TALK2REF: A Dataset for Reference Prediction from Scientific Talks

Frederik Broy, Maike Züfle, Jan Niehues
Karlsruhe Institute of Technology, Germany
frederik.broy@student.kit.edu,
{maike.zuefle, jan.niehues}@kit.edu

Abstract

Scientific talks are a growing medium for disseminating research, and automatically identifying relevant literature that grounds or enriches a talk would be highly valuable for researchers and students alike. We introduce Reference Prediction from Talks (RPT), a new task that maps long, and unstructured scientific presentations to relevant papers. To support research on RPT, we present TALK2REF, the first large-scale dataset of its kind, containing 6,279 talks and 43,429 cited papers (26 per talk on average), where relevance is approximated by the papers cited in the talk's corresponding source publication. We establish strong baselines by evaluating state-of-the-art text embedding models in zero-shot retrieval scenarios, and propose a dual-encoder architecture trained on TALK2REF. We further explore strategies for handling long transcripts, as well as training for domain adaptation. Our results show that fine-tuning on TALK2REF significantly improves citation prediction performance, demonstrating both the challenges of the task and the effectiveness of our dataset for learning semantic representations from spoken scientific content. The dataset and trained models are released under an open license to foster future research on integrating spoken scientific communication into citation recommendation systems.

Keywords: Scientific Talks, Citation Prediction, Spoken Language Processing

1. Introduction

In recent years, the number of recorded scientific talks across conferences, academic platforms, and educational settings has increased dramatically. These recordings represent a rapidly growing source of scientific communication, enabling researchers, students, and practitioners to revisit presentations, lectures, and discussions, and engage with scientific ideas beyond traditional written publications. However, identifying related or thematically relevant work from a scientific *talk* remains challenging and time-consuming, yet access to such related content is highly valuable for researchers who wish to explore prior work, follow up on mentioned ideas, and discover new connections.

Reference or citation prediction has been extensively studied in the context of written scientific text, in both *local* settings, where models predict in-text references based on the surrounding context (Zhang and Ma, 2020a; Gu et al., 2022a; Çelik and Tekir, 2025; Zhang and Ma, 2020b), and *global* settings, which recommend relevant references for an entire document (Bhagavatula et al., 2018; Li, 2024; Stergiopoulos et al., 2024). In both cases, methods assume clean, formal, and linguistically organized inputs that capture the semantics of scientific discourse, allowing models to rely on concise and well-structured representations, such as abstracts, titles, or surrounding sentences, without processing the full text.

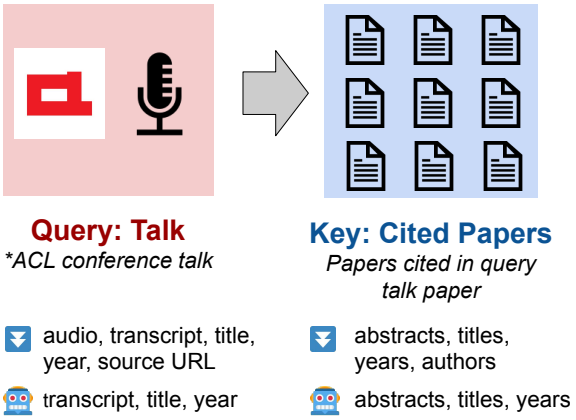

Figure 1: Illustration of the **TALK2REF** dataset and its use in the task of Reference Prediction from Scientific Talks (RPT), where query talks are paired with their cited papers. The ⏬ represents the information included in TALK2REF, and 🤖 represents the input used by our model for predicting cited papers.

By contrast, scientific talks are unstructured and often noisy, containing disfluencies, filler expressions, and spontaneous language (Züfle et al., 2025) that lacks the precision of academic prose. Moreover, transcripts are typically long and verbose, spanning tens of minutes, which poses additional challenges for effective representation, retrieval, and modeling. As a result, the entire talk must be processed to capture its content, unlike written papers where state-of-the-art methods rely on abstracts and section headings (Bhagavatula et al.,

[1]The dataset is available at https://huggingface.co/datasets/s8frbroy/talk2ref.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 357–371
11–16 May 2026.

2018; Li, 2024; Stergiopoulos et al., 2024).

Importantly, this introduces not only a domain shift from written to spoken language, but also an inherent mismatch between the query and retrieval spaces in style and domain: the queries are derived from spontaneous spoken content, while the targets are formal written papers. Capturing semantic correspondences across these distinct modalities and linguistic styles is therefore substantially more difficult, rendering traditional text-based citation prediction models poorly suited for this setting.

To address this gap, we introduce the novel task of Reference Prediction from Talks (RPT): given a scientific talk as a query, the *query talk*, the goal is to predict the set of papers that are relevant to the talk, the *relevant papers*. RPT extends the paradigm of citation recommendation from written to spoken scientific communication, opening new opportunities for integrating spoken research content into digital scholarly systems. An overview of the reference prediction setting is illustrated in Fig. 1.

To the best of our knowledge, no existing dataset supports research on this problem. We therefore construct and release TALK2REF, the first large-scale dataset that pairs scientific presentations with their corresponding relevant papers, modeling relevance using the papers referenced in each talk's source publication. TALK2REF includes 6,279 talks and 43,429 papers, with an average of 26 references per talk, providing a foundation for systematically studying reference prediction from spoken scientific content at scale.

To provide reference points for future research, we establish competitive baseline models. We begin by evaluating state-of-the-art sentence embedding models, (Reimers and Gurevych, 2019; Cohan et al., 2020) in a zero-shot retrieval setting, where both talks and papers are encoded into a shared embedding space, and the most semantically similar papers are retrieved for each talk.

We then fine-tune a dual-encoder model (Karpukhin et al., 2020) on TALK2REF to better align representations of talks and their cited papers. This adaptation enables the model to learn task-specific associations beyond those captured by pretrained embeddings. A key challenge in this setting is the considerable length and complexity of talk transcripts, which exceed the typical input limits of Transformer-based encoders (Devlin et al., 2019; Reimers and Gurevych, 2019; Cohan et al., 2020). To handle this, we design strategies that enable the model to represent long talks effectively while preserving their overall semantics.

Our results show that fine-tuning on TALK2REF yields substantial improvements over zero-shot performance, demonstrating that the dataset effectively supports model adaptation and generalization. Together, these findings indicate that TALK2REF not only provides a new research task, but also offers strong, competitive benchmarks that future models can build upon.

The main contributions of this work are threefold:

1. We introduce and publicly release the TALK2REF dataset for the new task of RPT under an open license (CC-BY-4.0).[1]

2. We establish strong baselines and propose a dual-encoder framework for modeling citation prediction from scientific talks, which we also release under an open license.[2]

3. We analyse different aggregation mechanisms and training scenarios for RPT.

2. Related Work

Citation recommendation is commonly divided into two tasks: *local* and *global* citation prediction. Local approaches predict in-text references based on surrounding context (Zhang and Ma, 2020a; Gu et al., 2022a; Çelik and Tekir, 2025; Zhang and Ma, 2020b), while global approaches suggest relevant references for an entire document (Bhagavatula et al., 2018; Li, 2024; Stergiopoulos et al., 2024). As our work targets reference prediction from entire talks, we focus exclusively on the global task.

