Claim2Source at CheckThat! 2025: Zero-Shot Style Transfer for Scientific Claim-Source Retrieval

Notebook for the CheckThat! Lab at CLEF 2025

Tobias Schreieder[1,2,*], Michael Färber[1,2]

[1]Dresden University of Technology (TUD), Dresden, Germany
[2]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany

Abstract

In this paper, we present our participation in the *CheckThat! 2025 Task 4b* on scientific claim-source retrieval. Our work systematically explores the impact of style transfer on performance in retrieving the scientific publication referenced by a COVID-19-related tweet. We apply seven distinct style transfer methods, distributed across claims and sources, to assess their impact on retrieval performance. These style transfer methods are evaluated across 15 retrieval systems, including 1 sparse, 7 dense, and 7 hybrid models, by testing each system with all combinations of claim and source styles. To guide the style transfer process, we employ a modular zero-shot prompting template with detailed instructions using a large language model (LLM). Our results show that GritLM-7B achieves the best performance without style transfer, suggesting strong robustness to informal text. In contrast, the majority of models, especially sparse and hybrid ones, benefit from applying a formal writing style to claims. We observe that hybrid retrieval models tend to outperform their dense counterparts. This highlights the potential advantage of integrating sparse and dense retrieval paradigms for scientific claim-source retrieval.

Keywords

Information Retrieval, Text Style Transfer, Large Language Model

1. Introduction

Nowadays, social media plays an increasingly important role in the communication and consumption of scientific information. Researchers and public institutions increasingly rely on platforms such as Twitter (now X), Bluesky, and Instagram to share findings, promote publications, and engage diverse audiences. Twitter, in particular, had stood out as a key channel for rapid scientific dissemination, especially in fast-moving fields like health and biomedicine [1, 2]. For instance, activity during academic conferences shows how these platforms foster broader discussion and spotlight emerging public health topics [3].

Automated accounts like bots significantly contribute to the circulation of scientific information on social media, complicating the identification of reliable sources [1]. Scientific claims shared online are frequently paraphrased into colloquial or simplified language, often obscuring their connection to original sources. These stylistic variations reduce the effectiveness of traditional information retrieval systems, which typically rely on lexical or semantic similarity measures. Moreover, the brevity and informal nature of social media posts, combined with their rapid spread, often strip these claims of the nuance and supporting evidence found in peer-reviewed literature. A systematic review by Suarez-Lledo and Alvarez-Galvez [4] found that health misinformation is highly prevalent on social media, particularly regarding vaccines, opioids, and noncommunicable diseases. During the COVID-19 pandemic, misinformation surged dramatically, and studies such as Sharma et al. [5] highlighted the urgent need for systems that can assess the credibility of claims shared online.

Given these challenges, there is an increasing need for automated fact verification systems that link social media claims to relevant peer-reviewed literature. The CLEF CheckThat! 2025 Lab [6] supports detecting and countering online disinformation across languages and platforms. In this work, we

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
tobias.schreieder@tu-dresden.de (T. Schreieder); michael.faerber@tu-dresden.de (M. Färber)
0009-0000-8268-4204 (T. Schreieder); 0000-0001-5458-8645 (M. Färber)

address Task 4b, *Scientific Claim-Source Retrieval*, which aims to retrieve the most relevant scientific paper corresponding to a given social media claim [7]. To tackle the retrieval challenges posed by stylistic divergence between claims and source texts, we implement a comprehensive suite of retrieval systems, testing a total of 15 models: 1 sparse, 7 dense, and 7 hybrid approaches. These configurations serve as baselines for evaluating the impact of style transfer on retrieval performance. Style transfer refers to the process of altering the style of text while preserving its original semantic content [8]. Specifically, we apply style transfer to both the claim texts and the source documents, investigating whether stylistic modifications help bridge the gap between informal, user-generated claims and the formal language of scientific publications. Our experiments involve four style transfer methods applied to claims and three to source documents, enabling a systematic evaluation of the interaction between style transformation and retrieval effectiveness. The results provide insights into how different retrieval architectures respond to stylistic adaptation and how these style transfer methods can support more robust claim verification in social media contexts. Overall, we make the following contributions:

1. We provide a comprehensive evaluation of 15 retrieval systems, covering sparse, dense, and hybrid models, for scientific claim-source retrieval in Task 4b of the CLEF CheckThat! 2025 Lab.
2. We present a systematic study of style transfer applied to both claims and source documents, using four claim styles and three source styles to assess their effect on retrieval performance.
3. We offer empirical insights into how different retrieval models respond to stylistic adaptation and identify effective combinations of style transfer methods that improve claim-source retrieval.

2. Related Work

Claim-Source Retrieval. The task has transitioned from simple ranking-based approaches to advanced pipelines leveraging LLMs. Initial approaches relied on traditional retrieval and learning-to-rank techniques, which for example have been applied to social media [9]. Soleimani et al. [10] employed BERT models for evidence retrieval and claim verification. Transformer-based systems improved retrieval and evidence selection. Some focused on dense representations and context-aware selection [11], while others employed hybrid strategies combining sparse and dense ranking methods for domain-specific misinformation detection [12]. To address retrieval across heterogeneous sources, Zuo et al. [13] proposed a cross-genre framework that bridges scientific and journalistic text, enhancing evidence alignment in COVID-19 misinformation scenarios. Beyond simple relevance, later systems modeled retrieval utility by incorporating feedback from verifiers [14] and introduced multi-step reasoning frameworks to capture evidence interdependency [15]. Generative retrieval models like GERE bypass document indexing by directly producing evidence identifiers, reducing memory and computational overhead [16]. Generative retrieval refers to methods where models generate document titles or sentence identifiers relevant to a claim, instead of retrieving documents from a static index, enabling more adaptive and efficient retrieval [16]. Building on these foundations, recent work has leveraged LLMs to improve retrieval performance and reasoning. Chen et al. [17] introduced a pipeline that incorporates question decomposition, breaking down complex claims into sub-questions to guide evidence search more effectively. Sriram et al. [18] proposed contrastive re-ranking with GPT-4 distillation, using multiple training signals such as answer correctness and sub-question alignment to fine-tune retrieval. Additionally, Churina et al. [19] developed techniques for generating enriched sub-questions with different reasoning styles to retrieve more diverse and relevant evidence. Retrieval-Augmented Generation has further enabled systems to evaluate both factual accuracy and relevance using LLM-generated summaries as reference points, enhancing performance and explainability [20].

