Beyond Retrieval: Topic-based Alignment of Scientific Papers to Research Proposal

Rudra Nath Palit[†*] **Manasi Patwardhan**[†*] **Lovekesh Vig**[†] **Gautam Shroff**[†]

[†]TCS Research

{rudra.palit, manasi.patwardhan, lovekesh.vig, gautam.shroff}@tcs.com

Abstract

The inception of a research agenda typically commences with the creation of a comprehensive research proposal. The efficacy of the proposal often hinges on its ability to connect with the existing scientific literature that supports its ideas. To effectively assess the relevance of existing articles to a research proposal, it is imperative to categorize these articles into high-level thematic groups, referred to as topics, that align with the proposal. This paper introduces a novel task of aligning scientific articles, relevant to a proposal, with researcher-provided proposal topics. Additionally, we construct a dataset to serve as a benchmark for this task. We establish human and Large Language Model (LLM) baselines and propose a novel three-stage approach to address this challenge. We synthesize and use pseudo-labels that map proposal topics to text spans from cited articles to train Language Models (LMs) for two purposes: (i) as a retriever, to extract relevant text spans from cited articles for each topic, and (ii) as a classifier, to categorize the articles into the proposal topics. Our retriever-classifier pipeline, which employs very small open-source LMs fine-tuned with our constructed dataset, achieves results comparable to a vanilla paid LLM-based classifier, demonstrating its efficacy. However, a notable gap of 23.57 F1 score between our approach and the human baseline highlights the complexity of this task and emphasizes the need for further research.

1 Introduction

Researchers frequently draft research proposals to present new ideas, define research agendas and seek funding grants. An integral part of the proposal writing process is reviewing relevant literature and relating it to different aspects of the proposal. Several existing approaches designed for automatic retrieval of scientific articles can be applied to identify articles relevant to a proposal, with a detailed description (abstract) of the proposal serving as the query (Cohan et al., 2020b). However, there is often a further need to formulate high-level thematic categories (henceforth referred to as topics) relevant to the proposal and map retrieved relevant scientific articles to these categories for fine-grained contextualization. Such fine-grained mapping can further facilitate motivating the research problem, identifying its novelty, establishing baselines, synthesizing methods, and automatic literature review generation.

Given a set of reference article abstracts relevant to a proposal, Zhu et al. (2023); Martin-Boyle et al. (2024) auto-generates the thematic categories in a hierarchical form (termed as a catalogue) and organize references. However, the results demonstrate that the auto-generated catalogue does not match with the original-author-defined catalogue, leading to discrepancies in downstream literature review generation. This is due to inability of state-of-the-art LMs as well as subjectivity of the task. As opposed to this, we consider a more realistic setting, where we assume the availability of not only the reference papers retrieved to be relevant to a proposal; but also high-level topics provided by the researcher, for further literature categorization. With this assumption, we focus on the novel task of alignment of these reference papers, to one or more of these topics, offering a comprehensive understanding of its distinct contributions to the target proposal. This ensures a more personalized approach, aligning the cataloguing process closely with the researcher's unique perspective representing their understanding of the field.

Figure 1 illustrates an example, showcasing proposal topics, relevant reference papers and their mappings to the topics, with distinct text spans from the reference paper representing the relevant context (henceforth termed as the reference text

[*] Equal contribution

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 70–83

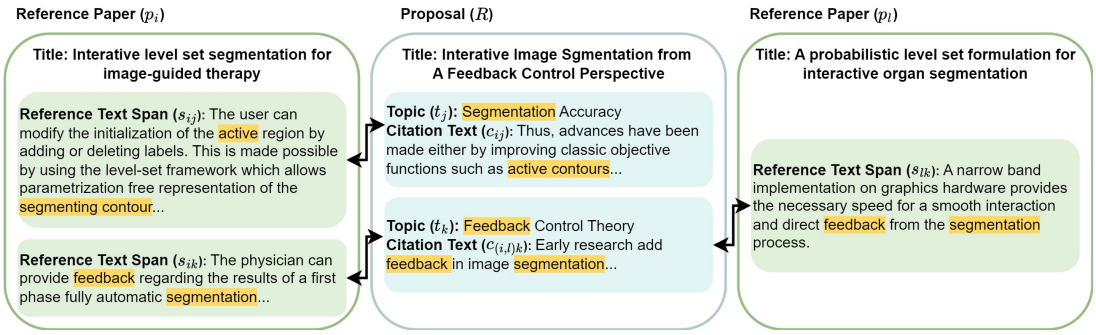

Figure 1: A sample from the dataset illustrating the many-to-many relationship between the proposal topics and reference papers. Reference Papers p_i and p_l are cited under Topic t_k, while Reference Paper p_i is cited by both Topics t_j and t_k with distinct reference text spans from the paper. The highlighted texts emphasize the similarities among the Topic, Citation Text, and Reference Text Spans.

span). Thus, a reference can be cited in more than one proposal topic with distinct context (reference text spans) (Li et al., 2023). Considering the token lengths of scientific articles, vanilla paid LLM-based approach for this task would be expensive. We consider the requirement of using smaller open-source LMs without compromising on the performance of LLMs. Hence, there is a need for a dataset to fine-tune domain-specific LMs for the task. As existing datasets for scientific document understanding (Clement et al., 2019; Lo et al., 2020; Saier et al., 2023; Kasanishi et al., 2023; Li et al., 2022) are not well-suited for this task, we extend the UnArXiv dataset (Saier et al., 2023), originally designed for literature review generation. We establish human and vanilla LLM baselines and define a novel three-stage approach for topic-based alignment of papers. Since ground truth labels for the reference text spans relevant to topics are not available, we augment the data by retrieving the reference text spans using the text around the article citation within the proposal topic (citation text), which is assumed to be available only for training. We use the synthesized pseudo pairs to train language models (LM) for retrieval of text spans from articles for a topic, and for classification of the retrieved articles to the topic. As opposed to all the prior approaches (Hoang and Kan, 2010; Hu and Wan, 2014; Chen and Zhuge, 2019; Wang et al., 2020; Pandey et al., 2022; Deng et al., 2021; Vajdecka et al., 2023), considering the realistic need of initial proposal writing stage, we neither assume the availability of citation text for the cited papers nor detailed description of the topics during inference. Our primary contributions are:

- We define the novel task of aligning reference articles relevant to a target proposal with user-defined proposal topics.
- We construct a dataset using a set of research papers as proposals, the subsection headings of the related work sections as the 'topics' and the papers cited under those sections as the reference papers relevant to those 'topics'.
- We establish human and LLM baselines and find that higher human F1 scores demonstrate task feasibility.
- We define a novel approach wherein we assume no citation text or detailed topic descriptions to retrieve reference text spans for a topic during inference. We devise a novel strategy using the citation text, available only for training, as a link between the topic and the reference text spans to create pseudo-labels for training a retriever and classifier pipeline.
- Our pipeline using much smaller LMs trained with pseudo-labels yields comparable performance to that of the LLM baseline, demonstrating the efficacy of the constructed dataset and the approach. In contrast, a significant gap (23.57 F1 score) exists between our approach and the human baseline indicating the need for more sophisticated solutions.