Datasets. Unlike local citation tasks—which benefit from standardized benchmarks such as FullTextPeerRead (Jeong et al., 2019) or RefSeer (Huang et al., 2014), global citation recommendation lacks widely used datasets. Most prior work instead builds on large-scale bibliographic corpora, especially the DBLP-Citation-Network and ACM-Citation-Network (Tang et al., 2008), both provided by AMiner[3]. These datasets contain millions of papers and citation links in the domain of computer science, enriched with metadata such as title, abstract, authorship, venue, publication year, and publisher.

Variants or subsets of these corpora are widely used to train and evaluate citation recommendation models. Prior studies rely on smaller subsets of the DBLP or ACM-Citation-Network containing tens of thousands of papers (Li, 2024; Bhagavatula et al., 2018), as well as larger-scale versions comprising hundreds of thousands of documents and citation links (Stergiopoulos et al., 2024; Zhang et al., 2024; Ali et al., 2021a).

[2] https://huggingface.co/s8frbroy/talk2ref_query_talk_encoder and https://huggingface.co/s8frbroy/talk2ref_ref_key_cited_paper_encoder.

[3] https://www.aminer.cn/citation

Other datasets also complement these large-scale bibliographic corpora. The ACL Anthology Network (Radev et al., 2009, AAN) provides a domain-specific benchmark for natural language processing research and is frequently reused for citation recommendation tasks (Zhang et al., 2024; Ali et al., 2021b). Broader, cross-domain resources such as the STM-Knowledge Graph (Brack et al., 2021; Brack et al., 2021, STM-KG), OpenCorpus (Bhagavatula et al., 2018), and combined datasets built from the Microsoft Academic Graph and CiteSeerX (Li et al., 2006; Sinha et al., 2015; Ayala-Gómez et al., 2018) have also been employed to explore large-scale citation and knowledge graph modeling.

Despite this variety, all existing datasets focus exclusively on written-scientific-text papers, abstracts, or titles, leaving spoken scientific content entirely unexplored. Our dataset, TALK2REF, addresses this gap by pairing scientific talks with their cited papers, enabling research on reference prediction from talks.

Modeling Approaches. Architectures for global citation recommendation differ primarily in the type of input they consider. Some rely only on document titles (Li, 2024), others use titles and abstracts (Bhagavatula et al., 2018; Stergiopoulos et al., 2024), and some leverage full-text papers (Chokkalingam, 2024). Additional information such as citation graphs (Li, 2024; Stergiopoulos et al., 2024) or user profiles (Stergiopoulos et al., 2024) has also been incorporated to improve prediction. These systems can be broadly categorized into classification-based and retrieval-based architectures. Classification-based models (Chokkalingam, 2024; Stergiopoulos et al., 2024) predict citation relevance for each query–candidate pair, while retrieval-based models (Li, 2024; Bhagavatula et al., 2018) precompute embeddings for all candidates and rank them by similarity to the query, allowing for more efficient large-scale recommendation.

3. Task Formulation and Dataset

Global Reference Prediction from Scientific Talks (RPT) aims to identify the set of relevant references for a talk, bridging unstructured spoken content and structured scientific literature. Formally, given a scientific presentation, the *query talk*, T, the objective is to retrieve a set of *relevant papers* $\{R_1, R_2, \ldots, R_n\}$, where n may differ across talks, reflecting the varying number of references associated with each presentation. To support this task, we introduce TALK2REF, which, to the best of our knowledge, is the first dataset of its kind.

3.1. Dataset Construction

To support Reference Prediction from Talks (RPT), a dataset must pair each query talk, the audio recording of a scientific presentation, with its relevant papers. In TALK2REF, we model relevance using the citations in the source publication associated with each talk. Each input sample includes the talk's audio, transcript, title, publication year, and source URL, while each cited paper is represented by its title, abstract, and metadata such as authorship and year.

The Association for Computational Linguistics (ACL) Anthology[1] provides a suitable source, offering talks linked to papers with accessible citation information. All talks are distributed under the Creative Commons Attribution 4.0 International License, allowing us to use them to construct a new dataset. To streamline construction, we build on the NUTSHELL dataset (Züfle et al., 2025), which aggregates ACL conference talk recordings with their corresponding paper abstracts and metadata. Using this foundation, we compile TALK2REF, a dataset that meets the requirements of RPT by pairing query talks with the references of their underlying papers.

However, while NUTSHELL provides talk recordings and paper metadata, it does not include the list of references from the query talk papers. We construct TALK2REF by extracting the references from the original papers to obtain the full set of cited papers for each talk. The dataset construction process is summarized as follows:

1. **PDF retrieval.** Each query talk in NUTSHELL provides a link to its ACL page; from there we obtain the PDF of the corresponding paper, discarding any corrupted files.
2. **Query talk paper parsing.** We parse the PDF of each query talk paper using GeneRation Of Bibliographic Data (GROBID Developers, 2008, GROBID). From the parsed document, we extract the paper title (for the query talk) as well as the structured metadata of the references, including titles, author lists, venues, years of publication, and Digital Object Identifiers (DOIs). The talk's audio and year are obtained from the original metadata.
3. **Transcript generation.** We transcribe the audio of each query talk to produce textual input for processing, using `whisper-large-v3` (Radford et al., 2023).
4. **Abstract retrieval for cited papers.** Since cited papers are not exclusively from ACL and coverage varies across sources, we use the metadata obtained in Step 2 for each cited

[1] https://aclanthology.org

Split	Confs.	Years	Talks	Avg. len. (min)	Avg. words/ transcript	Avg. papers	Total papers	Citation years	Avg. words/ abstract
Train	ACL, NAACL EMNLP	2017–2021	3971	12.1	1615.3	26.75	31,064	1948–2021	142.4
Dev	ACL	2022	882	9.9	1326.9	26.05	11,805	1967–2022	147.8
Test	EMNLP, NAACL	2022	1426	9.1	1186.1	25.66	16,935	1953–2022	149
Total	ACL, NAACL EMNLP	2017–2022	6279	11.1	1477.6	26.4	43,429	1948–2022	144.8

Table 1: Dataset statistics for our proposed dataset **Talk2Ref**, that includes query talks and cited papers. We show that citation years for years with at least 10 references, words are split at whitespace.

paper to query six bibliographic APIs and datasets to obtain abstracts, including Crossref[2], arXiv[3], OpenAlex[4], Semantic Scholar (Kinney et al., 2023), the Laion arXiv-abstract dataset[5], and the ACL OCL corpus (Rohatgi et al., 2022). We prioritize DOI-based queries, as DOIs provide a unique and unambiguous identifier for each paper; however, they are not always present in the parsed metadata. Therefore, we first perform title–author queries, which may return abstracts directly or allow us to recover a DOI. Once a DOI is obtained, it is used for subsequent API queries, reducing the risk of mismatches due to title variations.

5. **Post-processing.** We filter out incorrect or placeholder abstracts, such as entries containing only authorship, venue, or year information.

The resulting dataset links each query talk, represented by its audio, transcript, title, year, and source URL to its cited papers, which are enriched with abstracts and metadata such as title, authors, and publication year.

3.2. Dataset Statistics

Our dataset construction process results in 6,279 query talks linked to 43,429 cited papers in total. The query talks average 11.1 minutes in duration, and their transcripts contain on average 1,478 words. Each query talk is associated with an average of 26.4 cited papers, providing a rich set of references per talk.