Text Style Transfer. Early efforts in text style transfer (TST) emerged from information retrieval, where query rewriting through paraphrasing aimed to improve retrieval relevance. For example, Zukerman et al. [21, 22] explored lexical paraphrases using WordNet and thesauri to enhance document retrieval. Building on this, Apresjan et al. [23] developed rule-driven systems for synonymous and quasi-synonymous paraphrasing, applying these techniques to search engine optimization and information extraction. More broadly, these retrieval-focused methods treat paraphrase generation as a form of

style transfer to improve information access, setting the foundation for later work on style transfer. The field then shifted toward neural approaches and LLMs. Fu et al. [24] proposed adversarial networks to separate content and style representations using data that does not contain paired examples of the same text in different styles (i.e., non-parallel data), and introduced key evaluation metrics for transfer strength and content preservation. With the rise of LLMs, augmented zero-shot prompting was introduced as a method to guide style transfer through natural language instructions alone, requiring no training examples [25]. Gou et al. [26] bridged retrieval and generation by proposing RAST, a retrieval-augmented reinforcement learning framework that models question formulation as stylistic variation, balancing diversity and consistency. Subsequent studies expanded LLM-based TST in various directions. Mukherjee et al. [27] analyzed multilingual transfer tasks such as sentiment and detoxification, showing that fine-tuning outperforms zero- and few-shot prompting across English, Hindi, and Bengali. Zhang et al. [28] introduced CoTeX, leveraging chain-of-thought prompting to distill complex rewriting and reasoning into efficient style transfer models, particularly effective in low-resource settings. Lai et al. [29] presented sNeuron-TST, a neuron-level control method that identifies and manipulates style-specific neurons in LLMs to steer generation toward target styles, improving stylistic diversity while maintaining fluency via enhanced contrastive decoding. Finally, Aarnes et al. [30] applied TST with an LLM to adapt political debate texts into tweet-like formats for cross-domain claim detection, highlighting that style transfer alone cannot fully overcome domain mismatch challenges. Overall, this progression reflects a trajectory from early lexical paraphrasing in retrieval to sophisticated, LLM-driven methods incorporating reasoning, neuron-level control, and multilingual adaptation.

3. Dataset

The dataset for the CheckThat! 2025 Subtask 4b is designed for the retrieval of scientific papers implicitly referenced in social media posts. It comprises a query set of tweets and a collection set of candidate papers drawn from the CORD-19 corpus [31]. The query set includes 14,399 tweets with implicit references to scientific literature, partitioned into training, development (1,400 tweets), and test (1,446 tweets) subsets, each annotated with the unique identifier of the referenced paper. The collection set consists of metadata for 7,718 CORD-19 papers. For this task, we use only the title and abstract fields from each paper, which are concatenated into a single string to represent the paper. Our evaluation is conducted solely on the development and test sets, requiring retrieval systems to return the top five most likely referenced papers for each tweet.

4. Methodology

For scientific claim source retrieval, we have implemented a comprehensive suite of state-of-the-art retrieval systems, comprising eight distinct models and seven hybrid ranking approaches. These configurations serve as baselines to assess the impact of style transfer on retrieval performance. Specifically, we apply style transfer to both the claim texts and the source documents, thereby investigating the effects of stylistic transformations on retrieval efficacy. Our experiments involve the application of four style transfer methods to the claims and three style transfer methods to the source documents, enabling a systematic examination of the interactions between style transfer and retrieval performance.

4.1. Retrieval Systems

We distinguish between sparse, dense, and hybrid retrieval systems and include at least one system per category for comparison. Sparse retrieval systems utilize a high-dimensional vector representation of textual data, where each dimension corresponds to a specific term or feature, and the vector values are typically based on the frequency of words. This representation results in sparse vectors, characterized by a majority of zero values, which emphasizes the importance of exact term matches in retrieving relevant documents. The sparse nature of these vectors allows for efficient computation and storage, but may

limit the system's ability to capture nuanced semantic relationships between queries and documents. In contrast, dense retrieval systems use neural networks to convert text into low-dimensional, dense vectors that capture semantic meaning. These systems measure similarity based on vector closeness in this learned space, retrieving semantically similar documents even without word overlap. We also consider hybrid ranking approaches that are designed to extend the benefits of sparse and dense retrieval systems. All dense retrieval models have been implemented using a Faiss index, which enables a fast and scalable similarity search, allowing efficient retrieval of relevant documents from a large corpus [32].

BM25. Okapi BM25 is a widely adopted sparse retrieval model that relies on exact term matching, leveraging term frequency and inverse document frequency to rank documents based on their relevance to a given query [33, 34]. We selected BM25 because of its ability to accurately retrieve documents containing search terms, which allows style transfer methods to have a strong impact on retrieval performance. Nevertheless, Lv and Zhai [35] identified limitations of BM25 for long documents, which can negatively impact scientific claim-source retrieval, especially with very long source documents. We tuned the BM25 parameters b and k_1 to 1.0 and 1.2, respectively, by performing a grid search on a sample of the training dataset.

