# Evidence Retrieval for Fact Verification using Multi-stage Reranking

**Shrikant Malviya** and **Stamos Katsigiannis**
Department of Computer Science, Durham University, UK
{shrikant.malviya,stamos.katsigiannis}@durham.ac.uk

## Abstract

In the fact verification domain, the accuracy and efficiency of evidence retrieval are paramount. This paper presents a novel approach to enhance the fact verification process through a Multi-stage ReRanking (M-ReRank) paradigm, which addresses the inherent limitations of single-stage evidence extraction. Our methodology leverages the strengths of advanced reranking techniques, including dense retrieval models and list-aware rerankers, to optimise the retrieval and ranking of evidence of both structured and unstructured types. We demonstrate that our approach significantly outperforms previous state-of-the-art models, achieving a recall rate of 93.63% for Wikipedia pages. The proposed system not only improves the retrieval of relevant sentences and table cells but also enhances the overall verification accuracy. Through extensive experimentation on the FEVEROUS dataset, we show that our M-ReRank pipeline achieves substantial improvements in evidence extraction, particularly increasing the recall of sentences by 7.85%, tables by 8.29% and cells by 3% compared to the current state-of-the-art on the development set.

## 1 Introduction

The proliferation of false and misleading information, fuelled by the rapid progress in artificial intelligence (AI), poses a significant societal threat, as highlighted in the World Economic Forum's report (WEF, 2024). For example, the widespread mis/dis-information about COVID-19 vaccines has caused a surge in anti-vaccination sentiment online, leading to a low vaccination coverage (Islam et al., 2021). A recent study shows that low-veracity media-induced overconfidence exacerbates the adverse effects of widespread misinformation (i.e., fake news), especially in current global election scenarios (Kartal and Tyran, 2022). To combat this, researchers are focusing on developing automatic fact verification systems to prevent disinformation from spreading online (Guo et al., 2022).

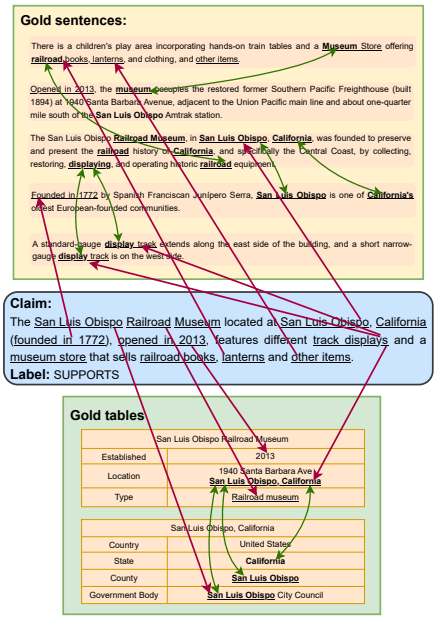



Figure 1: An example in FEVEROUS: The blue, yellow and green rectangle contains claim, sentence evidence, and table evidence, respectively. Arrows depict the interaction between two pieces of text. Keywords are underlined to show claim-evidence overlap and boldly highlighted to indicate intra-evidence interactions.

To answer the increasing demand for such systems, a number of datasets have been released, ranging from claims collected from fact-checking websites, e.g. LIAR (Wang, 2017), to complex collections of claims associated with proof-of-evidence, e.g. FEVER (Thorne et al., 2018), CLEF CheckThat! (Nakov et al., 2021), SemEval (Wang et al., 2021), FEVEROUS (Aly et al., 2021). In this paper, we focus on solving the FEVEROUS task, where the challenge is not only to extract evidence sentences/table cells from millions of passages (Wikipedia), but also to combine them to verify a given claim. Unlike other datasets, FEVEROUS proposes a real-world scenario where the evidence could be in both structured (e.g. Tables, lists) or unstructured format (e.g. sentences, passages).



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7295–7308

Key advancements on FEVEROUS task are not only on improving the claim verification procedure (Hu et al., 2022), but also focusing on evidence extraction in various formats (Hu et al., 2023; Wu et al., 2023). The DCUF, a fact-verification model introduced by (Hu et al., 2022), performs interaction of evidence in each format to improve the final verification accuracy, leaving the evidence extraction within each format separately. Recent works, e.g. UnifEE (Hu et al., 2023), SEE-ST (Wu et al., 2023), give attention to evidence extraction, focusing either on individual format or interaction across various formats. They mostly look for lexical (word-based) or semantic (meaning-based) overlaps between the claim and evidence pieces. They do not take into account how different pieces of evidence might relate to one another within the same format.

Figure 1 illustrates an example from FEVEROUS, where the goal is to extract both unstructured (e.g., sentences) and structured (e.g., tables or cells) evidence to verify a claim. The figure highlights two types of overlap: (1) between the claim and its associated evidence, and (2) among the evidence pieces themselves. Recognising interactions among evidence is crucial for determining the retrieval score of individual evidence. Critical evidence may not have obvious overlaps with the claim, but their relevance becomes clear when viewed in the context of other evidence. For instance, one piece of evidence might state the "Railroad Museum" is in "San Luis Obispo, California", while another mentions it opened in the year "2013". The underutilisation of interactions among these evidence pieces can lead to the omission of crucial information that could otherwise strengthen the verification process. Therefore, leveraging interactions between candidate evidence in each format is essential for effective evidence extraction.

In this paper, we propose the Multi-stage Reranking (M-ReRank) paradigm, which exploits overlapping among connected evidence candidates as collaborative filtering (Zhang et al., 2022b, 2023) to improve evidence extraction, thereby achieving higher accuracy in veracity prediction. To the best of our knowledge, this has been largely unexplored in the fact-verification domain. We design a novel pipeline, M-ReRank, which comprises a sequence of robust rerankers, e.g. Cross-Encoder (improved recall) (Humeau et al., 2019), HybRank (collaborative assessment) (Zhang et al., 2023), and HLATR (list-aware reranking) (Zhang et al., 2022b). It helps improve the first and second steps in the FEVEROUS pipeline, i.e. wiki-page retrieval and evidence extraction. Experiments on FEVEROUS show that our M-ReRank model significantly enhances evidence extraction performance and, consequently, boosts final fact verification scores. Detailed ablation experiments exhibit the effectiveness of M-ReRank in evidence extraction, showcasing how each component contributes to the overall improvement. A case study further highlights its role in accurately retrieving and utilising evidence for verification.

The contributions of this work can be summarised as follows: (i) We propose a Multi-stage ReRanking (M-ReRank) pipeline investigating how the retrieval and reranking architectures influence the evidence retrieval process. (ii) We show how evidence extraction can be improved by leveraging the relationships that exist among the evidence through collaborative filtering and list-aware reranking. (iii) Experiments show that our proposed multi-stage reranking significantly outperforms previous works on both the evidence extraction and the final verification accuracy. Detailed analysis reveals that our M-ReRank performs well in retrieving multi-hop evidence and combining evidence in both formats (e.g., sentences and tables).

## 2 Background

### 2.1 Multi-stage Text Retrieval

Traditionally, information retrieval has relied on lexical methods such as TFIDF and BM25 (Robertson and Zaragoza, 2009), treating queries and documents as sparse bag-of-words vectors and matching them at the token level. Recently, text retrieval systems armed with pre-trained language models have become a dominant paradigm to improve the overall performance where queries and documents are encoded into dense contextualised semantic vectors (Ren et al., 2021; Zhang et al., 2022a), and performing retrieval with optimised vector search algorithms (Johnson et al., 2021).