The data split follows the original NUTSHELL dataset (Züfle et al., 2025), which partitions talks chronologically by conference year. Talks in earlier years form the training set (2017-2021), while those from later years are used for development and testing (2022). By preserving this ordering in Talk2Ref, the test set consists of talks from more recent years than the training set, providing a realistic, temporally consistent evaluation scenario. Consequently, there are 3971 query talks in training, 882 query talks in development, and 1426 talks in the test split. Detailed dataset statistics are also provided in Table 1.

In general, the number of references increases after 2017, mirroring real-world publishing trends in this field. Further statistics about coverage over years can be found in Fig. 2 in Section 9. The ten most cited papers can be fond in Fig. 3.

4. Analysis

We now use our dataset, Talk2Ref, to benchmark baseline models for reference prediction from scientific talks. This analysis serves two purposes: first, to assess the difficulty of the task, and second, to evaluate whether models can be effectively trained on the dataset, thereby establishing its value as a resource for developing reference prediction systems.

4.1. Experimental Setting

4.1.1. Input Representations

Query talk representations. On the query talk side, we do not use raw audio directly in the tested models, but rely on long-form transcripts of the talks. Although the transcripts can be noisy and lengthy, they enable us to leverage the rich textual content of the talk, along with its title and publication year, while keeping the input modality consistent with the cited-paper side.

Transcribing first also offers practical benefits: automatic speech recognition systems are mature and widely available, and their intermediate outputs are easier to inspect than learned audio embeddings. While emerging models such as SONAR (Duquenne et al., 2023) could, in principle, support direct speech–text retrieval using similar chunking and pooling strategies, we leave this exploration to future work.

[2] https://www.crossref.org/documentation/retrieve-metadata/rest-api/

[3] https://info.arxiv.org/help/api/basics.html

[4] https://openalex.org/rest-api

[5] https://huggingface.co/datasets/laion/arXiv-abstract

However, working with transcripts introduces another challenge: common pretrained sentence encoders like BERT (Devlin et al., 2019), Sentence-BERT (Reimers and Gurevych, 2019), or Specter (Cohan et al., 2020) accept only a limited number of tokens (512), much shorter than the average transcript length of 1,478 words. To obtain fixed-size embeddings for these long transcripts, we compare several strategies:

1. *Truncation*: keeping only the first 512 tokens;
2. *Chunking with mean pooling*: splitting the transcript into 512-token segments and averaging their embeddings;
3. *Chunking with max pooling*: splitting the transcript into 512-token segments and taking the element-wise maximum of the segment embeddings;
4. *Chunking with learned weighted mean*: splitting the transcript into 512-token segments. A small feed-forward (linear) layer produces a scalar for each segment. The scalars are passed through a softmax layer to produce weights. These weights determine how much each segment contributes to the weighted mean.

We experiment with different input configurations, including using transcripts alone or appended to the query talk title and publication year.

Cited paper representations. On the output side, we rely on the title, abstract, and publication year of the cited papers. We experiment with each feature individually as well as their combinations. No truncation or aggregation is needed, as these inputs fall well within the encoder context length.

4.1.2. Training and Retrieval

We evaluate several models (detailed in Section 4.1.3) in both zero-shot and trained settings on our dataset. Our training strategy is detailed below.

Training. We employ a contrastive learning setup inspired by dense passage retrieval (Karpukhin et al., 2020, DPR). We do not initialize from pretrained DPR weights due to the mismatch in task structure: DPR models are trained for open-domain question answering. Instead we use pretrained sentence embedding models as detailed below. Each query talk T and cited paper R_i is mapped to a vector representation by separate encoders f_T and f_R, producing embeddings $f_T(T)$ and $f_R(R_i)$, respectively. The similarity between a talk and a candidate paper is computed as the dot product $s_i = f_T(T) \cdot f_R(R_i)$.

The model is optimized to assign higher similarity scores to correct talk–reference pairs and lower similarity scores to incorrect talk–reference pairs.

Unlike Karpukhin et al. (2020), where each query has a single positive, talks often have multiple relevant papers. We therefore replace the original softmax-based objective with a sigmoid-based binary classification loss, allowing multiple correct references per talk:

$$\mathcal{L} = -\sum_{i=1}^{N} \left[y_i \log \sigma(s_i) + (1 - y_i) \log(1 - \sigma(s_i)) \right],$$

where $y_i \in \{0, 1\}$ indicates whether R_i is a true citation, and $\sigma(\cdot)$ denotes the sigmoid function.

For efficient training, other talk–reference pairs within the same batch are used as negative examples. Specifically, for a query talk at index i in the batch, all cited papers from different talks $j \neq i$ are treated as negatives, unless a cited paper is shared across both query talks.

If f_T and f_R have different embedding dimensions, a linear projection layer is added to align them.

Domain adaptation stage. We also experiment with a task-specific domain adaptation stage of the pretrained encoders before the main training stage. This domain adaptation stage follows a contrastive objective similar to Karpukhin et al. (2020), however, each query talk T is paired with its own abstract instead of the cited papers' abstracts. The resulting one-to-one alignment enables the use of a standard softmax-based loss, encouraging the model to learn meaningful representations before finetuning on the main citation recommendation task.

The domain adaptation stage serves two purposes. First, it adapts the encoders to the domain of scientific talks, improving their ability to process noisy, long-form transcripts. Second, it provides a simpler learning objective, as mapping a talk to its own query paper's abstract is easier than retrieving cited papers, whose abstracts may only partially overlap with the talk.

Retrieval and inference. All candidate paper embeddings are precomputed and stored in a FAISS index (Johnson et al., 2021) for efficient retrieval. To ensure temporal consistency, we restrict retrieved papers to those published prior to the query talk paper, preventing the model from selecting "future" papers. At inference, the query talk transcript is embedded and compared against the candidate paper embeddings to identify the top-k most relevant references.

Query Input	Key Input	Precision	P@ 10	P@ 20	R@ 20	R@ 50	MAP@ 10	MAP@ 20
N/A	10-most cited papers	10.33	17.48	11.64	9.82	13.82	12.05	7.39
Transcript	Abstract	8.89	12.14	9.65	7.79	13.18	6.63	4.49
Transcript	Abstract + Title	9.32	12.82	10.02	8.16	13.74	7.12	4.77
Transcript	Abstract + Title + Year	9.32	12.79	10.07	8.21	13.76	7.12	4.78
Transcript + Title	Abstract + Title + Year	9.55	**13.32**	10.39	8.58	14.13	7.15	4.83
Transcript + Title + Year	Abstract + Title + Year	**9.57**	13.29	**10.42**	**8.59**	**14.21**	**7.33**	**4.94**
Abstract	Abstract + Title + Year	**13.19**	**19.02**	**14.54**	**12.05**	**18.99**	**11.05**	**7.50**
Abstract + Title	Abstract + Title + Year	13.06	18.73	14.40	11.97	18.83	10.95	7.45
Abstract + Title + Year	Abstract + Title + Year	13.03	18.70	14.29	11.86	18.82	10.88	7.39

Table 2: Results (in %) on the Talk2Ref dataset using zero-shot SBERT, truncating inputs to SBERT's maximum sequence length. We experiment with different features for both query talks and candidate papers (keys). Using abstracts on the key side is an unrealistic setting, as abstracts are not provided with the talk; we include them to contextualize the difficulty of the task.