MiniLM. The *all-MiniLM-L6-v2* model is employed as a retrieval system by leveraging a lightweight pre-trained language model with 23M parameters to generate dense vector representations of queries and documents. MiniLM follows the Sentence-BERT framework [36], operating as a sentence-level encoder that captures semantic relationships between textual inputs, thereby enhancing retrieval effectiveness through semantically meaningful embeddings.

MPNet. Likewise as MiniLM the *all-mpnet-base-v2* model, is employed as a dense retrieval encoder, differing from MiniLM primarily in scale and architectural design. MPNet is based on a pre-training framework that combines masked language modeling with permuted language modeling [37], enabling it to capture both bidirectional and dependency-aware representations. With 110M parameters, MPNet is substantially larger than MiniLM.

SciNCL. The dense retrieval model *malteos/scincl* with 110M parameters is tailored for scientific and academic domains [38]. Based on a transformer encoder architecture, SciNCL is trained using a contrastive learning objective with hard negative sampling on a large corpus of scientific papers, abstracts, and citation contexts. The neural contrastive learning methodology focuses on optimizing representations to distinguish between semantically similar and dissimilar scientific texts, making it effective for scientific information retrieval tasks.

Specter. Another scientific document embedding model with 125M parameters is *allenai/specter*, designed for scholarly information retrieval and citation recommendation [39]. Built on the RoBERTa transformer architecture, SPECTER is trained using a citation-informed contrastive learning objective. This training paradigm enables it to capture semantic relationships between scientific texts. Unlike general-purpose models, SPECTER is domain-specific and optimized for academic use cases.

E5-Large. The *intfloat/e5-large-v2* model, with 335M parameters, serves as a dense retrieval encoder grounded in a transformer architecture derived from BERT and RoBERTa. It is trained using a contrastive objective on a large corpus of question–answer and passage triplets, following a retrieval-oriented training methodology proposed by Wang et al. [40]. Unlike MiniLM and MPNet, which are general-purpose language models adapted for sentence embeddings, E5 is explicitly optimized for retrieval tasks through instruction tuning and large-scale dual-encoder training.

GritLM-7B. *GritLM/GritLM-7B* is a dense retrieval encoder built on a decoder-only transformer architecture, comprising 7B parameters [41]. Developed with a focus on retrieval-centric applications, GritLM-7B is instruction-tuned and trained on large-scale web and document corpora, enabling robust performance in zero-shot and few-shot information retrieval scenarios. Unlike traditional encoder-based models such as E5 or GTR-XL, GritLM-7B leverages a generative pretraining framework while maintaining competitive retrieval capabilities through embedding extraction from decoder representations. Ajith et al. [42] have outlined the strengths of GritLM-7B in scientific literature retrieval.

GTR-XL. The model *gtr-t5-xl*, introduced by Ni et al. [43] is a state-of-the-art dense retrieval encoder based on the T5 architecture, specifically optimized for retrieval tasks. It is trained using a contrastive learning objective on an extensive dataset of question-answer and passage triplets, employing a retrieval-

centric training approach similar to that of E5. However, GTR-XL distinguishes itself with its ability to handle large-scale text-to-text transformations and is designed to achieve superior performance in both retrieval and generative tasks. With 11B parameters, GTR-XL is the largest model used in our comparison and might therefore be particularly adept at handling complex retrieval queries.

Hybrid. Hybrid retrieval systems, which integrate both sparse and dense retrieval paradigms, have demonstrated superior retrieval effectiveness compared to systems relying solely on either approach [44, 45]. In our experimental setup for scientific claim-source retrieval, we employ BM25 as the sparse retrieval component and pair it once with each previously mentioned dense retrieval model to generate document relevance scores. To combine the strengths of both models, we compute a hybrid score for each candidate document based on a weighted linear combination of the individual scores. Let d denote a candidate document and $S_{\text{sparse}}(d)$, $S_{\text{dense}}(d)$ represent the relevance scores assigned to d by the sparse and dense retrievers, respectively. The hybrid score $S_{\text{hybrid}}(d)$ is computed as:

$$S_{\text{hybrid}}(d) = \alpha \cdot S_{\text{sparse}}(d) + (1 - \alpha) \cdot S_{\text{dense}}(d) \tag{1}$$

In our experiments, we fix the interpolation parameter to $\alpha = 0.5$, giving equal weight to both retrieval components. All scores are min-max normalized before ranking to ensure comparability across models.

4.2. Style Transfer

We apply style transfer to align the linguistic style of user-generated tweets and scientific source documents to improve their comparability for retrieval. This includes converting both claims and sources into a consistent style, such as scientific language, and testing alternative styles to assess their impact on retrieval performance. For all experiments, we use the LLaMA 3.3 70B Instruct model [46], optimized for instruction-following and capable of controlled style adaptation across domains. We designed a modular prompt template with four components: context, task, instructions, and output specification. The context describes the retrieval objective and is adapted based on whether the input is a claim or a source. The task defines the LLM's role, such as generating a scientific question from a tweet. The instructions guide the style transfer, including tone and structure. The output specification ensures consistent formatting. All prompts are shown in Table 2 and Table 3 in the appendix.

Claim Formal (C1). The first style transfer focuses on converting informal tweets into a more formal tone. This process involves removing informal elements such as hashtags and emojis, as well as enhancing the readability of the text. The LLM is instructed to preserve the original terminology used in the tweet and to maintain the core semantic content without alteration.

Claim Scientific (C2). C2 involves rewriting each tweet into a more formal scientific-claim format. This process requires transforming informal or casual language into precise academic language with domain-specific terminology, thereby creating clear, standalone scientific claims. Similar to C1 non-essential elements such as hashtags, emojis, and informal symbols are removed. The original meaning of the tweet must be preserved, and any existing scientific terminology should remain unchanged.

