ABCD-LINK: Annotation Bootstrapping for Cross-Document Fine-Grained Links

Serwar Basch[1], Ilia Kuznetsov[1], Tom Hope[2], Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI), TU Darmstadt
[2] Hebrew University of Jerusalem and The Allen Institute for AI (AI2)
www.ukp.tu-darmstadt.de

Abstract

Understanding fine-grained links between documents is crucial for many applications, yet progress is limited by the lack of efficient methods for data curation. To address this limitation, we introduce a domain-agnostic framework for bootstrapping sentence-level cross-document links from scratch. Our approach (1) generates and validates semi-synthetic datasets of linked documents, (2) uses these datasets to benchmark and shortlist the best-performing linking approaches, and (3) applies the shortlisted methods in large-scale human-in-the-loop annotation of natural text pairs. We apply the framework in two distinct domains – peer review and news – and show that combining retrieval models with LLMs achieves a 73% human approval rate for suggested links, more than doubling the acceptance of strong retrievers alone. Our framework allows users to produce novel datasets that enable systematic study of cross-document understanding, supporting downstream tasks such as media framing analysis and peer review assessment. All code, data, and annotation protocols are released to facilitate future research.[1]

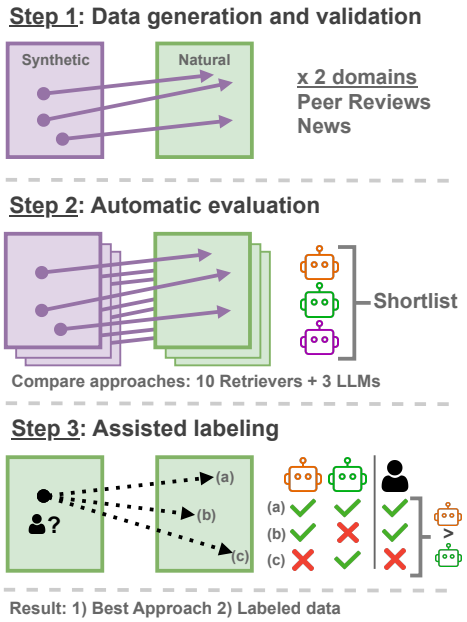


Figure 1: Framework overview.

1 Introduction

Documents rarely exist in isolation. In many practical scenarios, understanding one document requires reasoning over its relationships with others. Fact-checkers might trace a claim to the specific sentences across multiple articles that support or contradict it (Thorne et al., 2018; Wadden et al., 2020; Chen et al., 2024b). Authors and meta-reviewers reading peer reviews need to connect reviewer comments to the paper to decide on the course of action (Kuznetsov et al., 2022; D'Arcy et al., 2024). News analysts might use paraphrased or ideologically reframed sentences across different outlets to gain insights into reporting bias and source alignment (Giorgi et al., 2023). Yet finding such fine-grained relations can be cognitively demanding, motivating the need for machine assistance. While tasks like cross-document coreference resolution offer some support, cross-document relations are much more diverse, necessitating the development of general-purpose approaches to cross-document analysis.

Following Kuznetsov et al. (2022), we call fine-grained relations between documents *links*, and the general task of identifying them *linking*. While automatic linking holds great potential, the progress has been limited due to the manual effort required to annotate and evaluate links, and the diversity of link types across domains. Thus, labeled corpora of links are scarce, preventing the development and evaluation of automated linking approaches at scale.

To move beyond the bottleneck of full manual annotation, we propose a general framework for bootstrapping sentence-level cross-document link-

[1]https://github.com/UKPLab/eacl2026-abcd-link

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3399–3423

ing without the need for pre-existing labeled data (Figure 1). While links might be hard for humans to *detect*, linked documents are easier to *generate* with state of the art LLMs. Building upon this insight, we (1) propose a method to generate and validate semi-synthetic corpora of linked documents. We use this data to (2) automatically evaluate a wide range of approaches to arrive at a shortlist of best-performing linking approaches. Finally, we use the best-performing approaches to (3) assist humans in labeling natural, non-synthetic document pairs in a large-scale human-in-the-loop evaluation study. Our framework allows practitioners to efficiently select the best linking approach given a domain and link type, while producing manually labeled many-to-many cross-document linking datasets for downstream applications.

We assess our framework by applying it in two distinct domains, peer reviews and news articles, each with its own task-specific link type and underlying data. In total, we contribute:

- A novel framework for efficient human-in-the-loop evaluation and labeling of links;
- A data generation approach for creating semi-synthetic datasets of linked documents in the domain of interest, incl. human validation;
- A large-scale analysis of state-of-the-art linking approaches, incl. a novel human evaluation protocol;
- Two new manually annotated non-synthetic corpora of cross-document links in peer review and news.

We release the code, datasets, and annotation protocols to support future research on sentence-level cross-document linking and its applications.

2 Related Work

Numerous NLP tasks study or rely on **cross-document relations**, including cross-document coreference resolution (CDCR) (Miyabe et al., 2008; Cybulska and Vossen, 2014; Ravenscroft et al., 2021), fact-checking (Thorne et al., 2018; Wadden et al., 2020; Chen et al., 2024b), scholarly peer review analysis (Kuznetsov et al., 2022; D'Arcy et al., 2024), citation/source detection (Syed et al., 2023; Liang et al., 2024), and cross-document question answering (Lin et al., 2025). While prior work focuses on limited relation types and relies on manual labeling, we contribute a general, structured framework for bootstrapping linking from scratch in new domains and for new link types. We evaluate our framework in two domains: peer reviews and news, which target challenging and underrepresented forms of linking, such as subjective evaluation and ideological framing, as opposed to encyclopedic and factual relations found in Wikipedia (Feith et al., 2024). Crucially, unlike CDCR where the main goal is to resolve and cluster entity and event mentions, linking focuses on detecting a broad spectrum of sentence-level relationships, such as paraphrases, quotes and implicit references.

From a **machine-assisted annotation** perspective, prior work reduces labeling effort by pre-selecting candidate links with heuristics (Ravenscroft et al., 2021), or simple retrieval methods, e.g. cosine similarity with a fixed cutoff (Kuznetsov et al., 2022). Our framework generalizes from this idea: we automatically benchmark various zero-shot approaches, and validate the results through human evaluation to select the best approach for the task and domain at hand. In that, our work contributes to the study of **synthetic data** in NLP (Ding et al., 2024; Liu et al., 2022; Zhao et al., 2025; Josifoski et al., 2023; Veselovsky et al., 2023). Møller et al. (2024) and Kazemi et al. (2025) find that synthetic data can match or complement human-labeled corpora in low-resource settings. To the best of our knowledge, we are the first to use synthetic data to study cross-document links. The use of synthetic data carries risks (Van Breugel et al., 2023; Lu et al., 2023). To mitigate them, our framework only uses synthetic data to automatically evaluate different linking approaches, employs a human validation step, and supplements evaluations on synthetic data with focused human evaluation, contributing to best practices in the use of synthetic data in NLP.

Human judgment is essential for evaluating NLP systems, yet difficult to elicit reliably. There is growing interest in structured **human evaluation protocols**, particularly for tasks involving model-assisted annotation or subjective decision-making. Prior work examined inter-annotator calibration over time (Uma et al., 2021), preference-based assessment of model outputs (Clark et al., 2021), and the influence of interface design on annotation quality (Klie et al., 2018). In the context of LLM evaluation, human-in-the-loop setups have proven essential for capturing fine-grained distinctions and improving label quality (Zhou et al., 2023; Min et al., 2025). Here, we contribute a novel and efficient annotation protocol that combines candidate pre-

selection with a ranking-based comparison for a reliable assessment of top-performing approaches.

3 Setup

We focus on *sentence-level linking* cast as a sentence pair classification task, representing a trade-off between task complexity and utility: a more narrow unit of analysis restricts the available context, while the links between broader units can be reconstructed by aggregating consecutive sentences.[2] Given two documents, the source document $A = \{a_1, \ldots, a_N\}$ and the target document $B = \{b_1, \ldots, b_M\}$, the goal is to predict for each cross-document sentence pair (a_i, b_j) whether a relation r holds between them. The definition accommodates many-to-many links, both directed (e.g. a review sentence criticizing a claim in a paper) and undirected (e.g. sentences in two new articles addressing the same fact). Linking can be seen as a special case of information retrieval, where a_i serves as a query to retrieve relevant sentences from B. However, linking is constrained by a task-specific relationship holding between sentences, and might require documents A and B for contextualizing the linking decisions.

There are many potential approaches to perform automatic linking in a zero-shot fashion, from simple relation-agnostic cosine similarity to classification with LLMs. Yet, labeled datasets of links are scarce, and *evaluating and comparing* the performance of different linking approaches on a wide range of link types and domains via human evaluation is infeasible. Our framework addresses this limitation by generating and validating semi-synthetic data (Section 4) to arrive at a shortlist of best-performing linking approaches (Section 5) which are then used in a focused human evaluation and annotation study (Section 6).

The framework is designed to be domain-agnostic in that it does not depend on any domain-specific characteristics like writing style, relationship type or the presence of explicit references such as citations. To demonstrate this, we apply the framework in two distinct practical scenarios: in `REVIEWS`, we link peer reviews to their papers, and in `NEWS`, we link pairs of news articles. While the framework itself aims to be domain-agnostic, the particular link types (and cutoff values, see Section 5) are flexible and defined per domain. In `REVIEWS`, we define a link as a sentence in a review that comments on, critiques, or supports a specific sentence in the corresponding paper. In `NEWS`, a link connects two sentences from different articles that convey the same or closely related factual content, typically reflect semantic equivalence or paraphrastic overlap, even if the framing or tone differs between the two sources.

We showcase our framework on two distinct practical scenarios: in `REVIEWS`, we link peer reviews to their papers, and in `NEWS`, we link pairs of news articles. In `REVIEWS`, we define a link as a sentence in a review that comments on, critiques, or supports a specific sentence in the corresponding paper. In `NEWS`, a link connects two sentences from different articles that convey the same or closely related factual content, typically reflect semantic equivalence or paraphrastic overlap, even if the framing or tone differs between the two sources.

4 Synthetic data

4.1 Generation

We hypothesize that while links in existing document pairs can be time-consuming to annotate, they are relatively easy to generate. Based on this intuition, we construct a semi-synthetic linking dataset for each of our target application domains using a non-synthetic, natural document as target, and prompting an LLM to generate a synthetic source document that links to the target on the sentence level. For this experiment, we use DeepSeek-R1 (DeepSeek-AI et al., 2025) due to its strong instruction-following performance and ability to handle long context. To avoid potential bias (Liu et al., 2024), we explicitly exclude DeepSeek-R1 from downstream experiments, and solely use it for synthetic data generation. While strictly speaking only the source documents in our data are synthetic, for simplicity, we further refer to the resulting data as our *synthetic data*.