2 Related Work

2.1 Scientific Document Understanding

Literature Review Generation: Current approaches for automatic literature review generation either independently summarize articles (Hayashi et al., 2023; Urlana et al., 2022; Akkasi, 2022) or generate citation text (yan Wu et al., 2021; Jung

et al., 2022; Wang et al., 2022) for each article without considering their inter-relations (Li et al., 2022; Li and Ouyang, 2024). These approaches often produce monolithic extractive (Hu and Wan, 2014; Wang et al., 2018) or abstractive (Chen et al., 2021, 2022; Kasanishi et al., 2023; Chen et al., 2021; Liu et al., 2023) reviews lacking any structure. Kasanishi et al. (2023); Martin-Boyle et al. (2024) introduce a method for generating reviews with well-structured subsections, grouping relevant articles by specific topics. However, they either assume the availability of the mapping of research articles to proposal topics or take human inputs for the same. Our task serves as an upstream task for comprehensive literature review generation and focuses on the automatic alignment of relevant articles to a set of proposal topics.

Citation Text Generation: Citation text generation crafts sentences that cite reference articles based on their abstracts, aiming to integrate them into a literature review. However, existing methods (yan Wu et al., 2021; Jung et al., 2022; Wang et al., 2022) assume precise knowledge of intents (background, methodology, etc), behind citing a paper which is largely unavailable in the early stages of proposal writing. Moreover, these approaches rely solely on the reference paper's abstracts, potentially lacking adequate information for appropriate mapping to an intent or topic. In contrast, our approach of extracting reference text spans from the reference paper ensures the availability of comprehensive information for alignment. Prior approaches to text span extraction from research papers use citation texts (Pandey et al., 2022; Zerva et al., 2020; Vajdecka et al., 2023) or queries (Li et al., 2023). In contrast, our approach considers the more practical setting at the proposal writing stage and retrieves text spans without relying on the availability of citation text or detailed queries assuming only the availability of high-level topics.

Citation Intent Detection: Citation intent detection (Lahiri et al., 2023; Roman et al., 2021; Berrebbi et al., 2022) presumes the presence of citation text to classify papers into predefined categories such as motivation, background, etc. However, during the proposal writing stage, citation text is unavailable. Moreover, the classification categories for potential reference papers differ for each target proposal. In contrast, our work introduces a new task of mapping potentially relevant papers to proposal-specific topics defined by users without assuming the availability of the citation text.

Scientific Paper Retrieval: Several approaches for retrieving scientific papers from a corpus rely either on abstracts and titles of a target paper (Singh et al., 2023; Cohan et al., 2020a), detailed textual queries (Sesagiri Raamkumar et al., 2017; Anand et al., 2017; Parisot and Zavrel, 2022; Medic and Šnajder, 2023), or relevant aspects such as the problem or methodology (Mysore et al., 2022; Ostendorff et al., 2022; Singh et al., 2023). These approaches employ strategies to generate suitable embeddings for queries and research papers, often focusing on pre-defined 'aspects' common across the target papers. In contrast, our work assumes the availability of research articles relevant to a proposal and perfoms a more fine-granular mapping of these articles to a set of user-defined topics.

2.2 Existing Datasets

We evaluate existing datasets for our task. The SciReviewGen (Kasanishi et al., 2023) dataset is designed for literature review generation, is derived from S2ORC (Lo et al., 2020) and consists of only survey papers with their contents extracted from their PDFs, resulting in highly erroneous extractions. CORWA (Li et al., 2022), derived from ArXiv (Clement et al., 2019), extracts information from the LaTeX version of ACL conference papers, ensuring relatively error-free extractions. However, while it employs a tagger to classify transitions in text-forming paragraphs, it falls short of extracting the associated chapter heading (topic in our context) for each text span. (Medic and Šnajder, 2023) utilize the CORWA dataset for scientific paper retrieval, under the assumption that the first transition text spans of extracted paragraphs indirectly refers to the topic. Our analysis, however, reveals that this assumption holds only for a smaller subset of the samples. UnArXiv dataset (Saier et al., 2023), derived from ArXiv papers, struggles to correctly identify in-text citations for these papers required to derive the mapping of topics to the corresponding reference papers. We extend the UnArXiv dataset and overcome this limitation with our own parser (details in Section 4).

3 Task Definition

Consider a research proposal with its title and abstract R and with a set of topics forming the catalogue denoted as $T_R = t_1, ..., t_K$. We assume a corpus of reference research papers $P_R = p_1, ..., p_n$ retrieved using R via existing retrieval methods

