. Jin and P. Szolovits, "Hierarchical neural networks for sequential sentence classification in medical scientific abstracts," *arXiv preprint arXiv:1808.06161*, 2018.

[41] R. Pradeep, X. Ma, R. Nogueira, and J. Lin, "Scientific claim verification with VerT5erini," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, and F. Rinaldi, Eds., online: Association for Computational Linguistics, Apr. 2021, pp. 94–103. [Online]. Available: https://aclanthology.org/2021.louhi-1.11/.

[42] Z. Deng, M. Schlichtkrull, and A. Vlachos, "Document-level claim extraction and decontextualisation for fact-checking," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 11 943–11 954. DOI: 10.18653/v1/2024.acl-long.645. [Online]. Available: https://aclanthology.org/2024.acl-long.645/.

[43] E. Choi, J. Palomaki, M. Lamm, T. Kwiatkowski, D. Das, and M. Collins, "Decontextualization: Making sentences stand-alone," *Transactions of the Association for Computational Linguistics*, vol. 9, B. Roark and A. Nenkova, Eds., pp. 447–461, 2021. DOI: 10.1162/tacl_a_00377. [Online]. Available: https://aclanthology.org/2021.tacl-1.27/.

[44] A. Segura-Tinoco and I. Cantador, "Argael: Argument annotation and evaluation tool," *SoftwareX*, vol. 23, p. 101 410, 2023.

[45] K. Al Khatib, T. Ghosal, Y. Hou, A. de Waard, and D. Freitag, "Argument mining for scholarly document processing: Taking stock and looking ahead," in *Proceedings of the Second Workshop on Scholarly Document Processing*, I. Beltagy, A. Cohan, G. Feigenblat, *et al.*, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 56–65. DOI: 10.18653/v1/2021.sdp-1.7. [Online]. Available: https://aclanthology.org/2021.sdp-1.7/.

[46] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[47] N. Wies, Y. Levine, and A. Shashua, "The learnability of in-context learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 637–36 651, 2023.

[48] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "Enhancing argument structure extraction with efficient leverage of contextual information," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7563–7571. DOI: 10.18653/v1/2023.findings-emnlp.507. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.507/.

[49] J. Dagdelen, A. Dunn, S. Lee, *et al.*, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, p. 1418, 2024.

[50] A. de Wynter and T. Yuan, ""i'd like to have an argument, please": Argumentative reasoning in large language models," in *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18-20, 2014*, C. Reed, M. Thimm, and T. Rienstra, Eds., ser. Frontiers in Artificial Intelligence and Applications, vol. 388, IOS Press, 2024, pp. 73–84. DOI: 10.3233/FAIA240311. [Online]. Available: https://doi.org/10.3233/FAIA240311.

[51] R. M. Palau and M.-F. Moens, "Argumentation mining: The detection, classification and structure of arguments in text," in *Proceedings of the 12th international conference on artificial intelligence and law*, 2009, pp. 98–107.

[52] A. J. Yepes, J. G. Mork, and A. R. Aronson, "Using the argumentative structure of scientific literature to improve information access," in *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, 2013, pp. 102–110.

[53] S. Agarwal and H. Yu, "Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion," *Bioinformatics*, vol. 25, no. 23, pp. 3174–3180, 2009.

[54] T. Bao, H. Zhang, and C. Zhang, "Enhancing abstractive summarization of scientific papers using structure information," *Expert Systems with Applications*, vol. 261, p. 125 529, 2025.

[55] Z. Jiang, P. Yin, and G. Neubig, "Improving open information extraction via iterative rank-aware learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5295–5300. DOI: 10.18653/v1/P19-1523. [Online]. Available: https://aclanthology.org/P19-1523/.

[56] X. Li, G. Burns, and N. Peng, "Scientific Discourse Tagging for Evidence Extraction," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2550–2562.

[57] A. Brack, A. Hoppe, P. Buschermöhle, and R. Ewerth, "Cross-domain multi-task learning for sequential sentence classification in research papers," in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 2022, pp. 1–13.

