Multi-label Sequential Sentence Classification via Large Language Model

Mengfei Lan, Lecheng Zheng, Shufan Ming, Halil Kilicoglu*
School of Information Sciences
University of Illinois Urbana-Champaign
{mlan3, lecheng4, shufanm2, halil}@illinois.edu

Abstract

Sequential sentence classification (SSC) in scientific publications is crucial for supporting downstream tasks such as fine-grained information retrieval and extractive summarization. However, current SSC methods are constrained by model size, sequence length, and single-label setting. To address these limitations, this paper proposes LLM-SSC, a large language model (LLM)-based framework for both single- and multi-label SSC tasks. Unlike previous approaches that employ small- or medium-sized language models, the proposed framework utilizes LLMs to generate SSC labels through designed prompts, which enhance task understanding by incorporating demonstrations and a query to describe the prediction target. We also present a multi-label contrastive learning loss with auto-weighting scheme, enabling the multi-label classification task. To support our multi-label SSC analysis, we introduce and release a new dataset, BIORC800, which mainly contains unstructured abstracts in the biomedical domain with manual annotations. Experiments demonstrate LLM-SSC's strong performance in SSC under both in-context learning and task-specific tuning settings. We release BIORC800 and our code at: https://github.com/ScienceNLP-Lab/LLM-SSC.

1 Introduction

With the increasing number of published scientific papers today, researchers face significant challenges in quickly pinpointing needed information. To address this problem, organizing complex paper content according to the rhetorical roles of each sentence in a structured format has garnered interest (Teufel and Moens, 1998; Pradhan et al., 2003; Ruch et al., 2007). Since the rhetorical roles of each sentence are often informed by the context from neighboring sentences (e.g. appendix D), this task is referred to as sequential sentence classification (SSC) (Cohan et al., 2019). SSC can enable the fine-grained information retrieval (Jimeno Yepes et al., 2013), enhance extractive summarization (Agarwal and Yu, 2009), and improve other downstream tasks. For example, labeling objective sentences in scientific 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 (Agibetov et al., 2018; Jin and Szolovits, 2018; Jiang et al., 2019; Gonçalves et al., 2020; Yamada et al., 2020; Shang et al., 2021; Li et al., 2021; Brack et al., 2022). Other research utilizes the masked token objective or transformers, introducing special tokens to encode contextual information and using these tokens to predict sequential labels (Cohan et al., 2019).

Despite promising progress in the SSC task, 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 (Jin and Szolovits, 2018), SciBERT (Cohan et al., 2019; Li et al., 2021)), 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 (Cohan et al., 2019). 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 (Mollá, 2022). Moreover, the widely used

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16086–16104
November 12-16, 2024 ©2024 Association for Computational Linguistics

SSC dataset in the biomedical field, PubMed 200K RCT, is automatically generated from structured abstracts in PubMed (Dernoncourt and Lee, 2017). However, this dataset does not include unstructured abstracts with free-form writing styles, which may deviate from the common patterns found in structured abstracts (Cohan et al., 2019; Gonçalves et al., 2020).

To bridge these gaps, this paper explores the application of LLMs in multi-label SSC using manually created datasets. We propose LLM-SSC, a novel unified framework for in-context learning and parameter-efficient finetuning (PEFT) using Gemma-2b (Team et al., 2024) 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 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 1.

2.1 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 ($C_i = concate(S_{i1}, S_{i2}, ...S_{in})$), and Y_{ij} as the SSC label of S_{ij}. Our goal is to model the probability of generating the SSC label Y_{ij}.

2.2 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 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 (Gao et al., 2021)*. 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:

<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}:

*https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

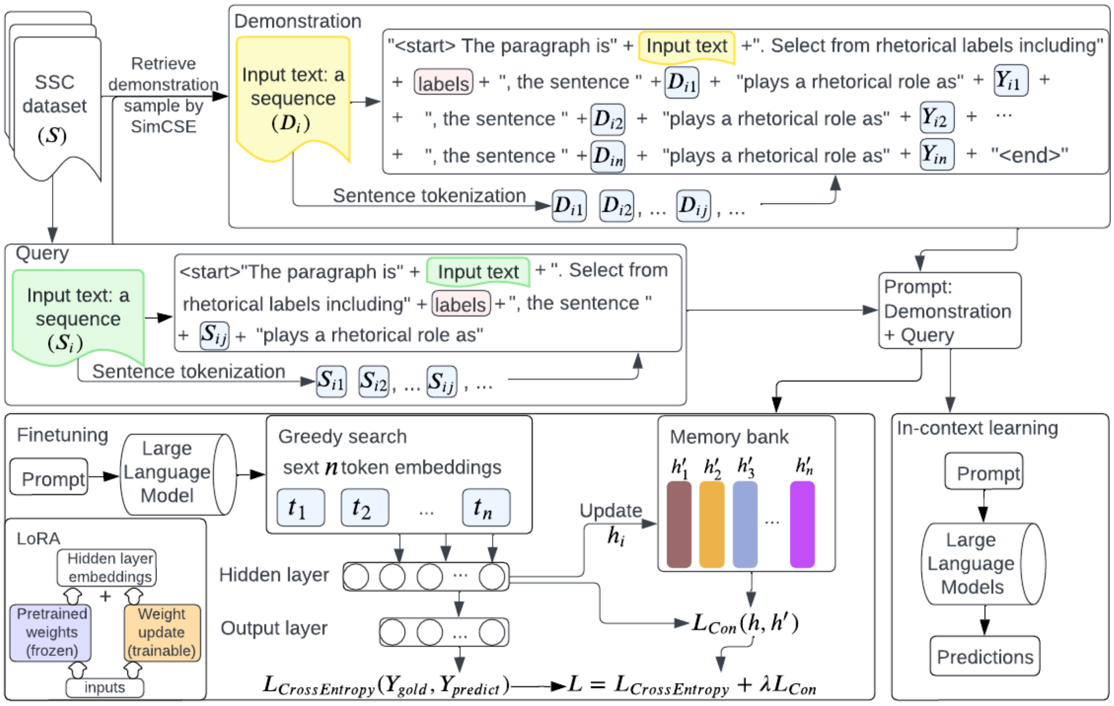

Figure 1: Structure of our LLM-based in-context learning and finetuning for SSC.

$$X_{ICL} = D_{prompt} || Q_{prompt} \quad (1)$$

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

$$Y_{predict} = \arg\max_{Y} P(Y|X_{ICL}) \quad (2)$$

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}. Some actual prompts are provided in Appendix A.