`REVIEWS-SYNTH` builds upon the NLPeer dataset (Dycke et al., 2023), which includes EMNLP24 papers pre-segmented into sentences. To mitigate long-context issues (Levy et al., 2024) and have a comparable size to `REVIEW-F1000`, we select the 200 shortest papers by sentence count and prompt the model to generate a peer review for each. The prompt instructs the LLM to write review sentences and link them to specific sentences in the paper,

[2] While broader argumentative or narrative relations that span multiple consecutive sentences do exist, in practice, many-to-many relations naturally emerge as clusters of consecutive sentence-level links. We provide an exploratory analysis of such clusters in Appendix A

Stat	NEWS ECB+	NEWS SYNTH	REVIEWS SYNTH	REVIEWS F1000
Doc Pairs	2505	346	200	211
Number of Links	9383	2323	2181	1205
Avg. Sents (Src)	17.7	12.0	10.5	22.4
Avg. Sents (Tgt)	18.2	13.7	130.8	145.0
Avg. Links (Src)	2.06	4.97	5.14	4.84
Avg. Links (Tgt)	3.75	6.71	10.9	5.71
Tgt/Src Ratio	1.82	1.17	2.14	1.18
Src/Tgt Ratio	1.86	1.07	1.03	1.15

Table 1: Dataset statistics: number of document pairs, number of links, average number of sentences in source and target documents, average number of linked sentences in source and target document, and link density.

simulating naturally occurring reviewer comments.

To match the number of links in REVIEWS-SYNTH, for **NEWS-SYNTH** we sample 346 news articles from the WikinewsSum dataset (Calizzano et al., 2022). To reduce noise in the documents, we clean each article using GPT-4o-mini to remove scraping artifacts. We then segment the text into sentences, and prompt the LLM to produce a related article on the same topic, written with a different tone or editorial perspective, and instructing the model to link sentences in the generated article to sentences in the original article, resulting in cross-document links. This simulates a use case in media analysis where the same events reported by different outlets are compared to detect bias and framing. See Appendix P for more details, prompt templates and example links for both datasets.

To contextualize our synthetic datasets and experimental results, we derive two additional datasets from existing cross-document corpora. **NEWS-ECB+** is derived from the ECB+ corpus (Cybulska and Vossen, 2014) designed for cross-document coreference resolution. We repurpose this dataset for sentence-level linking by aligning sentence pairs across documents that share event mentions in the same cluster. **REVIEWS-F1000** leverages the F1000RD dataset introduced by Kuznetsov et al. (2022), which links peer review sentences to their corresponding paper sentences. Although its annotation design restricts the choices of possible links, it is the most structurally similar resource to our task. Dataset conversion details are provided in Appendix B. Table 1 provides key statistics on the synthetic and converted datasets.

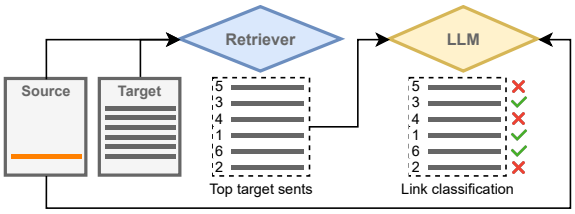


Figure 2: R+LLM Setup.

4.2 Validation

We ensure the quality of the synthetic data, and explore potential biases and hallucinations in machine-generated data, through automatic and manual inspection. At the document level, we compare lexical diversity, subjectivity, and complexity (Flesch-Kincaid score (Flesch, 1948)). In NEWS-SYNTH, synthetic articles are more lexically diverse, similarly subjective, and slightly more lexically complex than naturally occurring articles. In REVIEWS-SYNTH, synthetic reviews are less diverse but more complex, reflecting LLMs' preference for concise, formal phrasing. Subjectivity remains comparable across domains. Overall, synthetic documents align well with natural ones in most stylistic dimensions (more details in Appendix C)

To further assess the quality of the synthetic data, we conduct a human validation study. For NEWS-SYNTH, we randomly selected 30 natural news articles and their corresponding synthetic counterparts. Four annotators rated each article on a five-point Likert scale for fluency, coherence, realism (i.e., could the article be written by a human), and specificity (i.e., does the article contain specific facts). Synthetic articles match fluency/coherence but score lower on realism (as shown in Figure 14a in the Appendix) which we attribute to the models limitation in generating realistic sounding news articles. The synthetic data also shows lower specificity, likely due to LLMs producing more general statements compared to natural articles. While these limitations may affect journalistic authenticity, they do not hinder our linking task where coherence is more important, because it provides better context to reason over.

For REVIEWS-SYNTH, we randomly selected 30 synthetic reviews and retrieved natural reviews for the same papers from the NLPeer dataset. Three NLP PhD students rated each review on fluency, coherence, helpfulness (i.e., would this review be helpful for the paper's authors), and specificity (i.e., does the review address concrete parts of the paper).

Model	NEWS-ECB+		NEWS-SYNTH		REVIEWS-SYNTH		REVIEWS-F1000		Overall
	Avg. F1	R@10	Avg. F1	R@10	Avg. F1	R@20	Avg. F1	R@20	Avg. F1
Sparse Models									
BM25	37.26	89.40	32.58	94.35	18.91	66.39	29.53	93.90	29.57
SPLADEv3	39.83	93.23	33.04	95.45	19.46	69.83	29.10	95.59	30.36
BGE-M3-SPARSE	38.81	92.21	34.52	96.49	19.73	70.08	29.68	95.38	30.68
Bi-Encoders									
SFR	43.62	98.01	36.28	97.8	19.20	75.93	25.56	92.05	31.17
all-mpnet-base	42.18	96.52	36.73	98.97	18.09	68.69	24.54	89.16	30.38
BGE-M3-DENSE	42.33	96.23	37.32	99.79	21.52	77.14	29.80	96.14	32.74
Contriever	41.17	95.73	33.17	97.31	17.30	66.69	26.14	93.45	29.45
Dragon+	42.42	96.72	36.99	99.17	21.11	73.50	31.51	97.08	**33.01**
Cross-Encoders									
BGE-M3-Reranker	41.39	96.74	36.83	100.0	19.52	77.11	27.22	95.74	31.24
ms-marco-MiniLM	41.88	96.27	36.71	98.76	21.26	72.97	31.88	96.65	<u>32.93</u>

Table 2: Performance of retrieval models across synthetic and converted datasets. Metrics are: average F1 score across all cutoffs ($k \in \{1, 3, 5, 7, 10, 20\}$) and recall at a fixed cutoff (R@10 for news, R@20 for reviews). The overall average F1 is computed across all datasets. Recall cutoffs are selected based on domain-specific document lengths to maximize retrieval coverage. Detailed per-dataset and per-cutoff results are provided in Appendix E.

Synthetic reviews were rated higher in fluency and coherence, and lower in helpfulness and specificity due to a lack of concrete suggestions and feedback. While not perfect, the results align with our goal of generating reviews that are coherent and linkable. See Appendix D for further details on the setup.

In summary, while LLM-generated linked document pairs are sometimes less specific or natural, they offer a practical, scalable way to create sentence-level links that can be used for automatic evaluation.

5 Automatic Evaluation

5.1 Approaches

Using our synthetic data, we benchmark a broad range of retrievers in a zero-shot setup to shortlist strong candidates for human evaluation. We test sparse models (BM25 (Robertson and Zaragoza, 2009), SPLADEv3 (Lassance et al., 2024), BGE-M3 (Chen et al., 2024a)), bi-encoders (all-mpnet-base-v2[3], SFR-Embedding-Mistral (Meng et al., 2024), BGE-M3, Contriever (Izacard et al., 2022), Dragon+ (Lin et al., 2023)), and cross-encoders (ms-marco-MiniLM-L6-v2[4], BGE-M3). Each retriever ranks candidate sentences in the target document based on cosine similarity to a sentence from the source document. We convert ranks into binary decisions via a threshold: all sentences that appear within the top-k are treated as links.

We then extend this setup with LLMs to refine candidate links. The best-performing retriever supplies the top-k target sentences. We choose the k based on the performance of the retrieval models, and the length of the documents, to maximize the ($k = 10$ for NEWS domain, $k = 20$ for REVIEWS domain). Next, the LLM classifies the source-target pairs as linked or not linked (Figure 2). We use LLMs for their zero-shot ability to adapt to different link types via prompting only, making them well-suited for domains where link semantics vary and labeled data for fine-tuning is unavailable. We further refer to this approach as R+LLM for brevity.

We investigate two prompting setups. **Pairwise**, where each source–target pair is judged independently, and **Listwise**, where all k candidates are considered jointly in a single pass. In each setup, we prompt the LLM with both whole documents in four configurations: with a link description, with in-context examples, with both, and with no guidance. For evaluation, we use two open-source models, Phi-4 (14B) (Abdin et al., 2024) and Qwen2.5 (32B) (Yang et al., 2024), selected for their SOTA performance within their respective size classes. We also include GPT-4o[5], as the SOTA closed-source non-reasoning model at the time of writing, to serve as a high-performance reference point. Prompt templates are given in Appendix I.

[3] sentence-transformers/all-mpnet-base-v2

[4] cross-encoder/ms-marco-MiniLM-L6-v2

[5] Version gpt-4o-2024-08-06

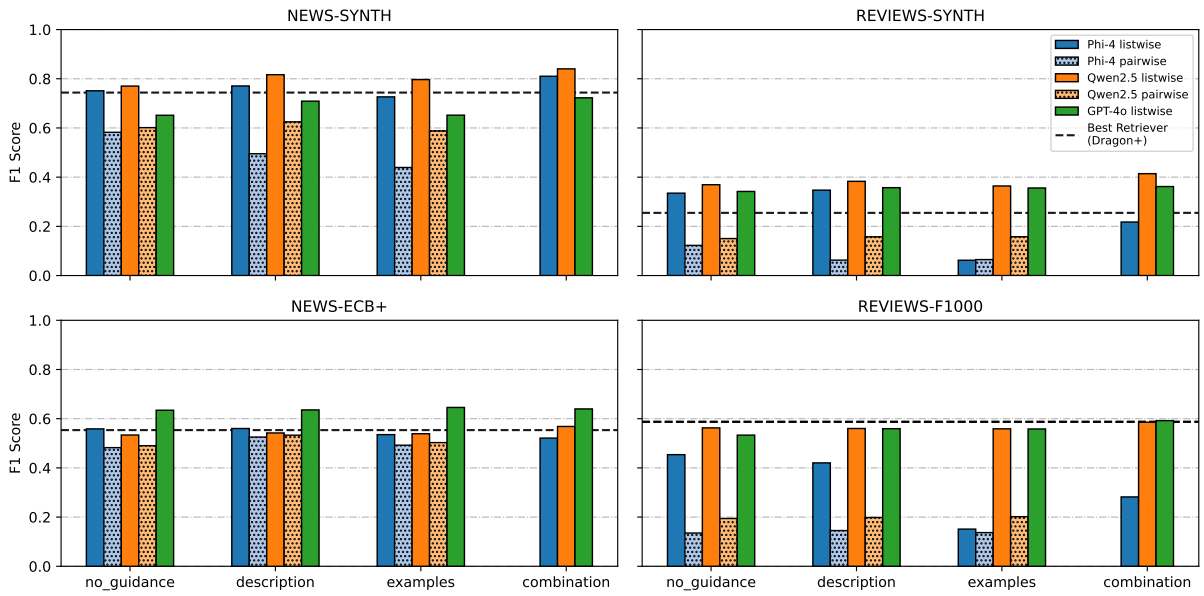


Figure 3: F1 scores across datasets using different prompt configurations and models (Phi-4, Qwen2.5, GPT-4o). Results show a consistent trend across domains with listwise prompting outperforming pairwise prompting. Furthermore, combination prompts (tested listwise only) achieve the highest F1 scores on synthetic datasets, while performance gaps narrow on the converted datasets. These results showcase the interaction between prompt design, model capabilities, and dataset complexity.