[58] D. Mollá, "Overview of the 2022 alta shared task: Piboso sentence classification, 10 years later," in *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, 2022, pp. 178–182.

[59] M. Lan, M. Cheng, L. Hoang, G. Ter Riet, and H. Kilicoglu, "Automatic categorization of self-acknowledged limitations in randomized controlled trial publications," *Journal of biomedical informatics*, vol. 152, p. 104 628, 2024.

[60] H. Kilicoglu, G. Rosemblat, M. Malički, and G. Ter Riet, "Automatic recognition of self-acknowledged limitations in clinical research literature," *Journal of the American Medical Informatics Association*, vol. 25, no. 7, pp. 855–861, 2018.

[61] G. Ter Riet, P. Chesley, A. G. Gross, *et al.*, "All that glitters isn't gold: A survey on acknowledgment of limitations in biomedical studies," *PloS one*, vol. 8, no. 11, e73623, 2013.

[62] G. Alvarez, R. Núñez-Cortés, I. Solà, *et al.*, "Sample size, study length, and inadequate controls were the most common self-acknowledged limitations in manual therapy trials: A methodological review," *Journal of Clinical Epidemiology*, vol. 130, pp. 96–106, 2021.

[63] B. Newman, L. Soldaini, R. Fok, A. Cohan, and K. Lo, "A question answering framework for decontextualizing user-facing snippets from scientific documents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3194–3212. DOI: 10.18653/v1/2023.emnlp-main.193. [Online]. Available: https://aclanthology.org/2023.emnlp-main.193/.

[64] Y. Wang, R. Gangi Reddy, Z. M. Mujahid, *et al.*, "Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 14 199–14 230. DOI: 10.18653/v1/2024.findings-emnlp.830. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.830/.

[65] A. Gunjal and G. Durrett, "Molecular facts: Desiderata for decontextualization in LLM fact verification," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3751–3768. DOI: 10.18653/v1/2024.findings-emnlp.215. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.215/.

[66] S. Teufel and M. Moens, "Sentence extraction and rhetorical classification for flexible abstracts," in *AAAI Spring Symposium on Intelligent Text summarization*, 1998, pp. 89–97.

[67] S. Pradhan, K. Hacioglu, W. Ward, J. H. Martin, and D. Jurafsky, "Semantic role parsing: Adding semantic structure to unstructured text," in *Third IEEE international conference on data mining*, IEEE, 2003, pp. 629–632.

[68] A. Jimeno Yepes, J. Mork, and A. Aronson, "Using the argumentative structure of scientific literature to improve information access," in *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 102–110. [Online]. Available: https://aclanthology.org/W13-1913.

[69] G. Team, T. Mesnard, C. Hardin, *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[70] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "Neural networks for joint sentence classification in medical paper abstracts," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 694–700. [Online]. Available: https://www.aclweb.org/anthology/E17-2110.

[71] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, BioMed Central, vol. 12, 2011, pp. 1–10.

[72] C. Stead, S. Smith, P. Busch, and S. Vatanasakdakul, "Emerald 110k: A multidisciplinary dataset for abstract sentence classification," in *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, 2019, pp. 120–125.

[73] C. Dayrell, A. Candido Jr, G. Lima, *et al.*, "Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora.," in *LREC*, 2012, pp. 1604–1609.

[74] B. Fisas, H. Saggion, and F. Ronzano, "On the discoursive structure of computer graphics research papers," in *Proceedings of the 9th linguistic annotation workshop*, 2015, pp. 42–51.

[75] M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor, "Corpora for the conceptualisation and zoning of scientific papers," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/644_Paper.pdf.

[76] L. McKnight and P. Srinivasan, "Categorization of sentence types in medical abstracts," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2003, 2003, p. 440.

[77] J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur, "Generative content models for structural analysis of medical abstracts," in *Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology*, 2006, pp. 65–72.

[78] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[79] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, Ieee, 2017, pp. 1–6.