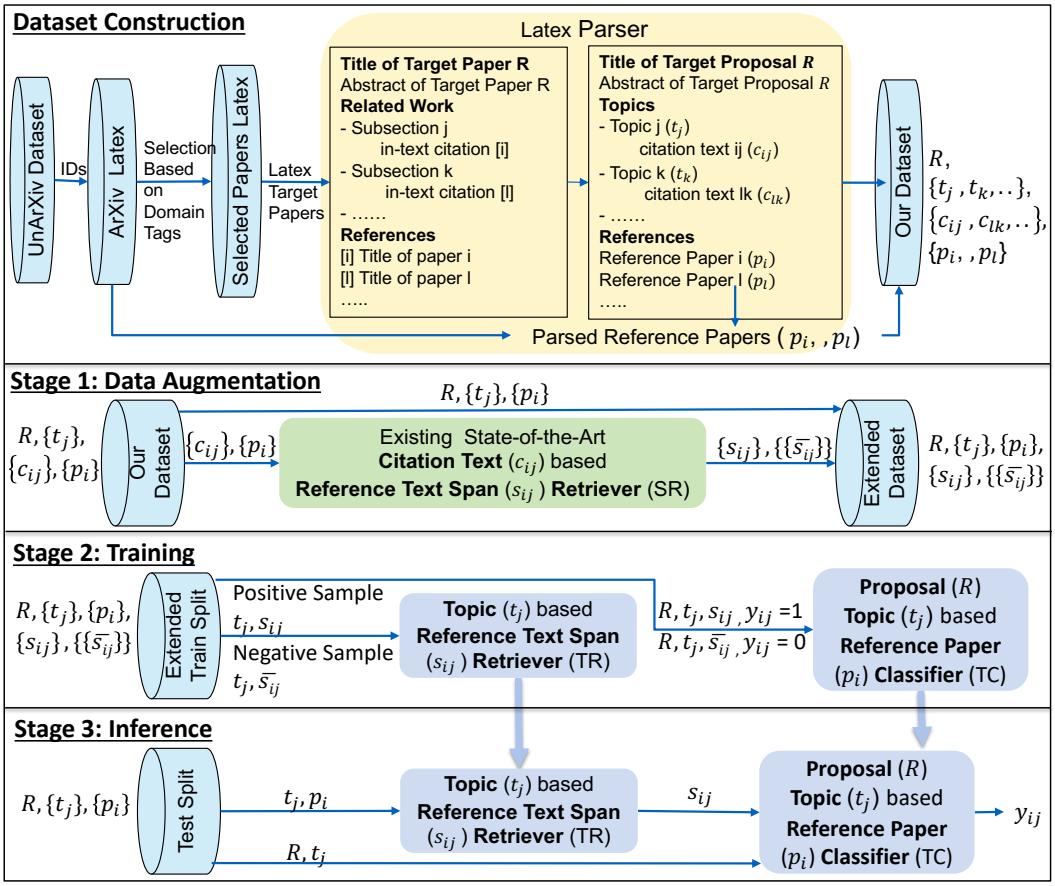

Figure 2: Stages of Proposal Topic based Reference Paper Alignment R: Proposal Title and Abstract, P_R: papers relevant to R, T_R: topics relevant to R, $p_i \in P_R$, $t_j, t_k \in T_R$, p_i, p_l are cited in t_j, t_k with citation text c_{ij}, c_{lk}, s_{ij}: reference text span from relevant paper p_i for topic t_j (Top-K chunks similar to query c_{ij}), $\{\overline{s_{ij}}\}$: text spans negatively sampled for topic t_j (Figure 3), y_{ij}, label set as 1 if paper p_i is aligned to topic t_j, otherwise 0.

such as Specter (Singh et al., 2023). The task aims to classify the mapping of each research paper $p_i \in P_R$ to each topic $t_j \in T_R$ using binary labels $y_{ij} \in 0, 1$. The dataset consists of positive tuples $< R, p_i, t_j, c_{ij} >$, where c_{ij} is the citation text citing the paper p_i under topic t_j and is only available during training. We define a process to augment the dataset with the reference text span s_{ij} by using the citation text c_{ij}. The reference text span refers to the contents of the referenced paper p_i, needed to determine whether p_i is relevant to the topic t_j. Again, we do not assume the availability of s_{ij} for training or inference in the dataset.

4 Dataset Construction

Figure 2 demonstrates the steps we have followed for the construction of our dataset. We utilize the papers in the UnArxiv dataset (Saier et al., 2023), with their titles and abstracts as target proposals R. We leverage the ArXiv tags of these papers to obtain their domains and focus on papers in AI and Computer Vision. This was to ensure high-quality annotations since the human annotator who facilitated the establishment of the baseline (Section 6.2) had expertise in these domains.

We use UnArxiv IDs to retrieve papers from ArXiv, obtain LaTeX sources, and employ a LaTeX parsing technique with Regex statements to extract in-text citations within our simulated proposal papers. Our parser identifies section headings from these target papers and extracts topics (t_j), from the subsection headings of the 'Related Work' or 'Literature Review' Sections. We selectively retain only those target proposals where multiple topics exist in the literature review. For each topic t_j thus identified, we further identify in-text citations for referenced papers $\{p_i\}$ and extract the corresponding citation texts $\{c_{ij}\}$. We then match the citation text with the corresponding entry in the references section of the paper to obtain the referenced paper title. We obtained the PDFs of the referenced papers from various sources like ArXiv, ACL, and

Semantic Scholar. The contents of the referenced papers are extracted using our PDF extractor.

Table 1: Dataset Statistics

Samples	Splits			Total
	Train	Validation	Test	
Target Proposals	1,934	241	242	2,417
Proposal Topics	5,680	720	723	7,123
Reference Papers	39,608	4,987	4,911	49,506
Avg. Topics / Proposal	2.94	2.99	2.99	2.95
Avg Papers / Topic	6.9	6.93	6.79	6.95
Topic-Paper Pairs	50,116	6,329	6,075	62,520

Thus, our final dataset consists of set-of research papers from selected domains from the UnArxiv dataset with each paper simulating a target proposal R. The dataset also includes the topics for each target proposal T_R and reference papers relevant to the target proposal P_R. Moreover, the ground truth labels of our task, in terms of alignments of reference papers to the topics are solicited from the in-text citations, explicitly specified by the authors of the target papers. For the training set, we assume the availability of the citation text c_{ij} for every pair t_j and p_i. Thus, a sample of the training data is depicted by the tuple $< R, p_i, t_j, c_{ij} >$. Whereas for the test set a sample is $< R, p_i, t_j >$. We split target proposals into train, validation, and test sets to prevent information leakage across splits. Dataset statistics are summarized in Table 1. The resulting corpus has target proposals pertaining to Artificial Intelligence (AI) (2.69 %), Machine Learning (ML) (15.56%), Computational Linguistics (CL) (7.28%), Computer Vision (CV) (73.23%) and a combination of CL and CV (1.24%). We make the dataset available at[1].

5 Approach

We break down our approach into three stages (Figure 2): (i) Augmenting the dataset with positive and negative reference text spans for a proposal topic, retrieved from papers relevant to the proposal, with the citation text as the query, using a state-of-the-art retriever model (SR) (ii) Training the Topic based reference text span Retriever (TR) and reference paper classifier (TC) using the augmented data (iii) Using TC to classify a reference paper for its relevance to a proposal topic in context of the proposal and the reference text span retrieved from the paper relevant to the topic using TR .

5.1 Stage 1: Data Augmentation

To train the retriever TR, we need positive and negative pairs of topics t_j and reference text spans s_{ij}. We do not have such pairs available in our training data. However, we do have the citation text c_{ij}, which we utilize as a 'link' to retrieve the reference text span s_{ij} from p_i, relevant to c_{ij} and consequently relevant to t_j. We assess the performance of existing retrieval models on equivalent tasks in the science domain, viz., (i) Retrieval of paragraphs from scientific documents given a question and (ii) Retrieval of a text span given the citation text. We identify a retriever model SR having the best zero-shot performance for both these tasks, making it generalizable for our task and dataset. Section 6.4, details the experiments performed to choose SR.