Recent approaches in reranking concatenate query-passage pairs and input them into a Transformer pre-trained on large corpora, allowing for more nuanced relevance estimation and improved retrieval outcomes through enhanced interaction (Humeau et al., 2019; Nogueira and Cho, 2020). However, these methods typically treat each candidate passage in isolation, neglecting valuable con-

textual information from other retrieved passages in the list. Some learning to rank techniques (Rahimi et al., 2016) and pseudo-relevance feedback approaches (Zhai and Lafferty, 2001; Zamani et al., 2016) leverage the ordinal relationship or list-wise context of retrieved documents for enhanced retrieval, a need corroborated in multi-stage retrieval systems (Liu et al., 2022). HybRank (Zhang et al., 2023) investigates collaboration among the candidate text in the retrieval lists and shows that collaborative filtering improves the precision of retrieval systems by exploiting linguistic aspects of sparse and dense retrieval methods. HLATR (Zhang et al., 2022b) has shown improved performance as a multi-stage text retrieval system by coupling features from both retrieval and reranking stages. We combine HybRank and HLATR in our M-ReRank pipeline.

### 2.2 Multi-stage Evidence Reranking for Fact-Verification

Multi-stage text retrieval can be highly beneficial for fact verification by enabling a more comprehensive and nuanced approach to rank the evidence and assess the veracity of claims or statements. Evidence in the same format also provide context information to each other. Past works on FEVEROUS mainly rely on using a single-stage evidence extraction (Aly et al., 2021; Bouziane et al., 2021; Saeed et al., 2021; Hu et al., 2022). Some methods propose to fuse evidence in different formats to leverage cross-format dependence but still leave the evidence extraction within each format separate (Hu et al., 2023; Wu et al., 2023). Utilising the collaboration that exists among candidate evidence has largely been unexplored for fact verification. Intuitively, for a specific claim, a set of evidence relevant to the claim tends to describe the same entities, events and relations (Lee et al., 2019), while irrelevant ones address a variety of unrelated topics.

### 2.3 FEVEROUS Task & Dataset

We use FEVEROUS[1] as the test bed for our approach because it is the only open-domain fact verification benchmark, to our knowledge, that integrates both unstructured and structured evidence. FEVEROUS has two main objectives: first, to extract sentences and table cells evidence from English Wikipedia and second, to predict the veracity of a given claim labelled as SUPPORTS, REFUTES, or NOT ENOUGH INFO (NEI). Each claim in the FEVEROUS dataset can be verified in multiple ways, represented by different evidence sets, each potentially comprising multiple pieces of evidence. For a response to be considered correct, participating systems only need to provide one complete evidence set. Hence, a prediction is considered correct only if at least one complete gold evidence set $E$ is a subset of the predicted evidence $\hat{E}$ and the predicted label is correct. Statistics for the FEVEROUS dataset are provided in Appendix A.

[1] https://fever.ai/dataset/feverous.html

## 3 Our Approach

The aim of the FEVEROUS open-domain fact verification (Aly et al., 2021) benchmark's task is to verify a claim $c$ based on content from Wikipedia. We follow the widely adopted three-step pipeline, which involves (i) retrieving relevant pages from the Wikipedia dump, (ii) extracting sentences and table cells as evidence from these pages, and (iii) predicting the veracity label of the given claim based on the compiled evidence set. In this work, we explore improving the first and second steps—wiki-page retrieval and evidence extraction—by employing our multi-stage retrieval pipeline.

In the three-step pipeline, as shown in Figure 2, the Wikipedia pages are *first* retrieved and refined by our M-ReRank approach. The top five pages are then used to extract evidence of both formats (e.g., sentences and tables) in the *second step*. We train the models in the M-ReRank pipeline separately for page, sentence and table retrieval. Combining the first five sentences and five tables, we use SEE-ST's (Wu et al., 2023) cell-retriever to extract potential cell evidence. Finally, at the *third* step, verification, we utilise DCUF (Hu et al., 2022), a method that converts evidence into dual-channel encodings to verify the claim.

### 3.1 Wikipedia Page Retrieval

Firstly, given a claim $c$, a set of relevant Wikipedia pages $\mathcal{P}$=$[p_1, p_2, ..., p_{n_p}]$ are retrieved from TFIDF and BM25-based retrievers to narrow down the search space from millions of pages to a few hundred (Robertson and Zaragoza, 2009). We combine the results from TFIDF and BM25, retaining the top $n_p$ documents. TFIDF is effective at capturing the importance of terms within a document, and across the corpus, while BM25 is a probabilis-

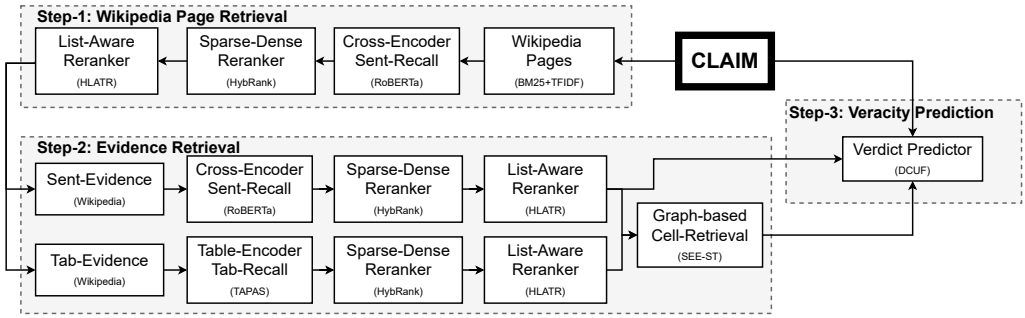



Figure 2: Overview of the M-ReRank pipeline for claim-based evidence retrieval and veracity prediction.

tic model that adjusts term weights based on term frequency saturation and document length normalisation. The retrieved pages are further reordered by robust upstream retrievers within the proposed M-ReRank pipeline, as shown in Figure 2 (Step-1).

### 3.2 Evidence Retrieval

Top five pages from the previous step are selected to extract the relevant evidence for veracity prediction. We use Cross-Encoder (Humeau et al., 2019) to extract $k$ sentences $S$=$\{s_i\}_{i=1}^{k}$ and TAPAS[2] based SEE-ST (Wu et al., 2023) model to extract $n$ tables $T$=$\{t_i\}_{i=1}^{n}$. The set of initial sentence and table evidence are then reordered by our M-ReRank (see Figure 2). All the models in the proposed multi-stage pipeline are trained separately using the FEVEROUS dataset's train and dev splits. Based on the extracted sentence/table evidence, we use the Graph-based cell retriever proposed by (Wu et al., 2023), which leverages the row and column semantics of tables to retrieve $r$ cell evidence $C$=$\{c_i\}_{i=1}^{r}$.

### 3.3 Multi-stage ReRanking (M-ReRank)

Once the initial set of documents, e.g. pages, sentences, tables, are retrieved, the proposed M-ReRank framework reorders them by prioritising their relevance to the given claim based on contextual understanding and semantic similarity. Initially, unstructured candidates like sentences, are reranked using a Cross-Encoder (Humeau et al., 2019). Subsequently, we utilise advanced rerankers HybRank (Zhang et al., 2023) and HLATR (Zhang et al., 2022b) in the pipeline. HybRank leverages both sparse and dense information to enhance reranking, while HLATR integrates retrieval and reranking features for hybrid list-aware reranking.

For tables, the reranking pipeline starts with the SEE-ST model (Wu et al., 2023), which is effective in capturing the row and column relevance of tables, thereby achieving a more precise extraction of structured candidates. As depicted in Figure 2, the retrieved tables are further reranked sequentially by HybRank and HLATR. Both rerankers take the flattened table (linearised as a single line) as input. After all reranking stages, the retrieved tables and sentences are utilised by SEE-ST's cell-retriever to extract relevant cells.

The following subsections provide detailed discussions of the proposed pipeline.