[80] R. Wang, X. Dai, *et al.*, "Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 672–679.

[81] Y. Zhang, Z. Shen, C.-H. Wu, *et al.*, "Metadata-induced contrastive learning for zero-shot multi-label text classification," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3162–3173.

[82] L. Zheng, Z. Chen, J. He, and H. Chen, "MULAN: multi-modal causal structure learning and root cause analysis for microservice systems," in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, T. Chua, C. Ngo, R. Kumar, H. W. Lauw, and R. K. Lee, Eds., ACM, 2024, pp. 4107–4116.

[83] S. S. S. Das, A. Katiyar, R. Passonneau, and R. Zhang, "CONTaiNER: Few-shot named entity recognition via contrastive learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6338–6353. DOI: 10.18653/v1/2022.acl-long.439. [Online]. Available: https://aclanthology.org/2022.acl-long.439.

[84] Y. Huang, K. He, Y. Wang, *et al.*, "Copner: Contrastive learning with prompt guiding for few-shot named entity recognition," in *Proceedings of the 29th International conference on computational linguistics*, 2022, pp. 2515–2527.

[85] X. Zhang, J. Yuan, L. Li, and J. Liu, "Reducing the bias of visual objects in multimodal named entity recognition," in *Proceedings of the Sixteenth ACM international conference on web search and data mining*, 2023, pp. 958–966.

[86] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552.

[87] P. P. Liang, Z. Deng, M. Q. Ma, J. Y. Zou, L.-P. Morency, and R. Salakhutdinov, "Factorized contrastive learning: Going beyond multi-view redundancy," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[88] Q. Dong, L. Li, D. Dai, *et al.*, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[89] X. Sun, X. Li, J. Li, *et al.*, "Text classification via large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8990–9005. DOI: 10.18653/v1/2023.findings-emnlp.603. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.603.

[90] S. Wadhwa, S. Amir, and B. Wallace, "Revisiting relation extraction in the era of large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 15 566–15 589. DOI: 10.18653/v1/2023.acl-long.868. [Online]. Available: https://aclanthology.org/2023.acl-long.868.

[91] S. Goyal, Z. Ji, A. S. Rawat, A. K. Menon, S. Kumar, and V. Nagarajan, "Think before you speak: Training language models with pause tokens," *arXiv preprint arXiv:2310.02226*, 2023.

[92] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[93] P. Khosla, P. Teterwak, C. Wang, *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[94] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[95] Q. Chen, R. Zhang, Y. Zheng, and Y. Mao, "Dual contrastive learning: Text classification via label-aware data augmentation," *arXiv preprint arXiv:2201.08702*, 2022.

[96] S. Xie, C. Hou, H. Yu, Z. Zhang, X. Luo, and N. Zhu, "Multi-label disaster text classification via supervised contrastive learning for social media data," *Computers and Electrical Engineering*, vol. 104, p. 108 401, 2022.

[97] L. Zheng, Y. Zhu, and J. He, "Fairness-aware multi-view clustering," in *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, S. Shekhar, Z. Zhou, Y. Chiang, and G. Stiglic, Eds., SIAM, 2023, pp. 856–864.

[98] L. Zheng, J. Xiong, Y. Zhu, and J. He, "Contrastive learning with complex heterogeneity," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2594–2604.

[99] L. Zheng, B. Jing, Z. Li, H. Tong, and J. He, "Heterogeneous contrastive learning for foundation models and beyond," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, R. Baeza-Yates and F. Bonchi, Eds., ACM, 2024, pp. 6666–6676.

[100] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[101] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[102] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *EMNLP*, Association for Computational Linguistics, 2019. [Online]. Available: https://www.aclweb.org/anthology/D19-1371.

[103] H. Else, "How a torrent of COVID science changed research publishing-in seven charts.," *Nature*, pp. 553–553, 2020.

[104] C. Watson, "Rise of the preprint: how rapid data sharing during COVID-19 has changed science forever," *Nature Medicine*, vol. 28, no. 1, pp. 2–5, 2022.