We chunk the reference paper p_i using a sliding-window approach choosing 7 sentences as a chunk with a stride of 3. The top-k chunks from the reference paper most relevant to the citation text c_{ij} retrieved using the best-performing retriever SR, serve as the retrieved reference text span $\hat{s_{ij}}$ of article p_i for the topic t_j. The similarity score between the citation text and a chunk of the reference paper is calculated on the lines of (Nogueira et al., 2020), by taking log softmax over 'True' and 'False' tokens to get the probability of 'True' token as the score and rank the chunks. The retrieved top-k chunks may or may not be contiguous, and we use a higher k value for good recall, avoiding information loss. The retrieved reference text span $\hat{s_{ij}}$ functions as a pseudo-positive pair for the topic t_j. We consider three distinct types of pseudo-negative reference text spans $\overline{s_{ij}}$ for topic t_j (Figure 3): (i) **Type 1**: The bottom-k chunks retrieved from p_i, with citation text c_{ij} of topic t_j as query, using SR serve as easy negatives for the topic t_j in proposal R. (ii) **Type 2**: Text spans (top-k chunks) retrieved from the reference paper p_l, **NOT** cited in topic t_j, but cited in t_k where $k \neq j$, with the citation text c_{lk} as query using SR, serve as easy negatives for the topic t_j in proposal R. (iii) **Type 3**: With each c_{ij} for topic t_j citing p_i as the query, we retrieve top-k chunks from reference articles p_l, **NOT** cited in topic t_j, but cited in t_k where $l \neq i$ and $k \neq j$ using SR. The top-k chunks demonstrating maximum similarity with one of the c_{ij} serve as hard negatives for the topic t_j in proposal R.

We augment our training dataset with these

[1]https://github.com/NeuralNimbus/Beyond-Retrieval under license: GNU GPL v3

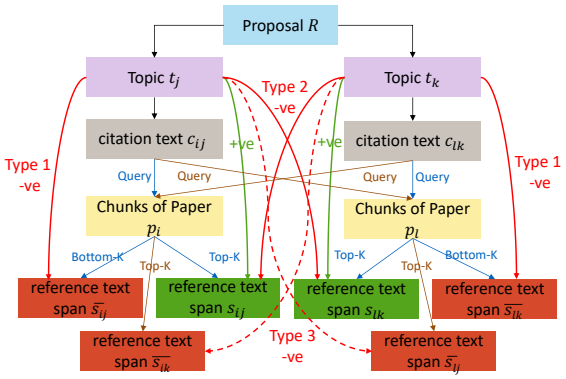

Figure 3: Negative sampling reference text spans for topics. Blue represents querying,Brown represents cross-querying,Green represents positive sample pairs, Red represents negative sample pairs

pseudo positive and negative reference text spans for each topic leading to the resultant dataset, where a sample is depicted by the tuple $< R, p_i, t_j, c_{ij}, \hat{s_{ij}}, \{\overline{s_{ij}}\} >$. The statistics of the augmented dataset are illustrated in Table 2.

Table 2: Augmented Dataset Statistics

Samples	**Splits**			**Total**
	Train	**Test**	**Validation**	
Positives	50,116	6,329	6,075	62,520
Easy Negatives	1,42,637	19,382	17,635	1,79,654
Hard Negatives	1,10,164	15,561	13,492	1,39,217
Total	3,02,917	41,272	37,202	3,81,391

5.2 Stage 2: Training

Topic based Reference Text Span Retriever (TR): We train TR, which is a Language Model (LM) using the positive and negative topic and reference text span pairs. Following the methodology discussed in (Nogueira et al., 2020), we try to maximize the probability $Pr(True|t_j, s_{ij})$ and $Pr(False|t_j, \overline{s_{ij}})$ using the cross-entropy loss. Due to the significantly higher number of negative samples compared to positives in the dataset, we include all positives but randomly sub-sample the negatives of each domain uniformly to maintain a balanced training dataset. Sampling is performed with replacement for every epoch to ensure that the model sees all negatives.

Proposal Topic based Reference Paper Classifier (TC): We train the TC using the augmented training dataset. We form samples $< R, t_j, \hat{s_{ij}} >$ with label $y_{ij} = 1$ and $< R, t_j, \overline{s_{ij}} >$ with label $y_{ij} = 0$. The model learns to classify if the paper p_i can be assigned to the topic t_j in the context of R and s_{ij}, with the supervision of the label y_{ij}. We try to maximize the probability $Pr(y = True|R, t_j, s_{ij})$ and $Pr(y = False|R, t_j, \overline{s_{ij}})$ using cross-entropy loss.

5.3 Stage 3: Inference

During inference, we feed the proposal title and abstract R and a topic t_j from the test set to TR along with chunks ch_q of a reference article p_i in the test set tagged as relevant to R. The model TR provides us with the a similarity score $sim_{jq} = TR(t_j, ch_q)$ for each ch_q. The score is calculated by applying a softmax on the logits of the 'True' and 'False' tokens and taking into consideration the probability $P(True|t_j, ch_q)$ (Nogueira et al., 2020). We rank ch_q for the given R and t_j based on sim_{jq} and use the top-k ranked ch_q of paper p_i as the retrieved reference text spans $\hat{s_{ij}}$ from paper p_ij, for topic t_j in R. We further feed $< R, t_j, \hat{s_{ij}} >$ to TC If probability $P(y = True|R, t_j, \hat{s_{ij}}) \geq 0.5$, then only we consider p_i to be aligned with t_j.

6 Experimentation and Results

6.1 Evaluation Metric

To assess the performance of SR on evidence and reference text span retrieval tasks (Section 6.4), we compute the evidence F1 score (Dasigi et al., 2021). For the evaluation of our pipeline (TR followed by TC), we utilize the binary ground truth labels depicting the alignment of each proposal topic and each reference paper relevant to that proposal and the labels predicted by TC in the context of retrieved reference text spans to compute the confusion matrix for the binary classification task. We choose the F1 score as the metric, given the label imbalance.

6.2 Baselines

To establish baselines we evaluate the 'Reference Paper Topic Classification' task with human and LLM-generated annotations. Our human annotator is a researcher with expertise in AI and allied fields. The details of the annotation interface are discussed in Appendix B. Considering the complexity, cognitive load and LLM cost for the task, the baselines are developed only on a smaller evaluation subset

of the test set. We utilize GPT-3.5 Turbo[2] LLM. To ensure deterministic and reproducible results, we set the temperature to 0. The structure of the prompt is detailed in Appendix A. Acknowledging the limitation of over-representation of CV papers in the dataset (Section 4), we construct the evaluation subset by uniformly sampling random four proposals from each of the five domains from the test set and thus ensuring balanced domain representation. This results in selecting 20 proposals with 52 topics citing 362 reference papers, forming 378 positive topic-reference paper pairs.