Baselines. Two baselines are considered, addressing different aspects of the task: (i) Task-level baseline: predicting the top-k most frequently cited papers across all talks, providing a simple frequency-based reference for the overall difficulty of the citation prediction task; and (ii) Model-level baseline: using only the first x tokens of each transcript without any aggregation, serving as a lower bound for our transcript-encoding strategies.

4.1.3. Models

To assess the effect of different pretrained encoders on the task, we evaluate four models:

- **BERT** (Devlin et al., 2019): a general-purpose language model not trained for retrieval tasks. A sentence representation is obtained by averaging over the embeddings of each token.
- **Longformer** (Beltagy et al., 2020): a general-purpose language model capable of processing sequences of up to 4,096 tokens, thereby reducing the need for aggressive truncation.
- **SPECTER2** (Cohan et al., 2020): a model pretrained on scientific citation data and designed to encode scientific papers, building on SciBERT (Beltagy et al., 2019), a variant of BERT.
- **Sentence-BERT** (Reimers and Gurevych, 2019, SBERT): a model optimized for producing fixed-size sentence embeddings using a siamese network architecture based on BERT.

All models have a maximum sequence length of 512 tokens, except Longformer, which can handle up to 4,096 tokens. For BERT and Longformer, sentence embeddings are obtained by averaging token embeddings, following the approach recommended by Beltagy et al. (2019). In the same way, SBERT (Reimers and Gurevych, 2019) uses mean pooling over tokens (Reimers and Gurevych, 2019), and SPECTER2 (Cohan et al., 2020) uses the `[CLS]` token as its sentence representation, as done in the original papers.

Further details about the models, including number of parameters, are provided in Table 4 in Appendix B.

Training details. We train for up to 72 hours on a single GPU H100, with early stopping with patience of four epochs on validation performance. We use a batch size of 24 for the models with sequence length of 512, and a smaller batch size of 3 for Longformer. Detailed hyperparameters are given in Table 5 in Section 9.

4.1.4. Evaluation

Following prior work on citation recommendation (Bhagavatula et al., 2018; Chokkalingam, 2024; Gu et al., 2022b), we evaluate retrieval quality using Precision, Precision at cutoff k (P@k), Recall, Recall at cutoff k (R@k), and Mean Average Precision (MAP) at cutoff k (MAP@k). These metrics capture complementary aspects of retrieval quality. P@k and MAP@k focus on the quality of the top-ranked recommendations, and R@k measures the coverage of all relevant references.

We retrieve the top-k most similar papers for each query talk and compute the metrics with respect to the gold set of cited papers.

We report P@k and MAP@k for $k \in \{10, 20\}$, R@k for $k \in \{10, 20, 50\}$ following conventions in previous work (Bhagavatula et al., 2018) and well aligned with the characteristics of our Talk2Ref, where each query talk cites approximately 26 papers on average.

4.2. Results

In the following, we report results on Talk2Ref and analyze different aspects of the task, including

Strategy	Encoder	Max Seq. Len	Aggregation	Prec.	P@ 10	P@ 20	R@ 20	R@ 50	MAP@ 10	MAP@ 20
Top-k most freq.	–	–	–	10.33	17.48	11.63	9.82	13.81	**12.04**	7.39
Zero-Shot	BERT	512	Truncated	0.56	0.67	0.57	0.49	0.82	0.24	0.17
	Longformer	4096	–	0.20	0.23	0.21	0.18	0.34	0.08	0.05
	Specter2	512	Truncated	8.31	11.78	9.10	7.37	12.43	6.15	4.09
	SBERT	512	Truncated	9.57	13.29	10.42	8.59	14.21	7.33	4.94
	SBERT	512	Mean	10.60	14.83	11.71	9.54	15.76	8.21	5.56
Finetuned	SBERT	512	Truncated	12.94	17.23	13.85	11.49	19.59	9.76	6.91
	SBERT	512	Mean	13.34	17.75	14.28	11.92	20.35	10.13	7.15
	SBERT	512	Learned mean	13.95	18.81	15.12	12.59	21.00	10.86	7.69
	Longformer/ SBERT	4096	–	8.47	10.75	9.15	7.32	14.06	5.12	3.57
Dom. Adapt. + Finetuned	SBERT	512	Learned mean	**14.18**	**19.14**	**15.32**	**12.71**	**21.60**	10.98	**7.72**
	Longformer/ SBERT	4096	–	8.90	11.25	9.57	7.28	14.16	5.43	3.78

Table 3: Results (in %) on the Talk2Ref dataset. Query inputs consist of transcript, title, and year, while cited papers (keys) are represented by title, year, and abstract. We report different aggregation mechanisms for handling long query transcripts apart from Longformer, where the transcript fits completely, and show results for finetuning the best zero-shot model on Talk2Ref.

which input features to use, the impact of using full talk transcripts versus only abstracts, and the effects of different fine-tuning strategies.

Selecting input features. We begin by exploring various input configurations in the zero-shot setting to assess which features contribute most to retrieval quality, using SBERT as the encoder. The results are reported in Table 2. Starting with transcripts for the query and abstracts for the candidate papers (keys), we incrementally add title and year information on both sides. We find that including these additional features consistently improves performance, and adopt the final configuration for subsequent experiments. However, this zero-shot model is still surpassed by the simple but strong baseline of always predicting the most frequently cited papers in the dataset (first row). Results on more input combinations and higher k for recall can be found in Table 6 in Appendix C.

Contextualizing Task Difficulty: Spoken vs. Textual Content We evaluate an alternative setting in which the long-form, noisy transcript is replaced by the abstract of the talk's corresponding paper (Table 2). While this is an unrealistic scenario, since talks often do not provide abstracts, it helps contextualize the relative difficulty of retrieving references from spoken versus textual content. We find that using abstracts instead of transcripts significantly improves performance, showing that state-of-the-art zero-shot models struggle with transcripts as input. This underscores the need for our newly introduced Talk2Ref dataset.

Evaluating different SOTA models. Next, we explore which encoder model is most suitable for our task and can be used for subsequent training, avoiding the need to train all models for efficiency reasons. These results can be found in the top lines in Table 3. Unsurprisingly, a clear difference can be observed between models specifically trained to produce meaningful representations for scientific documents, such as SPECTER2 and SBERT, and general-purpose language models. Among the specialized models, SBERT outperforms SPECTER2 and is therefore chosen for the remaining experiments. We also include Longformer due to its ability to handle longer input sequences, eliminating the need for truncation or aggregation to fit the transcript. While its out-of-the-box performance is lower, we expect it to benefit from fine-tuning on our retrieval task.

Results for higher recall k can be found in Table 7 in Section 9, these do not change the ranking of the models.

Performance of finetuned models. We now finetune SBERT and Longformer on Talk2Ref. Since talk transcripts exceed SBERT's input limits, we apply chunking and aggregation strategies on the query side as detailed in Section 4.1.1: truncation, mean pooling, and a learnable weighted mean. On the key side (paper abstracts, titles, and years), no aggregation is required due to the shorter lengths.

For Longformer, by contrast, the full transcript fits within the model's context window, so no chunking or aggregation is needed on the query side. However, on the key side, we continue to use SBERT as

the encoder, since it provides substantially stronger representations for abstracts than Longformer as discussed above.