#### 3.3.1 Cross-Encoder with Contrastive Learning

(Humeau et al., 2019) showed that cross-encoders typically outperform bi-encoders on sentence-scoring tasks by enabling rich interactions between the claim and candidate evidence. In this stage, the claim and evidence are jointly encoded using a transformer architecture into a single vector as $E_s$=RoBERTa(claim, cand), "cand" represents the candidate evidence. The scoring mechanism involves reducing this embedding through multiple layers, including dropout ($D$), linear layers ($L_1$, $L_2$), and activation functions (relu $R$, sigmoid $\sigma$) to obtain a final score $S(\text{claim}, \text{cand}) = \sigma(L_2(R(L_1(D(E_s)))))$. The network is trained using contrastive learning criteria, aiming to minimise margin ranking loss between pairs of positive $x_1$ and negative $x_2$ candidate evidence:

$$\text{MRL}(x_1, x_2, y) = \max(0, -y \cdot (x_1 - x_2)) \quad (1)$$

where $x_1$ and $x_2$ are the predicted scores of positive and negative evidence, respectively. $y$ is set to 1, indicating a positive candidate ranked higher than the negative.

#### 3.3.2 Table Parser Contrastive Learning

SEE-ST (Wu et al., 2023) showed that leveraging both row and column semantics significantly improves the recall of structured evidence, e.g. tables, table-cells. SEE-ST begins by extracting ta-

[2]TAPAS: Table Parser (Herzig et al., 2020)

bles from selected Wikipedia pages targeting the most relevant rows and columns for the given claim, thereby minimising confusion from irrelevant cells. First, the claim and table pair are fed to TAPAS, a pre-trained table model aware of table structures (Herzig et al., 2020), to generate table embedding. In parallel, TAPAS tokenizer also provides row ($R_{pool}$) and column ($C_{pool}$) pooling matrix as $E_t, Row_{pool}, Col_{pool}$=TAPAS(claim,table) which are later used for estimating table, row and column losses $L_r$, $L_c$, respectively, and final loss $L$:

$$\begin{aligned} L_r &= CrE(R(L(R_{pool}E_t))) \\ L_c &= CrE(R(L(C_{pool}E_t))) \\ Lt &= \sigma(R(L(E_t))) \\ L_t &= \text{MRL}(Lt_{pos}, Lt_{neg}, 1) \\ L &= \alpha_t L_r + \beta_t L_c + \gamma_t L_t \end{aligned} \tag{2}$$

Since a cell represents the intersection of a row and a column, its relevance can be determined by analysing both dimensions. During inference, the table score is estimated based on criteria such as $L_r + L_c$, $L_r \times L_c$, $L_r$ or $L_c$. For the Table retrieval task, $L_r \times L_c$ provides higher retrieval accuracy:

$$S(\text{claim}, \text{table}) = L_r \times L_c \tag{3}$$

#### 3.3.3 HybRank

HybRank (Zhang et al., 2023) utilises the strategy of collaborative filtering (Goldberg et al., 1992) by incorporating lexical and semantic properties of both sparse and dense retrievers in reranking. We utilise BM25 as sparse and RoBERTa as dense retriever to rerank the candidates for a given claim through a 3-stage process:

**(a) Retrieval Stage:**

**Sparse Retrieval:** Given the claim $c$ and the candidate $d$, the BM25 score is obtained by summing the BM25 weights over the terms that co-occur in $c$ and $d$. Refer to (Robertson and Zaragoza, 2009) for more details about BM25.

**Dense Retrieval:** The relevance score is estimated as the dot product of encoded claim $c$ and candidate $d$, as $S_d(c, d) = E(c)^\top E(d)$, where $E(\cdot)$ denotes the encoder (RoBERTa), which generates the embeddings for the claim and candidate text.

**(b) Collaborative Filtering Stage:** The collaborative filtering stage leverages the sparse and dense scores between candidates, distinguishing positive ones in the retrieval list. For each candidate and claim, a sequence of similarity scalars $x_{d_i}$=$[s_{i1}, s_{i2}, ..., s_{iL}] \in \mathbb{R}$ is estimated with a set of Top-$L$ anchors from both sparse and dense scores. After applying softmax and min-max normalisation, the sparse and dense scores are stacked in a dual channel manner $x_{ij} = [s_{ij}^{\text{sparse}}, s_{ij}^{\text{dense}}] \in \mathbb{R}^2$. Thus, the similarity sequence vector becomes like $X_{d_i} = [x_{i1}, x_{i2}, ..., x_{iL}] \in \mathbb{R}^{L\times 2}$. This dual-channel similarity vector is transformed to $D$ dimensions with a trainable projection layer $e_{ij} = x_{ij}W$, where $W \in \mathbb{R}^{2\times L}$ is a learnable parameter and $e_{ij} \in \mathbb{R}^D$ are embedded similarities. Thereafter, candidate $d_i$ becomes a sequence of similarity embeddings $E_{d_i} = [e_{i1}, e_{i2}, ..., e_{iL}] \in \mathbb{R}^{L\times D}$, which consists of candidate $d_i$ similarity information with anchor list. As a result, we obtain a total of $N_d + 1$ collaborative sequences, where each sequence corresponds to either a candidate or a query and incorporates both lexical and semantic similarity information with respect to $L$ anchors.

**(c) Aggregation Reranking Stage:** To perform anchor-wise interaction, we gather the $j$-th similarity embedding $e_j^*$ from the claim sequence and all candidate sequences, refining them using a Transformer encoder as:

$$e'_{cj}, e'_{1j}, \ldots, e'_{N_dj} = \text{Trans}_{\text{inter}}(e_{cj}; e_{1j}; \ldots; e_{N_dj}) \tag{4}$$

where, $e'_{*j} \in \mathbb{R}^D$. This transforms the similarity embedding sequence $E_*$ to $E'_*$. We transform these sequences into dense vectors by consolidating the refined similarity embeddings. Specifically, we add a [CLS] token at the beginning of the collaborative sequence, process it through another Transformer encoder, and take the output of the [CLS] token as the representation of candidate $d_i$ and claim $c$ as:

$$h_{d_i} = \text{Trans}_{\text{aggr}}([\text{CLS}] \oplus E'_{d_i})[\text{CLS}] \tag{5}$$

$$h_c = \text{Trans}_{\text{aggr}}([\text{CLS}] \oplus E'_c)[\text{CLS}] \tag{6}$$

where [CLS] $\in \mathbb{R}^{1\times D}$ and $\oplus$ denotes the concatenation operation. Finally, the dot product between encoded vector $h_{d_i}$ of candidate and claim vector $h_c$ determines the similarity score.

#### 3.3.4 HLATR

HLATR (Zhang et al., 2022b) improves text retrieval by combining retrieval and reranking features using a lightweight transformer encoder. As a retrieve-then-reranking architecture, HLATR follows a three-stage pipeline: (a) the *Retrieval Stage*

identifies potentially relevant documents, (b) the *Reranking Stage* refines the relevance scores of the retrieved documents, and (c) the *HLATR Stage* consists of a multi-stage feature fusion layer and a transformer encoder to further improve the ranking:

**(a) Retrieval Stage:** In the Retrieval Stage, we consider the candidate documents retrieved from previous modules in our pipeline, e.g. HybRank, Cross-Encoder/SEE-ST, instead of using a separate dense retrieval model, as the original HLATR algorithm suggests.