5.2 Results

`Retriever-only`. Table 2 reports performance across synthetic and converted datasets. Sparse models perform weakest overall, while dense models (`SFR`, `Dragon+`, `BGE-M3`) demonstrate stronger F1 scores across all datasets. Overall, `Dragon+` achieves the highest average F1 across datasets and cutoffs. This ranking is calculated by averaging the F1 scores computed at each cutoff ($k \in \{1, 3, 5, 7, 10, 20\}$) for each dataset, followed by averaging across all datasets. While the performance gap to the cross-encoders model is not large, due to their higher inference costs, we select `Dragon+` as the base retriever for subsequent experiments.

`R+LLM`. Building on `Dragon+`, we next experiment with using LLMs as classifiers to refine the results. Figure 3 shows the F1 scores of the different `R+LLM` setups. On the synthetic datasets (`NEWS-SYNTH` and `REVIEWS-SYNTH`), applying the LLM as a filter leads to consistent improvements, particularly for `REVIEWS-SYNTH`, which benefits from the LLM's ability to capture nuanced feedback and critique relations. On the converted datasets (`NEWS-ECB+` and `REVIEWS-F1000`), however, the gains are marginal. We attribute this to the mismatch between these datasets and our linking task: `NEWS-ECB+` focuses on coreference links, while `REVIEWS-F1000` is derived from `F1000RD` whose papers are on average longer than papers in `REVIEWS-SYNTH`, and include domains (biology, medicine) where key evidence appears in figures, limiting LLM effectiveness in our text-only setup. Significance tests (Appendix F) support these observations. Overall, LLMs add value across domains by capturing relations that are implicit or span multiple lines of reasoning. Thus, while both domains benefit from `R+LLM`, the nature of their linking tasks determines how large the performance gap over the baseline is. Manual error analysis reveals that most failures stem from scattered or ambiguous source sentences, which make fine-grained linking difficult (more details in Appendix G). Finally, to verify that the improvements of `R+LLM` are not specific to `Dragon+`, we also tested the pipeline with other retrievers and observed similar performance gains (see Appendix H for details).

Choice of LLM Across the three LLMs tested (`Phi-4`, `Qwen2.5`, and `GPT-4o`), we find that `Phi-4` performs on average worse than the others. `Qwen2.5` is the strongest overall, especially on the synthetic datasets, while `GPT-4o` slightly outperforms the others on `NEWS-ECB+`. We attribute this to properties of the dataset itself: `NEWS-ECB+` links are based on coreferent event mentions, which could

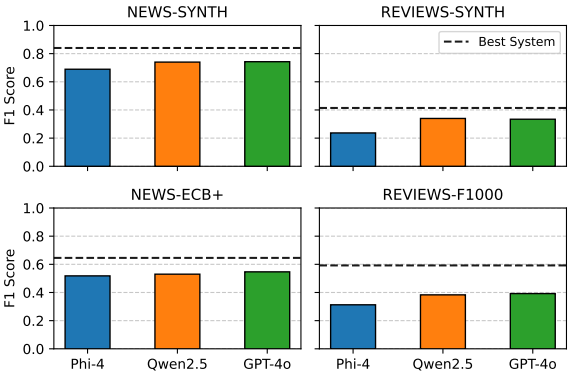


Figure 4: LLM-only ablation. Models were prompted with both full documents (source and target), the specific source sentence, a link description, and in-context examples. While this setup captures more context and task-specific information, it still underperforms compared to the combination of Retriever and LLM. This highlights the importance of retrieval for narrowing the candidate space and reducing distractors, especially in long documents.

be easier for stronger models like `GPT-4o` to detect. Despite these relative differences, all three models achieve only modest absolute F1 scores on the peer review datasets, highlighting the intrinsic difficulty of sentence-level linking in long, technical documents. Nevertheless, across all datasets, the best prompting configuration is clear: combining a link description (i.e., a description of what it means for two sentences to be related) with in-context examples leads to the strongest performance. In addition to the prompt content, we find that the way candidate sentences are presented to the LLM matters. In particular, *listwise prompting*, where all top-k candidate sentences are shown together, outperforms *pairwise prompting*, where the LLM considers each candidate in isolation. We hypothesize that listwise prompting allows the LLM to condition its decisions on the global context rather than making them in isolation.

LLM-only ablation. To evaluate the necessity of retrieval, we ablate the retriever and use LLMs alone for classification. We use the aforementioned best prompt setup and instruct the models to classify every sentence in the target document as linked or not, (more details in Appendix J). As shown in Figure 4, across all datasets, the LLM-only approach underperforms the `R+LLM` approach, confirming that retrieval is crucial for narrowing the search space and filtering distractors. The performance gap is especially large on the longer more complex peer review documents. For example, `GPT-4o` reaches only 33.42 and 39.22 F1 on `REVIEWS-SYNTH` and `REVIEWS-F1000`, compared to 41.42 and 59.18 with `R+LLM`. Even on the shorter news articles, the LLM-only setup falls short (e.g., 74.27 vs. 84.02 on `NEWS-SYNTH`). While LLMs can capture some cross-document signals without retrieval, their high inference cost makes them inefficient as standalone solutions compared to their performance, making retrieval as a first step essential for both efficiency and accuracy.

In summary, our results suggest that the strongest approach for both domains is `R+LLM` using `Dragon+` and `Qwen2.5`, prompting listwise with both a link description and in-context examples.

6 Assisted labeling

The final step of our framework applies the best-performing `R+LLM` approach established on synthetic data to help humans find links in naturally occurring text pairs, and compares it to `Dragon+` as a baseline in a human evaluation experiment.

6.1 Setup

We randomly sample 20 review-paper pairs and news article pairs. **`REVIEWS-HE`** consists of short conference papers and their corresponding peer reviews, drawn from the ARR22 subset of the NLPeer dataset. We focus on short papers to avoid long-context limitations during model inference, and select the shortest available review when multiple are present. **`NEWS-HE`** comprises article pairs from the SPICED dataset (Shushkevich et al., 2024), sampled from the *Politics*, *Sports*, and *Culture* categories to ensure topical diversity. In both domains, we segment documents into sentences, followed by manual corrections to improve segmentation accuracy and remove non-content elements such as social media links or boilerplate text. See Appendix K for further details on data preparation. To support the annotation process, we used the INCEpTION platform (Klie et al., 2018), which provides a side-by-side document view and allows annotators to accept or reject pre-highlighted candidate links with access to full sentence and document context; see Appendix L for further details.

6.2 Protocol

Annotators were presented with the source and target documents side-by-side, and asked to evaluate candidate links from the source document to

	REVIEWS-HE	NEWS-HE	Avg.
Annotation Results			
R+LLM only	56.4%	57.0%	56.7%
Retriever only	42.7%	17.8%	30.3%
Both	**77.7%**	**68.6%**	**73.1%**
Random	4.3%	7.7%	6.0%
Statistics			
# Doc. Pairs	20	20	
# Labeled Links	1022	804	

Table 3: Acceptance rate between the two annotators on candidate links, broken down by link suggestion method. Also statistics on the number of document pairs and total labeled links for each domain.

the target using domain-specific guidelines (Appendix O). In REVIEWS-HE, up to 8 candidate targets were shown for each source sentence: top-3 from R+LLM, top-3 from Dragon+, and 2 random distractors. In NEWS-HE, which contains shorter documents, the pool was limited to 5 candidates per source sentence: top-2 from R+LLM, top-2 from Dragon+, and one random distractor. To reduce workload, we excluded source sentences that were too short (three words or fewer), as well as explicit links (e.g. Line.X, Figure.Y) which can be resolved trivially. If a candidate target is suggested by both R+LLM and Dragon+, it was only displayed once but is counted toward both during the analysis. Annotators could accept any number of target sentences as linked to a given source, incl. no matches.

6.3 Annotation study

We recruited 15 annotators via Prolific[6] and conducted a qualification task using one "gold" document pair with 30 cross-document links labeled by the study authors for each domain (Selection criteria detailed in Appendix M). Annotators were provided with annotation guidelines for each domain. We measured agreement with the gold labels using Cohen's κ (Cohen, 1960) and selected the two annotators with the highest scores: $\kappa = 0.68$ for REVIEWS-HE and $\kappa = 0.72$ for NEWS-HE, both indicating substantial agreement. The selected annotators then labeled the full evaluation set in two batches of 10 document pairs per domain. After the first batch, we conducted a feedback round to clarify guideline interpretations and highlight common issues, after which the annotation resumed independently. Final inter-annotator agreement across both batches was $\kappa = 0.59$ for REVIEWS-HE and $\kappa = 0.60$ for NEWS-HE, reflecting substantial agreement given the subjective and open-ended nature of the task. On reviews, this exceeds the reported agreement in machine-assisted annotation by Kuznetsov et al. (2022), despite our use of crowd annotators with minimal training.

[6] https://www.prolific.com/

6.4 Results and Analysis

Table 3 summarizes the acceptance rates across link suggestion methods. Links suggested by both Dragon+ and the R+LLM achieve the highest acceptance rates: 77.7% in REVIEWS-HE and 68.6% in NEWS-HE. This is expected, as mutual agreement between both indicates higher confidence.

More importantly, we observe a consistent pattern across domains: links selected only by the R+LLM approach are accepted at substantially higher rates than those selected only by Dragon+, namely 56.7% vs. 30.3% on average. The contrast is especially pronounced in the news domain (57.0% vs. 17.8%), where retrieval alone appears less effective. We attribute this to the retriever's reliance on semantic similarity which fails to capture the nuanced, context-sensitive understanding required for accurate linking. In contrast, LLMs can reason about discourse and topic structure, drawing on prior knowledge to identify implicit connections. This highlights the LLM's strength as a task-aware filter to refine surface-level retrieval results.

Crucially, this finding echoes our results on synthetic data, where adding LLM classification consistently improved performance. The human evaluation confirms that these gains hold on natural text, demonstrating that the LLM contributes meaningful additional value beyond the retriever's capabilities. The low acceptance rate for random distractor candidates (6.0% on average) confirms that the task is non-trivial and that valid cross-document links are unlikely to arise by chance.

Our study results in substantially-sized human-labeled annotated datasets in peer reviews, and news articles, comparable in size to datasets from the literature (Kuznetsov et al., 2022; D'Arcy et al., 2024). While our framework was not designed for precise cost tracking, D'Arcy et al. (2024) report approx. 30 minutes per pair using fully manual linking on paper edits, in contrast, our annotators completed each full document pair in roughly 15 minutes on average, while annotating 10 times more links. This demonstrates the effectiveness of our candidate filtering in reducing annotation effort. Although the precise savings depend on

	Recall	Precision	F1
NEWS			
R+LLM	**0.77**	**0.93**	**0.82**
Retriever	0.57	0.70	0.61
Both	0.54	0.76	0.61
Random	0.02	0.05	0.02
REVIEWS			
R+LLM	**0.59**	**0.62**	**0.55**
Retriever	0.28	0.26	0.25
Both	0.26	0.37	0.27
Random	0.00	0.00	0.00

Table 4: True recall estimation results under exhaustive evaluation. While "Both" achieves higher precision, its recall is substantially reduced, limiting its overall F1. Metrics are macro-averaged across source sentences.

annotation tools, annotator training, and task familiarity, our results suggest substantial gains in annotation throughput without sacrificing quality.