[105] K. A. Bramstedt, "The carnage of substandard research during the COVID-19 pandemic: a call for quality," *Journal of Medical Ethics*, vol. 46, no. 12, pp. 803–807, 2020.

[106] M. Zdravkovic, J. Berger-Estilita, B. Zdravkovic, and D. Berger, "Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study," *PLoS One*, vol. 15, no. 11, e0241826, 2020.

[107] T. J. Quinn, J. K. Burton, B. Carter, *et al.*, "Following the science? Comparison of methodological and reporting quality of COVID-19 and other research from the first wave of the pandemic," *BMC Medicine*, vol. 19, no. 1, pp. 1–10, 2021.

[108] R. G. Jung, P. Di Santo, C. Clifford, *et al.*, "Methodological quality of covid-19 clinical research," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[109] J. P. Ioannidis, "Limitations are not properly acknowledged in the scientific literature," *Journal of Clinical Epidemiology*, vol. 60, no. 4, pp. 324–329, 2007.

[110] P. T. Ross and N. L. Bibler Zaidi, "Limited by our limitations," *Perspectives on medical education*, vol. 8, pp. 261–264, 2019.

[111] M. A. Puhan, E. A. Akl, D. Bryant, F. Xie, G. Apolone, and G. t. Riet, "Discussing study limitations in reports of biomedical studies-the need for more transparency," *Health and quality of life outcomes*, vol. 10, no. 1, pp. 1–4, 2012.

[112] D. Lahav, J. S. Falcon, B. Kuehl, *et al.*, "A search engine for discovery of scientific challenges and directions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 11 982–11 990.

[113] G. ter Riet, P. Chesley, A. G. Gross, *et al.*, "All That Glitters Isn't Gold: A Survey on Acknowledgment of Limitations in Biomedical Studies," *PLOS ONE*, vol. 8, no. 11, pp. 1–6, Nov. 2013. DOI: 10.1371/journal.pone.0073623.

[114] D. Moher, S. Hopewell, K. F. Schulz, *et al.*, "CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials," *BMJ*, vol. 340, 2010. DOI: 10.1136/bmj.c869.

[115] A. Bhide, P. S. Shah, and G. Acharya, "A simplified guide to randomized controlled trials," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 97, no. 4, pp. 380–387, 2018.

[116] I. Chalmers and P. Glasziou, "Avoidable waste in the production and reporting of research evidence," *The Lancet*, vol. 374, no. 9683, pp. 86–89, 2009. DOI: 10.1016/s0140-6736(09)60329-9.

[117] S. N. Goodman, J. Berlin, S. W. Fletcher, and R. H. Fletcher, "Manuscript quality before and after peer review and editing at Annals of Internal Medicine," *Annals of Internal Medicine*, vol. 121, no. 1, pp. 11–21, 1994.

[118] J. H. Price and J. Murnan, "Research limitations and the necessity of reporting them," *American Journal of Health Education*, vol. 35, no. 2, p. 66, 2004.

[119] K. F. Schulz, D. G. Altman, and D. Moher, "Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials," *Journal of Pharmacology and pharmacotherapeutics*, vol. 1, no. 2, pp. 100–107, 2010.

[120] S. N. Goodman, J. Berlin, S. W. Fletcher, and R. H. Fletcher, "Manuscript quality before and after peer review and editing at Annals of Internal Medicine," *Annals of internal medicine*, vol. 121, no. 1, pp. 11–21, 1994.

[121] L. Turner, L. Shamseer, D. G. Altman, K. F. Schulz, and D. Moher, "Does use of the consort statement impact the completeness of reporting of randomised controlled trials published in medical journals? a cochrane review a," *Systematic reviews*, vol. 1, pp. 1–7, 2012.

[122] N. Pandis, L. Shamseer, V. G. Kokich, P. S. Fleming, and D. Moher, "Active implementation strategy of CONSORT adherence by a dental specialty journal improved randomized clinical trial reporting," *Journal of Clinical Epidemiology*, vol. 67, no. 9, pp. 1044–1048, 2014.