2.3 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 (Dong et al., 2022; Sun et al., 2023; Wadhwa et al., 2023). To further explore the LLM application in SSC, we design a parameter-efficient fine-tuning framework of LLM. Figure 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 2.2). The tuning process modifies how the model adjusts the given demonstration and query within the prompt to generate appropriate token sequence $\hat{t}$:

$$\hat{t} = \arg\max_{t} P(t|X_{finetune}) \quad (3)$$

$$\hat{t} = \{t_1, t_2, ..., t_i, ...\} \quad (4)$$

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 (Goyal et al., 2024). In our preliminary analysis, we observe a similar phenomenon. When employing the ICL approach outlined in Subsection 2.2, 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 (Goyal et al., 2024) 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 3 and concatenated:

$$e_i = concat(t_1, t_2, ..., t_n) \tag{5}$$

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

$$h_i = ReLU(w_1 e_i + b_1) \tag{6}$$

$$p_{i,predict} = \sigma(w_2 h_i + b_2) \tag{7}$$

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:

$$L_{CrossEntropy} = -\sum_{i=1}^{N} y_{i,gold} \log(p_{i,predict}) \tag{8}$$

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 (Hu et al., 2022). 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.

2.4 Auto-weighting Multi-label Contrastive Learning

Supervised Contrastive Learning (Khosla et al., 2020) has been widely employed in fine-tuning language models (Gunel et al., 2021; Chen et al., 2022b; Xie et al., 2022). These methods typically construct positive and negative pairs based on the equivalence of label vectors in multi-class classification problems (Zheng et al., 2023). 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 (Zheng et al., 2021, 2022, 2024b), 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:

$$L_{con} = -\sum_{c=1}^{m} \mathbb{E}_i \, \mathbb{E}_{j \in P_i(c)} \frac{\alpha_{ij} \text{sim}(h_i, h_j)}{\sum_k (1 - \alpha_{ik}) \text{sim}(h_i, h_k)} \tag{9}$$

$$\alpha_{ij} = \sigma(MLP(y_i, y_j))$$

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. 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-scale parameters of the LLM, we incorporate supervised contrastive learning supported by

BIORC800 Statistics		
Overall	7911 (Sentences)	800 (Docs)
Multi-label Sentences	452 (Sentences)	Percentage: 5.7%
Doc Length (sentence)	Avg: 9.89	Std: 2.68
Sentence Length (word)	Avg: 23.04	Std: 11.10
Label Distribution	BACKGROUND	1252, 15.8%
	OBJECTIVE	827, 10.5%
	METHODS	2319, 29.3%
	RESULTS	2757, 34.9%
	CONCLUSIONS	1114, 14.1%
	OTHER	112, 1.4%

Table 1: BIORC Statistics

a memory bank (He et al., 2020) 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):

$$L = L_{CrossEntropy} + \lambda L_{Con} \tag{10}$$

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

3.1 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 (Cohan et al., 2019; Gonçalves et al., 2020) (e.g. appendix E). Therefore, we re-annotated those sentences from 100 structured abstracts to more accurately reflect their category. The collected RCT abstracts were sampled from PubMed Central Open Access subset[†] using a modified version of Cochrane's sensitivity and precision-maximizing query for RCTs. 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 Cohen's κ 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 (guideline final version: Appendix B), 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 splited 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 1. Additional dataset information is provided in Appendix C.

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

CS-ABSTRACT (Gonçalves et al., 2020) 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[‡].

PUBMED 20K RCT (Dernoncourt and Lee, 2017) 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[§].

ART-CORESC (Liakata et al., 2010) is a multi-domain dataset, containing sentence-level scientific

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

[‡]https://github.com/sergiog95/csabstracts

[§]https://github.com/Franck-Dernoncourt/pubmed-rct/tree/master/PubMed_20k_RCT discourse annotation for 265 full papers selected from physics, chemistry, and biochemistry fields. In this SSC task, only the abstracts of these papers are used. The sentences in abstracts are annotated as background, hypothesis, motivation, objective, goal, methods, results, observation, experiment, or conclusion[¶].

3.2 Baselines

We selected SSC methods that achieved state-of-the-art performance on SSC datasets and had publicly accessible code as baselines for testing on our BIORC800 dataset. To adapt these methods, which were originally designed for single-label settings, to multi-label prediction, we modified the code provided by the authors 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) (Jin and Szolovits, 2018) 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 [‖].

Sequential Sentence Classification (SSC) (Cohan et al., 2019) employs BERT model (Devlin et al., 2019) 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[**].

Scientific Discourse Tagging (SDT) (Li et al., 2021) uses token embeddings from SciBERT (Beltagy et al., 2019), 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[††].

SciBERT-HSLN (Brack et al., 2022) is built upon the HSLN model with SciBERT (Beltagy et al., 2019) as word embeddings[‡‡].

We also include a contrastive learning baseline:

HeroCon (Zheng et al., 2022) 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[§§].

3.3 Experimental Setup

Implementation and Evaluation We select Gemma-2b (Team et al., 2024) 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 we adopt 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, we remove the demonstration part of the input and only input the query part. To evaluate the proposed in-context learning method, we use the training set samples 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. We use PEFT [¶¶] package to tune model using LoRA. Our 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.

3.4 Results and Discussion

3.4.1 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 presents the performance of our in-context learning approach across all datasets. Specifically, we have the following observations:

[¶] https://live.european-language-grid.eu/catalogue/corpus/972/download/

[‖] https://github.com/jind11/HSLN-Joint-Sentence-Classification

[**] https://github.com/allenai/sequential_sentence_classification

[††] https://github.com/jacklxc/ScientificDiscourseTagging

[‡‡] https://github.com/arthurbra/sequential-sentence-classification

[§§] https://github.com/Leo02016/HeroCon

[¶¶] https://huggingface.co/docs/peft/en/index

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

(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 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.