6.5 Estimating True Recall

Similar to other assisted labeling settings, our setup does not provide an estimate of true recall, since only candidate links were annotated. To measure true recall, we conducted a supplementary experiment where we exhaustively labeled *all* links on a small subset of approx. 10% of REVIEWS-HE and NEWS-HE, resulting in 102 links for reviews and 80 for news. We use this subset to estimate the true performance of the link suggestion approaches. As Table 4 demonstrates, across both domains, R+LLM substantially improves recall and precision over retrieval-only baselines. As expected for an intersection strategy, the Both method demonstrates a drop in recall, and consequently does not outperform R+LLM on overall F1. This demonstrates that while Both can serve as a conservative high-confidence filter in annotation settings, it is not a balanced solution when considering recall. Note that in the main human study (Table 3), higher acceptance rates are due to annotating top-k candidates only, whereas Table 4 measures full-corpus metrics. We provide details and additional analysis in Appendix R. In sum, LLMs capture links that retrieval alone misses, and validates our framework's effectiveness in high-recall scenarios.

7 Conclusion

High annotation cost and diversity of link types hinder progress in cross-document linking. We present a domain-agnostic framework for sentence-level linking that bootstraps annotation with no labeled data upfront. By generating synthetic linked documents, we enable automatic evaluation of linking approaches and identify the best-performing ones for assisted human annotation.

Applied in peer review and news, our framework shows that combining retrieval with LLM-based classification yields high-quality links, doubling the recall and precision of retrieval alone. It generalizes across domains, lowers annotation effort, and produces reusable datasets. Our results highlight the framework's potential for scalable, real-world cross-document analysis.

Ethical considerations

Our work does not carry substantial additional risks and contributes to better technological support for many socially relevant application areas such as academic quality control, journalism, fake news detection and propaganda analysis. As with any AI technology, we call for additional testing and oversight if the proposed method is deployed in a sensitive domain such as law, medicine or critical infrastructure. Some general ethical risks include bias due to the use of synthetic data, lower performance in under-represented languages and domains, and dual use, for example using links between news articles and social media commentary to spot dissent. Moreover, if the generated links are used without evaluation, then incorrect links could lead to potential harms like misinformation (news domain) and wrong assumptions (peer-review domain). The participants of the synthetic data validation study in the news domain contributed on a voluntary basis; the participants of the synthetic data validation study on peer reviews conducted it as a part of their employment at the authors' institution. The crowdworkers on Prolific have been fairly compensated with 13 Euro per hour. The NLPeer dataset has a CC BY 4.0 license, and we release our corresponding derivative dataset under the same license. The ECB+ corpus has a CC BY 3.0 license which allows us to release the corresponding derivative dataset under CC BY 4.0 license. The SPICED dataset has a CC BY 4.0 license, but the licensing for its source material (news) is unclear. To work around this, we release a script for reconstructing the dataset along with our added data. All datasets were used according to their intended research purpose.

Limitations

Generality vs. domain specificity. We propose a general framework for linking annotation and demonstrate its effectiveness across two distinct domains. We note that while the overall framework is agnostic to domain and link type, some steps (data generation, prompting, evaluation) need to be instantiated with domain-specific configurations (e.g., link definitions, sampling strategies, prompt templates). We view this as a modest cost compared to the human effort required for fully manual annotation, as it balances general applicability with minimal task-specific tuning. Domain-specific characteristics (e.g., subjective critique in reviews vs. factual alignment in news) can affect how well a general approach performs out-of-the-box. Thus, more specialized models or task formulations might yield stronger results in single-domain settings. Future work could explore adaptive modules or domain-specialized tuning within our general framework, as well as evaluate the applicability of our framework in further domains.

Synthetic data. Our approach to generating synthetic data relies on prompting an LLM to produce source documents conditioned on a target document. While results show this is effective, coverage could be improved through more targeted generation strategies—such as prompting for individual sentence-level links or applying multi-pass generation. That could also improve the realism of the links as discussed in Section 4.2. We also note that small human-annotated gold sets could in principle support retrieval model selection, but as shown in Section 6.5 creating them is costly and often impractical at scale. Our synthetic approach is designed for the common case where no labeled data exists upfront; when domain-specific gold data is available, it can be incorporated as a complementary signal in the evaluation stage. Finally, LLM hallucinations can lead to synthetic data that does not follow the intended goal. While we ensure data quality through human validation, targeted evaluation of LLM hallucinations in synthetic linking data generation lies beyond our scope and is left for future work.

Model coverage and training. Due to cost constraints, we trade off the number of LLMs against the depth of evaluation for each. We prioritize coverage of representative retrievers and LLMs, rather than exhaustive prompt variants or multiple runs. Expanding these dimensions would offer deeper insight into robustness and prompt sensitivity. We do not fine-tune smaller models like BERT, as our goal is fast iteration without task-specific supervision. Although we could fine-tune on our generated synthetic data, this type of training on synthetic data is known to limit generalization in subjective tasks (Li et al., 2023).

Document modalities. Our framework currently operates on text only. However, in domains like scientific writing, non-textual elements (e.g., tables, figures) often contain essential linkable information. Supporting multi-modal linking is an important direction for future extensions of our method.

Link coverage. A core limitation of assisted annotation is that recall cannot be fully measured without exhaustive manual labeling. This is a known challenge in linking and retrieval tasks more broadly: any system that filters or ranks candidate links will inevitably leave some valid links unexamined. To mitigate this, we first tune the retrieval cutoff on synthetic data to maximize recall, ensuring that high-coverage candidates are passed to the LLM. Second, we conduct a targeted manual annotation study with full link coverage to estimate true recall on a subset of document pairs.

Acknowledgments

This work has been co-funded by the German Federal Ministry of Research, Technology and Space (BMFTR) under the promotional reference 01ZZ2314H (GeMTeX), and by the European Union (ERC, InterText, 101054961). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We gratefully acknowledge the support of Microsoft with a grant for access to OpenAI GPT models via the Azure cloud (Accelerate Foundation Model Academic Research). We thank Simone Balloccu for insightful feedback, and Thy Thy Tran, Hassan Soliman and Shivam Sharma for helpful comments on an earlier draft of this paper.

References

Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero

Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *CoRR*, abs/2412.08905.

Rémi Calizzano, Malte Ostendorff, Qian Ruan, and Georg Rehm. 2022. Generating extended and multilingual summaries with pre-trained transformers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1640–1650, Marseille, France. European Language Resources Association.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. 2024b. MetaSumPerceiver: Multimodal multi-document evidence summarization for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8742–8757, Bangkok, Thailand. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024. ARIES: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6985–7001, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Tomás Feith, Akhil Arora, Martin Gerlach, Debjit Paul, and Robert West. 2024. Entity insertion in multilingual linked corpora: The case of Wikipedia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22796–22819, Miami, Florida, USA. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Wang, and Arman Cohan. 2023. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8177–8199, Singapore. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.

Arefeh Kazemi, Sri Balaaji Natarajan Kalaivendan, Joachim Wagner, Hamza Qadeer, Kanishk Verma, and Brian Davis. 2025. Synthetic vs. gold: The role of LLM generated labels and data in cyberbullying detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language*

Processing - Natural Language Processing in the Generative AI Era, pages 531–540, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. Splade-v3: New baselines for SPLADE. *CoRR*, abs/2403.06789.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yuan Liang, Massimo Poesio, and Roonak Rezvani. 2024. A fine-grained citation graph for biomedical academic papers: the finding-citation graph. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 416–426, Bangkok, Thailand. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Teng Lin, Yuyu Luo, Honglin Zhang, Jicheng Zhang, Chunlin Liu, Kaishun Wu, and Nan Tang. 2025. MEBench: Benchmarking large language models for cross-document multi-entity question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1494, Suzhou, China. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Yingzhou Lu, Huazheng Wang, and Wenqi Wei. 2023. Machine learning for synthetic data generation: a review. *CoRR*, abs/2302.04062.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfrembedding-mistral: enhance text retrieval with transfer learning. *Salesforce AI Research Blog*, 3:6.

Qingkai Min, Zitian Qu, Qipeng Guo, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2025. Multi-document event extraction using large and small language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19265–19296, Suzhou, China. Association for Computational Linguistics.

Yasunari Miyabe, Hiroya Takamura, and Manabu Okumura. 2008. Identifying cross-document relations between sentences. In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 141–148. The Association for Computer Linguistics.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. Association for Computational Linguistics.

James Ravenscroft, Amanda Clare, Arie Cattan, Ido Dagan, and Maria Liakata. 2021. CD^2CR: Co-reference resolution across documents and domains. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 270–280, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Elena Shushkevich, Long Thanh Mai, Manuel V. Loureiro, Steven Derby, and Tri Kurniawan Wijaya. 2024. SPICED: News similarity detection dataset with multiple topics and complexity levels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15181–15190, Torino, Italia. ELRA and ICCL.

Shahbaz Syed, Ahmad Dawar Hakimi, Khalid Al-Khatib, and Martin Potthast. 2023. Citance-contextualized summarization of scientific papers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8551–8568, Singapore. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Boris Van Breugel, Zhaozhi Qian, and Mihaela Van Der Schaar. 2023. Synthetic data, real errors: How (Not) to publish and use synthetic data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34793–34808. PMLR.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *CoRR*, abs/2305.15041.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Jin Zhao, Jingxuan Tu, Bingyang Ye, Xinrui Hu, Nianwen Xue, and James Pustejovsky. 2025. Beyond benchmarks: Building a richer cross-document event coreference dataset with decontextualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3499–3513, Albuquerque, New Mexico. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Group-level link analysis

To assess the extent of group-level relations, we analyzed clusters of adjacent links, defined as sets of two or more links where source and/or target sentences are consecutive. Such clusters approximate larger argumentative or narrative units.

The results in Table 5 confirm that group-level structures are common-particularly in news, where shorter documents promote multi-sentence alignments. Importantly, the analysis also shows that sentence-level links compose naturally into such clusters, providing atomic building blocks for broader structures. Future extensions of our framework may incorporate explicit span-level or hierarchical modeling.

B Datasets Conversion

In **NEWS-ECB+**, for each document pair, we label a sentence pair as linked if they contain event mentions in the same cluster. While this conversion provides a useful signal, it introduces two key limitations: (1) the resulting labels are incomplete, as only coreferent events are annotated, and (2) some event mentions occur in headlines or titles, which leads to spurious links from titles to many other sentences. In **REVIEWS-F1000**, the authors didn't adjudicate the annotations, so we only select examples where both expert annotators agreed on the label. This reduces the number of links we can use, but ensures we have a comparable dataset to **REVIEWS-SYNTH**.

C Quantitative Criteria

To assess how closely our synthetic documents resemble natural ones, we selected three document-level metrics that capture distinct stylistic dimensions: (1) lexical diversity, as a proxy for richness

Domain	Dataset	#Pairs	#Links	%Links in Clusters	Mean / Max Size
News	Synthetic	346	2323	70%	3.1 / 7
Reviews	Synthetic	200	2181	44%	2.4 / 7
News	Human	20	296	54%	3.1 / 12
Reviews	Human	20	334	21%	2.2 / 4

Table 5: Analysis of clusters of consecutive links. Group-level structures are frequent, especially in news.

of vocabulary; (2) subjectivity[7], to assess differences in tone or stance; and (3) lexical complexity, using the Flesch-Kincaid Reading Ease score, to estimate linguistic difficulty. These criteria were chosen to reflect qualities that may affect linkability and document coherence, while remaining agnostic to content. Figure 13 plots the results for synthetic and natural documents in both domains.