Claim Abstract (C3). This style transfer requires rewriting each tweet as a scientific abstract of approximately 150 words, using domain-specific terminology while preserving existing scientific terms. The abstract should present key aspects such as motivation, methodology, and findings in a coherent narrative. Unlike C2, which transforms tweets into clear scientific claims, C3 involves more extensive restructuring to produce a continuous scientific summary. The approach aims to improve comparability by providing a uniform outline, as both claims and sources are expressed in similar language and structure. Since the model has no access to the original paper and it might not be included in the LLM's training data, C3 risks generating hallucinated content. To reduce this risk, we explicitly prompt the model not to introduce any information beyond what is present in the original tweet. This setup tests whether strong stylistic alignment alone can bridge the domain gap and improve retrieval performance.

Claim Question (C4). This style transfer involves formulating a single, precise scientific question for each tweet that ideally can be answered exclusively by the paper referenced in the tweet. The question must be clearly stated as a unified inquiry, without being split into multiple sub-questions, and must retain the original terminology used in the tweet without modification. The formulation should

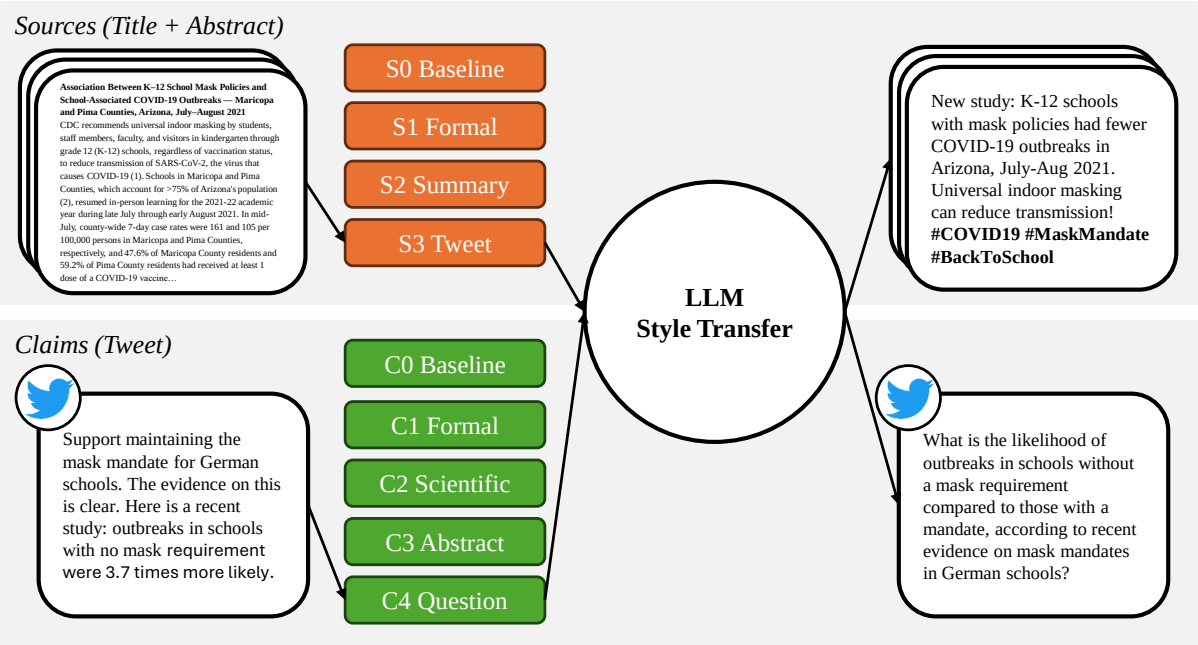

Figure 1: Overview of the style transfer pipeline. The pipeline applies style transfer independently to sources and claims. Source documents undergo style transfer with three distinct styles before indexing, including a baseline without transformation. During inference, style transfer can be applied to claims using one of four styles. This enables individual style transfer of sources and claims.

be carefully optimized to maximize retrieval performance by aligning with the input expectations of retrieval systems designed for question-answering tasks, where a question serves as the query for document retrieval. The motivation for generating well-structured scientific questions lies in the potential for improved performance in modern question-centric retrieval models such as GTR or E5, which are specifically optimized to handle question-based inputs more effectively.

Source Formal (S1). Analogous to the style transfer applied to claims, we perform a structured transformation of the source texts. The objective of S1 is to improve clarity and readability through minimal, targeted edits while rendering the abstract in a more formal, standardized tone. All key details and factual content must be preserved, and original terminology and core meaning remain unchanged. The transformation involves removing unnecessary filler and overly verbose academic phrasing, omitting non-essential self-referential statements, and correcting unclear grammar or excessively long sentences. This process follows the same formalization principles as the C1 transformation for claims, adapted to the linguistic and structural characteristics of scientific source texts.

Source Summary (S2). The objective of S2 is to generate a more concise and retrieval-oriented summary of each scientific abstract. While an abstract is itself a summary, it often includes broad motivation, general background, or other information that is not directly relevant for retrieval or scientific claim support. This style transfer focuses on extracting the core findings, relevant methodology, and key contextual details while strictly preserving the original terminology and core meaning. Content that does not directly contribute to understanding the scientific contribution or supporting document retrieval is omitted, resulting in a shorter, more focused version of the abstract.

Source Tweet (S3). Analogous to the C3 style transfer, which generates scientific abstracts from tweets, the S3 style transfer performs the inverse transformation by deriving a concise tweet from a scientific abstract. The objective is to produce a brief, accurate, and engaging summary that emphasizes the key research findings while maintaining fidelity to the original content. Each tweet is constrained to a maximum of 280 characters and is formulated to be suitable for dissemination by the paper's authors or interested third parties on social media platforms. To enhance discoverability and outreach, relevant hashtags are included. Although the format is informal and highly compressed, the transformation preserves the core scientific message without introducing distortion or oversimplification.