3.4.2 Task-specific Model Tuning

Table 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 understand 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.

3.4.3 Ablation Studies

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 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

	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 (Jin and Szolovits, 2018)	0.849	0.826	0.723	0.652	0.919	0.870	0.400	0.163
SSC (Cohan et al., 2019)	0.905	0.892	**0.780**	0.714	0.924	0.859	0.470	0.253
SDT (Li et al., 2021)	-	-	0.767	0.653	**0.940**	**0.903**	**0.534**	0.270
SciBERT-HSLN (Brack et al., 2022)	0.902	0.897	0.765	0.712	0.931	0.882	0.467	0.246
LLM-SSC with HeroCon (Zheng et al., 2022)	0.902	0.906	0.767	0.714	0.921	0.871	0.503	0.260
LLM-SSC with WeighCon (Ours)	**0.908**	**0.912**	0.768	**0.716**	0.925	0.879	0.524	**0.282**

Table 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.

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 4: Ablation study.

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

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.

3.4.4 Think before Speak Analysis

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

4 Related Works

SSC datasets SSC datasets are from various domains. PUBMED 20K RCT (Dernoncourt et al., 2017) and NICTA-PIBOSO (Kim et al., 2011) are two datasets generally used in biomedical domain. CSABSTRUCT (Cohan et al., 2019) and CS-ABSTRACTS (Gonçalves et al., 2020) are datasets utilizing abstracts from computer science papers. EMERALD 100K (Stead et al., 2019) and MAZEA (Dayrell et al., 2012) contains samples from multi-domains. In addition to these abstract-based datasets, some others use the full paper, such as DR. INVENTOR (Fisas et al., 2015) collecting samples from the computer graphics domain and ART-CORESC (Liakata et al., 2010) from physics, chemistry, and biochemistry domains.

SSC methods Before the deep learning paradigm, machine learning algorithms are applied to SSC (Ruch et al., 2007; McKnight and Srinivasan, 2003; Lin et al., 2006). These methods rely on hand-selected features and the classification performance is limited to the annotation amount and quality (Brack et al., 2022). Inspired by transfer learning and deep learning bringing pre-learned knowledge from external large datasets and simulates human-like thinking, recent SSC works leverage neural networks have been reported (Agibetov et al., 2018; Jin and Szolovits, 2018; Li et al., 2021;

Shang et al., 2021; Brack et al., 2022, 2024). The current SoTA methods commonly follow a hierarchical framework (Brack et al., 2022), including an encoding layer to represent word tokens (e.g. Word2Vec (Mikolov et al., 2013)) or embed sentences (e.g., CNN (Albawi et al., 2017)), followed by a context interaction layer to enrich the embedding using the surrounding context (e.g. Bi-LSTM (Shang et al., 2021)), and a labeling optimization layer to output the optimized sequential labels (e.g. CRF (Yamada et al., 2020)). 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 (Cohan et al., 2019).

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 (Chen et al., 2022a; Pan et al., 2022; Wang et al., 2022; Zhang et al., 2022; Liao et al., 2024; Zhang et al., 2024; Zheng et al., 2024a). For named entity recognition, leveraging labeled entity types to create positive and negative pairs helps the model distinguish between different entities more effectively (Das et al., 2022; Huang et al., 2022; Zhang et al., 2023b,a; Mo et al., 2024). In semantic similarity evaluation, supervised contrastive learning improves the model's ability to recognize subtle semantic differences, thereby boosting task performance (Gao et al., 2021; Liang et al., 2024).

5 Conclusion

In this paper, 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.

6 Limitations

First, the LLM-SSC requires a vast amounts of computational power and time to train due to the LLM size and complexity. For example, it takes over three hours to train a LLM-SSC model using BIORC800 on NVIDIA A100 GPU (5 epochs). Furthermore, the WeighCon mechanism introduced for multi-label SSC in this study could potentially be applied to other multi-label classification tasks; however, due to space limitations in this paper, a comprehensive exploration of the generalizability of this multi-label contrastive objective was not feasible.

7 Acknowledgement

This work was partially supported by the National Library of Medicine of the National Institutes of Health under the award number R01LM014079. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funder had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Asan Agibetov, Kathrin Blagec, Hong Xu, and Matthias Samwald. 2018. Fast and scalable neural embedding models for biomedical sentence classification. *BMC bioinformatics*, 19(1):1–9.

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Arthur Brack, Elias Entrup, Markos Stamatakis, Pascal Buschermöhle, Anett Hoppe, and Ralph Ewerth. 2024. Sequential sentence classification in research papers using cross-domain multi-task learning. *International Journal on Digital Libraries (2024), online first*.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–13.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022a. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10492–10500.

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. 2022b. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv preprint arXiv:2201.08702*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. 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)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. 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)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Carmen Dayrell, Arnaldo Candido Jr, Gabriel Lima, Danilo Machado Jr, Ann A Copestake, Valéria Delisandra Feltrim, Stella EO Tagnin, and Sandra M Aluísio. 2012. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. In *LREC*, pages 1604–1609.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 308–313. Asian Federation of Natural Language Processing.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. 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*, pages 694–700, Valencia, Spain. Association for Computational Linguistics.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th linguistic annotation workshop*, pages 42–51.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*, 32:6793–6807.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. 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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International conference on computational linguistics*, pages 2515–2527.

Xinyu Jiang, Bowen Zhang, Yunming Ye, and Zhenhua Liu. 2019. 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*, pages 78–89. Springer.

Antonio Jimeno Yepes, James Mork, and Alan Aronson. 2013. Using the argumentative structure of scientific literature to improve information access. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 102–110, Sofia, Bulgaria. Association for Computational Linguistics.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. 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*.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. 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).