Results for these trained models are shown in the middle section of Table 3. Unsurprisingly, training on TALK2REF substantially improves performance for both SBERT and Longformer, with SBERT now outperforming the top-k most frequent baseline. Among the different aggregation strategies, the learned weighted mean performs best for SBERT, surpassing truncation and simple mean pooling.

For Longformer, finetuning yields a substantial improvement, though the model still performs slightly worse than zero-shot SBERT, likely because SBERT has been explicitly trained for representation learning, whereas Longformer has not. We additionally experimented with splitting the input and applying aggregation mechanisms for Longformer as well; however, these configurations did not lead to further gains. The corresponding results are reported in Table 8 in Appendix C.

Effect of the domain adaptation stage. Finally, we add a task-specific domain adaptation stage to our models before finetuning, as described in Section 4.1.2. This stage serves two purposes: it provides a simpler learning task and helps the model adapt to the scientific domain. We include two configurations in our evaluation: SBERT (learned mean), the best-performing model after fine-tuning, and Longformer/SBERT, which shows significant gains from fine-tuning.

The bottom rows of Table 3 show that domain adaptation further improves performance, with only small gains for Longformer/SBERT but significant improvements for SBERT (learned mean). In comparison to SBERT's finetuned model with learned mean (but without domain adaptation stage), it gains improvements across all metrics. In fact, it achieves the highest scores across all metrics except MAP@10, where the frequency-based baseline remains superior. These results indicate that the adaptation stage enhances the model's ability to capture semantic relevance beyond surface-level similarity.

4.3. Discussion

Challenges of TALK2REF. The TALK2REF dataset presents several challenges that make reference prediction from talks a non-trivial task. First, reference prediction from talks is significantly harder than from papers or their abstracts, as shown in Table 2. Second, transcripts are long and often exceed the input window of standard encoders, making the production of fixed-size embeddings challenging. Aggregation strategies, such as mean or weighted pooling over chunks, help capture information from longer inputs, as illustrated in Table 3, though they provide only an approximate representation.

Our experiments demonstrate that finetuning on TALK2REF substantially improves retrieval performance, showing that the dataset provides rich and meaningful training signals. The results also highlight the benefits of both aggregation strategies and the domain adaptation stage, confirming that TALK2REF supports effective learning for the task of reference prediction from scientific talks.

Comparison to prior work. Prior work on citation recommendation, such as Bhagavatula et al. (2018), achieves competitive results using clean, structured text (titles and abstracts) on both the query and key sides. For example, their NNSelect model reaches $P@10 = 28.7$, $P@20 = 23.0$, and $R@20 = 36.3$ on the DBLP dataset (Tang et al., 2008). In comparison, our best finetuned model on *TALK2REF* achieves $P@10 = 19.1$, $P@20 = 15.3$, and $R@20 = 12.7$. While these numbers are lower, they remain in a comparable range, highlighting that our task, predicting references from long, noisy spoken-language transcripts, is substantially more challenging. Nonetheless, the results demonstrate that our dual-encoder models can still produce reasonable and meaningful predictions in this difficult setting.

5. Conclusion

This work introduces the task of Reference Prediction from Talks (RPT). RPT is practically relevant, as researchers, lecturers, and students could benefit from automated recommendations of related works for scientific talks. At the same time, it is highly challenging, requiring models to map long, noisy, and unstructured spoken content to the corresponding relevant papers.

To support this task, we present TALK2REF, the first large-scale benchmark for RPT. To our knowledge, it is also the first dataset to incorporate the spoken modality on the query side of a citation prediction task.

Despite the challenges, our experiments show that RPT is tractable. By training dual-encoder architectures inspired by dense passage retrieval, we demonstrate that models can effectively learn semantic representations aligning scientific talks with their cited references. Finetuning these architectures on TALK2REF significantly improves performance over zero-shot and heuristic baselines, confirming that dedicated training is essential for this modality.

We will release the dataset and the trained models under an open license, providing the community with ready-to-use tools for this task. We hope that

TALK2REF and the accompanying models will serve as a foundation for further research at the intersection of speech, language, and scholarly retrieval.

Future work may explore audio-based representations of talks, multimodal fusion of speech and text, and the development of encoders capable of handling longer input sequences, improving the ability to accurately predict relevant citations from spoken scientific content.

6. Ethical Considerations

The TALK2REF dataset is constructed from publicly available scientific talks and papers, and does not include any private or sensitive information. As such, we do not anticipate significant ethical risks associated with its release. All content is already in the public domain and intended for scholarly use.

Potential considerations include:

- **Biases in scientific citations:** Like all citation datasets, TALK2REF reflects the citation practices of the underlying field, which may underrepresent certain authors, institutions, or geographic regions. Models trained on this data may inadvertently perpetuate these biases.
- **Misuse for evaluation:** the dataset is intended for research on reference prediction and related tasks. Misuse for ranking or evaluating researchers or institutions should be avoided.

Overall, we consider TALK2REF suitable for research use, while encouraging users to remain aware of the inherent limitations and potential biases of citation-based datasets.

7. Acknowledgements

Part of this work received support from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People).

8. Bibliographical References

Zafar Ali, Guilin Qi, Khan Muhammad, Pavlos Kefalas, and Shah Khusro. 2021a. Global citation recommendation employing generative adversarial network. *Expert Syst. Appl.*, 180(C).

Zafar Ali, Guilin Qi, Khan Muhammad, Asim Khalil, Inam Ullah, and Amin Khan. 2021b. Global citation recommendation employing multi-view heterogeneous network embedding. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6.

Frederick Ayala-Gómez, Bálint Daróczy, András Benczúr, Michael Mathioudakis, Aristides Gionis, David Pinto, Vivek Kumar Singh, Aline Villavicencio, Philipp Mayr-Schlegel, and Efstathios Stamatatos. 2018. Global citation recommendation using knowledge graphs. *J. Intell. Fuzzy Syst.*, 34(5):3089–3100.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Arthur Brack, Anett Hoppe, and Ralph Ewerth. 2021. Citation recommendation for research papers via knowledge graphs. In *Linking Theory and Practice of Digital Libraries - 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13-17, 2021, Proceedings*, volume 12866 of *Lecture Notes in Computer Science*, pages 165–174. Springer.

Balasubramanian Chokkalingam. 2024. Improving citation recommendation accuracy using an svm-based model. *Journal of Information Systems Engineering and Management*, 2025:2468–4376.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. 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.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations.

GROBID Developers. 2008. Grobid. https://github.com/kermitt2/grobid.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022a. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022b. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 274–288, Berlin, Heidelberg. Springer-Verlag.

Jeff Johnson, Matthijs Douze, and Herve Jegou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Weijuan Li. 2024. Scientific paper recommender system using deep learning and link prediction in citation network. *Heliyon*, 10.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Vaios Stergiopoulos, Michael Vassilakopoulos, Eleni Tousidou, and Antonio Corral. 2024. An academic recommender system on large citation data based on clustering, graph modeling and deep learning. *Knowledge and Information Systems*, 66:4463–4496.

Xiaojuan Zhang, Shuqi Song, and Yuping Xiong. 2024. Personalized global citation recommendation with diversification awareness. *Scientometrics*, 129(7):3625–3657.