5. Evaluation

In this section, we present a comprehensive evaluation. We assess the impact of the 7 distinct style transfer methods, compared to a no-transfer baseline, across 15 retrieval models. We also report the performance of our best-performing method on the CheckThat! 2025 Task 4b benchmark. Additionally, we conducted a qualitative analysis, which can be found in the appendix in Table 4.

5.1. Evaluation of Style Transfer for Scientific Claim Source Retrieval

We evaluated all style transfer methods across each retrieval system using the 1,400 tweets from the development dataset. To support style-specific retrieval, we constructed a separate Faiss index for each source document style, resulting in four distinct indices. All experiments were evaluated using Mean Reciprocal Rank@5 (MRR@5). The results for each configuration are reported in Table 1 and Table 2.

The evaluation results demonstrate that the highest retrieval performance is achieved using GritLM-7B without style transfer, yielding an MRR@5 of 0.7115 and outperforming the BM25 baseline (0.5575) by approximately 0.15. Among models without style transfer, only E5-Large and most hybrid models, excluding H-Specter, surpass the BM25 baseline. Notably, scientific embedding models such as SciNCL (0.3735) and Specter (0.0728) perform significantly worse, indicating limited effectiveness in this context.

Nevertheless, applying style transfer generally improves retrieval performance, particularly when using a formal style for claims (C1) and no style transfer for source documents (S0). Under this configuration, BM25, E5-Large, and most hybrid models including H-MiniLM, H-MPNet, H-Specter, H-GritLM-7B, and H-GTR-XL show modest improvements around 0.01 in MRR@5, while H-SciNCL achieves a slightly larger gain around 0.02. In contrast, H-E5-Large shows minimal improvement of 0.001. Additionally, MiniLM, MPNet, SciNCL, and Specter benefit more noticeably when scientific style is applied to claims (C2), although optimal style for sources varies. MiniLM performs best with formal-style source documents (S1), whereas Specter benefits most from tweet-like source documents (S3). The largest relative improvement from style transfer is observed with Specter, where using summary-style sources and scientific-style claims yields an increase of approximately 0.11 in MRR@5. However, its absolute retrieval performance remains low (0.1870), substantially lagging behind all other models.

Table 1
Evaluation of style transfer for claim-source retrieval *(S0–S1)*. This table reports MRR@5 for each retrieval model described in Section 4.2. Hybrid retrieval models, denoted as "H-", combine the sparse BM25 retriever with one of the dense retrieval models. We evaluate retrieval performance across two source document styles: Original (S0) and Formal (S1). Each is paired with five claim styles: Original (C0), Formal (C1), Scientific (C2), Abstract (C3), and Question (C4). S0–S1 serves as the baseline for comparison across all style transfer configurations.

	S0					S1				
Model	C0	C1	C2	C3	C4	C0	C1	C2	C3	C4
BM25	0.5575	0.5800	0.5549	0.4579	0.4934	0.5071	0.5337	0.5287	0.4587	0.4837
MiniLM	0.4897	0.4895	0.4947	0.4622	0.4794	0.4937	0.4988	0.5111	0.4733	0.4910
MPNet	0.5052	0.5146	0.5185	0.5151	0.5143	0.4939	0.5026	0.5178	0.5080	0.5030
SciNCL	0.3735	0.3766	0.3793	0.3640	0.3557	0.3525	0.3549	0.3769	0.3558	0.3455
Specter	0.0728	0.0795	0.0865	0.1272	0.0682	0.1134	0.1263	0.1455	0.1732	0.1190
E5-Large	0.6568	0.6578	0.6557	0.6112	0.6317	0.6045	0.6149	0.6295	0.5744	0.6012
GritLM-7B	**0.7115**	0.7090	0.7049	0.6649	0.6687	0.6859	0.6795	0.6760	0.6339	0.6598
GTR-XL	0.5322	0.5268	0.5113	0.4661	0.5089	0.5305	0.5296	0.5159	0.4669	0.4934
H-MiniLM	0.6248	0.6370	0.6238	0.5396	0.5724	0.5886	0.6055	0.6087	0.5423	0.5651
H-MPNet	0.6323	0.6457	0.6344	0.5601	0.5851	0.5935	0.6150	0.6218	0.5565	0.5760
H-SciNCL	0.6027	0.6182	0.6093	0.5101	0.5426	0.5632	0.5794	0.5814	0.5089	0.5371
H-Specter	0.5523	0.5645	0.5361	0.4052	0.4849	0.5217	0.5498	0.5410	0.4529	0.5005
H-E5-Large	0.6642	0.6727	0.6623	0.6119	0.6091	0.6207	0.6339	0.6378	0.5875	0.5930
H-GritLM-7B	0.6898	0.6982	0.6872	0.6381	0.6461	0.6537	0.6679	0.6732	0.6282	0.6379
H-GTR-XL	0.6370	0.6457	0.6257	0.5284	0.5770	0.6008	0.6084	0.6050	0.5340	0.5647

Table 2
Evaluation of style transfer for claim-source retrieval *(S2–S3)*. This table reports MRR@5 for each retrieval model described in Section 4.2. Hybrid retrieval models, denoted as "H-", combine the sparse BM25 retriever with one of the dense retrieval models. We evaluate retrieval performance across two source document styles: Summary (S2) and Tweet (S3). Each is paired with five claim styles: Original (C0), Formal (C1), Scientific (C2), Abstract (C3), and Question (C4).