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2024. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Yuzhong Chen, Xi Jiang, David Liu, Dajiang Zhu, et al. 2024. Mask-guided bert for few-shot text classification. *Neurocomputing*, page 128576.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology*, pages 65–72.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *AMIA annual symposium proceedings*, volume 2003, page 440. American Medical Informatics Association.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Ying Mo, Jian Yang, Jiahao Liu, Qifan Wang, Ruoyu Chen, Jingang Wang, and Zhoujun Li. 2024. mcl-ner: Cross-lingual named entity recognition via multi-view contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18789–18797.

Diego Mollá. 2022. 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*, pages 178–182.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H Martin, and Daniel Jurafsky. 2003. Semantic role parsing: Adding semantic structure to unstructured text. In *Third IEEE international conference on data mining*, pages 629–632. IEEE.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2-3):195–200.

Xichen Shang, Qianli Ma, Zhenxi Lin, Jiangyue Yan, and Zipeng Chen. 2021. 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)*, pages 198–203.

Connor Stead, Stephen Smith, Peter Busch, and Savanid Vatanasakdakul. 2019. Emerald 110k: a multidisciplinary dataset for abstract sentence classification. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 120–125.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Simone Teufel and Marc Moens. 1998. Sentence extraction and rhetorical classification for flexible abstracts. In *AAAI Spring Symposium on Intelligent Text summarization*, pages 89–97.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Ran Wang, Xinyu Dai, et al. 2022. 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)*, pages 672–679.

Shaorong Xie, Chunning Hou, Hang Yu, Zhenyu Zhang, Xiangfeng Luo, and Nengjun Zhu. 2022. Multi-label disaster text classification via supervised contrastive learning for social media data. *Computers and Electrical Engineering*, 104:108401.

Kosuke Yamada, Tsutomu Hirao, Ryohei Sasano, Koichi Takeda, and Masaaki Nagata. 2020. Sequential span classification with neural semi-markov crfs for biomedical abstracts. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 871–877.

Ruohong Zhang, Yau-Shian Wang, and Yiming Yang. 2024. Generation-driven contrastive self-training for zero-shot text classification with instruction-following LLM. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 659–673, St. Julian's, Malta. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023a. Optimizing bi-encoder for named entity recognition via contrastive learning. *ICLR 2023*.

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023b. 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*, pages 958–966.

Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022. Metadata-induced contrastive learning for zero-shot multi-label text classification. In *Proceedings of the ACM Web Conference 2022*, pages 3162–3173.

Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. 2024a. 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*, pages 4107–4116. ACM.

Lecheng Zheng, Baoyu Jing, Zihao Li, Hanghang Tong, and Jingrui He. 2024b. 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*, pages 6666–6676. ACM.

Lecheng Zheng, Jinjun Xiong, Yada Zhu, and Jingrui He. 2022. Contrastive learning with complex heterogeneity. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2594–2604.

Lecheng Zheng, Yada Zhu, and Jingrui He. 2023. 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*, pages 856–864. SIAM.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. 2021. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051.

A In-context Learning / Task-specific Tuning Prompts

The one-shot in-context learning prompts, which are also utilized as task-specific model tuning instructions, are created for each dataset as figures 2, 3, 4, and 5 show. In these figures, the demonstration portion of the prompt is highlighted in green, labels are marked in brown, and the query section is shown in blue.

B Annotation Guideline of BIORC800

Introduction We aim to annotate PubMed abstracts for sequential rhetorical categories: BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS (BOMRC schema). Existing research on recognizing rhetorical categories from Randomized Controlled Trail (RCT) abstracts often use normalized labels derived from NLM mappings as ground truth label (Jin and Szolovits, 2018; Cohan et al., 2019; Li et al., 2021; Brack et al., 2022). This allows generating large amounts of training data for free, but our analysis shows that these labels are somewhat noisy. Additionally, though existing research have explored annotating the sequential sentences in other domain (e.g. computer science), no previous work attempted to extend the

<Start> The paragraph is "Rifapentine is a highly active antituberculosis antibiotic with treatment-shortening potential; however, exposure-response relations and the dose needed for maximal bactericidal activity have not been established. We used pharmacokinetic/pharmacodynamic data from 657 adults with pulmonary tuberculosis participating in treatment trials to compare rifapentine (n = 405) with rifampin (n = 252) as part of intensive-phase therapy. Population pharmacokinetic/pharmacodynamic analyses were performed with nonlinear mixed-effects modeling. Time to stable culture conversion of sputum to negative was determined in cultures obtained over 4 months of therapy. Rifapentine exposures were lower in participants who were coinfected with human immunodeficiency virus, black, male, or fasting when taking drug. Rifapentine exposure, large lung cavity size, and geographic region were independently associated with time to culture conversion in liquid media. Maximal treatment efficacy is likely achieved with rifapentine at 1,200 mg daily. Patients with large lung cavities appear less responsive to treatment, even at high rifapentine doses.". Select from rhetorical labels including background, objective, method, result and conclusion, the sentence "Rifapentine is a highly active antituberculosis antibiotic with treatment-shortening potential; however, exposure-response relations and the dose needed for maximal bactericidal activity have not been established." plays rhetorical role in the paragraph as <background>, the sentence "We used pharmacokinetic/pharmacodynamic data from 657 adults with pulmonary tuberculosis participating in treatment trials to compare rifapentine (n = 405) with rifampin (n = 252) as part of intensive-phase therapy." plays rhetorical role in the paragraph as <objective, method>, the sentence "Population pharmacokinetic/pharmacodynamic analyses were performed with nonlinear mixed-effects modeling." plays rhetorical role in the paragraph as <method>, the sentence "Time to stable culture conversion of sputum to negative was determined in cultures obtained over 4 months of therapy." plays rhetorical role in the paragraph as <method>, the sentence "Rifapentine exposures were lower in participants who were coinfected with human immunodeficiency virus, black, male, or fasting when taking drug." plays rhetorical role in the paragraph as <result>, the sentence "Rifapentine exposure, large lung cavity size, and geographic region were independently associated with time to culture conversion in liquid media." plays rhetorical role in the paragraph as <result>, the sentence "Maximal treatment efficacy is likely achieved with rifapentine at 1,200 mg daily." plays rhetorical role in the paragraph as <result>, the sentence "Patients with large lung cavities appear less responsive to treatment, even at high rifapentine doses." plays rhetorical role in the paragraph as <result> <End>
<Start> The paragraph is "The purpose of this meta-analysis is to investigate whether statin is a key therapy for myocardial infarction (MI) by comparing all randomized controlled trials that appraised the effects of statin on risk of MI. Pubmed, Embase, and Medline databases (up to December 2016) were used to search all related articles. Using the data from 18 available publications, we examined the efficacy in treating or reducing the risk of MI by using random-effects models of odds ratio (OR) comparing the highest with the lowest category. Statins have demonstrated efficacy in treating or reducing the risk of MI (OR = 0.73, 95\% confidence interval = 0.58-0.93, P = .010). This meta-analysis suggests that statin have light efficacy in treating or reducing the risk of MI patients.". Select from rhetorical labels including background, objective, method, result and conclusion, the sentence "The purpose of this meta-analysis is to investigate whether statin is a key therapy for myocardial infarction (MI) by comparing all randomized controlled trials that appraised the effects of statin on risk of MI." plays rhetorical role in the paragraph as