D Rationale for Evaluation Criteria

We selected domain-specific evaluation criteria to reflect both general text quality and the particular requirements of our linking task. For `NEWS-SYNTH`, we assessed:

- **Fluency** assesses grammatical correctness and naturalness of the language, ensuring the text is readable.
- **Coherence** measures logical flow and topic consistency, which is critical for linking tasks that depend on understanding article structure and progression.
- **Realism** evaluates whether the article could plausibly have been written by a human journalist, and ensure the generated article is not clearly AI-generated.
- **Specificity** checks for the inclusion of concrete facts (e.g., names, dates, events, etc.) to ensure the AI-generated review is not general and vague.

For `REVIEWS-SYNTH`, we assessed:

- **Fluency** and **coherence**, as above, ensure the review reads naturally and follows a logical structure.
- **Helpfulness** replaces realism in this context, measuring whether the review provides meaningful feedback that could assist authors, which is central to the function of peer reviews.
- **Specificity** captures whether the review addresses concrete elements of the paper, such as specific sections, claims, or experiments, helping differentiate generic summaries from insightful critique.

These criteria were chosen to ensure that our synthetic datasets maintain linguistic quality and task relevance, even if they do not replicate all the nuanced characteristics of human writing. Figure 14 illustrates detailed validation results.

E Retrievers' Results

Tables 11, 12, 13 and 14 provide the full retrieval results for all models for all datasets individually, broken down by cutoff values ($k \in \{1, 3, 5, 7, 10, 20\}$).

F Statistical significance testing

We conducted bootstrap resampling (10k iterations) to test whether improvements from adding LLM classification (`R+LLM`) over the best retriever baseline (`Dragon+`) are statistically significant. Results are shown in Table 6.

The results confirm that improvements on the synthetic datasets are statistically significant, whereas gains on the converted datasets are not, which aligns with our earlier discussion that these datasets are less aligned with our task definition.

G Error analysis

To better understand the limitations of our approach, we manually analyzed 100 failure cases across datasets. We identified three frequent patterns:

1. Scattered target sentences: this happens when the target sentences are not concentrated in a specific location in the target document, but are spread out. This is especially a problem in the reviews domain where some review sentences address something in the paper that's presented listwise over multiple pages (e.g. steps of a methodology). This problem is present in the news domain too, but due to the shorter document sizes, it's less prominent.
2. Ambiguous source sentences: Some source sentences lack sufficient signals or specificity

[7] https://huggingface.co/cffl/bert-base-styleclassification-subjective-neutral

Dataset	Retriever F1	R+LLM F1	ΔF1	95% CI	p-value
NEWS-SYNTH	0.74	0.84	+0.096	[0.045, 0.140]	<0.001
REVIEWS-SYNTH	0.25	0.41	+0.159	[0.127, 0.188]	<0.001
NEWS-ECB+	0.55	0.57	+0.015	[-0.011, 0.042]	0.72
REVIEWS-F1000	0.59	0.59	-0.002	[-0.020, 0.032]	0.22

Table 6: Bootstrap significance test of F1 improvements from R+LLM over Dragon+.

for reliable linking (e.g., short, vague.) Similar to the point above, this is a problem in the reviews domain where reviews are shorter documents with more condensed sentences. However, even in the news domain, depending on the article's source and its editing style, some sentences can lack sufficient signals necessary for linking.

3. Too many distractors: Because we use the retriever to reduce the search space, we end up passing a list of very semantically similar target sentences to the LLM. In the reviews domain, this is less of a problem, because the more semantically similar sentences are more concentrated in the same sections on the paper, but this in itself leads to problem 1 mentioned above. In the news domain, due to the shorter documents, there are similar sentences that get passed to the LLM requiring more reasoning to filter them out.

H R+LLM across different retrievers

To test whether the benefits of LLM-classification depend on the choice of base retriever, we applied the `R+LLM` step (Qwen2.5, listwise prompting with description + examples) to four additional retrieval models: `SFR`, `ms_marco_MiniLM`, `BM25`, and `BGE-M3-sparse`. Results are shown in Table 7.

On the synthetic datasets, the R+LLM setup yields consistently positive and similarly sized improvements on REVIEWS-SYNTH (≈+9–16 F1) compared to the other synthetic dataset NEWS-SYNTH (≈+8–14 F1). On converted datasets (NEWS-ECB+ and REVIEWS-F1000), improvements are marginal, which we attribute to task mismatch (Section 4). Interestingly, on REVIEWS-F1000, BM25 shows the largest positive improvement among the retrievers when combined with the LLM, also achieving the highest F1 in the R+LLM setup. This matches our explanation that BM25's lexical matches are particularly helpful in this biomedical/technical domain due to domain-specific terminology in biology and medicine. On the other datasets BM25 starts from a lower base performance than the dense retrievers, hence the LLM has more room to correct errors, yielding larger relative gains.

I LLM Prompting Techniques

We created two prompt templates corresponding to the *pairwise* and *listwise* classification setups (Figures 5 and 6). Each prompt includes a system message that frames the task: determining whether a candidate target sentence from one document is related to a source sentence from another document, using full document context. In the pairwise setup, the model classifies one sentence pair at a time; in the listwise setup, it classifies a ranked list of candidate targets in a single pass. Both formats support four prompting configurations: no guidance, link description only, in-context examples only, and both. The prompts are designed to output binary classifications and return structured JSON outputs, facilitating scalable evaluation across top-k retrieved candidates. For `Phi-4` and `Qwen2.5`, we used vLLM (Kwon et al., 2023) with the structured outputs function to ensure output consistency. The experiments with the open-source models were run on A100 GPUs. For `GPT-4o`, we used the OpenAI API with structured outputs too. For all models, we set `temperature = 0.3` and `top-p = 0.9`.

J LLM-only Ablation

Figure 7 shows the prompt template used in the LLM-only setup, where the model classifies all sentences in the target document without prior retrieval. The prompt includes full document context and a single source sentence, along with a link description, and in-context examples, and instructs the LLM to evaluate each target sentence for its relevance. As in the other setups, we support four prompt configurations: no guidance, with in-context examples, with a link description, and with both. We used structured outputs here too, with `temperature = 0.3` and `top-p = 0.9`

Dataset	Retriever	Retriever Only	R+LLM	Diff
NEWS-ECB+	SFR	58.19	53.88	-4.31
	ms_marco_MiniLM	54.63	54.40	-0.23
	BM25	45.03	53.44	+8.41
	BGE-M3-sparse	48.30	53.65	+5.35
	Dragon+	*55.36*	*56.84*	+1.48
NEWS-SYNTH	SFR	70.04	78.18	+8.14
	ms_marco_MiniLM	77.69	80.56	+2.87
	BM25	61.78	76.63	+14.85
	BGE-M3-sparse	67.56	77.61	+10.05
	Dragon+	*74.38*	*84.02*	+9.64
REVIEWS-SYNTH	SFR	23.09	36.16	+13.07
	ms_marco_MiniLM	27.60	37.12	+9.52
	BM25	22.21	35.53	+13.32
	BGE-M3-sparse	25.79	38.12	+12.33
	Dragon+	*25.48*	*41.42*	+15.94
REVIEWS-F1000	SFR	42.44	42.07	-0.37
	ms_marco_MiniLM	61.03	57.38	-3.65
	BM25	54.75	59.14	+4.39
	BGE-M3-sparse	54.11	57.97	+3.86
	Dragon+	*58.73*	*58.56*	-0.17

Table 7: Performance of R+LLM using Qwen2.5 across different retrievers. Scores are F1 on synthetic and converted datasets.

K Data Preparation for Human Evaluation

For REVIEWS-HE, we randomly sampled 20 paper-review pairs from the ARR22 subset of NLPeer from the shortest 100 ones to minimize long-context issues during model inference. In cases with multiple reviews, we selected the shortest available. Peer reviews were segmented into sentences using spaCy [8] (Honnibal et al., 2020) and manually corrected for sentence boundary errors. The paper texts came pre-segmented via NLPeer. For NEWS-HE, we sampled 20 article pairs from the SPICED dataset across the *Politics*, *Sports*, and *Culture* categories to ensure diversity. The SPICED dataset is a multi-document summarization dataset that pairs news articles of different sources that talk about the same topic or event, so the document pairs are naturally related, and suitable for our task.

L Annotation interface

We used the INCEpTION annotation platform (Klie et al., 2018), which supports side-by-side viewing of two documents and interactive annotation (Figure 15). Source sentences, along with their candidate target sentences were pre-highlighted in the interface, along with visual cues, in the form of arrows pointing from the source sentence to each target sentence, to reinforce the linking aspect of the task. Annotators could view the full sentence and surrounding context and were asked to accept or reject each candidate. Candidates from both the retriever and R+LLM approach were visually indistinguishable and shown in the normal order they show up in the document.

[8] model: en_core_web_md

M Annotator Selection Criteria

Annotators were screened based on the following eligibility criteria: Native English speakers, residing in United States, United Kingdom, Canada, Australia, or Ireland, with at least a Master's or PhD degree in Computer Science, a Prolific approval rate of at least 90%, and had completed a minimum of 30 prior studies on the platform.

N Agreement Trends and Annotation Feedback

We observed a learning curve in annotator agreement. In REVIEWS-HE, agreement started relatively low, improved steadily toward the middle of the batch, and declined slightly toward the end, likely due to annotator fatigue. In NEWS-HE, agreement rose more gradually and plateaued at a slightly lower level overall. First-batch inter-annotator agreement was $\kappa = 0.55$ for REVIEWS-HE and $\kappa = 0.54$ for NEWS-HE, indicating moderate agreement despite the subjective nature of the task.

After the first batch, we reviewed disagreement patterns. One annotator tended to reject links that

LLM Classification Prompt (Pairwise)

System Message:
You are an AI assistant specialized in evaluating sentence relations.
You will get two related documents, along with a sentence from Document 1 (source) and a sentence from Document 2 (target). Your task is to determine if the target sentence is related to the source sentence.

Prompt Format:
- Full Document 1: [text of Document 1]
- Full Document 2: [text of Document 2]
- Source Sentence from Document 1: [source sentence]
- Target Sentence from Document 2: [target sentence]

Prompt Variants:
- **Mode 1 (No Guidance):** No additional information; model is directly instructed to decide if the pair is related.
- **Mode 2 (Examples Only):** A few positive example sentence pairs are provided to guide the model.
- **Mode 3 (Description Only):** A link description is provided to define what counts as a "link."
- **Mode 4 (Description + Examples):** Both the link description and examples are included.

Response Format:
- A JSON object of the form: {"related": true} or {"related": false}

Figure 5: Prompt template for LLM-based pairwise sentence classification.