Yang Zhang and Qiang Ma. 2020a. Dual attention model for citation recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3179–3189, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yang Zhang and Qiang Ma. 2020b. Dual attention model for citation recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3179–3189. International Committee on Computational Linguistics.

Ege Yiğit Çelik and Selma Tekir. 2025. Citebart: Learning to generate citations for local citation recommendation.

9. Language Resource References

Chandra Bhagavatula and Sergey Feldman and Russell Power and Waleed Ammar. 2018. *Content-Based Citation Recommendation*. Association for Computational Linguistics. PID https://doi.org/10.18653/v1/n18-1022.

Arthur Brack and Daniel Uwe Müller and Anett Hoppe and Ralph Ewerth. 2021. *Coreference Resolution in Research Papers from Multiple Domains*. Springer. PID https://doi.org/10.1007/978-3-030-72113-8_6.

Huang, Wenyi and Zhaohui Wu and Mitra, Prasenjit and Giles, C. Lee. 2014. *RefSeer: A citation recommendation system*. Pennsylvania State University, IEEE/ACM Joint Conference on Digital Libraries. PID https://api.semanticscholar.org/CorpusID:951281.

Chanwoo Jeong and Sion Jang and Hyuna Shin and Eunjeong L. Park and Sungchul Choi. 2019. *A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks*. arXiv repository.

Rodney Michael Kinney and Chloe Anastasiades and Russell Authur and Iz Beltagy and Jonathan

Bragg and Alexandra Buraczynski and Isabel Cachola and Stefan Candra and Yoganand Chandrasekhar and Arman Cohan and Miles Crawford and Doug Downey and Jason Dunkelberger and Oren Etzioni and Rob Evans and Sergey Feldman and Joseph Gorney and David W. Graham and F.Q. Hu and Regan Huff and Daniel King and Sebastian Kohlmeier and Bailey Kuehl and Michael Langan and Daniel Lin and Haokun Liu and Kyle Lo and Jaron Lochner and Kelsey MacMillan and Tyler C. Murray and Christopher Newell and Smita R Rao and Shaurya Rohatgi and Paul Sayre and Zejiang Shen and Amanpreet Singh and Luca Soldaini and Shivashankar Subramanian and A. Tanaka and Alex D Wade and Linda M. Wagner and Lucy Lu Wang and Christopher Wilhelm and Caroline Wu and Jiangjiang Yang and Angele Zamarron and Madeleine van Zuylen and Daniel S. Weld. 2023. *The Semantic Scholar Open Data Platform*. ArXiv. PID https://api.semanticscholar.org/CorpusID:256194545.

Li, Huajing and Councill, Isaac and Lee, Wang-Chien and Giles, C. Lee. 2006. *CiteSeerx: an architecture and web service design for an academic document search engine*. Association for Computing Machinery, WWW '06. PID https://doi.org/10.1145/1135777.1135926.

Radev, Dragomir R. and Muthukrishnan, Pradeep and Qazvinian, Vahed. 2009. *The ACL Anthology Network Corpus*. Association for Computational Linguistics. PID https://aclanthology.org/.

Shaurya Rohatgi and Yanxia Qin and Benjamin Aw and Niranjana Unnithan and Min-Yen Kan. 2022. *The ACL OCL Corpus: advancing Open science in Computational Linguistics*. arXiv. PID https://huggingface.co/datasets/ACL-OCL/ACL-OCL-Corpus.

Sinha, Arnab and Shen, Zhihong and Song, Yang and Ma, Hao and Eide, Darrin and Hsu, Bo-June (Paul) and Wang, Kuansan. 2015. *An Overview of Microsoft Academic Service (MAS) and Applications*. Association for Computing Machinery, WWW '15 Companion. PID https://doi.org/10.1145/2740908.2742839.

Tang, Jie and Zhang, Jing and Yao, Limin and Li, Juanzi and Zhang, Li and Su, Zhong. 2008. *ArnetMiner: extraction and mining of academic social networks*. Association for Computing Machinery, KDD '08. PID https://doi.org/10.1145/1401890.1402008.

Maike Züfle and Sara Papi and Beatrice Savoldi and Marco Gaido and Luisa Bentivogli and Jan Niehues. 2025. *NUTSHELL: A Dataset for Abstract Generation from Scientific Talks*. arXiv repository. PID https://huggingface.co/datasets/maikezu/nutshell.

A. Dataset Statistics

The temporal coverage of references is concentrated in recent years: the majority of all cited works were published between 2015 and 2022. A clear growth trend is visible, with particularly sharp increases in references from 2018 onwards, mirroring the surge of research activity in natural language processing, with newer publications dominating the dataset as cited papers. This temporal skew ensures alignment with current research but naturally reduces representation of older foundational works. The distribution over years is shown in Fig. 2. The ten most cited papers in Talk2Ref are shown in Fig. 3.

B. Encoder Models

Detailed information to the encoder models used in this work can be found in Table 4. Hyperparmaters used to train these models are given in Table 5.

Model	HF-ID	Ref.	# Params	Transformer Version
SBERT	`sentence-transformers/all-MiniLM-L6-v2`	Reimers and Gurevych (2019)	22.7M	4.51.3
Longformer	`allenai/longformer-base-4096`	Beltagy et al. (2020)	148.8M	4.51.3
Specter	`allenai/specter2_base`	Cohan et al. (2020)	110M	4.51.3
BERT	`bert-base-uncased`	Devlin et al. (2019)	110M	4.51.3

Table 4: Encoder models used in our experiments. All models use Transformers v4.51.3. For zero-shot experiments, the same encoder is used for both query and key encoding without any aggregation or projection layers. During training, SBERT (Reimers and Gurevych, 2019) or Longformer (Beltagy et al., 2020) serve as query encoders, while the key encoder remains fixed to SBERT. A linear projection layer adds approximately 0.3M parameters when applied, and a learned weighted mean aggregation layer adds about 0.04M (SBERT) or 0.15M (Longformer) parameters on the query side.

C. Results and Discussion

This section provides an extended overview of the experimental results and additional analyses that complement the findings presented in the main paper.

C.1. Input representations

To determine the most effective input configuration, we evaluate all combinations of query and key inputs using SBERT (Reimers and Gurevych, 2019) as encoder (Table 6). On the key and query side, the combination of *abstract, title, and year* yields the best results across nearly all metrics and is therefore fixed for the remaining experiments.

However, the non-realistic scenario of using text abstracts instead of transcripts, remain markedly stronger: Abstract+Title+Year outperforms Transcript+Title+Year by $\sim$5.43 P@10 points (18.72 vs. 13.29) and $\sim$5.44 R@100 points (25.43 vs. 19.99). This gap is consistent with transcripts being long, noisy, and truncated to 512 tokens, whereas abstracts are concise, well-structured summaries. Since our goal is citation recommendation from the talk, we treat abstract-based results as a strong upper-bound baseline and focus modeling efforts on transcript-based inputs because talks rarely provide abstracts or summaries.

Parameter		
Query Encoder	SBERT	Longformer
Key Encoder	SBERT	SBERT
Batch size	24	3
Max. Epochs	100	100
Freeze layers (Key side)	2	4
Freeze layers (Query side)	2	8
Gradient accumulation	3	3
Learning rate (base)	6e-6	6e-6
Learning rate (head)	2e-4	2e-4
Weight decay	0.01	0.01
Dropout rate	0.05	0.05
Adam epsilon	1e-8	1e-8
Avg. inference time (s / example)	0.0083	0.0286
Early Stopping in Epochs	4	4

Table 5: Training configurations for finetuning SBERT (Reimers and Gurevych, 2019) and Longformer(Beltagy et al., 2020) on the Talk2Ref dataset.