We have the following baselines: (i) Random: We randomly assign Yes/No labels to the test samples, (ii) TR + Human: The human annotator is provided with information about the target proposal's title, abstract (R), topic t_j, reference paper title p_i, and the text span of the reference paper s_{ij}, retrieved using the topic t_j with our trained TR model. The task is to assess the relevance of the reference paper to the given topic, labelling it 1 if the reference paper is aligned to the topic; and 0 otherwise, (iii) LLM: For vanilla LLM baseline we provide the complete reference paper p_i to the LLM along with information about the target proposal's title, abstract (R), topic t_j as the part of the prompt (Appendix A) and ask the LLM to classify the reference paper's relevance to the topic in the context of the proposal. We truncate the tail of the paper if the token length of the paper exceeds the maximum token length of the LLM. This baseline does not take into consideration text span from the reference paper, which is relevant to the topic and hence can be treated as is topic-agnostic reference text span based method, (iv) TR + LLM: This is the same has TR + Human except the relevant information mentioned above is provided to LLM as opposed to the human annotator. In this case, the LLM performs the job of classification of reference paper alignment to topics in the context of the proposal and retrieved reference text spans from the paper, as opposed to the complete paper, which is the case in vanilla LLM. Note that, on similar lines to the LLM baseline we could have had only the Human baseline, where the human gets to read the complete reference paper to perform the classification task. However, due to the very high cognitive load of this task, we skip this baseline.

[2]https://platform.openai.com/docs/models/gpt-3-5-turbo

6.3 Models and Hyperparameter Setting

Topic-Based Reference Text Span Retriever (TR): We fine-tune a T5-base model(Raffel et al., 2019), consisting of 223M parameters with a batch size of 120, learning rate of 3×10^{-4} with the AdamW optimizer, on Nvidia A100 for 20 epochs taking 84 hours, evaluated every 500 steps on the validation split. The model with the best validation F1 is obtained after fine-tuning for 1000 steps.

Research Paper Topic Classifier (TC): We chose RoBERTa (Liu et al., 2019) and Flan-T5 (Chung et al., 2022) base consisting of 125M and 248M parameters respectively, for their robust reasoning capabilities, as TC. Both models are fine-tuned on a class-balanced dataset with the batch sizes 240 and 120, using the AdamW optimizer and learning rates 5×10^{-5} and 3×10^{-4}, respectively. Fine-tuning is performed for 50 epochs on an Nvidia A100 taking approximately 120 hours each. The models with the best validation F1 are obtained at 33 and 35 epochs for RoBERTa and Flan-T5, respectively.

6.4 Results

Table 3: Results on Evidence Retrieval Task on QASPER and Citation Text Span Retrieval Task on Cl-SciSumm. The numbers indicate F1 scores for top-k retrieved chunks where k=3

Models	Datasets	
	QASPER	Cl-SciSumm
CGSN* (Nie et al., 2022)	53.98	NA
CitRet* (Pandey et al., 2022)	NA	19.79
Specter 2 (Singh et al., 2023)	13.68	19.28
CoSentBert (Mysore et al., 2022)	17.28	15.08
MPNET (Song et al., 2020)	15.81	14.14
Mono T5 (Nogueira et al., 2020)	25.90	25.98

* State-of-the-Art supervised models

Citation Text based Reference Text Span Retrieval (SR): As discussed in Section 5.1, our dataset lacks ground truth reference text span labels s_{ij} from a reference paper p_i, which are relevant to a topic t_j. Given the availability of citation text c_{ij} belonging to a topic t_j for the train set, we perform experiments to identify a state-of-the-art model to retrieve the reference text spans s_{ij} for that topic. We evaluate existing retrieval models on the test sets of QASPER (Dasigi et al., 2021) for evidence retrieval and CLScisumm (Chandrasekaran et al., 2019) for reference text span retrieval. We assess

the performance of models depicted in Table 3 in a zero-shot setting, to select the model most generalizable for our task and dataset. For the evidence retrieval task with QASPER, Specter 2 is used with the AdHoc Query Adapter for questions and the Proximity Adapter for paragraphs (Singh et al., 2022). For CLScisumm, Specter 2's AdHoc Query Adapter is employed to embed both the citation text and candidate citation text spans. Top-3 sentences are retrieved for each model, following the approach in (Pandey et al., 2022). We observe that Mono T5 achieves the highest F1 Score among pre-trained models for both datasets (Table 3). While not as good as the supervised state-of-the-art model CGSN (Nie et al., 2022) for QASPER, it surpasses the SOTA model CitRet (Pandey et al., 2022) in CL-SciSumm. Hence, we select Mono T5 as SR.

Table 4: F1 scores for the reference paper topic alignment task on the subset of test data (Section 6.2).

Annotation	Precision	Recall	F1 Score
Random	37.47	47.88	42.04
TR + Human	80.67	89.42	84.82
LLM [GPT-3.5 Turbo]	56.00	74.07	63.78
TR + LLM [GPT-3.5 Turbo]	57.77	72.75	64.40
TR+ TC [RoBERTa]*	52.27	65.60	61.16
TR+ TC [Flan-T5]*	56.17	72.22	63.19

* Our pipeline: models fine-tuned with our dataset. TR: T-5-Base Model fine-tuned with our dataset

Topic based Reference Text Span Retriever (TR) and Reference Paper Classifier (TC): Table 4 illustrates the evaluation of our inference pipeline discussed in section 5.3 on the subset of test data (Section 6.2). A high F1 score achieved by the human baseline establishes the task's feasibility. Higher F1 of our supervised models as compared to Random confirms task learnability. Higher F1 scores of TR + LLM as compared to LLM demonstrate the benefit of using our retrieval model. More importantly, comparable scores of our fine-tuned pipeline with very small (248M) LMs to that of pre-trained Large LM (175B), suggest the efficacy of utilizing our synthesized training data for task modelling. However, the noticeable difference in the human baseline concerning both the zero-shot LLM-based approach and our fine-tuned LM-based pipeline highlights the task's inherent difficulty, indicating a need for more sophisticated techniques. For a comprehensive evaluation, we evaluate the pipeline on our complete test set. RoBERTa and Flan-T5 yield F1 scores of 63.40% and 62.82%, respectively.

We conducted a detailed error analysis (Appendix C) of the errors made by the expert and identified author subjectivity, insufficient information in the retrieved reference text spans and lack of in-depth subject knowledge, as the challenges encountered by the expert. We further analyze the errors made by LLM and fine-tuned TC models for samples correctly classified by experts, detailed in Appendix C. The majority of errors are due to lexical overlap between topics and reference text spans with no semantic alignment causing False Positives or the models failing to perform complex reasoning required for the task.