LLM Classification Prompt (Listwise)

System Message:
You are an AI assistant specialized in evaluating sentence relations.
You will get two related documents, along with a sentence from Document 1 (source) and a list of sentences from Document 2 (targets). The targets are ranked based on their similarity to the source sentence. Your task is to determine for each target sentence if it is related to the source sentence. This will help filter out irrelevant sentences and improve the quality of the ranked sentences.

Prompt Format:
- Document 1: [text of Document 1]
- Document 2: [text of Document 2]
- Source Sentence from Document 1: [source sentence]
- Ranked Target Sentences from Document 2 (Sentence_ID: Sentence_text):
0: "..."
1: "..."
...

Prompt Variants:
- **Mode 1 (No Guidance):** Direct classification with no additional context.
- **Mode 2 (Examples Only):** Positive examples are provided before classification.
- **Mode 3 (Description Only):** A definition of what constitutes a "related" sentence is given.
- **Mode 4 (Description + Examples):** Both the description and examples are included.

Response Format:
- A JSON object with sentence IDs as keys and true or false as values.
e.g., {"0": true, "1": false, "2": true}

Figure 6: Prompt template for LLM-based listwise sentence classification.

relied on broader document context, favoring self-contained links. The other applied stricter interpretations of the guidelines, occasionally rejecting links that only partially aligned. Based on these observations, we gave targeted feedback, clarified borderline cases, and encouraged consistent application of the criteria. This led to improved agreement in the second batch ($\kappa = 0.62$ for REVIEWS-HE, 0.64 for NEWS-HE). However, this could also be attributed to the annotators getting used to the task, and thus understanding the guidelines better.

O Annotation Guidelines

Table 8 and 9 contain the criteria provided to the annotators to decide whether a sentence pair should be labeled as linked or not. Because the concept of a "link" can be broad, we provided domain-specific examples to clarify what qualifies as a meaningful connection in the context of peer reviews and news.

P Synthetic Data Generation Prompts

To generate synthetic peer reviews and news articles, we designed detailed prompting templates that reflect realistic domain practices while introducing controlled variation. The peer review prompt (Figure 9) instructs the model to simulate a structured review grounded partially in the source text, while maintaining a natural and critical tone consistent with academic peer reviews. It emphasizes abstraction, editorial judgment, and partial grounding to avoid simple paraphrasing. Because the news articles from the WikinewsSum dataset were scraped directly from the webpage, they contain some artefacts that are not content of the article. Thus, we

LLM-only Classification Prompt

```
System Message:
You are an AI assistant specialized in evaluating sentence relations.
You will get two related documents, along with a sentence from Document 1 (source). Your task is to determine for each sentence in the target document (Document 2) if it is related to the source sentence. This will help filter out irrelevant sentences.

Prompt Format:
- Document 1: [text of Document 1]
- Document 2: [text of Document 2]
- Source Sentence from Document 1: [source sentence]

Prompt Variants:
- Mode 1 (Minimal): No additional context—just classify based on the input documents and source sentence.
- Mode 2 (Examples Only): A few examples of positive sentence pairs are provided.
- Mode 3 (Description Only): A link description is provided to define what counts as a link.
- Mode 4 (Description + Examples): Both the link description and in-context examples are included.

Response Format:
- A JSON object where each key is a target sentence ID and each value is either true or false.
e.g., {"11": true, "4": false, "9": true}
- Ensure that every sentence in Document 2 is classified.
```

Figure 7: Prompt template for LLM-only sentence classification across an entire target document.

first clean that using GPT-4o-mini with the prompt template shown in Figure 8. We chose GPT-4o-mini because it reliably removed social media links and scraping artifacts (based on manual checks), while being cost-effective. Next, we segment the cleaned articles using spaCy. For the choice of model used in synthetic document generation, we had three criteria: (1) strong general performance, (2) strong instruction-following (due to the specificity of our prompts), and (3) a long-context window (to handle long docs like papers). We selected DeepSeek-R1, as reasoning models showed strong prompt adherence. Importantly, we chose DeepSeek-R1 over models like OpenAI-o1 to avoid overlap with the OpenAI models used in the downstream classification task, thereby minimizing potential model-specific biases in evaluation (Liu et al., 2024). DeepSeek-R1 also offers comparable performance to OpenAI models on standard benchmarks (DeepSeek-AI et al., 2025). The generation prompt (Figure 10) guides the model to produce synthetic news article covering the same topic, but differing in structure, tone, and emphasis to simulate diverse editorial styles. Both generation prompts include format specifications and grounding strategies. While these prompts may not be the most optimal, they produce sufficiently high-quality and controllable outputs for our specific goal of in the linking task. Table 15 provides illustrative examples of generated sentences and their corresponding linked natural sentences in both domains

Link	No Link
The review sentence critiques or praises the same claim or result as the paper sentence.	The review mentions grammar, structure, or other writing aspects not discussed in the candidate sentence.
The review analyzes or extends a method/result introduced in the paper sentence.	The review proposes future work or experiments not mentioned in the paper sentence.
The review addresses the same technical method, component, or result.	The review discusses different topics or experiments from the candidate sentence.

Table 8: Linking Criteria for Peer Reviews

Link	No Link
Both sentences refer to the same event or topic.	Sentences refer to different events, even if they happen at the same place/time.
One sentence is a follow-up or elaboration of the other.	The target sentence provides unrelated opinion, editorial, or background.
One is a paraphrase or restatement of the other (even if phrased differently).	The anchor and target describe distinct stories with no direct overlap.

Table 9: Linking Criteria for News

Q Human Evaluation Results

Figures 11 and 12 illustrate the results of the human evaluation in detail.

R Manual Recall Estimation Details

To estimate recall, we sampled document pairs from REVIEWS-HE and NEWS-HE. For each domain, we selected random source sentences and exhaustively searched the target document for related sen-

tences according to the domain-specific link definition. This was repeated until we labeled approximately 10% of the total links: 102 in reviews and 80 in news. These links were used to compute precision, recall, and F1 for the top-performing R+LLM, retriever-only (Dragon+), the "Both" intersection strategy, and a random baseline. The manual annotation was done by one of the authors of this paper.

The annotation of the 102 links in reviews took approx. 11 hours. The majority of the time was spent understanding the paper and the review, and ensuring that all possible target sentences were assessed. The 80 links in news took less time at approx. 5 hours due to the shorter documents, and the majority of the time was spent carefully reading the content of each sentence to judge whether it was related. This annotation further highlighted the sizable effort needed for full manual annotation of cross-document links, and the need for automatic methods to speed up and scale up the process.

R.1 Alternative Evaluation under Candidate-Pool Constraints

As noted in Section 6, the main human evaluation (Table 3) was conducted on top-k candidate pools shown to annotators. To reconcile the difference between these acceptance rates and the exhaustive recall analysis in Table 4, we repeated the recall estimation under the same top-k constraints. The results are shown in Table 10.

	Recall	Precision	F1
NEWS			
R+LLM	0.77	0.93	0.82
Retriever	0.57	0.70	0.61
Both	**0.68**	**0.95**	0.76
Random	0.02	0.05	0.03
REVIEWS			
R+LLM	0.43	0.64	0.49
Retriever	0.18	0.28	0.21
Both	0.24	0.52	0.32
Random	0.00	0.00	0.00

Table 10: Recall estimation under candidate-pool constraints, mimicking the human evaluation setup. "Both" regains the high precision observed in Table 3, but at the cost of recall.

These results confirm that "Both" is best understood as a conservative high-precision filter: it produces links that are very likely to be correct, but its lower recall limits standalone utility. R+LLM remains the superior balanced approach when considering overall F1 under exhaustive conditions.

News Cleaning Prompt

```
This is a scraped article from Wikinews.
Due to the scraping, it may contain
sentences that are image captions and
social media links. Remove such sentences,
but do not change the content of the article.
Output the cleaned article in JSON format,
where the key is 'cleaned_article' and the
value is the cleaned article text.
Input: {INPUT}
```

Figure 8: Prompt template for corrected scraped news articles.

R.2 Practical Guidelines for Practitioners

Based on our experiments, our framework supports multiple operating modes depending on annotation budget and risk profile:

- **Single best method (R+LLM)**: Balanced precision/recall with moderate annotator load. Default choice when no strong constraints exist.
- **Intersection of top-N ("Both")**: Conservative high-precision filter with reduced recall. Suitable when annotator time is very limited.
- **Union of top-N**: Maximizes recall at the cost of lower precision, producing larger candidate pools. Useful when coverage is critical and annotators can handle more candidates.

These modes can be selected to match application priorities, e.g., efficiency versus coverage.

Synthetic Peer Review Generation Prompt

```
Task Overview
You are a peer reviewer evaluating a research paper in NLP. Your task is to write a realistic and well-rounded peer review for the paper.
Guidelines:
- Your review should be structured and natural, consisting of 8-12 sentences.
- 3 to 5 sentences should be implicitly grounded in specific ideas from the paper. These should express relevant critiques, observations, or praises without directly quoting or referencing the original text.
- The remaining sentences should address broader aspects such as clarity, methodology, contributions, generalization, writing quality, or suggestions for improvement.
- Do NOT explicitly cite, reference, or quote any sentence from the paper. The review should not be a direct commentary on specific lines.
How to Structure the Review:
- Rephrase, summarize, or abstract ideas: When addressing parts of the paper, reword and generalize instead of copying.
- Introduce new considerations: Some comments should reflect editorial judgment, unanswered questions, or high-level concerns rather than being tied to specific sentences.
- Omit some possible links: Not every review sentence should directly correspond to a sentence from the paper. Aim for a balanced mix of specific and general feedback.
- Rearrange information: The review's flow should be different from the order of the paper to reflect a natural peer review process.
Input Format
You will receive a JSON object where:
- Each key is a sentence index from a paper section(s).
- Each value is the corresponding sentence text.
Output Format
- A peer review as a JSON object where:
    - Each key is a sentence index in the review (starting from 0).
    - Each value is the corresponding sentence text.
- A sentence mapping that links review sentences to the paper as a JSON object where:
    - Keys: Indices from the peer review.
    - Values: A list of corresponding indices from the paper section(s), or null if the sentence is not directly linked.
Example Input:{INPUT_EXAMPLE}
Example Output:{OUTPUT_EXAMPLE}
Input:{INPUT}
Output:
```

Figure 9: Prompt template for generating REVIEWS-SYNTH.

Model	K=1			K=3			K=5			K=7			K=10			K=20		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bm25	56.71	34.12	40.91	37.21	63.17	45.03	28.75	76.74	40.28	24.17	83.59	35.95	21.17	89.40	32.43	18.72	96.40	28.95
splade	62.03	37.47	44.87	41.60	70.11	50.25	30.74	81.93	43.09	25.44	88.37	37.92	21.88	93.23	33.60	18.90	98.26	29.27
bgem3-sparse	59.20	35.30	42.47	39.94	67.59	48.30	30.07	80.30	42.17	25.07	86.97	37.35	21.70	92.21	33.31	18.88	98.07	29.24
sfr	69.14	41.03	49.50	48.34	80.72	58.19	34.51	91.97	48.38	27.49	95.85	41.04	22.81	98.01	35.13	19.02	99.68	29.50
all-mpnet	65.08	39.40	47.12	45.81	77.36	55.35	33.23	88.90	46.62	26.76	93.52	39.95	22.48	96.52	34.60	18.99	99.36	29.44
bgem3-dense	67.65	41.05	49.06	45.55	76.86	55.01	33.03	88.13	46.32	26.59	92.78	39.69	22.44	96.23	34.53	18.95	98.95	29.38
contriever	64.13	38.85	46.46	43.38	73.45	52.46	32.16	85.82	45.10	26.26	91.65	39.18	22.35	95.73	34.38	18.99	99.34	29.45
dragon_plus	66.67	40.31	48.21	45.82	77.36	55.36	33.27	89.02	46.68	26.86	93.87	40.10	22.53	96.72	34.67	19.01	99.50	29.48
bge-reranker	61.29	36.29	43.78	44.86	75.38	54.10	33.03	88.05	46.29	26.78	93.33	39.96	22.56	96.74	34.72	19.02	99.69	29.50
ms_marco_minilm	65.06	39.03	46.83	45.24	76.24	54.63	32.84	87.76	46.07	26.66	93.07	39.80	22.45	96.27	34.55	18.98	99.25	29.43

Table 11: Results on NEWS-ECB+. Precision (P), Recall (R), and F1 at different cutoffs (K).