C.2. Model Selection

Having fixed the input representation, we next compare encoder architectures to identify the most suitable model for subsequent training (Table 7). Across all configurations, SBERT (Reimers and Gurevych, 2019) consistently outperforms other encoders by a clear margin, achieving the highest precision and recall values for both abstract and transcript inputs (e.g., P@10 = 18.70 vs. 15.42 for SPECTER2(Cohan et al., 2020) and 13.29 for transcript-based SBERT). SPECTER2 performs second-best, followed by Longformer (Beltagy et al., 2020) and BERT (Devlin et al., 2019), which struggle under the truncated input. Based on these results, we fix SBERT with the full input configuration (transcript + title + year on the query side; title + year + abstract on the key side) for all subsequent training experiments to ensure efficiency and consistency.

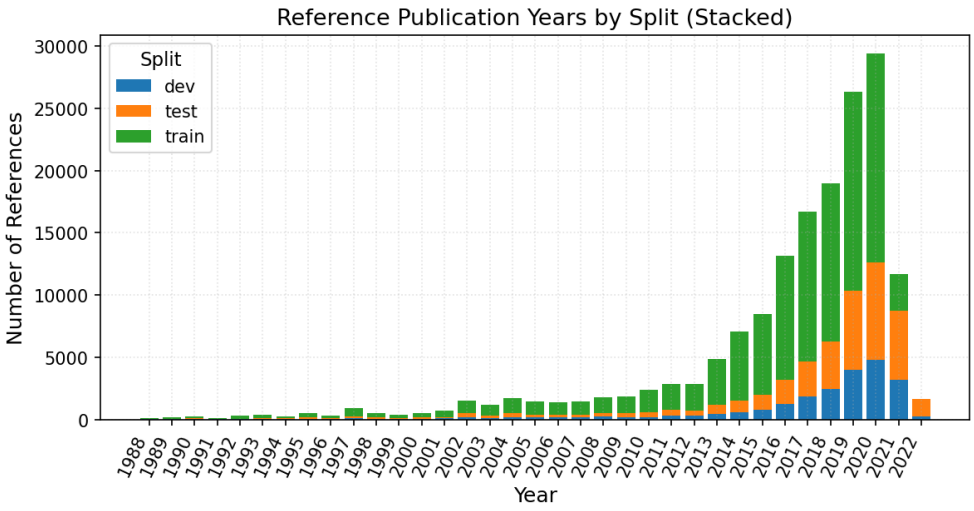

Figure 2: Temporal distribution of cited works and abstracts in the dataset. The majority of references are concentrated between 2015 and 2022, with a marked increase from 2018 onward, reflecting the surge of research in natural language processing. This distribution ensures alignment with current research trends but underrepresents older foundational work.

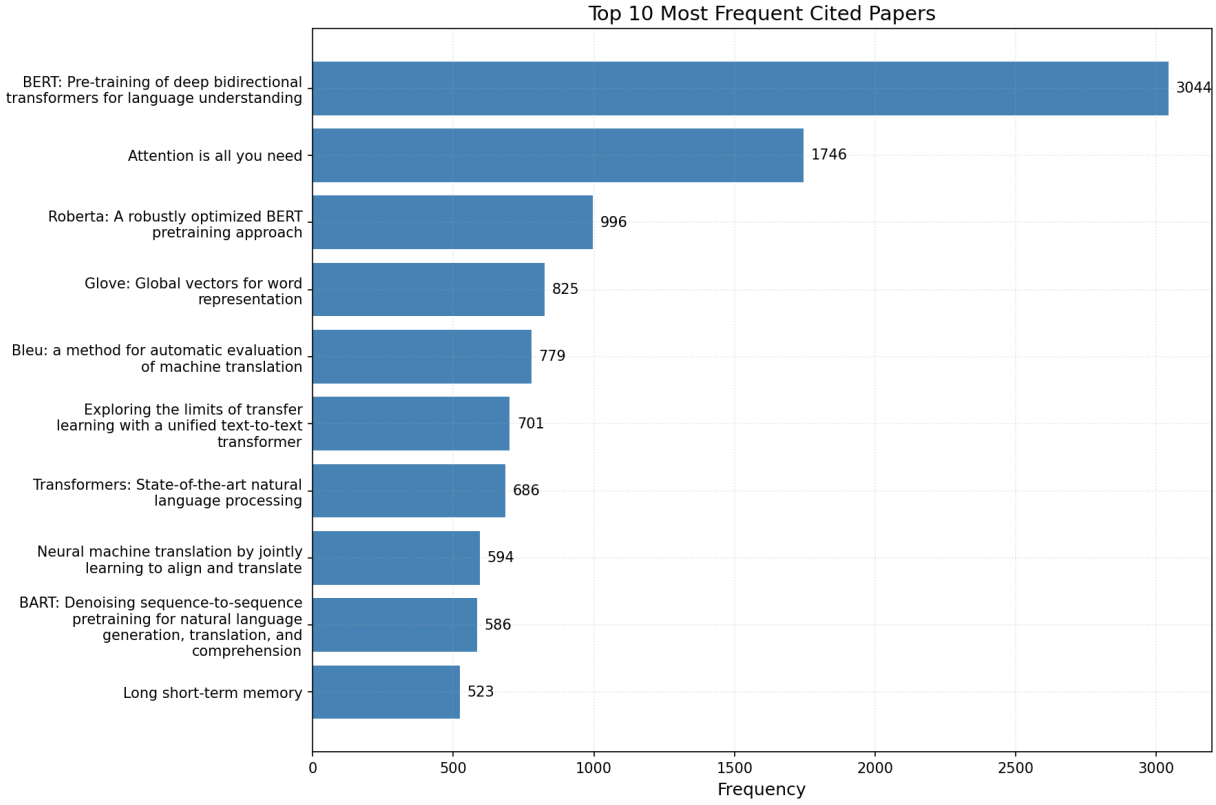

Figure 3: Top 10 most frequently cited papers in the dataset.

C.3. Finetuned Models

After identifying SBERT (Reimers and Gurevych, 2019) as the most effective encoder and fixing the input representation (*Transcript* + *Title* + *Year* on the query side, *Title* + *Year* + *Abstract* on the key side), we trained dual-encoder models under different regimes: on training task, on the domain adaptation stage, and with subsequent training on domain adapted checkpoints. We also include Longformer (Beltagy et al., 2020) due to its capacity for longer sequences; although its out-of-the-box performance is lower, we expect it to benefit from fine-tuning on our retrieval task. Results are summarized in Table 8.