	S2					S3				
Model	C0	C1	C2	C3	C4	C0	C1	C2	C3	C4
BM25	0.4541	0.4799	0.4670	0.3785	0.4391	0.2489	0.2673	0.2369	0.1398	0.2125
MiniLM	0.4970	0.5007	0.5077	0.4788	0.4892	0.4427	0.4409	0.4420	0.4098	0.4284
MPNet	0.4878	0.4994	0.5131	0.4978	0.5068	0.4567	0.4550	0.4525	0.4176	0.4351
SciNCL	0.3576	0.3629	0.3757	0.3415	0.3577	0.3199	0.3161	0.3171	0.3053	0.3165
Specter	0.1224	0.1353	0.1461	0.1491	0.1320	0.1588	0.1689	0.1870	0.1612	0.1661
E5-Large	0.6169	0.6302	0.6304	0.5793	0.6135	0.5533	0.5583	0.5536	0.5363	0.5124
GritLM-7B	0.6744	0.6795	0.6760	0.6339	0.6598	0.5975	0.6795	0.6760	0.6339	0.6598
GTR-XL	0.4904	0.4843	0.4710	0.4305	0.4697	0.4854	0.4882	0.4696	0.4412	0.4503
H-MiniLM	0.5568	0.5657	0.5673	0.4957	0.5282	0.3732	0.3809	0.3719	0.3387	0.3520
H-MPNet	0.5595	0.5695	0.5700	0.5059	0.5338	0.3695	0.3793	0.3765	0.3281	0.3602
H-SciNCL	0.5237	0.5438	0.5280	0.4439	0.4949	0.3350	0.3450	0.3308	0.2794	0.3084
H-Specter	0.4822	0.5020	0.4864	0.3784	0.4564	0.3058	0.3239	0.3123	0.2588	0.2871
H-E5-Large	0.5828	0.5961	0.5895	0.5342	0.5578	0.3882	0.4041	0.3998	0.3750	0.3750
H-GritLM-7B	0.6270	0.6320	0.6354	0.5757	0.5919	0.4250	0.4747	0.4780	0.4552	0.4442
H-GTR-XL	0.5600	0.5641	0.5461	0.4668	0.5225	0.3873	0.3946	0.3812	0.3511	0.3578

5.2. Evaluation of ChecktThat! 2025 Task 4b

For the CheckThat! 2025 Lab, we submitted our best-performing model from the development phase (Section 5.1): GritLM-7B, without applying style transfer to claims or sources. Participating systems achieved MRR@5 scores ranging from 0.00 to 0.68. Our model obtained an MRR@5 of 0.59, ranking 12th out of 31 teams and outperforming the BM25 baseline (0.43) by +0.16. These results indicate that the test set was substantially more challenging, as reflected by the performance drops for both BM25 (-0.13) and GritLM-7B (-0.12) compared to their scores on the development set.

Table 3
Comparison of the BM25 baseline model and our submitted model GritLM-7B without style transfer on the development (Dev) and final test (Test) datasets of the CheckThat! 2025 Lab.

Model	**Dev**	**Test**
BM25 *(Baseline)*	0.56	0.43
GritLM-7B *(Ours)*	0.71	0.59

6. Discussion

We conducted a systematic comparison of 15 retrieval models and 7 style transfer methods. This section discusses our key research questions (RQs).

RQ1: How do sparse, dense, and hybrid retrieval systems compare in their effectiveness for retrieving scientific sources for social media claims?

Our evaluation reveals important findings regarding the effectiveness of sparse, dense, and hybrid retrieval systems for scientific source retrieval supporting social media claims. The sparse retrieval model BM25 establishes a strong baseline, demonstrating robust performance across the retrieval tasks. Among dense retrieval models, only a few, notably E5-Large and GritLM-7B, were able to outperform BM25. Regarding hybrid models, all evaluated hybrid models except H-GritLM-7B showed improved

performance compared to their dense-only counterparts, indicating that combining sparse and dense retrieval signals generally enhances retrieval effectiveness. Nevertheless, GritLM-7B stands out as the strongest retrieval model overall for claim-source retrieval, a finding consistent with results reported by Ajith et al. [42]. These results suggest that while traditional sparse methods remain competitive, state-of-the-art dense and hybrid approaches offer superior retrieval capabilities in this context.

RQ2: What is the impact of applying style transfer to claims, source documents, or both on retrieval performance across different retrieval systems?

Our findings show that applying a formal style transfer to claims tends to improve retrieval performance, particularly when source documents remain unaltered. Most retrieval systems benefit from this transformation, except for GritLM-7B and GTR-XL, which appear robust to stylistic noise or sensitive to content shifts. The improvements suggest informal tweet language hinders retrieval and that formalization, such as removing hashtags and correcting grammar, better aligns claims with scientific abstracts. In contrast, modifying source documents typically harms performance, likely due to loss of essential content and added noise, though exceptions occur with models such as MiniLM and Specter. The sparse retrieval model shows the strongest improvements from formalized claims, likely due to reliance on lexical overlap. Dense and hybrid systems also benefit, though less consistently. Overall, formalizing claims enhances compatibility with scientific language, especially in sparse retrieval settings.

RQ3: Which combinations of style transfer methods applied to claims and source documents are most effective in improving scientific claim-source retrieval?

We evaluated all combinations of claim and source style transfer methods to identify the most effective configuration. Across models, the combination of original source documents with a formalized claim style (S0–C1) consistently yielded strong performance, achieving the best results for BM25, E5-Large, and all hybrid retrieval systems. Applying a scientific style to claims (S0-C2) also led to notable improvements, particularly for MPNet and SciNCL. These results suggest that formal and scientific claim styles align more closely with the linguistic structure of source documents, enhancing retrieval effectiveness. In contrast, source documents appear sufficiently formal by default, so applying style transfer to them might introduce noise rather than improving retrieval. While the overall trend favors keeping source documents in their original form, certain models perform better with alternative configurations. For instance, MiniLM achieved its highest retrieval scores when claims were transferred into scientific style and sources into formal style (S1–C2). Similarly, Specter performed best when claims used scientific style and sources were transformed into tweet-like text (S3–C2). These exceptions highlight that optimal style transfer combinations can vary depending on model architecture.