Figure 2: BIORC800 1-shot Prompt

<Start>The paragraph is "Communities typically capture homophily as people of the same community share many common features. This paper is motivated by the problem of community detection in social networks, as it can help improve our understanding of the network topology and the spread of information. Given the selfish nature of humans to align with like-minded people, we employ game theoretic models and algorithms to detect communities in this paper. Specifically, we employ coordination games to represent interactions between individuals in a social network. We represent the problem of community detection as a graph coordination game. We provide a novel and scalable two phased approach to compute an accurate overlapping community structure in the given network. We evaluate our algorithm against the best existing methods for community detection and show that our algorithm improves significantly on benchmark networks (real and synthetic) with respect to standard Normalised Mutual Information measure.". Select from rhetorical labels including background, objective, method, result and conclusion, the sentence "Communities typically capture homophily as people of the same community share many common features." plays rhetorical role in the paragraph as <background>, the sentence "This paper is motivated by the problem of community detection in social networks, as it can help improve our understanding of the network topology and the spread of information." plays rhetorical role in the paragraph as <background>, the sentence "Given the selfish nature of humans to align with like-minded people, we employ game theoretic models and algorithms to detect communities in this paper." plays rhetorical role in the paragraph as <background>, the sentence "Specifically, we employ coordination games to represent interactions between individuals in a social network." plays rhetorical role in the paragraph as <method>, the sentence "We represent the problem of community detection as a graph coordination game." plays rhetorical role in the paragraph as <result>, the sentence "We provide a novel and scalable two phased approach to compute an accurate overlapping community structure in the given network." plays rhetorical role in the paragraph as <result>, the sentence "We evaluate our algorithm against the best existing methods for community detection and show that our algorithm improves significantly on benchmark networks (real and synthetic) with respect to standard Normalised Mutual Information measure." plays rhetorical role in the paragraph as <result> <End>
<Start> The paragraph is "On social networks, while nodes bear rich attributes, we often lack the `semantics' of why each link is formed-- and thus we are missing the `road signs' to navigate and organize the complex social universe. How to identify relationship semantics without labels? Founded on the prevalent homophily principle, we propose the novel problem of Attribute-based Relationship Profiling (ARP), to profile the closeness w.r.t. the underlying relationships (e.g., schoolmate) between users based on their similarity in the corresponding attributes (e.g., education) and, as output, learn a set of social affinity graphs, where each link is weighted by its probabilities of carrying the relationships. As requirements, ARP should be systematic and complete to profile every link for every relationship-- our challenges lie in effectively modeling homophily. We propose a novel reverse smoothness principle by observing that the similarity-closeness duality of homophily is consistent with the well-known smoothness assumption in graph-based semi-supervised learning-- only the direction of inference is reversed. To realize smoothness over noisy social graphs, we further propose a novel holistic closeness modeling approach to capture `high-order' smoothness by extending closeness from edges to paths. Extensive experiments on three real-world datasets demonstrate the efficacy of ARP.". Select from rhetorical labels including background, objective, method, result and conclusion, the sentence "Extensive experiments on three real-world datasets demonstrate the efficacy of ARP." plays rhetorical role in the paragraph as