[123] Y. Jin, N. Sanger, I. Shams, *et al.*, "Does the medical literature remain inadequately described despite having reporting guidelines for 21 years?–a systematic review of reviews: An update," *Journal of multidisciplinary healthcare*, vol. 11, p. 495, 2018.

[124] H. Kilicoglu, G. Rosemblat, L. Hoang, *et al.*, "Toward assessing clinical trial publications for reporting transparency," *Journal of Biomedical Informatics*, vol. 116, p. 103 717, 2021.

[125] T. Weissgerber, N. Riedel, H. Kilicoglu, *et al.*, "Automated screening of covid-19 preprints: Can we help authors to improve transparency and reproducibility?" *Nature Medicine*, vol. 27, no. 1, pp. 6–7, 2021.

[126] R. Schulz, A. Barnett, R. Bernard, *et al.*, "Is the future of peer review automated?" *BMC Research Notes*, vol. 15, no. 1, pp. 1–5, 2022.

[127] H. Kilicoglu, "Biomedical text mining for research rigor and integrity: tasks, challenges, directions," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1400–1414, 2018.

[128] J. Menke, M. Roelandse, B. Ozyurt, M. Martone, and A. Bandrowski, "The rigor and transparency index quality metric for assessing biological and medical science methods," *Iscience*, vol. 23, no. 11, p. 101 698, 2020.

[129] K. Keserlioglu, H. Kilicoglu, and G. Ter Riet, "Impact of peer review on discussion of study limitations and strength of claims in randomized trial reports: A before and after study," *Research integrity and peer review*, vol. 4, no. 1, p. 19, 2019.

[130] Y. Gu, R. Tinn, H. Cheng, *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[131] D. Demner-Fushman and J. Lin, "Answering clinical questions with knowledge-based and statistical techniques," *Computational Linguistics*, vol. 33, no. 1, pp. 63–103, 2007.

[132] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," *BMC Bioinformatics*, vol. 12, no. 2, pp. 1–10, 2011.

[133] B. C. Wallace, J. Kuiper, A. Sharma, M. Zhu, and I. J. Marshall, "Extracting PICO sentences from clinical trial reports using supervised distant supervision," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 4572–4596, 2016.

[134] D. Jin and P. Szolovits, "Advancing PICO element detection in biomedical text via deep neural networks," *Bioinformatics*, vol. 36, no. 12, pp. 3856–3862, 2020.

[135] Y. Hu, V. K. Keloth, K. Raja, Y. Chen, and H. Xu, "Towards precise PICO extraction from abstracts of randomized controlled trials using a section-specific learning approach," *Bioinformatics*, vol. 39, no. 9, btad542, 2023.

[136] B. Nye, J. J. Li, R. Patel, *et al.*, "A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 197–207. [Online]. Available: https://www.aclweb.org/anthology/P18-1019.

[137] T. Kang, S. Zou, and C. Weng, "Pretraining to recognize PICO elements from randomized controlled trial literature," *Studies in Health Technology and Informatics*, vol. 264, p. 188, 2019.

[138] N. Stylianou, G. Razis, D. G. Goulis, and I. Vlahavas, "EBM+: advancing evidence-based medicine via two level automatic identification of populations, interventions, outcomes in medical literature," *Artificial Intelligence in Medicine*, vol. 108, p. 101 949, 2020.

[139] F. Mutinda, K. Liew, S. Yada, S. Wakamiya, and E. Aramaki, "PICO Corpus: A Publicly Available Corpus to Support Automatic Data Extraction from Biomedical Literature," in *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, 2022, pp. 26–31.

[140] I. J. Marshall, J. Kuiper, and B. C. Wallace, "RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials," *Journal of the American Medical Informatics Association*, vol. 23, no. 1, pp. 193–201, 2016.

[141] S. Kiritchenko, B. De Bruijn, S. Carini, J. Martin, and I. Sim, "ExaCT: automatic extraction of clinical trial characteristics from journal publications," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, pp. 1–17, 2010.