7. Conclusion

In this paper we have demonstrated that the submitted GritLM-7B model, without fine-tuning or style transfer, achieves competitive retrieval performance with an MRR@5 of 0.59, ranking 12th out of 30 participants in the *CheckThat! 2025 Task 4b* on scientific claim-source retrieval. Although BM25 establishes a robust sparse baseline, certain dense models such as GritLM-7B and E5-Large surpass its performance, while hybrid models typically demonstrate enhanced effectiveness by integrating both sparse and dense retrieval architectures. Applying a formal writing style to claims improves retrieval performance for most tested models by aligning informal social media language with the structured style of scientific abstracts. In contrast, applying style transfer to source documents typically results in decreased effectiveness, although a few exceptions were observed. These findings are limited by evaluation on a single, comparatively challenging dataset focused solely on COVID-19 research, which also exhibits strong performance variations between the development and test sets. The study also relies on a single LLM for style transfer. Therefore, future research should validate the results across multiple diverse datasets and models and further investigate style transfer methods to enhance robustness and generalizability. By integrating style transfer through zero-shot prompting using LLMs, this study provides valuable insights into effective strategies for scientific claim-source retrieval and emphasizes potential pathways for advancing evidence discovery associated with social media claims.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig“, project identification number: ScaDS.AI.

The authors also acknowledge computing resources provided by the NHR Center at TU Dresden, supported by the Ministry of Research, Technology and Space and the participating state governments within the NHR framework.

Availability

Reference code for all experiments, as well as all style transfer datasets, is available in our repository at https://github.com/faerber-lab/Claim2Source.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and LLaMA 3.3 70B Instruct for grammar, spelling correction, and rephrasing assistance. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Haunschild, L. Bornmann, D. Potnis, I. Tahamtan, Investigating dissemination of scientific information on twitter: A study of topic networks in opioid publications, Quantitative Science Studies 2 (2021) 1486–1510. doi:10.1162/qss_a_00168.

[2] L. Guenther, C. Wilhelm, C. Oschatz, J. Brück, Science communication on twitter: Measuring indicators of engagement and their links to user interaction in communication scholars' tweet content, Public Understanding of Science 32 (2023) 860–869. doi:10.1177/09636625231166552.

[3] C. G. Allen, B. Andersen, D. A. Chambers, J. Groshek, M. C. Roberts, Twitter use at the 2016 conference on the science of dissemination and implementation in health: analyzing #discience16, Implement. Sci. 13 (2018). doi:10.1186/s13012-018-0723-z.

[4] V. Suarez-Lledo, J. Alvarez-Galvez, Prevalence of health misinformation on social media: Systematic review, J Med Internet Res 23 (2021) e17187. doi:10.2196/17187.

[5] M. S. Al-Rakhami, A. M. Al-Amri, Lies kill, facts save: Detecting covid-19 misinformation in twitter, IEEE Access 8 (2020) 155961–155970. doi:10.1109/ACCESS.2020.3019600.

[6] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478. doi:10.1007/978-3-031-88720-8_68.

[7] S. Hafid, Y. S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! lab task 4 on scientific web discourse, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.

[8] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, R. Mihalcea, Deep learning for text style transfer: A survey, Computational Linguistics 48 (2022) 155–205. doi:10.1162/coli_a_00426.

[9] W. Ma, W. Chao, Z. Luo, X. Jiang, Claim retrieval in twitter, in: WISE'18, Springer International Publishing, Cham, 2018, pp. 297–307. doi:10.1007/978-3-030-02922-7_20.

[10] A. Soleimani, C. Monz, M. Worring, Bert for evidence retrieval and claim verification, in: ECIR'20, Springer-Verlag, Berlin, Heidelberg, 2020, p. 359–366. doi:10.1007/978-3-030-45442-5_45.

[11] C. Samarinas, W. Hsu, M. L. Lee, Improving evidence retrieval for automated explainable fact-checking, in: NAACL'21, ACL, Online, 2021, pp. 84–91. doi:10.18653/v1/2021.naacl-demos.10.

[12] M. Sundriyal, G. Malhotra, M. S. Akhtar, S. Sengupta, A. Fano, T. Chakraborty, Document retrieval and claim verification to mitigate COVID-19 misinformation, in: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 66–74. doi:10.18653/v1/2022.constraint-1.8.

[13] C. Zuo, C. Wang, R. Banerjee, Cross-genre retrieval for information integrity: A covid-19 case study, in: ADMA'23, Springer Nature Switzerland, Cham, 2023, pp. 495–509. doi:10.1007/978-3-031-46677-9_34.

[14] H. Zhang, R. Zhang, J. Guo, M. de Rijke, Y. Fan, X. Cheng, From relevance to utility: Evidence retrieval with feedback for fact verification, in: Findings of EMNLP'23, ACL, Singapore, 2023, pp. 6373–6384. doi:10.18653/v1/2023.findings-emnlp.422.

[15] H. Liao, J. Peng, Z. Huang, W. Zhang, G. Li, K. Shu, X. Xie, Muser: A multi-step evidence retrieval enhancement framework for fake news detection, in: SIGKDD'23, ACM, New York, NY, USA, 2023, p. 4461–4472. doi:10.1145/3580305.3599873.