Figure 3: CS-ABSTRACT 1-shot Prompt

<Start> The paragraph is "The aim of this study was to assess the efficacy of a @-month prevention program conducted in @ community pharmacies in reducing the risk for diabetes . In a cluster-randomized controlled trial in @,@ participants , mean change in the risk for diabetes (indicated by the Finnish Diabetes Risk Score -LSB- FINDRISC -RSB-) between intervention and control groups was calculated . In the intervention program GLICEMIA , three appointments with individual counseling and five educational group sessions were combined , whereas in the control group , only information about the participants ' health was obtained in three assessments . After adjusting for cluster structure and differences in baseline characteristics , improvement in FINDRISC in the intervention group was @ points (@ \% CI @-@ @) above the control group . The GLICEMIA program shows the feasibility of a pharmacy-based intervention and leads to a significant modest reduction in diabetes risk score but does not reduce the rate of diabetes progression over @ year .". Select from rhetorical labels including background, objective, method, result, and conclusion, the sentence "The aim of this study was to assess the efficacy of a @-month prevention program conducted in @ community pharmacies in reducing the risk for diabetes ." plays rhetorical role in the paragraph as <objective>, the sentence "In a cluster-randomized controlled trial in @,@ participants , mean change in the risk for diabetes (indicated by the Finnish Diabetes Risk Score -LSB- FINDRISC -RSB-) between intervention and control groups was calculated ." plays rhetorical role in the paragraph as <method>, the sentence "In the intervention program GLICEMIA , three appointments with individual counseling and five educational group sessions were combined , whereas in the control group , only information about the participants ' health was obtained in three assessments ." plays rhetorical role in the paragraph as <method>, the sentence "After adjusting for cluster structure and differences in baseline characteristics , improvement in FINDRISC in the intervention group was @ points (@ \% CI @-@ @) above the control group ." plays rhetorical role in the paragraph as <result>, the sentence "The GLICEMIA program shows the feasibility of a pharmacy-based intervention and leads to a significant modest reduction in diabetes risk score but does not reduce the rate of diabetes progression over @ year ." plays rhetorical role in the paragraph as <conclusion> <End>
<Start> The paragraph is "To assess the effects of a patient oriented decision aid for prioritising treatment goals in diabetes compared with usual care on patient empowerment and treatment decisions . Pragmatic randomised controlled trial . @ general practices in the north of the Netherlands . @ patients with type @ diabetes aged @ years at the time of diagnosis and managed in primary care between April @ and August @ : @ were allocated to the intervention group and @ to the usual care group . The intervention comprised a decision aid for people with diabetes , with individually tailored risk information and treatment options for multiple risk factors . The aid was intended to empower patients to prioritise between clinical domains and to support treatment decisions . It was offered to participants before a regular diabetes check-up and to their healthcare provider during the consultation . Four different formats of the decision aid were included for additional explorative analyses . The primary outcome was the effects on patient empowerment for setting and achieving goals . The secondary outcomes were changes in the prescribing of drugs to regulate glucose , blood pressure , lipids , and albuminuria. Data were collected through structured questionnaires and automated data extraction from electronic health records during six months before and after the intervention . Of all intervention participants , @ (@ \%) reported to have received the basic elements of the intervention . For the primary outcome analysis , @ intervention and @ control patients with sufficient baseline and follow-up data could be included . The mean empowerment score increased @ on a @ point scale in the overall intervention group , which was not significantly different from that of the control group (mean difference after adjusting for baseline @ , @ \% confidence interval -@ to @) . Lipid regulating drug treatment was intensified in @ \% of intervention and @ \% of control participants with increased cholesterol levels , which did not reach significance when the intervention was compared with the usual care group (odds ratio @ , @ \% confidence interval @ to @) . Prespecified explorative analyses showed that this effect was significant for the printed version of the decision aid in comparison to usual care (@ , @ to @) . No relevant or significant changes were seen for other treatments . We found no evidence that the patient oriented treatment decision aid improves patient empowerment by an important amount . The aid was not used to its full extent in a substantial number of participants .". Select from rhetorical labels including background, objective, method, result and conclusion, the sentence "The aid was not used to its full extent in a substantial number of participants ." plays rhetorical role in the paragraph as

Figure 4: PUBMED 20K RCT 1-shot Prompt

<Start>The paragraph is "We present calculations and simulations to investigate different theories describing phase transitions in thin films, with special emphasis on the growth of the new phase after nucleation. In particular, models with geometric and statistical growth rules are compared. It is demonstrated that the commonly employed geometrical approach, which assumes nucleation and subsequent radial growth of the newly formed phase, has distinct limitations for thin film systems. More realistic statistical Monte Carlo simulations that are governed by statistical growth rules, predict that a non-spherical (prolate) shape may develop after nucleation at or near a surface or interface. In addition, the predicted kinetics of the phase transformation is notably different for the geometric vs. the statistical model, for similar parameters. The simulation results are compared to recent experiments on the crystallization of thin amorphous solid water films.". Select from rhetorical labels including background, motivation, hypothesis, goal, objective, method, result, experiment, conclusion, the sentence "We present calculations and simulations to investigate different theories describing phase transitions in thin films, with special emphasis on the growth of the new phase after nucleation." plays rhetorical role in the paragraph as <objective>, the sentence "In particular, models with geometric and statistical growth rules are compared." plays rhetorical role in the paragraph as <objective>, the sentence "It is demonstrated that the commonly employed geometrical approach, which assumes nucleation and subsequent radial growth of the newly formed phase, has distinct limitations for thin film systems." plays rhetorical role in the paragraph as <observation>, the sentence "More realistic statistical Monte Carlo simulations that are governed by statistical growth rules, predict that a non-spherical (prolate) shape may develop after nucleation at or near a surface or interface." plays rhetorical role in the paragraph as <observation>, the sentence "In addition, the predicted kinetics of the phase transformation is notably different for the geometric vs. the statistical model, for similar parameters." plays rhetorical role in the paragraph as <observation>, the sentence "The simulation results are compared to recent experiments on the crystallization of thin amorphous solid water films." plays rhetorical role in the paragraph as <objective> <End>
<Start> The paragraph is "The formation of particles by the rapid expansion of supercritical solutions (RESS) of four organic substances (n-undecane, naphthalene, trans -stilbene and benzoic acid) dissolved in supercritical CO 2 has been investigated employing two complementary on-line in situ detection techniques. For the first time, a laser-based shadowgraphy (LABS) setup was applied, which allows the recording of the diameter, morphology, size distribution and concentration of particles down to sizes of roughly 8 μm. In addition, laser-based three-wavelength extinction measurements (3-WEM) have been employed to determine the average particle diameter, the width of the particle size distribution and the concentration. The average particle sizes determined from 3-WEM for n-undecane, naphthalene, trans -stilbene and benzoic acid are 2660, 1260, 550 and 430 nm, respectively, and the standard deviation of the (assumed) logarithmic normal distributions 0.37, 0.42, 0.78 and 0.60. The particle size distributions at large diameters from LABS compare favourably with the tails of the 3-WEM distributions. An analysis of the morphology of trans -stilbene particles and n-undecane droplets reveals an, on average, slightly elliptical shape.". Select from rhetorical labels including background, motivation, hypothesis, goal, objective, method, observation, result, experiment, conclusion, the sentence "An analysis of the morphology of trans -stilbene particles and n-undecane droplets reveals an, on average, slightly elliptical shape." plays rhetorical role in the paragraph as

Figure 5: ART-CORESC 1-shot Prompt

sequential sentence annotation from single label to multi-label. To address these issues, we will manually annotate 800 RCT abstracts using multi-label schema. The selected 800 RCT abstracts contain both structured (with NLM mapping; 100) and unstructured abstracts (without NLM mapping; 700). The structured abstracts with NLM mapping are involved to re-annotate since we hope to explore the noise introduced by auto-generation labeling.