Synthetic News Generation Prompt

```
Task Overview:
You are generating a news article that covers the same event or topic as an original article
presenting it from a different editorial angle. The goal is to simulate how separate news
organizations might independently report on the same subject, differing in structure, tone,
detail, and emphasis. The new article should be a plausible alternative version of coverage on
the same topic, not a direct rephrasing or summary of the original.
Guidelines:
- The article should be realistic, coherent, and reflective of a distinctive voice or editorial
style.
- 3 to 5 sentences in the new article should reflect content from the original. These sentences
may describe similar facts, events, or issues, but using different wording, tone, or framing.
- Do not replicate the original article's sentence-by-sentence structure or closely paraphrase
its content.
- The remaining sentences should introduce: New but plausible perspectives, context, or editorial
framing. Additional background or expert input. A different narrative structure or omission of
certain original points.
- The new article must have a different length from the original, but not be much shorter than
the original.
Variation Strategies:
- Rephrase and shift style: Change vocabulary, sentence structure, or writing tone to reflect a
different editorial voice.
- Frame the topic differently: Adjust emphasis or viewpoint—, for example, highlighting
controversy, local impact, or long-term implications.
- Add or omit information: Introduce plausible context, background, or expert input, or skip
less relevant details from the original.
- Reorganize the narrative: Present the information in a different order to create a new logical
or rhetorical flow.
Input Format:
You will receive a JSON object where:
- Each key is a sentence index from the original article.
- Each value is the corresponding sentence text.
Output Format:
- The generated news articlea as a JSON object:
- Keys: Indices of sentences in the new article (starting from 0).
- Values: The text of each sentence.
A sentence mapping that links sentences in the new article to sentences in the original article
as a JSON object where:
- Keys: Indices of sentences in the new article.
- Values: A list of sentence indices from the original that the new sentence relates to, or null
if it is not directly linked to any original sentence.
Example input: {INPUT_EXAMPLE}
Example Output: {OUTPUT_EXAMPLE}
Input: {INPUT}
Output:
```

Figure 10: Prompt template for generating NEWS-SYNTH.

Model	K=1			K=3			K=5			K=7			K=10			K=20		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bm25	65.29	60.26	61.78	29.75	78.60	42.25	19.75	86.12	31.52	14.64	89.23	24.73	11.14	94.35	19.63	8.63	99.59	15.56
splade	67.36	61.57	63.36	29.48	78.39	42.00	20.17	87.16	32.08	15.05	91.43	25.41	11.26	95.45	19.85	8.61	99.17	15.52
bgem3-sparse	71.49	65.85	67.56	31.96	84.86	45.50	20.41	88.95	32.56	15.35	93.03	25.89	11.39	96.49	20.07	8.63	99.59	15.56
sfr	74.79	67.99	70.04	34.71	90.11	48.91	22.31	95.23	35.35	16.29	97.18	27.37	11.59	97.80	20.42	8.65	100.00	15.60
all-mpnet	76.86	70.47	72.38	34.57	90.79	48.99	22.15	96.17	35.30	16.35	97.85	27.50	11.72	98.97	20.64	8.65	100.00	15.60
bgem3-dense	79.34	72.53	74.59	35.26	92.11	49.83	22.23	95.43	35.30	16.53	99.04	27.80	11.80	99.79	20.79	8.65	100.00	15.60
contriever	65.70	60.48	62.05	30.58	81.16	43.51	19.92	86.68	31.76	15.35	92.27	25.84	11.51	97.31	20.27	8.65	100.00	15.60
dragon_plus	78.51	72.52	74.38	34.71	91.54	49.29	21.57	93.90	34.42	16.41	98.47	27.61	11.72	99.17	20.64	8.65	100.00	15.60
bge-reranker	76.03	70.12	71.90	34.85	90.66	49.17	22.40	95.85	35.51	16.65	99.17	27.96	11.84	100.00	20.86	8.65	100.00	15.60
ms_marco_minilm	82.64	75.56	77.69	33.88	88.79	47.98	21.90	94.46	34.83	16.12	96.75	27.13	11.72	98.76	20.63	8.65	100.00	15.60

Table 12: Results on NEWS-SYNTH. Precision (P), Recall (R), and F1 at different cutoffs (K).

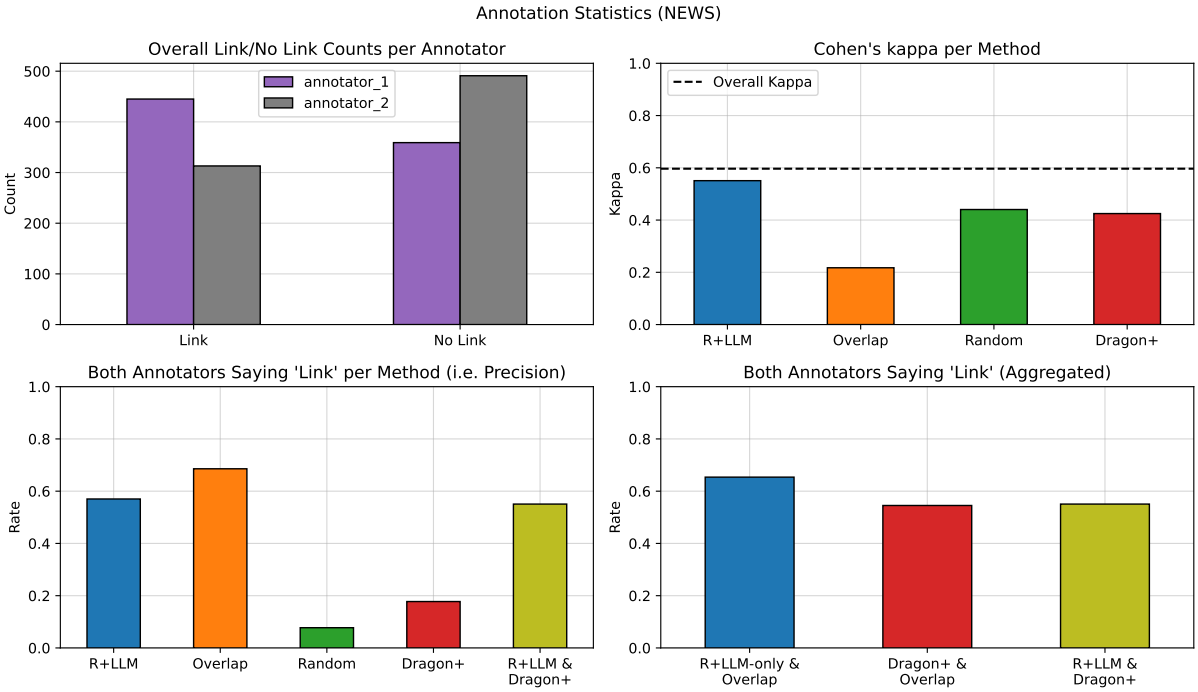


Figure 11: Annotation statistics in NEWS-HE.

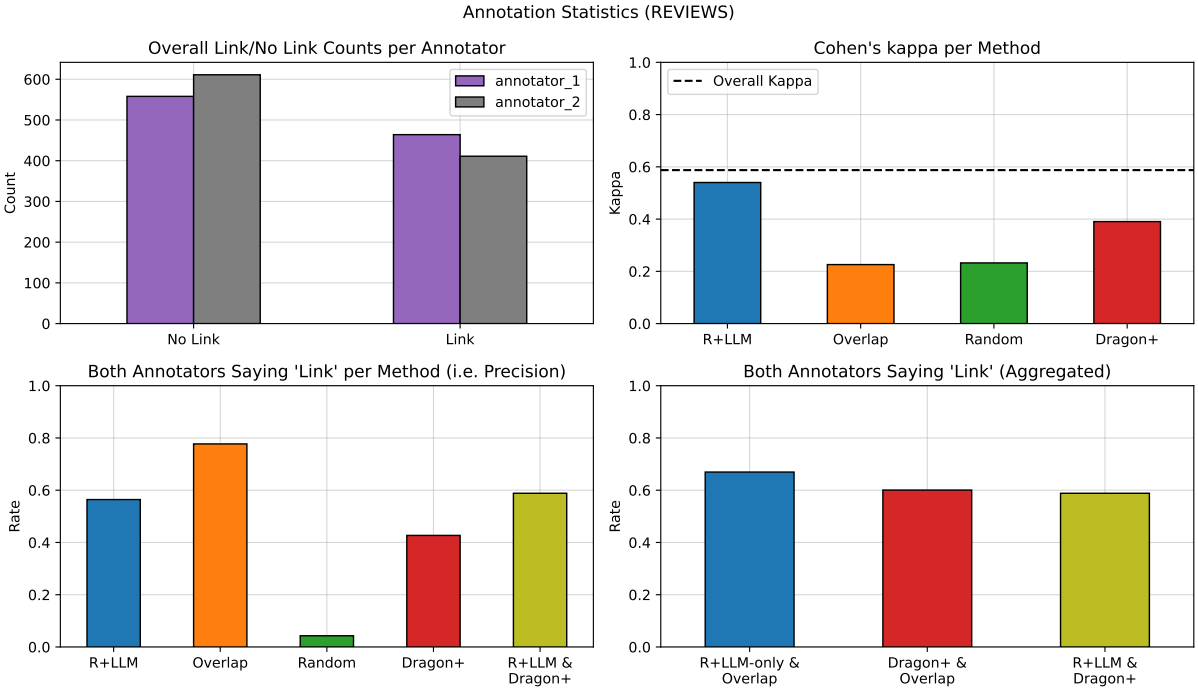
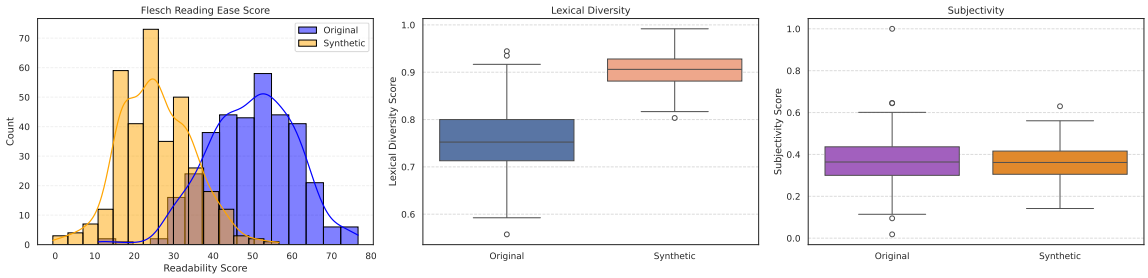
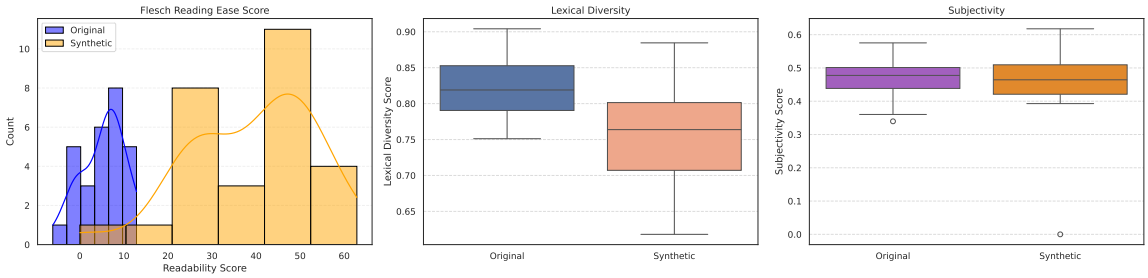


Figure 12: Annotation statistics in REVIEWS-HE.