[142] L. Hoang, Y. Guan, and H. Kilicoglu, "Methodological information extraction from randomized controlled trial publications: a pilot study," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2022, 2022, p. 542.

[143] L. Hoang, L. Jiang, and H. Kilicoglu, "Investigating the impact of weakly supervised data on text mining models of publication transparency: A case study on randomized controlled trials," in *AAMIA Informatics Summit Proceedings*, American Medical Informatics Association, vol. 2022, 2022, pp. 254–263.

[144] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *The VLDB Journal*, vol. 29, no. 2, pp. 709–730, 2020.

[145] S. Y. Feng, V. Gangal, J. Wei, *et al.*, "A Survey of Data Augmentation Approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 968–988.

[146] J. Wei and 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)*, 2019, pp. 6382–6388.

[147] T. Kang, A. Perotte, Y. Tang, C. Ta, and C. Weng, "UMLS-based data augmentation for natural language processing of clinical research literature," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 812–823, 2021.

[148] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, *et al.*, "Do not have enough data? Deep learning to the rescue!" In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 7383–7390.

[149] Y. Yang, C. Malaviya, J. Fernandez, *et al.*, "Generative data augmentation for commonsense reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1008–1025. DOI: 10.18653/v1/2020.findings-emnlp.90. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.90.

[150] Y. Wang, C. Xu, Q. Sun, *et al.*, "PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.292.

[151] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[152] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059.

[153] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining: Applications and Theory*, pp. 1–20, 2010.

[154] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[155] V. P. Bhapkar, "A note on the equivalence of two test criteria for hypotheses in categorical data," *Journal of the American Statistical Association*, vol. 61, no. 313, pp. 228–235, 1966.

[156] R. Artstein and M. Poesio, "Survey article: Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008. DOI: 10.1162/coli.07-034-R2. [Online]. Available: https://aclanthology.org/J08-4004.

[157] G. Alvarez, I. Sola, M. Sitja-Rabert, *et al.*, "A methodological review revealed that reporting of trials in manual therapy has not improved over time," *Journal of clinical epidemiology*, vol. 121, pp. 32–44, 2020.

[158] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[159] M. Grootendorst, *Keybert: Minimal keyword extraction with bert.* Version v0.3.0, 2020. DOI: 10.5281/zenodo.4461265. [Online]. Available: https://doi.org/10.5281/zenodo.4461265.

[160] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, PMLR, 2017, pp. 3319–3328.

[161] M. A. Hernán and J. M. Robins, *Causal inference*. CRC Boca Raton, FL, 2010.

[162] Z. Yang, J. Ma, H. Chen, H. Lin, Z. Luo, and Y. Chang, "A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, *et al.*, Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2608–2621. [Online]. Available: https://aclanthology.org/2022.coling-1.230/.

[163] J. Jang, S. Ye, S. Yang, *et al.*, "Towards continual knowledge learning of language models," in *International Conference on Learning Representations*, 2021.

[164] B. Y. Lin, Y. Sheng, N. Vo, and S. Tata, "Freedom: A transferable neural architecture for structured information extraction on web documents," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1092–1102.

[165] Y. Zhang, M. Zhong, S. Ouyang, *et al.*, "Automated mining of structured knowledge from text in the era of large language models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6644–6654.

[166] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, The Association for Computer Linguistics, 2015, pp. 344–354. DOI: 10.3115/V1/P15-1034. [Online]. Available: https://doi.org/10.3115/v1/p15-1034.

[167] S. Zhou, B. Yu, A. Sun, C. Long, J. Li, and J. Sun, "A survey on neural open information extraction: Current status and future directions," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L. D. Raedt, Ed., ijcai.org, 2022, pp. 5694–5701. DOI: 10.24963/IJCAI.2022/793. [Online]. Available: https://doi.org/10.24963/ijcai.2022/793.