7 Conclusion and Future Work

We have defined a novel task of mapping relevant scientific articles to research proposal topics as a precursor to the Literature Review Generation task for a new research proposal. We introduce a large-scale dataset for the task and establishment of competitive baselines by an expert and an LLM, underscoring task feasibility. We define a novel approach for the task and are the first ones to simulate the real-life scenario, at the early stage of proposal writing, of unavailability of citation text or detailed topic descriptions to retrieve topic-wise reference text spans from relevant articles. We came up with a novel strategy of using citation text (available for the training data) as a link between the topic and the text spans to create pseudo-labels for training a retriever. Our pipeline using much smaller open-source LMs trained with pseudo-labels yields comparable performance to that of the paid LLM baseline, demonstrating the efficacy of constructed data and designed pipeline. In contrast, a noticeable gap of 23.57 F1 score between our approach and the human baseline underscores the task's complexity, demanding more sophisticated solutions.

While organizing reference articles by topic, the hierarchy among these topics (catalogue), is crucial for a comprehensive understanding. However, we found it difficult to capture this hierarchy due to the PDF parsing challenges, treating every topic as a standalone entity. In future, with more sophisticated PDF parsing and semantic paragraph segmentation techniques, we plan to capture the topic hierarchy for task completeness.

References

Abbas Akkasi. 2022. Multi perspective scientific document summarization with graph attention networks (GATS). In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 268–272, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ankesh Anand, Tanmoy Chakraborty, and Amitava Das. 2017. Fairscholar: Balancing relevance and diversity for scientific paper recommendation. In *European Conference on Information Retrieval*.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 779–783. ACM.

Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir R. Radev, Dayne Freitag, and Min-Yen Kan. 2019. Overview and results: Cl-scisumm shared task 2019. In *BIRNDL@SIGIR*.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, 31.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6068–6077, Online. Association for Computational Linguistics.

Xiuying Chen, Hind Alamro, Li Mingzhe, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Target-aware abstractive related work generation with contrastive learning. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. *ArXiv*, abs/1905.00075.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020a. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020b. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Zekun Deng, Zixin Zeng, Weiye Gu, Jiawen Ji, and Bolin Hua. 2021. Automatic related work section generation by sentence extraction and reordering. In *AII@iConference*.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2023. What's new? summarizing contributions in scientific literature. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1019–1031, Dubrovnik, Croatia. Association for Computational Linguistics.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Coling 2010: Posters*, pages 427–435, Beijing, China. Coling 2010 Organizing Committee.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633, Doha, Qatar. Association for Computational Linguistics.

Shing-Yun Jung, Ting-Han Lin, Chia-Hung Liao, Shyan-Ming Yuan, and Chuen-Tsai Sun. 2022. Intent-controllable citation text generation. *Mathematics*, 10(10).

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. SciReviewGen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6695–6715, Toronto, Canada. Association for Computational Linguistics.

Avishek Lahiri, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Citeprompt: Using prompts to identify citation intent in scientific papers. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–55.

Xiangci Li, Yi-Hui Lee, and Jessica Ouyang. 2023. Cited text spans for citation text generation. *ArXiv*, abs/2309.06365.

Xiangci Li, Biswadip Mandal, and Jessica Ouyang. 2022. CORWA: A citation-oriented related work annotation dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5426–5440, Seattle, United States. Association for Computational Linguistics.

Xiangci Li and Jessica Ouyang. 2024. Explaining relationships among research papers. *ArXiv*, abs/2402.13426.

Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, Liang Hu, and Ivor Tsang. 2023. Causal intervention for abstractive related work generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2148–2159, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Anna Martin-Boyle, Aahan Tyagi, Marti A Hearst, and Dongyeop Kang. 2024. Shallow synthesis of knowledge in gpt-generated texts: A case study in automatic related work composition. *arXiv preprint arXiv:2402.12255*.

Zoran Medic and Jan Šnajder. 2023. Paragraph-level citation recommendation based on topic sentences as queries. *ArXiv*, abs/2305.12190.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Yuxiang Nie, Heyan Huang, Wei Wei, and Xian-Ling Mao. 2022. Capturing global structural information in long document question answering with compressive graph selector network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5036–5047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized document embeddings for aspect-based similarity of research papers. *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–12.

Amit Pandey, Avani Gupta, and Vikram Pudi. 2022. CitRet: A hybrid model for cited text span retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4528–4536, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mathias Parisot and Jakub Zavrel. 2022. Multi-objective representation learning for scientific document retrieval. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 80–88, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubâa, and Lisu Yu. 2021. Citation intent classification using word embedding. *IEEE Access*, 9:9982–9995.

Tarek Saier, John T. Krause, and Michael Färber. 2023. unarxive 2022: All arxiv publications pre-processed for nlp, including structured full-text and citation network. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 66–70.

Aravind Sesagiri Raamkumar, Schubert Foo, and Natalie Pang. 2017. Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing & Management*, 53(3):577–594.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *ArXiv*, abs/2004.09297.

Ashok Urlana, Nirmal Surange, and Manish Shrivastava. 2022. LTRC @MuP 2022: Multi-perspective scientific document summarization using pre-trained generation models. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 279–284, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Peter Vajdecka, Elena Callegari, Desara Xhura, and Atli Ásmundsson. 2023. Predicting the presence of inline citations in academic text using binary classification. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 717–722, Tórshavn, Faroe Islands. University of Tartu Library.

Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang. 2020. Toc-rwg: Explore the combination of topic model and citation information for automatic related work generation. *IEEE Access*, 8:13043–13055.

Yifan Wang, Yiping Song, Shuai Li, Chaoran Cheng, Wei Ju, Ming Zhang, and Sheng Wang. 2022. Disencite: Graph-based disentangled representation learning for context-specific citation generation. In *AAAI Conference on Artificial Intelligence*.

Yongzhen Wang, Xiaozhong Liu, and Zheng Gao. 2018. Neural related work summarization with a joint context-driven attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786, Brussels, Belgium. Association for Computational Linguistics.

Jia yan Wu, Alexander Te-Wei Shieh, Shih Ju Hsu, and Yun-Nung Chen. 2021. Towards generating citation sentences for multiple references with intent control. *ArXiv*, abs/2112.01332.

Chrysoula Zerva, Minh-Quoc Nghiem, Nhung T. H. Nguyen, and Sophia Ananiadou. 2020. Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics*, 125:3109–3137.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical catalogue generation for literature review: A benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6790–6804, Singapore. Association for Computational Linguistics.