Each sentence needs to be annotated with one or more of the labels from [BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS]. Insert all appropriate labels for a sentence in the "Annotated_multi_labels" column of the annotation spreadsheet and separate the labels by ",". If a sentence is assigned with more than one labels, you should split the sentence by the rhetorical label and add split result in the column "Sentence_split" following the format as:

<OBJECTIVE>: We tested three models of this regularity <METHODS>: originally formulated for primate cerebral cortex, using quantitative data on the relative supragranular layer origins (SGN%) of 151 projections from 19 areas (approximately 145,000 neurons) to four areas of cat extrastriate cortex

Note that the sentences in an abstract are usually in the order of BACKGROUND – OBJECTIVE-METHODS- RESULTS- CONCLUSIONS, though some categories may be missing. For each abstract, they are not necessarily standardized to the 5 categories above. The Excel spreadsheet has the sentence ID for each sentence. To see the full abstract in context, you can click on the link in the third column, which will show the abstract in PubMed. If the abstract is structured, the "stru_unstru" column value is "stru", otherwise, "unstru". Also note that some sentences do not fit into any of these categories. You can mark these as OTHER. For example, a sentence about study funding can be marked as OTHER.

Each annotator will be assigned 50 abstracts in the first round and another 50 abstracts in the second round. The abstracts assigned to each annotator in the first and second round will be the same, based on which to calculate the inter-annotator agreement. If the inter-annotator agreement (calculated by Cohen's Kappa statistic) is lower than expected (0.85), a meeting will be scheduled for all annotators to discuss the cases that cause disagreement. The annotation created in the first and second rounds will be reconciled by one author to determine the golden standard. In the third round, the rest 700 abstracts will be evenly assigned to every annotators. Finally, the annotation results from the first round (50 abstracts), the second round (50), and the third round (225 abstracts from each of the four annotators) will be aggregated to the BIORC-1000 dataset.

Points to Note There might be some situations that cause bias. Some points to follow to decrease the disagreements:

- With regards to the Objective/Background confusion, the presence of interrogative verbs, such as 'evaluate', 'investigate', 'assess', 'aim', or their derivational forms are strong clues for the Objective category.
- Background sentences are often general statements about a disease, intervention, etc. (e.g., Diabetes mellitus is the common chronic metabolic disease', 'Regularity of laminar origin and termination of projections appears to be a common feature of corticocortical connections.').
- With regards to the Results/Conclusion confusion, we need to distinguish results obtained from the study (which often contain numbers, confidence intervals, etc.) from the implication/conclusion that can be drawn based on that result. One example we saw was a sentence about future work based on the results ('The next step is external validation upon existence of independent trial data.'), which qualifies as Conclusion as it does not say anything about the results of this study.
- One outlier was an abstract about a disease guideline. I expect we may see some meta-analysis/review abstracts that may be similar. In this case, I annotated sentences summarizing the guidelines/recommendations as Conclusion sentences, and the sentence that describes what the guidelines are about as the Objective sentence.
- We try to annotate multi-label for each sentence. For example, if it is a sentence that sets the objective of the study but also mentions the method used, this should qualify as "OBJECTIVE, METHODS" sentence. (e.g., 'We tested three models of this regularity, originally formulated for primate cerebral cortex,

using quantitative data on the relative supragranular layer origins (SGN%) of 151 projections from 19 areas to four areas of cat extrastriate cortex.' is an "OBJECTIVE, METHODS" sentence).

- There may be some sentence splitting errors. Annotate such sentences, with the label of the actual sentence they would belong in.

- In reviews/surveys, the sentences that use the evidence from other articles should be annotated as RESULTS. A sample for review annotation (PMC26237111):

  Pressure ulcers (PUs) in individuals with spinal cord injury (SCI) present a persistent and costly problem. BACKGROUND

  Continuing effort in developing new technologies that support self-managed care is an important prevention strategy. BACKGROUND

  Specifically, the aims of this scoping review are to review the key concepts and factors related to self-managed prevention of PUs in individuals with SCI and appraise the technologies available to assist patients in self-management of PU prevention practices. OBJECTIVE

  There is broad consensus that sustaining long-term adherence to prevention regimens is a major concern. RESULTS

  Recent literature highlights the interactions between behavioral and physiological risk factors. RESULTS

  We identify four technology categories that support self-management: computer-based educational technologies demonstrated improved short-term gains in knowledge (2 studies), interface pressure mapping technologies demonstrated improved adherence to pressure-relief schedules up to 3 mo (5 studies), electrical stimulation confirmed improvements in tissue tolerance after 8 wk of training (3 studies), and telemedicine programs demonstrated improvements in independence and reduced hospital visits over 6 mo (2 studies). RESULTS

  ...

  However, if a sentence in the review abstract discusses a common finding shared by multiple previous papers or outlines the potential next steps for a topic, it should be annotated as "CONCLUSIONS"(PMC26237111):

  Overall, self-management technologies demonstrated low-to-moderate effectiveness in addressing a subset of risk factors. CONCLUSIONS

  ...

  However, the effectiveness of technologies in preventing PUs is limited due to a lack of incidence reporting. CONCLUSIONS

  In light of the key findings, we recommend developing integrated technologies that address multiple risk factors. CONCLUSIONS

- If the sentence starts with "we show that ..." followed by some discussion of the experiment results, apply "OBJECTIVE, RESULTS", or "RESULTS" to the sentence. Here is one example to illustrate the appropriate usage of "OBJECTIVE, RESULTS" (PMC23258531):

  ...

  Aberrant expression of miR-31 has been found in various cancers, including colorectal cancer. BACKGROUND

  Here, we show that miR-31 is upregulated in human colon cancer tissues and cell lines, and that repression of miR-31 inhibited colon cancer cell proliferation and colony formation in soft agarose. OBJECTIVE, RESULTS

  To further elucidate the mechanism underlying the role of miR-31 in promoting colon cancer, we used online miRNA target prediction databases and found that the tumor suppressor RhoTBT1 may be a target of miR-31. METHODS

  ...