When finetuning, SBERT-based models consis-

Query Input	Key Input	Prec.	P@ 10	P@ 20	R@ 20	R@ 50	R@ 100	R@ 200	MAP@ 10	MAP@ 20
N/A	most cited papers	10.33	17.48	11.64	9.82	13.82	17.25	21.83	**12.05**	7.39
Transcript	Abstract	8.89	12.14	9.65	7.79	13.18	18.87	25.49	6.63	4.49
Transcript	Title	8.32	11.40	8.94	7.33	12.30	17.41	23.70	6.10	4.10
Transcript	Title + Abstract	9.32	12.82	10.02	8.16	13.74	19.56	26.08	7.12	4.77
Transcript	Abstract + Year	8.87	12.12	9.59	7.74	13.12	18.77	25.36	6.68	4.50
Transcript	Title + Year	7.61	10.60	8.22	6.78	11.22	15.95	21.72	5.46	3.60
Transcript	Title + Year + Abstract	9.32	12.79	10.07	8.21	13.76	19.46	26.10	7.12	4.78
Transcript + Title	Title + Year + Abstract	9.55	**13.32**	10.39	8.58	14.13	19.87	26.65	7.15	4.83
Transcript + Year	Title + Year + Abstract	9.45	13.07	10.27	8.42	14.04	19.81	26.60	7.25	4.86
Transcript + Title + Year	Title + Year + Abstract	**9.57**	13.29	**10.42**	**8.59**	**14.21**	**19.99**	**26.72**	**7.33**	**4.94**
Title	Title + Year + Abstract	9.95	13.56	10.69	9.09	14.60	19.92	26.56	7.52	5.21
Title + Year	Title + Year + Abstract	9.68	12.92	8.81	10.28	14.22	19.50	25.99	7.19	5.02
Abstract	Title + Year + Abstract	**13.19**	**19.02**	**14.54**	**12.05**	**18.99**	**25.75**	**33.06**	**11.05**	**7.50**
Abstract + Title	Title + Year + Abstract	13.06	18.73	14.40	11.96	18.83	25.61	32.80	10.95	7.45
Abstract + Title + Year	Title + Year + Abstract	13.03	18.70	14.30	11.86	18.82	25.43	32.67	10.88	7.39

Table 6: Results on the Talk2Ref dataset using zero-shot SBERT, truncating inputs to SBERT's maximum sequence length. We experiment with different features for both query talks and candidate papers (keys). Using abstracts on the key side is an unrealistic setting, as abstracts are not provided with the talk; we include them to contextualize the difficulty of the task.

Strategy	Encoder	Max Seq. Len	Aggregation	Prec.	P@ 10	P@ 20	R@ 20	R@ 50	R@ 100	R@ 200	MAP@ 10	MAP@ 20
Top-k most freq.	–	–	–	10.33	17.48	11.64	9.82	13.82	17.25	21.83	12.05	7.39
Zero-Shot	SBERT	512	Truncated	**9.57**	**13.29**	**10.42**	**8.59**	**14.21**	**19.99**	**26.72**	**7.33**	**4.94**
	Specter2	512	Truncated	8.31	11.78	9.10	7.37	12.43	17.39	23.81	6.15	4.09
	Longformer	4096	–	0.17	0.21	0.20	0.16	0.31	0.49	0.97	0.10	0.06
	BERT	512	Truncated	0.56	0.67	0.57	0.49	0.82	1.43	2.32	0.24	0.17

Table 7: Results on the Talk2Ref dataset for zero-shot retrieval. Query inputs consist of transcript, title, and year, while cited papers (keys) are represented by title, year, and abstract.

tently outperform all baselines. Among aggregation strategies, the *learned weighted mean* yields the strongest results ($P@20 = 15.12$, $R@200 = 38.21$), clearly surpassing both simple mean pooling and truncation. This indicates that weighting informative segments within long transcripts allows the model to better capture semantic relations between talks and their cited papers. Compared to the most-cited baseline ($R@200 = 21.82$), recall improves by more than 70%, while precision gains remain moderate—suggesting that the model retrieves semantically related papers even when not all are explicitly cited.

Finetuning on the continual pretrained SBERT (Reimers and Gurevych, 2019) checkpoint further improves results, with the learned weighted mean configuration achieving the overall best performance ($P@20 = 15.31$, $R@200 = 39.04$). This confirms that the domain adaptation stage facilitates better representation learning for long-form transcripts and that finetuning adapts these representations effectively to the citation retrieval task.

Models using the Longformer (Beltagy et al., 2020) encoder perform notably worse, limited by GPU memory constraints and small batch sizes (3 vs. 24 for SBERT). Mean aggregation slightly outperforms the learned weighted mean, implying that the model struggles to learn attention weights effectively across extended contexts. Although recall at larger cutoffs (e.g., $R@200 \approx 29.82$) exceeds the frequency baseline, Precision and MAP values remain significantly lower. These results suggest that, despite its architectural suitability for long sequences, Longformer (Beltagy et al., 2020) requires larger datasets and more compute to fully leverage its capacity.

Strategy	Query Enc.	Key Enc.	Max Seq.	Aggregation	Prec.	P@ 10	P@ 20	R@ 20	R@ 50	R@ 100	R@ 200	MAP@ 10	MAP@ 20
Top-k most freq.	–	–	–	10-most cited papers	10.33	17.48	11.64	9.82	13.82	17.25	21.83	12.05	7.39
Zero-shot	SBERT	SBERT	512	Truncated	9.57	13.29	10.42	8.59	14.21	19.99	26.72	7.33	4.94
Zero-shot	Longformer	Longformer	4096	–	0.20	0.23	0.21	0.18	0.34	0.59	0.109	0.08	0.05
Finetuned	SBERT	SBERT	512	Truncated	12.94	17.23	13.85	11.49	19.59	27.34	35.88	9.76	6.91
	SBERT	SBERT	512	Learned mean	**13.95**	**18.81**	**15.12**	**12.59**	**20.97**	**29.18**	**38.21**	**10.86**	**7.69**
	SBERT	SBERT	512	Mean	13.34	17.75	14.28	11.92	20.35	28.01	36.65	10.13	7.15
Finetuned	Longformer	SBERT	1024	Truncated	8.89	11.19	9.53	7.67	14.56	22.06	31.65	**5.35**	**3.73**
	Longformer	SBERT	2048	Truncated	**8.95**	**11.20**	**9.68**	**7.74**	**14.58**	**22.21**	**31.72**	5.32	3.72
	Longformer	SBERT	4096	–	8.47	10.75	9.15	7.32	14.06	21.41	30.68	5.12	3.57
	Longformer	SBERT	512	Mean	8.35	10.37	8.90	7.16	13.70	21.04	29.82	4.57	3.27
	Longformer	SBERT	1024	Mean	8.14	10.18	8.79	7.00	13.34	20.44	29.32	4.61	3.27
	Longformer	SBERT	512	Learned mean	7.70	9.59	8.19	6.57	12.78	19.89	28.86	4.45	3.08
Domain Adapt.	SBERT	SBERT	512	Truncated	**11.36**	**16.06**	**12.36**	**10.37**	**16.52**	**22.41**	**29.30**	**9.04**	**6.18**
	SBERT	SBERT	512	Learned mean	8.79	12.34	9.67	7.90	12.76	17.93	24.03	6.65	4.49
Domain Adapt. + Finetuned	SBERT	SBERT	512	Learned mean	**14.18**	**19.14**	**15.32**	**12.71**	**21.60**	**29.92**	**39.04**	10.98	**7.72**

Table 8: Results on the Talk2Ref dataset. Query inputs consist of transcript, title, and year, while cited papers (keys) are represented by title, year, and abstract. We report different aggregation mechanisms for handling long query transcripts and show results for finetuning the best zero-shot model on Talk2Ref.