**(b) Reranking Stage:** The Reranking Stage further refines the retrieval scores using an interaction-based model, e.g. Cross-Encoder. Each claim-candidate pair $(c, d)$ is rescored as $\text{score}(c, d) = f(E(c, d))$, where, $E(\cdot,\cdot)$ denotes the encoder (RoBERTa), and $f$ is the score function, e.g. $\sigma$ (sequence classifier). Training involves a contrastive learning objective ($L_c$), optimising the model with groups of $(c, d)$ pairs consisting of one positive candidate $d^+$ and multiple negatives as:

$$L_c = -\log \frac{\exp(\text{score}(c, d^+))}{\sum_{p \in G_d} \exp(\text{score}(c, d))} \quad (7)$$

**(c) HLATR Stage:** The core of this reranking paradigm is the HLATR component, which features a multi-stage fusion layer and a transformer encoder. It enhances the reranking results by combining features from both retrieval and reranking stages, creating a comprehensive representation. The combined features are processed through a lightweight transformer encoder, which models the interactions among all candidates, highlighting mutual relationships. The combined relevance score in HLATR is formulated as follows:

$$\text{score}_{\text{HLATR}}(c, D_r) = f_{\text{HLATR}}(E_{\text{HLATR}}(c, D_r)) \quad (8)$$

where $D_r$ represents a candidates list to be reranked, $E_{\text{HLATR}}$ is the encoder that processes the combined features, and $f_{\text{HLATR}}$ is the final relevance estimation function. Like the previous stage, this stage is also optimised with a list-wise contrastive loss, as defined by Eq 7.

# 4 Experimental Evaluation

## 4.1 Evaluation Metrics

In the FEVEROUS task, two primary official metrics are employed: accuracy (Acc.) and the FEVEROUS-score (F.S). Accuracy measures the proportion of instances for which the model correctly predicts the veracity label. The FEVEROUS-score evaluates not only the correctness of the final veracity label but also the adequacy of the extracted evidence set. It quantifies the proportion of instances where the extracted evidence aligns with one of the gold sets, and the predicted veracity label matches the gold standard. Three additional official metrics are utilised to assess the quality of extracted evidence sets in the FEVEROUS task: Evidence Precision (E-P), Evidence Recall (E-R), and Evidence F1 (E-F1). It also provides multiple gold evidence sets for each claim, and a piece of extracted evidence is deemed correct only if it is included in any of the gold evidence sets. For each instance, evidence precision represents the proportion of correctly predicted evidence. The overall evidence precision is determined by averaging this score across all instances. Evidence recall measures the proportion of instances with a correctly extracted evidence set, where correctness is defined by covering any of the gold evidence sets. Lastly, evidence F1 is the harmonic mean of evidence precision and recall, providing a balanced assessment of precision and recall in evidence extraction.

| Models | Page | Sentence | Table | Cell | Evidence |
|---|---|---|---|---|---|
| Baseline | 63 | 53 | 56 | 29 | 30 |
| FaBULOUS | 63 | 56.6 | - | 34.2 | 40.4 |
| DCUF | 85.20 | 62.54 | 75.59 | 58.41 | 43.22 |
| UnifEE | 85.20 | 75.59 | 75.36 | 67.44 | 55.08 |
| SEE-ST | 85.20 | 75.50 | 80.86 | 77.16 | 61.43 |
| **M-ReRank (ours)** | **93.63** | **83.35** | **89.15** | **80.16** | **66.69** |

Table 1: Recall of different formats of evidence on the development set.

## 4.2 Implementation Details

Implementation details for all the algorithms used, as well as training hyperparameters, are provided and discussed in Appendix B. To understand complexity and scalability impacts, we also perform computational analysis, included in Appendix C

## 4.3 Main Results

**Evidence Extraction Results:** Table 1 presents the evidence extraction results of our M-ReRank pipeline on the development set and compares it with the recent state-of-the-art. Previous methods, such as the official baseline (Aly et al., 2021) and FaBULOUS (Bouziane et al., 2021), employ a weaker document retrieval module, i.e. BM25/TFIDF, leading to error propagation and lower evidence recall. Recent methods,

| Models | Development set | | | | | Test set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F.S | Acc. | E-P | E-R | E-F1 | F.S | Acc. | E-P | E-R | E-F1 |
| Official Baseline | 19 | 53 | 12 | 30 | 17 | 17.73 | 48.48 | 10.17 | 28.78 | 15.03 |
| EURECOM | 19 | 53 | 12 | 29 | 17 | 20.01 | 47.79 | 13.73 | 33.73 | 19.52 |
| Z team | - | - | - | - | - | 22.51 | 49.01 | 7.76 | 42.64 | 13.12 |
| CARE | 26 | 63 | 7 | 37 | 12 | 23 | 53 | 7 | 37 | 11 |
| NCU | 29 | 60 | 10 | 42 | 17 | 25.14 | 52.29 | 9.91 | 39.07 | 15.81 |
| Papelo | 28 | 66 | - | - | - | 25.92 | 57.57 | 7.16 | 34.60 | 11.87 |
| FaBULOUS | 30 | 65 | 8 | 43 | 14 | 27.01 | 56.07 | 7.73 | 42.58 | 13.08 |
| DCUF | 35.77 | 72.91 | 15.06 | 43.22 | 22.34 | 33.97 | 63.21 | 14.79 | 44.10 | 22.15 |
| UnifEE | 44.86 | 73.67 | **19.04** | 55.08 | **28.30** | 41.50 | 65.04 | **18.35** | 53.87 | **27.37** |
| SEE-ST | 49.73 | 74.68 | 10.60 | 61.43 | 18.07 | 44.75 | 65.16 | 9.81 | 60.01 | 16.89 |
| **M-ReRank (ours)** | **60.57** | **87.58** | 10.68 | **66.69** | 18.40 | **47.13** | **65.24** | 10.35 | **63.71** | 17.81 |

Table 2: Model performance on the development set and test set. F.S is FEVEROUS-score, and Acc. is the accuracy of veracity labels. E-R, E-P, and E-F1 are recall, precision, and F1, which are computed based on the evidence set.

| **Retriever** | **T-150** | **T-5** |
|---|---|---|
| Baseline TFIDF | 91.43 | 62.71 |
| TFIDF (T) | 92.33 | 69.46 |
| BM25 (B) | 90.53 | 71.40 |
| Ensemble$_{(T,B)}$ (E) | 94.87 | 73.98 |
| E+Cross-Encoder (C) | 94.87 | 87.14 |
| E+HybRank (Hy) | 94.87 | 90.83 |
| E+HLATR (Hl) | 94.87 | 92.90 |
| E+C+Hy | 94.87 | 91.39 |
| E+C+Hy+Hl | 94.87 | **93.63** |

Table 3: Wikipedia page retrieval results with rerankers in our M-ReRank pipeline in Top-150/5 settings.

| **T-5 Pages** | **Retriever** | **T-100** | **T-20** | **T-5** |
|---|---|---|---|---|
| E | TFIDF | 71.56 | 67.11 | 54.38 |
| | RoBERTa (R) | 88.13 | 80.35 | 77.13 |
| | R+HybRank (HyS) | 88.13 | 86.65 | 78.76 |
| | R+HLATR (HlS) | 88.13 | 86.99 | 79.09 |
| | R+HyS+HlS | 88.13 | 87.02 | 80.06 |
| E+C+Hy+Hl | TFIDF | 90.03 | 82.65 | 67.30 |
| | RoBERTa (R) | 92.36 | 90.15 | 80.33 |
| | R+HybRank (HyS) | 92.36 | 90.50 | 81.73 |
| | R+HLATR (HlS) | 92.36 | 89.92 | 82.49 |
| | R+HyS+HlS | 92.36 | 90.60 | **83.35** |

Table 4: Sentence retrieval results with various rerankers in our M-ReRank pipeline in Top-100/20/5 settings.