[168] K. Gashteovski, M. Yu, B. Kotnis, C. Lawrence, M. Niepert, and G. Glavas, "Benchie: A framework for multi-faceted fact-based open information extraction evaluation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, 2022, pp. 4472–4490. DOI: 10.18653/V1/2022.ACL-LONG.307. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.307.

[169] A. Alamri and M. Stevenson, "A corpus of potentially contradictory research claims from cardiovascular research abstracts," *Journal of biomedical semantics*, vol. 7, pp. 1–9, 2016.

[170] A. Parikh, X. Wang, S. Gehrmann, *et al.*, "ToTTo: A controlled table-to-text generation dataset," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 1173–1186. DOI: 10.18653/v1/2020.emnlp-main.89. [Online]. Available: https://aclanthology.org/2020.emnlp-main.89/.

[171] Z. Wang, J. Araki, Z. Jiang, M. R. Parvez, and G. Neubig, "Learning to filter context for retrieval-augmented generation," *arXiv preprint arXiv:2311.08377*, 2023.

[172] L. Gao, Z. Dai, P. Pasupat, *et al.*, "RARR: Researching and revising what language models say, using language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 16 477–16 508. DOI: 10.18653/v1/2023.acl-long.910. [Online]. Available: https://aclanthology.org/2023.acl-long.910/.

[173] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.

[174] S. Zhou, B. Yu, A. Sun, *et al.*, "A survey on neural open information extraction: Current status and future directions," *arXiv preprint arXiv:2205.11725*, 2022.

[175] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni, "Open language learning for information extraction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, J. Tsujii, J. Henderson, and M. Pasca, Eds., ACL, 2012, pp. 523–534. [Online]. Available: https://aclanthology.org/D12-1048/.

[176] J. Christensen, Mausam, S. Soderland, and O. Etzioni, "Semantic role labeling for open information extraction," in *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, R. Mulkar-Mehta, J. Allen, J. Hobbs, E. Hovy, B. Magnini, and C. Manning, Eds., Los Angeles, California: Association for Computational Linguistics, Jun. 2010, pp. 52–60. [Online]. Available: https://aclanthology.org/W10-0907/.

[177] F. de Sá Mesquita, J. Schmidek, and D. Barbosa, "Effectiveness and efficiency of open relation extraction," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, ACL, 2013, pp. 447–457. [Online]. Available: https://aclanthology.org/D13-1043/.

[178] L. D. Corro and R. Gemulla, "Clausie: Clause-based open information extraction," in *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, D. Schwabe, V. A. F. Almeida, H. Glaser, R. Baeza-Yates, and S. B. Moon, Eds., International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 355–366. DOI: 10.1145/2488388.2488420. [Online]. Available: https://doi.org/10.1145/2488388.2488420.

[179] K. Gashteovski, R. Gemulla, and L. D. Corro, "Minie: Minimizing facts in open information extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, M. Palmer, R. Hwa, and S. Riedel, Eds., Association for Computational Linguistics, 2017, pp. 2630–2640. DOI: 10.18653/V1/D17-1278. [Online]. Available: https://doi.org/10.18653/v1/d17-1278.

[180] X. Wang, Y. Zhang, Q. Li, Y. Chen, and J. Han, "Open information extraction with meta-pattern discovery in biomedical literature," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2018, Washington, DC, USA, August 29 - September 01, 2018*, A. Shehu, C. H. Wu, C. Boucher, J. Li, H. Liu, and M. Pop, Eds., ACM, 2018, pp. 291–300. DOI: 10.1145/3233547.3233594. [Online]. Available: https://doi.org/10.1145/3233547.3233594.

[181] G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan, "Supervised open information extraction," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 885–895.

[182] J. Tang, Y. Lu, H. Lin, *et al.*, "Syntactic and semantic-driven learning for open information extraction," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 782–792. DOI: 10.18653/v1/2020.findings-emnlp.69. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.69/.