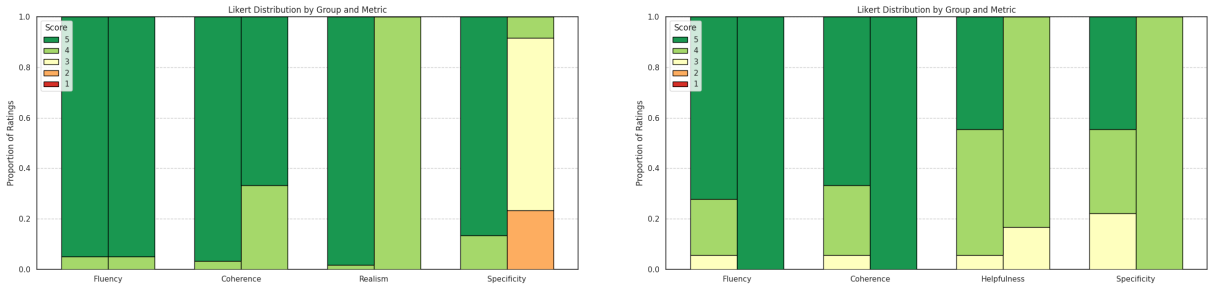


(a) Quantitative comparison of synthetic news to natural news data.



(b) Quantitative comparison of synthetic reviews to natural reviews.

Figure 13: Comparison of synthetic and natural documents in terms of stylistic and structural metrics, across two domains.



(a) NEWS-SYNTH Human Validation.

(b) REVIEWS-SYNTH Human Validation.

Figure 14: Human evaluation results for NEWS-SYNTH (left) and REVIEWS-SYNTH (right).

Model	K=1			K=3			K=5			K=7			K=10			K=20		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bm25	28.63	16.70	19.93	19.24	30.57	22.21	16.12	41.64	22.03	13.09	47.14	19.57	10.97	55.01	17.64	6.83	66.39	12.11
splade	25.11	14.49	17.27	20.85	33.58	24.16	16.56	42.65	22.66	14.10	51.12	21.12	11.72	60.04	18.92	7.14	69.83	12.65
bgem3-sparse	26.87	14.79	17.93	22.17	36.03	25.79	16.74	44.47	23.09	13.91	49.90	20.79	11.32	56.08	18.15	7.11	70.08	12.62
sfr	24.67	13.19	16.20	20.12	31.34	23.09	16.56	43.01	22.64	13.97	49.70	20.81	11.59	58.52	18.64	7.80	75.93	13.82
all-mpnet	27.75	15.10	18.40	17.77	27.94	20.31	15.07	38.93	20.58	12.96	46.37	19.36	10.88	54.90	17.52	6.96	68.69	12.36
bgem3-dense	29.96	16.56	20.15	22.32	35.17	25.73	18.68	47.51	25.48	15.61	55.51	23.31	12.78	64.00	20.53	7.84	77.14	13.91
contriever	24.67	13.27	16.34	18.21	28.16	20.88	14.10	36.51	19.33	12.59	45.04	18.86	10.18	51.32	16.43	6.74	66.69	11.98
dragon_plus	30.84	17.78	21.28	21.73	36.03	25.48	18.33	48.69	25.24	15.10	55.14	22.65	11.63	59.28	18.74	7.49	73.50	13.29
bge-reranker	23.35	12.34	15.18	20.41	34.16	23.98	16.12	43.83	22.33	14.73	53.11	22.06	12.16	60.70	19.54	7.91	77.11	14.02
ms_marco_minilm	33.92	20.31	23.96	23.64	38.66	27.60	18.59	48.73	25.52	15.10	53.58	22.53	11.94	59.51	19.19	7.42	72.97	13.17

Table 13: Results on REVIEWS-SYNTH. Precision (P), Recall (R), and F1 at different cutoffs (K).

Model	K=1			K=3			K=5			K=7			K=10			K=20		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bm25	58.28	53.15	54.75	28.21	73.84	40.00	18.75	81.30	29.95	14.15	85.36	23.91	10.29	88.51	18.22	5.51	93.90	10.33
splade	54.36	49.33	50.91	28.31	73.68	40.06	18.94	81.04	30.15	14.30	85.51	24.12	10.66	90.47	18.82	5.64	95.59	10.57
bgem3-sparse	57.49	52.53	54.11	28.57	74.61	40.49	18.94	81.68	30.23	14.31	85.99	24.17	10.49	89.20	18.53	5.62	95.38	10.54
sfr	45.94	40.83	42.44	24.39	62.65	34.34	16.96	72.01	26.94	13.08	77.95	22.05	9.82	82.82	17.34	5.46	92.05	10.23
all-mpnet	43.98	39.56	40.94	23.21	60.00	32.76	16.16	69.13	25.71	12.59	75.13	21.23	9.46	80.34	16.71	5.26	89.16	9.86
bgem3-dense	57.30	51.52	53.33	28.80	74.30	40.62	19.28	82.02	30.62	14.64	87.07	24.66	10.72	91.02	18.94	5.67	96.14	10.63
contriever	46.91	42.36	43.79	24.49	63.80	34.68	17.43	74.47	27.73	13.40	79.65	22.58	10.06	85.49	17.77	5.50	93.45	10.31
dragon_plus	63.08	56.75	58.73	30.33	77.79	42.67	20.20	86.17	32.11	15.10	89.80	25.44	10.95	93.02	19.35	5.73	97.08	10.74
bge-reranker	47.50	43.42	44.71	26.15	68.26	37.06	18.20	78.30	29.02	13.92	83.32	23.50	10.43	88.60	18.43	5.65	95.74	10.58
ms_marco_minilm	64.94	59.24	61.03	30.43	79.37	43.09	20.12	86.37	32.04	15.00	89.68	25.30	10.84	92.41	19.16	5.68	96.65	10.65

Table 14: Results on REVIEWS-F1000. Precision (P), Recall (R), and F1 at different cutoffs (K).

Domain	Source Sentence (Generated)	Target Sentence (Natural)
NEWS	In Southern Europe, the ripple effects of the disrupted flight schedules are being felt widely, with airlines cautioning that recovery from this crisis could take days or even weeks.	Some flights from Spain and Portugal, together with upwards of 4,000 flights across Northern Europe, have been affected, and the knock-on effect of aircraft and crews out of position could disrupt air travel worldwide for up to 72 hours.
NEWS	Newsweek's Steven Levy has pointed out that top-ranked blogs remain overwhelmingly white and male, underscoring a digital divide as real as any glass ceiling.	Steven Levy, a senior editor of Newsweek, has recently written a column addressing concerns about the over-representation of white males among top bloggers on the Internet.
PEER-REVIEWS	The inclusion of negative samples in the supervised fine-tuning pipeline appears to be non-trivial, as this variant underperforms relative to other supervised strategies.	Among all supervised methods, the SFT with negatives performs the worst, showing that using negative feedback in supervised training analogically to preference optimization is non-trivial.
PEER-REVIEWS	Coverage limitations of some rule-based components mean that a sizable fraction of utterances remain unhandled, which may impact the overall robustness.	P1.1 and P2 have to meet certain conditions to be applied, therefore they do not have full coverage of CS data: 36% and 31% for SEAME, 60% and 58% for Miami.

Table 15: Examples of sentence-level links in the NEWS-SYNTH and REVIEWS-SYNTH datasets. Source sentences are generated using an LLM conditioned on the target document, with sentence-level links specified during generation.

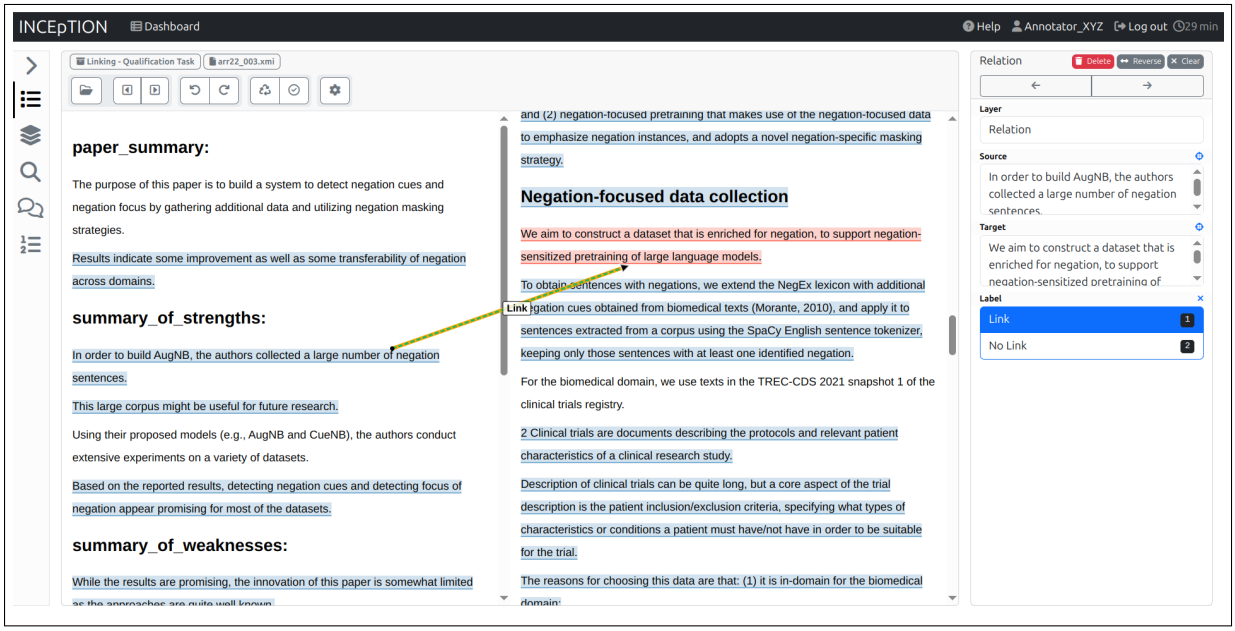


Figure 15: Annotation Interface.