DCUF, UnifEE, SEE-ST, utilise ensemble of Cross-Encoder[3] and BM25, improving page recall by 85.20%. However, limited page retrieval limits the overall evidence recall and, consequently, low accuracy in veracity prediction. Our multi-stage reranking improves the page recall by 8.43%. Notably, M-ReRank extracts 36% more gold-standard evidence compared to the official baseline and 5.26% compared to the best model SEE-ST. Through M-ReRank, we obtain substantial recall jump in all formats of evidence retrieval. This is also proved by our ablation study in (§4.4).

**Overall Results:** Our primary results, summarised in Table 2, demonstrate significant performance improvement in evidence extraction compared to the previous best models, i.e. DCUF, UnifEE, SEE-ST, thereby improving FEVEROUS-score (F.S) overall. Specifically, our model shows improvements of 5.26%/3.70% in evidence recall on the development/test set, respectively. Adopting the verification approach from (Hu et al., 2022), we achieved accuracy rates of 87.58% on the development set and 65.24% on the test set. These gains indicate that by leveraging context information from other evidence in the candidate list, our multi-stage reranking (M-ReRank) enhances the accuracy of evidence extraction.

Following the constraint on selecting the maximum number of sentences and cells, there are two ways to construct an evidence set. One way is to apply a threshold when selecting evidence to prioritise high precision, slightly sacrificing recall. For example, a former SOTA method, UnifEE, follows the same criteria for high precision, but the label accuracy remains largely unaffected by changes in the evidence set. We employ the maximum number of sentences and cells as constraints, keeping higher evidence recall as another way of constructing the evidence set (an ablation study on precision/recall tradeoff is presented in Appendix D). Demonstrating the effectiveness of our approach, an example of evidence extraction in both formats is presented in Appendix E.

[3] cross-encoder/ms-marco-MiniLM-L-12-v2

| T-5 Pages | Retriever | T-20 | T-5 | T-3 |
|---|---|---|---|---|
| E | TFIDF | 82.17 | 80.84 | 76.89 |
| | SEE-ST (S) | 88.84 | 86.27 | 83.99 |
| | S+HybRank (HyT) | 88.84 | 87.33 | 84.29 |
| | S+HLATR (HlT) | 88.84 | 87.45 | 85.23 |
| | S+HyT+HlT | 88.84 | 87.52 | 85.35 |
| E+C+Hy+Hl | TFIDF | 89.30 | 85.75 | 79.83 |
| | SEE-ST (S) | 93.40 | 88.44 | 86.87 |
| | S+HybRank (HyT) | 93.40 | 90.81 | 88.54 |
| | S+HLATR (HlT) | 93.40 | 90.83 | 88.65 |
| | S+HyT+HlT | 93.40 | 91.61 | **89.15** |

Table 5: Tables retrieval results with various rerankers in our M-ReRank pipeline in Top-20/5/3 settings.

The test set accuracy is typically lower than the development set accuracy. This discrepancy is primarily due to the unequal distribution of NEI (Not Enough Information) claims across the different splits. Our analysis of verdict prediction results reveals that DCUF underperforms on NEI instances, which accounts for the accuracy gap between the development and test sets.

### 4.4 Ablation Study

To evaluate the effectiveness of M-ReRank, we conducted a series of *ablation experiments* focusing on three aspects: i) Wikipedia page retrieval, ii) sentence extraction, and iii) table extraction. We first examined the impact of each reranker in M-ReRank by applying them individually. Subsequently, we applied them in a multi-stage manner, prioritising the order based on their individual performance to understand the cumulative effect. Since M-ReRank obtains the maximum number of Wikipedia pages, we also experiment with extracting sentence and table evidence solely from pages retrieved by $\text{Ensemble}_{(\text{T,B})}$, excluding M-ReRank for page retrieval as shown in Table 4 and Table 5. This enables a fairer comparison of rankers in the M-ReRank pipeline for sentence and table retrieval.

**Wikipedia Page Retrieval:** Table 3 presents the recall of various methods for *Wikipedia page retrieval*, ranging from FEVEROUS's baseline TFIDF to all rerankers in the M-ReRank pipeline. The FEVEROUS baseline achieves a 91.43% recall in the Top-150 setting but struggles to maintain relevance in the Top-5. By converting Unicode characters to their nearest ASCII equivalents during pre-processing, we observe a 6.75%/8.69% improvement by TFIDF(T) and BM25(B), respectively, in Top-5 recall. Further improvements are seen by applying ensemble reranking on the T and B results, increasing the page recall to 94.87% for Top-150 and 73.98% for Top-5 settings. We see a significant jump in page recall specific to Top-5 retrieval on applying Neural rankers, e.g. Cross-Encoder, HybRank, and HLATR, by 13.16%, 16.85%, and 18.92%, respectively. When combined (E+C+Hy+Hl), they achieve the highest page recall of 93.63% in the Top-5 setting.

**Sentence Extraction:** Table 4 depicts the ablation results on *sentence retrieval*. To show the effectiveness of M-ReRank based rerankers, we perform ablation with the Top-5 pages retrieved by earlier step via both $\text{Ensemble}_{\text{T,B}}$ and E+C+Hy+Hl settings. M-ReRank performs well for sentence retrieval in both scenarios. RoBERTa-based Cross-Encoder improves sentence recall in both cases by 22.75% and 13.03%. Using the RoBERTa results, the other rankers, HybRank, and HLATR, consistently achieve higher recall. In the E+C+Hy+Hl setting, M-ReRank achieves the highest recall by 83.35% for sentence retrieval, which is 7.85% higher than the previous SOTA methods.

**Table Extraction:** Table 5 display the effectiveness of M-ReRank on *table retrieval*. Like the ablation experiments of sentence extraction, we again choose the Wikipedia pages retrieved via both $\text{Ensemble}_{\text{T,B}}$ and E+C+Hy+Hl settings to fairly compare the rerankers' strengths. The retrievers' performance is compared on Top-3/5/20 recall. SEE-ST (Wu et al., 2023) has shown a significant recall improvement of 3-7% compared to the TFIDF baseline by incorporating row and column semantics. M-ReRank retrievers reorder the table candidates in flattened form. For retrieved pages in both $\text{Ensemble}_{\text{T,B}}$ and E+C+Hy+Hl setting, M-ReRank consistently improves the table recall, similar to that found in the sentence extraction. We observe a jump of 1.36% and 2.28% table recall in E and E+C+Hy+Hl settings, respectively.

In conclusion, M-ReRank performs well on evidence reranking, which is crucial for fact-checking systems. It demonstrates superior performance in reranking unstructured evidence, e.g., sentences and passages, compared to structured evidence. The reason is that structured evidence retrieval requires row and column semantics information, which is crucial for structured evidence retrieval. On the other hand, M-ReRank performs retrieval on the flattened table. However, it is still able to perform collaborative filtering by exploiting interaction among table candidates. Further analysis of the errors of M-ReRank is provided in Appendix F.

## 5 Conclusion

In this paper, we presented M-ReRank, a multi-stage reranking framework designed to enhance the evidence retrieval process for fact verification tasks. Our experiments with the FEVEROUS dataset demonstrate that M-ReRank significantly improves the recall of evidence extraction, achieving a FEVEROUS-Score jump of 10.84%/2.38% on development/test data compared to previous state-of-the-art methods. M-ReRank pipeline is comprised of a sequence rerankers, e.g. Cross-Encoder/SEE-ST, HybRank, HLATR. By leveraging the contextual interactions among multiple evidence pieces and incorporating both lexical and semantic similarities, M-ReRank effectively addresses the challenges of retrieving relevant evidence in both unstructured formats, e.g. sentences, and structured formats, e.g. tables or cells. Our ablation studies further validate the efficacy and contribution of each reranking stage, showcasing the robustness and adaptability of our approach. Overall, M-ReRank sets a new benchmark in the domain of fact verification, paving the way for more accurate and reliable verification systems.