[183] K. Kolluru, S. Aggarwal, V. Rathore, Mausam, and S. Chakrabarti, "IMoJIE: Iterative memory-based joint open information extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 5871–5886. DOI: 10.18653/v1/2020.acl-main.521. [Online]. Available: https://aclanthology.org/2020.acl-main.521/.

[184] Y. Ro, Y. Lee, and P. Kang, "Multi^2OIE: Multilingual open information extraction based on multi-head attention with BERT," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 1107–1117. DOI: 10.18653/v1/2020.findings-emnlp.99. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.99/.

[185] K. Dong, Z. Yilin, A. Sun, J. J. Kim, and X. Li, "Docoie: A document-level context-aware dataset for openie," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2377–2389.

[186] C. Ling, X. Zhao, X. Zhang, *et al.*, "Improving open information extraction with large language models: A study on demonstration uncertainty," *CoRR*, vol. abs/2309.03433, 2023. DOI: 10.48550/ARXIV.2309.03433. arXiv: 2309.03433. [Online]. Available: https://doi.org/10.48550/arXiv.2309.03433.

[187] L. Wang, N. Yang, X. Huang, *et al.*, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533*, 2022.

[188] X. Wang, Z. Wang, X. Gao, *et al.*, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17 716–17 736. DOI: 10.18653/v1/2024.emnlp-main.981. [Online]. Available: https://aclanthology.org/2024.emnlp-main.981/.

[189] G. Rosemblat, M. Fiszman, D. Shin, and H. Kilicoglu, "Towards a characterization of apparent contradictions in the biomedical literature using context analysis," *Journal of biomedical informatics*, vol. 98, p. 103 275, 2019.

[190] D. N. Sosa and R. B. Altman, "Contexts and contradictions: A roadmap for computational drug repurposing with knowledge inference," *Briefings in bioinformatics*, vol. 23, no. 4, bbac268, 2022.

[191] J. Li, V. Raheja, and D. Kumar, "ContraDoc: Understanding self-contradictions in documents with large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 6509–6523. DOI: 10.18653/v1/2024.naacl-long.362. [Online]. Available: https://aclanthology.org/2024.naacl-long.362/.

[192] F. S. Yazi, W.-T. Vong, V. Raman, P. H. H. Then, and M. J. Lunia, "Towards automated detection of contradictory research claims in medical literature using deep learning approach," in *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, IEEE, 2021, pp. 116–121.

[193] D. Makhervaks, P. Gillis, and K. Radinsky, "Clinical contradiction detection," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 1248–1263.

[194] J. Devlin, M.-W. Chang, K. Lee, and 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)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423/.

[195] G. Bekoulis, C. Papagiannopoulou, and N. Deligiannis, "A review on fact extraction and verification," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–35, 2021.

[196] I. Augenstein, T. Baldwin, M. Cha, *et al.*, "Factuality challenges in the era of large language models and opportunities for fact-checking," *Nature Machine Intelligence*, vol. 6, no. 8, pp. 852–863, 2024.

[197] X. Li, G. A. Burns, and N. Peng, "A paragraph-level multi-task learning model for scientific fact-verification," in *Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021*, A. P. B. Veyseh, F. Dernoncourt, T. H. Nguyen, W. Chang, and L. A. Celi, Eds., ser. CEUR Workshop Proceedings, vol. 2831, CEUR-WS.org, 2021. [Online]. Available: https://ceur-ws.org/Vol-2831/paper8.pdf.

[198] S. Bhardwaj, S. Aggarwal, and Mausam, "CaRB: A crowdsourced benchmark for open IE," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6262–6267. DOI: 10.18653/v1/D19-1651. [Online]. Available: https://aclanthology.org/D19-1651/.

[199] G. Stanovsky and I. Dagan, "Creating a large benchmark for open information extraction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2300–2305.

[200] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu, "Biored: A rich biomedical relation extraction dataset," *Briefings in Bioinformatics*, vol. 23, no. 5, bbac282, 2022.

[201] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[202] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[203] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1586–1596. DOI: 10.18653/v1/D18-1187. [Online]. Available: https://aclanthology.org/D18-1187/.