Multilabel Samples Some of the multilabel annotations are presented as follows. Notice that these samples don't cover every type of multilabel case. Your annotations should adapt to the specific content of real cases, being more flexible as needed.

a. OBJECTIVE-METHODS:

We investigated this hypothesis using functional MRI and covariance analysis in 43 healthy skilled readers.

<OBJECTIVE>: We investigated this hypothesis

<METHODS>: using functional MRI and covariance analysis in 43 healthy skilled readers.

b. RESULTS-CONCLUSION

The similarity of developmental features among the direct-developers suggests a correlation with mode of life history.

<RESULTS>: The similarity of developmental features among the direct-developers

<CONCLUSION>: suggests a correlation with mode of life history.

c. BACKGROUND-OBJECTIVE

Since the mPFC also projects heavily to NAc, we examined whether NAc-projecting pyramidal neurons also express 5-HT2A-R.

<BACKGROUND>: Since the mPFC also projects heavily to NAc

<OBJECTIVE>: we examined whether NAc-projecting pyramidal neurons also express 5-HT2A-R

d. OBJECTIVE-METHODS

Here, we describe the features of MAP(2.0)3D server by analyzing, as an example, the cytochrome P450BM3 monooxygenase (CYP102A1).

<OBJECTIVE>: Here, we describe the features of MAP(2.0)3D server

<METHODS>: by analyzing, as an example, the cytochrome P450BM3 monooxygenase (CYP102A1)

e. Three-label

Although p38 MAPK activation usually promotes apoptosis, pharmacologic inhibition of p38 MAPK exacerbated OA-induced DNA fragmentation and loss of delta psi(m) in T leukemia cells, suggesting that, in this instance, the p38 MAPK signaling pathway promoted cell survival.

<BACKGROUND>: p38 MAPK activation usually promotes apoptosis

<RESULTS>: pharmacologic inhibition of p38 MAPK exacerbated OA-induced DNA fragmentation and loss of delta psi(m) in T leukemia cells

<CONCLUSIONS>: suggesting that, in this instance, the p38 MAPK signaling pathway promoted cell survival

C BIORC800 Detailed Statistics

The detailed information of BIORC800 is shown in table 6.

D Contextual Dependence Analysis

We use two samples from BIORC800 to illustrate the sequential sentence contextual dependencies.

- PMID: 17448455

  Based on this biological attribute we gain the possibility by means of using MSCs as the donors to develop a future cell therapy in clinical application. BACKGROUND

  But using MSCs as donor cells inevitably raises ***the question as to whether these donor cells would be immunogenic, and if so, would they be rejected after transplantation. OBJECTIVE***

  To investigate this*, human MSCs were cultured in vitro and induced to differentiate along neuronal lineage. OBJECTIVE, METHODS*

  The expression of human leukocyte antigen (HLA) class I and class II molecules and the co-stimulatory protein CD80 were increased on the surface of MSCs in the course of neuronal differentiation. RESULTS

  But neither of the co-stimulatory proteins, CD40 or CD86, was expressed. RESULTS

  ...

  From the sample above, the OBJECTIVE of the work is indicated by the phrase "To investigate this" in the third sentence, with the pronoun "this" referring to "the question as to ..." mentioned in the previous sentence. Consequently, considering the context, the second sentence should be classified as OBJECTIVE rather than BACKGROUND.

- PMID: 7231017

  ...

  Anserine supplementation increased superoxide dismutase (SOD) by 50% *(p < 0.001, effect size d = 0.8 for both ANS-LD and ANS-HD), and preserved catalase (CAT) activity suggesting* ***an improved antioxidant activity****. RESULTS, CONCLUSIONS*

  ...

  There were slight but significant elevations in glutamate pyruvate transaminase (GPT) and creatine kinase isoenzyme (CKMB), especially in ANS-HD (p < 0.05) compared with ANS-LD or PLA. RESULTS

  Haematological biomarkers were largely unaffected by anserine*, its dose, and without interaction with post exercise time-course. RESULTS*

	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 6: BIORC800 Detailed Statistics

However, compared with ANS-LD and PLA, ANS-HD increased the mean cell volume (MCV), and decreased the mean corpuscular haemoglobin concentration (MCHC) (p < 0.001). RESULTS

Anserine preserves cellular homoeostasis through enhanced antioxidant activity and protects cell integrity in healthy men, which is important for chronic disease prevention. CONCLUSIONS

However, ***anserine temporal elevated exercise-induced cell-damage, together with enhanced antioxidant activity and haematological responses*** *suggest an augmented exercise-induced adaptative response and recovery. RESULTS, CONCLUSIONS*

From the sample above, the sentence "However, anserine temporal elevated exercise..." should be labeled as "RESULTS, CONCLUSION" instead of "CONCLUSION" given the context of the whole paragraph. It is seen that "anserine temporal elevated exercise-induced cell-damage, together with enhanced antioxidant activity and haematological responses" is a summary of the previous RESULTS sentences and should be assigned with "RESULTS" label.

E Semantic Coherence in Structured Abstracts Auto-generated Labels

We start from two examples from PubMed 20K RCT to analyze the semantic coherence of the labels for structured abstracts.

- PMID: 25165090

  ...

  Although working smoke alarms halve deaths in residential fires , many households do not keep alarms operational. BACKGROUND

  We tested whether theory-based education increases alarm operability*. BACKGROUND*

  Randomised multiarm trial , with a single arm randomly selected for use each day , in low-income neighbourhoods in Maryland , USA . METHODS

  Intervention arms : (1) Full Education combining a health belief module with a social-cognitive theory module that provided hands-on practice installing alarm batteries and using the alarm 's hush button ; (2) Hands-on Practice social-cognitive module supplemented by typical fire department education ; (3) Current Norm receiving typical fire department education only. METHODS

  ...

From the sample above, the sentence "We tested whether theory-based education increases alarm operability" should be labeled as "OBJECTIVES" instead of "BACKGROUND" given the context of the whole paragraph.