Future work could explore more efficient multi-stage reranking architectures that reduce computational overhead while maintaining accuracy. This might involve developing lightweight rerankers or incorporating parallel processing techniques (Sanh et al., 2020; Leonhardt et al., 2024). Moreover, further investigation is needed into handling the underrepresented NEI category, possibly by adopting advanced data augmentation strategies or rebalancing techniques to improve the model's performance on rare labels (Tayyar Madabushi et al., 2019; Cao et al., 2019). Additionally, expanding the model's capacity to better utilise inter-evidence interactions in more complex evidence structures, such as multi-modal data, would also be a valuable direction.

## 6 Limitations

Despite the promising results, our multi-stage reranking approach has several limitations that need addressing in future work. One significant challenge is the computational complexity introduced by the multi-stage process, which can lead to increased processing time and resource consumption, making real-time applications challenging. Additionally, scalability issues arise when handling large-scale datasets like the extensive Wikipedia corpus, potentially impacting the system's performance. The model's dependence on high-quality data means incomplete or noisy data can significantly reduce retrieval and verification performance.

Another limitation arises from the imbalance in the distribution of the three veracity labels. Specifically, as detailed in Appendix A, The NEI (Not Enough Information) label accounts for only 3% of the training data, making it difficult for models to predict accurately.

## Acknowledgements

The authors in this project have been funded by UK EPSRC grant "AGENCY: Assuring Citizen Agency in a World with Complex Online Harms" under grant EP/W032481/2.

## References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics. 1, 3, 6

Mostafa Bouziane, Hugo Perrin, Amine Sadeq, Thanh Nguyen, Aurélien Cluzeau, and Julien Mardas. 2021. FaBULOUS: Fact-checking Based on Understanding of Language Over Unstructured and Structured information. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 31–39, Dominican Republic. Association for Computational Linguistics. 3, 6

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 9

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank aggregation methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 613–622, New York, NY, USA. Association for Computing Machinery. 11

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70. 5

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206. 1

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics. 4, 5

Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. 2022. Dual-Channel Evidence Fusion for Fact Verification over Texts and Tables. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5232–5242. Association for Computational Linguistics. 2, 3, 7

Nan Hu, Zirui Wu, Yuxuan Lai, Chen Zhang, and Yansong Feng. 2023. UnifEE: Unified Evidence Extraction for Fact Verification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1150–1160. Association for Computational Linguistics. 2, 3, 12

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *International Conference on Learning Representations*. 2, 4

Md Saiful Islam, Abu-Hena Mostofa Kamal, Alamgir Kabir, Dorothy L. Southern, Sazzad Hossain Khan, S. M. Murshid Hasan, Tonmoy Sarkar, Shayla Sharmin, Shiuli Das, Tuhin Roy, Md Golam Dostogir Harun, Abrar Ahmad Chughtai, Nusrat Homaira, and Holly Seale. 2021. COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. *PLOS ONE*, 16(5):e0251605. 1

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547. 2

Melis Kartal and Jean-Robert Tyran. 2022. Fake News, Voter Overconfidence, and the Quality of Democratic Choice. *American Economic Review*, 112(10):3367–3397. 1

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics. 3

Jurek Leonhardt, Henrik Müller, Koustav Rudra, Megha Khosla, Abhijit Anand, and Avishek Anand. 2024. Efficient Neural Ranking Using Forward Indexes and Lightweight Encoders. *ACM Trans. Inf. Syst.*, 42(5):117:1–117:34. 9

Weiwen Liu, Yunjia Xi, Jiarui Qin, Fei Sun, Bo Chen, Weinan Zhang, Rui Zhang, and Ruiming Tang. 2022. Neural Re-ranking in Multi-stage Recommender Systems: A Review. In *Thirty-First International Joint Conference on Artificial Intelligence*, volume 6, pages 5512–5520. 3

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *Advances in Information Retrieval*, pages 639–649. Springer. 1

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. (arXiv:1901.04085). 2

Razieh Rahimi, Azadeh Shakery, Javid Dadashkarimi, Mozhdeh Ariannezhad, Mostafa Dehghani, and Hossein Nasr Esfahani. 2016. Building a multi-domain comparable corpus using a learning to rank method. *Natural Language Engineering*, 22(4):627–653. 3

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. PAIR: Leveraging Passage-Centric Similarity Relation for Improving Dense Passage Retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183, Online. Association for Computational Linguistics. 2

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389. 2, 3, 5

Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphael Troncy, and Paolo Papotti. 2021. Neural Re-rankers for Evidence Retrieval in the FEVEROUS Task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 108–112, Dominican Republic. Association for Computational Linguistics. 3

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. (arXiv:1910.01108). 9

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China. Association for Computational Linguistics. 9

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. 1

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics. 1

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics. 1

WEF. 2024. Global Risks Report 2024. https://www.weforum.org/publications/global-risks-report-2024/. 1

Zirui Wu, Nan Hu, and Yansong Feng. 2023. Enhancing Structured Evidence Extraction for Fact Verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6631–6641, Singapore. Association for Computational Linguistics. 2, 3, 4, 8, 12

Hamed Zamani, Javid Dadashkarimi, Azadeh Shakery, and W. Bruce Croft. 2016. Pseudo-Relevance Feedback Based on Matrix Factorization. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1483–1492, New York, NY, USA. Association for Computing Machinery. 3

Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410, New York, NY, USA. Association for Computing Machinery. 3

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022a. Adversarial Retriever-Ranker for Dense Text Retrieval. In *International Conference on Learning Representations*. 2

Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. 2022b. HLATR: Enhance Multi-stage Text Retrieval with Hybrid List Aware Transformer Reranking. *arXiv preprint arXiv:2205.10569*. 2, 3, 4, 5, 12

Zongmeng Zhang, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. Hybrid and Collaborative Passage Reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14003–14021, Toronto, Canada. Association for Computational Linguistics. 2, 3, 4, 5, 12

# A FEVEROUS Dataset Statistics

FEVEROUS is based on English Wikipedia, which contains a vast collection of 95.6 million sentences and 11.8 million tables. Within this dataset, there are 87,026 distinct claims, each with an average length of 25.3 words. On average, verifying each claim requires referencing 1.4 sentences and 3.3 cells (equivalent to 0.8 tables). Notably, evidence for verification is exclusively text-based in 34,963 cases, solely table-based in 28,760 cases, and a combination of both formats in 24,667 instances. Among these claims, 49,115 are classified as SUPPORTS, 33,669 as REFUTES, and the remaining 4,242 are categorised as NEI. Table 6 provides a detailed breakdown of label and evidence distributions across different splits.

| | Train | Dev | Test |
|---|---|---|---|
| Supported | 41,835(59%) | 3,908(50%) | 3,372 (43%) |
| Refuted | 27,215(38%) | 3,481(44%) | 2,973 (38%) |
| NEI | 2,241 (3%) | 501 (6%) | 1,500 (19%) |
| Total | 71,291 | 7,890 | 7,845 |
| Sentences | 31,607(41%) | 3,745(43%) | 3,589 (42%) |
| Cells | 25,020 (32%) | 2,738(32%) | 2,816 (33%) |
| Sentence+Cells | 20,865 (27%) | 2,468 (25%) | 2,062 (24%) |

Table 6: Details of each split in FEVEROUS. The first three rows highlight the distribution of classes across the splits, and the last three rows represent the distribution of claims in each split requiring only sentence evidence, cell evidence, or both, respectively.