A Prompt Structure for LLM Annotation

As outlined in Section 6.2, we employ GPT-3.5 Turbo to set a benchmark on the Annotated split of the dataset. After experimenting with various prompts, we find that the prompt depicted in Listing 1 yields the best results.

Listing 1: Prompt Structure for TR + LLM baseline. For vanilla LLM we replace the Text Span of paper 'B' with the actual reference paper.

```
system_prompt = '''
You are a Research Paper Relevance
    Classifier. Given the 'Title' and '
    Abstract' of Paper `A' and a 'Topic'
     of Paper 'A', and a text span from
    Paper 'B' predict whether Paper 'B'
    is directly related to the topic of
    Paper 'A', in context of the 'Title'
     and 'Abstract' of Paper 'A'. Output
     the model's prediction as:"

Output: [1] (for 'yes') or [0] (for 'no
    ')'''

user_prompt = '''
Title of Paper 'A':
<Target Proposal Title>
Abstract of Paper 'A':
<Target Proposal Abstract>

Topic of Paper 'A':
<Topic>

Text Span of Paper 'B':
<Reference Paper Title>
<Citation Text Span>

Is the content of Paper 'B' directly
    related to the topic of Paper 'A' (<
    Topic>)?
'''
```

B Annotation Interface

We have developed an Annotation tool aimed at offering a user-friendly interface to engage the human expert for annotating samples. The Annotator tool provides information on the target proposal's title, topic, reference paper title, abstract and the retrieved citation text span of the reference paper, as illustrated in Figure 6 (b). Moreover, we mandate the expert to review the disclaimer prior to the annotation process, ensuring that they are well informed about the context and ethical guidelines associated with the samples they are annotating, as reflected in Figure 6 (a).

C Error Analysis

We perform error analysis for the task of reference paper to topic mapping (classification). We use the sub-sampled test set discussed in Section 6.2 for the analysis.

C.1 Human Baseline

We analyze the False Positive and Negative samples with the annotations provided by the human expert.

The primary objective is to gain a profound understanding of the challenges inherent in the task. The identified 121 erroneous annotations provided by human are categorized into the following groups:

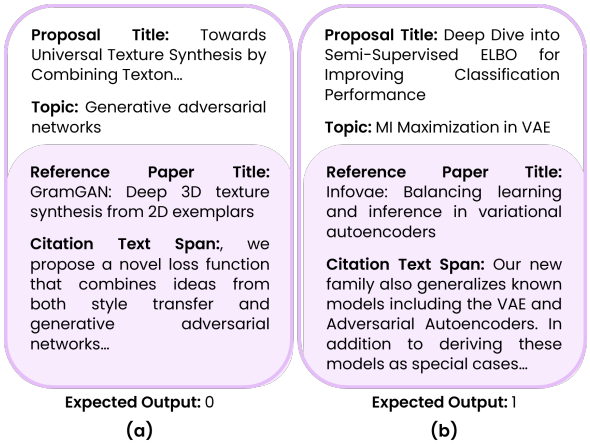

Figure 4: Samples illustrating Human errors due to (a) Subjectivity and (b) Insufficiency of information in the retrieved citation text span.

- **Subjectivity:** These errors, constituting 77.69% of total erroneous cases, stem from a misalignment between the thought processes of the author and the annotator. As shown in Figure 4 (a), the reference paper is not cited by the author under the given topic, however human annotator tends to map the reference paper to the topic due to the match between the semantics of the citation text span and the topic. In these instances, the annotator correctly deems the reference article relevant to the topic, while the author holds a contrasting view. This emphasizes the inherent subjectivity embedded in the task, where interpretations of relevance can diverge between individuals.

- **Insufficiency:** Errors in this category are 4.96% of total erroneous cases and are attributed to the insufficient information present in the retrieved citation text span. The experts face challenges in making well-informed decisions due to the lack of comprehensive details, impeding their ability to accurately assess the relevance of the reference article to the given topic. For example, as highlighted in Figure 4 (b), the citation text span does not carry enough information to let the author infer its mapping to the topic. We observe in some of these cases the citation text span may carry information semantically similar to the citation text and thus the topic, but the information is not sufficient to logically reason about the possible alignment of the reference paper to the topic. This problem can be addressed by having a better mechanism for citation text span retrieval given citation text, which would indirectly help us improve the results of topic based citation text span retrieval model to retrieve more relevant citation text spans. This can be improved by augmenting the pipeline with an entailment task following the retrieval, to evaluate which of the top-k retrieved citation text spans entails the citation text to accommodate the reasoning component. We leave this enhancement for future work.

- **Inadequacy:** Errors categorized under inadequacy, accounting for 7.44% of total erroneous cases, emerge due to the lack of subject matter knowledge of the expert annotator. This knowledge gap prevents her from accurately understanding the relevance of the reference article to the proposed topic.

- **True Errors:** This category, constitutes 9.92% of total erroneous cases, encompassing genuine mistakes made by human annotators. Despite concerted efforts to maintain accuracy, errors of this nature occur, underscoring the complexity and human cognitive load of the task.

The systematic categorization of these errors showcases the prevalence of subjectivity, arising from differences in the author's writing style and thought process which constitutes the majority of mis-classifications by human annotators.

C.2 LLM baseline and fine-tuned models against Human baseline

We analyze instances where human annotators correctly classified the samples that are misclassified by the LLM and fine-tuned models, as outlined in Table 5. We majorly categorize the error into two categories:

- **Requirement of complex reasoning:** These errors are due to the absence of direct alignment between the topics and the citation text span of the reference paper. It requires multi-hop complex reasoning for the mapping task. As demonstrated in Figure 5 (a), the reference paper is not directly relevant to the topic

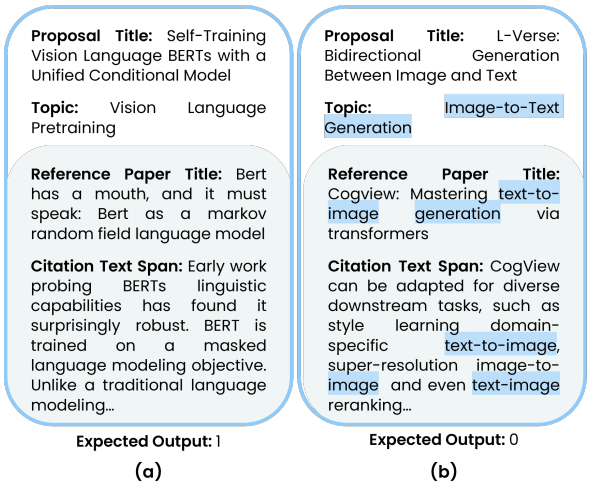

Figure 5: Samples correctly classified by humans but erroneously classified by LLM and fine-tuned models due to (a) Requirement of complex reasoning and (b) Misleading Lexical overlap between topic and retrieved citation text span but no semantic match

Table 5: Error Analysis of the Mis-classifications of LLM and fine-tuned models against Human baseline

Models	Samples	Complex Reasoning	Lexical Overlap
LLM [GPT-3.5 Turbo]	196	38.89%	61.11%
TC [RoBERTa]*	234	42.22%	57.78%
TC [Flan-T5]*	224	33.33%	66.67%

* models fine-tuned on our dataset.