# B Implementation Details

In the document retrieval step, we retrieve $n_p = 5$ pages from the Wikipedia dump for each claim. As a first step, 150 pages per claim are extracted by TFIDF and BM25 separately and merged together by ensemble reranking (Dwork et al., 2001) to retrieve a final set of 150 pages per claim. We keep the top 5 pages for evidence extraction after Multi-stage reranking. For the evidence retriever, the $n_k$=5 sentences and $n_t$=5 tables are extracted from the retrieved pages, and the sentence and table evidence are combined to extract $n_r = 25$ cells.

For Cross-Encoder, we use a RoBERTa-base[4] model, finetune it with contrastive learning criteria where for each positive example, a negative example is selected to determine MarginRanking loss as explained in (§3.3.1). The hyperparameters are set as batch size of 16 and learning rate $10^{-5}$.

[4] RoBERTa-base

For table extraction, we use SEE-ST (Wu et al., 2023) that encodes the claim-table pair by TAPAS-base[5] model. The hyperparameters are set to their default values as mentioned in (Wu et al., 2023), i.e. batch size of 8, learning rate $10^{-7}$ for TAPAS and $10^{-7}$ for the classifier, $\alpha_t = 1$ and $\beta_t = 1$.

For cell extraction, we use SEE-ST's evidence graph approach, which forms a graph of sentences and cells evidence and then scores each cell on the basis of row and column semantics. RoBERTa-base and TAPAS-base are used to encode sentence nodes and cell nodes in the graph. The hyperparameters are set as batch size of 4, learning rate $10^{-6}$, $\alpha_c = 2$, $\beta_c = 2$, and $\gamma_c = 1$.

In HybRank, the outputs from earlier step are used to extract sparse features by BM25 and dense features by a finetuned RoBERTa model[6]. Number of anchors are set to 100 for page/sentence retrieval and 20 during table retrieval. The remaining hyperparameters are set to default as mentioned in (Zhang et al., 2023).

In HLATR, retrieved candidates from the earlier step are used for reranking. We finetune a transformer model[7] to be used as reranker in the second step. Finetuning hyperparameters are batch size 4, train group size 16, learning rate $10^{-5}$, and number of epochs 5. In HLATR's third step, we finetune a lightweight RoBERTa-base model with reduced hidden_size as 128, num_attention_heads 2, and num_hidden_layers 4, with a learning rate $10^{-3}$, batch size 256, and 30 epochs.

| Model | Page | Sentence | Table |
|---|---|---|---|
| Base-Model† | 18 hrs | 12hrs 30min | 1hrs 46min |
| HybRank | 62s | 46s | 4s |
| HLATR | 45min+122s | 10min+90s | 3min+11s |

Table 7: Inference time on dev data using various models for Page, Sentence and Table retrieval. †Base-Model represents Cross-Encoder, RoBERTa and SEE-ST used respectively for {Page,Sentence,Table}-Retrieval.

## C Computational Analysis

To test the efficiency, we compare the inference time for different ranking models used in the pipeline, e.g. Base-Model, HybRank, and HLATR. All experiments were conducted on NVIDIA RTX 4090 24GB type GPUs.

In Table 7, we can observe that major time is taken by the Base-Model to retrieve the top candidate evidence, e.g. page, sentence or tables. The other models in the pipeline, HybRank (Zhang et al., 2023) and HLATR (Zhang et al., 2022b), take comparatively less time during inference. In contrast, the SOTA models, e.g. UnifEE (Hu et al., 2023), SEE-ST (Wu et al., 2023), use graph-based methods for joint evidence rescoring, which is more time-consuming.

[5]TAPAS-base

[6]sentence-transformers/msmarco-bert-base-dot-v5

[7]CoROM-Reranking

| Evidence | Recall | Precision | FEVEROUS-Score |
|---|---|---|---|
| Sentences=5, Cells=25 | 0.6669 | 0.1067 | 0.6058 |
| Sentences=5, Cells=15 | 0.6300 | 0.1462 | 0.5752 |
| Sentences=5, Cells=5 | 0.5258 | 0.2294 | 0.4833 |

Table 8: Ablation study illustrating the precision-recall tradeoff based on the number of evidence pieces considered, showing its impact on the FEVEROUS-score.

## D Precision/Recall Tradeoff

Our model has been optimised to achieve the highest FEVEROUS-score possible, as this is the best indicator of overall best performance. However, this indeed leads to a precision/recall tradeoff targeting a high FEVEROUS-score. We conducted an ablation study by changing the number of evidence considered, as shown in Table 8. By reducing the number of evidence considered, precision increases, but both recall and FEVEROUS-score decrease. We observe that keeping the maximum number of sentences and cells helps achieve higher evidence recall, leading to the best performance.

## E Case Study

A case is shown for evidence extraction of both sentence and table types in Figure 3. For the claim on *San Luis Obispo Railroad Museum*, our M-ReRank successfully retrieves sentences and tables of evidence by reordering them from the initial retrievals. We use RoBERTa (Cross-Encoder) and TAPAS (SEE-ST) retrieval results, respectively, for unstructured and structured evidence extraction. The main challenge in this case is multi-hop evidence extraction, where evidence is to be extracted from multiple sources to verify the claim. For sentence extraction, we observe that the initial retrieval could only retrieve three pieces of evidence in Top-5. Through M-ReRank, the pieces of evidence are rescored, retrieving the Top-5 from them. For instance, sentences with evidence ID *San Luis Obispo, California_sentence_6* and *San Luis Obispo Railroad Museum_sentence_6*, were earlier ranked six and

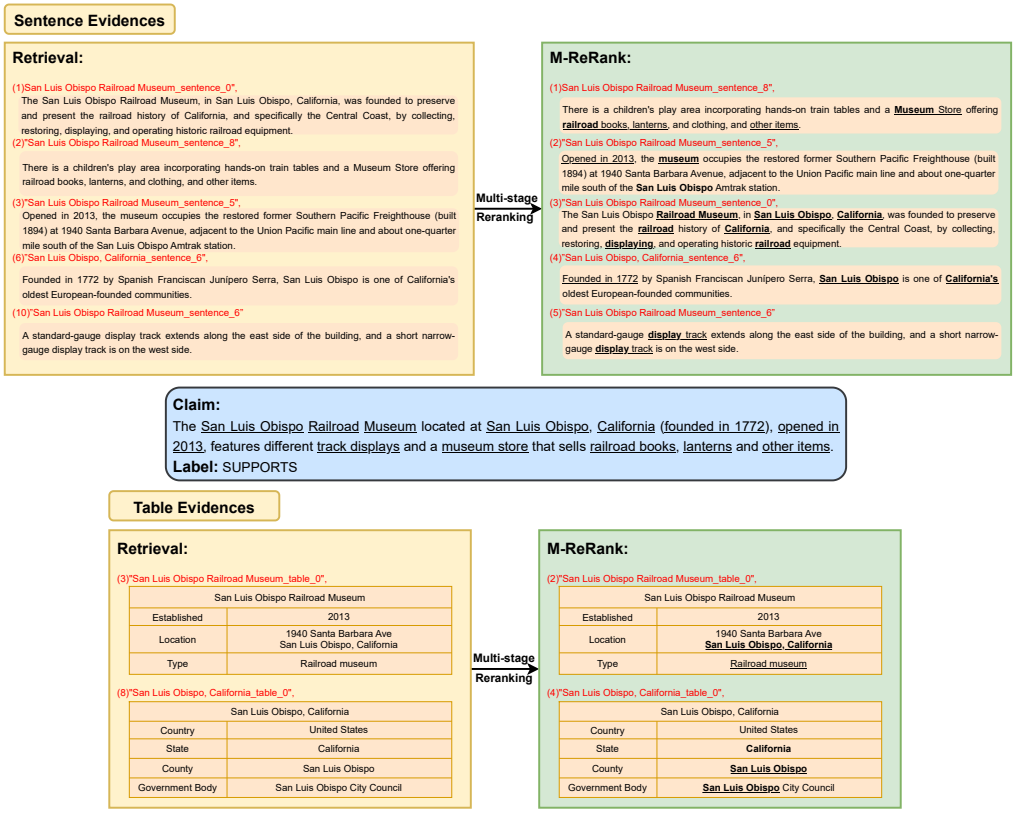



Figure 3: An example in FEVEROUS. The blue rectangle contains the claim. The yellow rectangle highlights the initially retrieved evidence (Retrieval), while the green rectangle depicts the reranked evidence set by our Multi-stage reranking (M-ReRank) paradigm. Text in red for each piece of evidence shows the order number (parenthesised) followed by its ID in the dataset. Words and phrases are underlined to show interactions between the claim and evidence, while bold text indicates inter-evidence interactions in the group, e.g. sentences or tables.

ten respectively, however, M-ReRank reranks them as four and five. Without this, the fact-verification model would be unable to confirm when *San Luis Obispo* was founded and what kind of *display track* the Railway Museum offers.

In structured evidence, the initial retrieval is unable to retrieve *San Luis Obispo, California_table_0* in Top-5, but M-ReRank reorders it to be included in Top-5 tables. It helps identify *San Luis Obispo* as a county in *California* state. This demonstrates M-ReRank's robustness in leveraging interactions among evidence to reorder them effectively, thereby improving overall evidence extraction in each format.

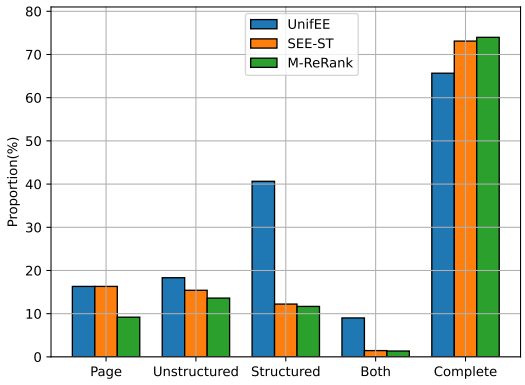


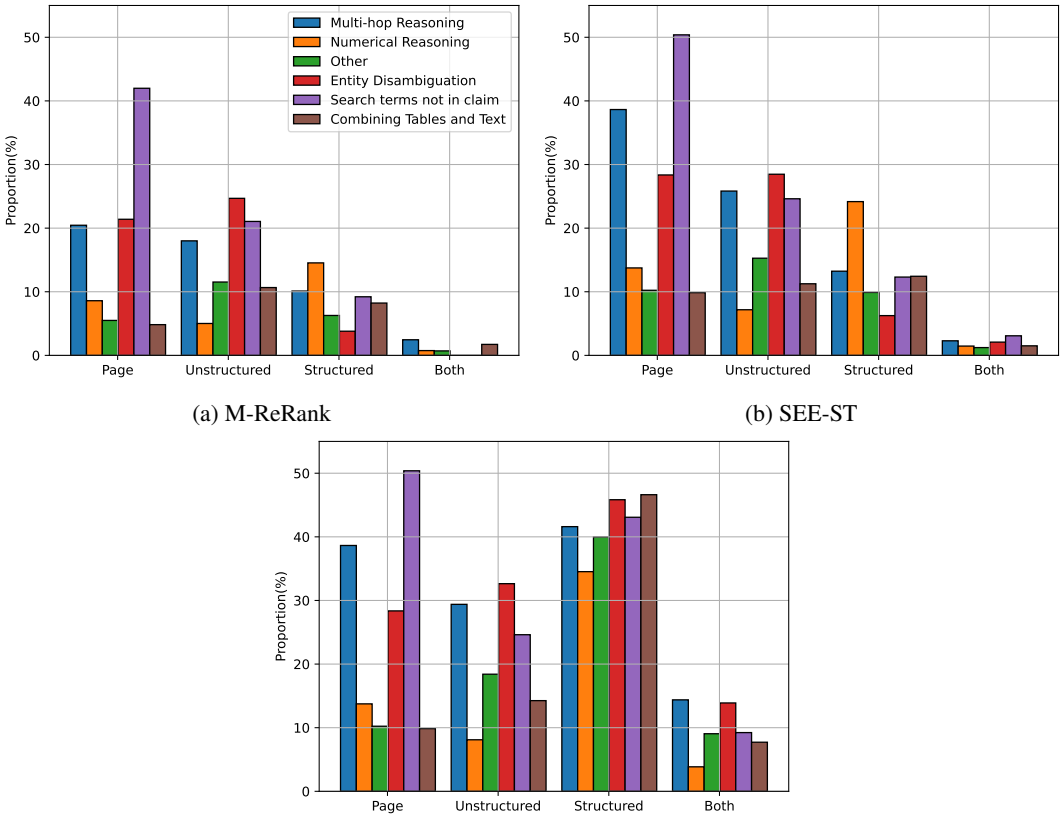

Figure 4: Overall error source analysis of extracted evidence set for the development set.

## F Error Analysis

To investigate error propagation within the FEVEROUS pipeline, we conduct a thorough error source analysis for both page and evidence retrieval stages. We also perform the error analysis on the challenge types to show M-ReRank's strengths and weaknesses.

### F.1 Error Source Analysis

Candidates not retrieved at any stage can lead to error propagation through the pipeline. In the three-step pipeline, the *Page* source error is determined by instances that fail to retrieve all pages containing evidence. Furthermore, sources of error can also arise when a specific evidence format is not




(a) M-ReRank

(b) SEE-ST

(c) UnifEE

Figure 5: Proportion of errors per source in relation to challenge type on the development set.

fully extracted. For instances with a complete document set, errors are categorised by the format of evidence that fails to be retrieved: *Unstructured* (sentences), *Structured* (tables or cells), and *Both*. Figure 4 displays the proportion of instances with failed evidence retrieval. We also show the percentage of instances with complete evidence set as *Complete*. Comparing the results with recent models, i.e. UnifEE and SEE-ST, our proposed M-ReRank approach performs well on each evidence type. On page retrieval, M-ReRank decreases the error from 15.8% to 9.2%. The decrement is also observed in proportion to source error on structured and unstructured evidence retrieval. These results demonstrate the effectiveness of M-ReRank in evidence extraction.

### F.2 Analysis Based on Challenge Types

In the FEVEROUS challenge, the samples are also categorised into various challenge categories. A fact-checker system's strength should also be analysed based on challenge types. These challenges encompass *Multi-hop Reasoning* (MR), performing *Numerical Reasoning* (NR), *Entity Disambiguation* (ED), dealing with *Search terms not present in claim* (ST), and *Combining Tables and Text* (CT). Any challenges outside these five categories are classified as *Other* (OT). We evaluate M-ReRank's performance to demonstrate its capability in retrieving evidence for claims with various challenges. M-ReRank achieves higher performance on almost all challenges with a major improvement on *Multi-hop Reasoning* and *Combining Tables and Text* challenges comparing SEE-ST and UnifEE as shown in Figure 5. M-ReRank achieves evidence extraction recall rates of 65.43%, 57.89%, 79.66%, 71.52%, 71.05%, and 76.75% for MR, NR, OT, ED, ST, and CT, respectively, showing that the collaborative filtering and modelling of inter-evidence context can effectively improve the evidence retrieval.

Our multi-stage reranking approach shows enhanced evidence retrieval capabilities, particularly in complex, challenging scenarios. M-ReRank decreases the error rate for Unstructured evidence by 4.71% against UnifEE and 2.92% against SEE-ST. For Structured evidence, it reduces the errors significantly by 28.97% against UnifEE, while less margin of 0.54% against SEE-ST as SEE-ST does well in structured evidence retrieval.