'Vision Language Pretraining. However, the BERT model can be used as the text encoder for the vision language model. The fine-tuned models and LLM sometimes find it difficult to perform such complex reasoning to come up with the expected mapping.

- **Lexical:** Errors in this category stem from the presence or absence of identical words in the extracted citation text span when compared to those in the topic, leading to misclassification. A representative example of a lexical error is illustrated in Figure 5 (b), where there is lexical overlap between the topic name and citation text span, however, the semantics do not match.

The primary factor contributing to misclassifications of models is lexical similarity or dissimilarity between the topic and the retrieved citation text span. This highlights the model's challenge in comprehending contextual nuances, leading to misinterpretations.

D Limitations

Our pipeline is fine-tuned with our constructed dataset containing papers in the AI (majorly CV) domain. The results are demonstrated with a test set containing papers in the AI domain only, leading to comparable results with the zero-shot vanilla LLM baseline. Our pipeline may not perform well in cross-domain settings. However, our dataset construction technique can be easily applied to the scientific articles belonging to any domain, based on the availability of the articles. Extrapolating the current results demonstrated for the AI - CV domain, the retrieval-classifier pipeline fine-tuned with such domain-specific synthesized dataset should lead us to comparable results with the zero-shot vanilla LLM baseline, for any domain.

We create pseudo-topic-reference text span labels for the training data by considering the text under that topic (section) which is written to cite a reference article (citation text) as an explicit link to connect the topic and the reference article. It clearly carries the information of the relevance of the scientific article with regards to the proposal, in the context of the given topic. The citations to the same reference article but at other places in the proposal may not carry information on the relevance of the reference article in the context of the given topic to which the reference article is to be aligned. However, we are not completely denying the possibility of some other sections carrying information about the relevance of the reference article to the proposal in the context of the given topic. However as there is no explicit signal that we can exploit to derive this connection, we cannot consider that information while establishing the connection.

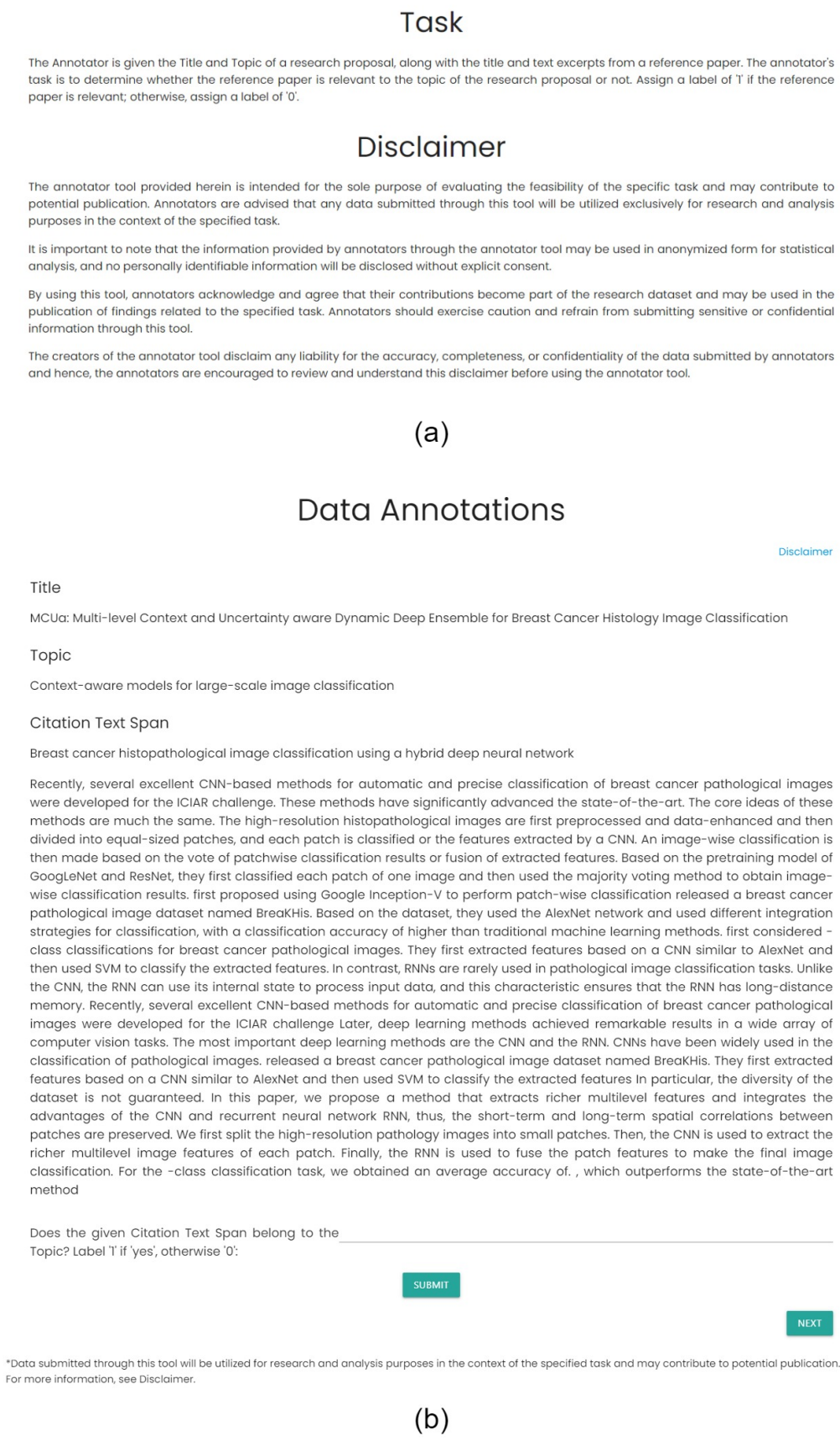

Figure 6: The Annotator Tool illustrating the (a) task description and disclaimer presented to the human expert before the annotation process and, (b) the annotation interface showcasing a sample from the dataset.