[16] J. Chen, R. Zhang, J. Guo, Y. Fan, X. Cheng, Gere: Generative evidence retrieval for fact verification, in: SIGIR'22, ACM, New York, NY, USA, 2022, p. 2184–2189. doi:10.1145/3477495.3531827.

[17] J. Chen, G. Kim, A. Sriram, G. Durrett, E. Choi, Complex claim verification with evidence retrieved in the wild, in: NAACL'24, ACL, Mexico City, Mexico, 2024, pp. 3569–3587. doi:10.18653/v1/2024.naacl-long.196.

[18] A. Sriram, F. Xu, E. Choi, G. Durrett, Contrastive learning to improve retrieval for real-world fact checking, in: FEVER'24, ACL, Miami, Florida, USA, 2024, pp. 264–279. doi:10.18653/v1/2024.fever-1.28.

[19] S. Churina, A. M. Barik, S. R. Phaye, Improving evidence retrieval on claim verification pipeline through question enrichment, in: FEVER'24, ACL, Miami, Florida, USA, 2024, pp. 64–70. doi:10.18653/v1/2024.fever-1.6.

[20] R. Upadhyay, M. Viviani, Enhancing health information retrieval with RAG by prioritizing topical relevance and factual accuracy, Discover Computing 28 (2025) 27. doi:10.1007/s10791-025-09505-5.

[21] I. Zukerman, B. Raskutti, Y. Wen, Experiments in query paraphrasing for information retrieval, in: AI'02, Springer-Verlag, Berlin, Heidelberg, 2002, p. 24–35. doi:10.1007/3-540-36187-1_3.

[22] I. Zukerman, B. Raskutti, Lexical query paraphrasing for document retrieval, in: COLING'02, 2002, p. 1–7. URL: https://aclanthology.org/C02-1161/.

[23] J. D. Apresjan, I. M. Boguslavsky, L. L. Iomdin, L. L. Cinman, S. P. Timoshenko, Semantic paraphrasing for information retrieval and extraction, in: FQAS'09, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 512–523. doi:10.1007/978-3-642-04957-6_44.

[24] Z. Fu, X. Tan, N. Peng, D. Zhao, R. Yan, Style transfer in text: exploration and evaluation, in: AAAI'18, AAAI Press, 2018, pp. 663–670. doi:10.5555/3504035.3504117.

[25] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, J. Wei, A recipe for arbitrary text style transfer with large language models, in: ACL'22, ACL, Dublin, Ireland, 2022, pp. 837–848. doi:10.18653/v1/2022.acl-short.94.

[26] Q. Gou, Z. Xia, B. Yu, H. Yu, F. Huang, Y. Li, N. Cam-Tu, Diversify question generation with retrieval-augmented style transfer, in: EMNLP'23, ACL, Singapore, 2023, pp. 1677–1690. doi:10.18653/v1/2023.emnlp-main.104.

[27] S. Mukherjee, A. K. Ojha, O. Dusek, Are large language models actually good at text style transfer?, in: INLG'24, ACL, Tokyo, Japan, 2024, pp. 523–539. URL: https://aclanthology.org/2024.inlg-main.42/.

[28] C. Zhang, H. Cai, Y. Li, Y. Wu, L. Hou, M. Abdul-Mageed, Distilling text style transfer with self-explanation from LLMs, in: NAACL-SRW'24, ACL, Mexico City, Mexico, 2024, pp. 200–211. doi:10.18653/v1/2024.naacl-srw.21.

[29] W. Lai, V. Hangya, A. Fraser, Style-specific neurons for steering LLMs in text style transfer, in: EMNLP'24, ACL, Miami, Florida, USA, 2024, pp. 13427–13443. doi:10.18653/v1/2024.emnlp-main.745.

[30] P. R. Aarnes, V. Setty, P. Galuščáková, Iai group at checkthat! 2024: Transformer models and data augmentation for checkworthy claim detection, 2024. arXiv:2408.01118.

[31] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, ACL, Online, 2020. URL: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

[32] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, 2025. arXiv:2401.08281.

[33] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: ACM SIGIR'94, Springer-Verlag, Berlin, Heidelberg, 1994, p. 232–241. doi:10.5555/188490.188561.

[34] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: TREC'94, volume 500-225, NIST, 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[35] Y. Lv, C. Zhai, When documents are very long, bm25 fails!, in: SIGIR'11, ACM, New York, NY, USA, 2011, p. 1103–1104. doi:10.1145/2009916.2010070.

[36] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: EMNLP-IJCNLP'19, ACL, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[37] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: masked and permuted pre-training for language understanding, in: NeurIPS'20, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 16857–16867. URL: https://arxiv.org/abs/2004.09297.

[38] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, G. Rehm, Neighborhood contrastive learning for scientific document representations with citation embeddings, in: EMNLP'22, ACL, Abu Dhabi, United Arab Emirates, 2022, pp. 11670–11688. doi:10.18653/v1/2022.emnlp-main.802.

[39] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. Weld, SPECTER: Document-level representation learning using citation-informed transformers, in: ACL'20, ACL, 2020, pp. 2270–2282. doi:10.18653/v1/2020.acl-main.207.

[40] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, 2024. arXiv:2212.03533.

[41] N. Muennighoff, H. SU, L. Wang, N. Yang, F. Wei, T. Yu, A. Singh, D. Kiela, Generative representational instruction tuning, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=BC4lIvfSzv.

[42] A. Ajith, M. Xia, A. Chevalier, T. Goyal, D. Chen, T. Gao, LitSearch: A retrieval benchmark for scientific literature search, in: EMNLP'24, ACL, Miami, Florida, USA, 2024, pp. 15068–15083. doi:10.18653/v1/2024.emnlp-main.840.

[43] J. Ni, C. Qu, J. Lu, Z. Dai, G. Hernandez Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.-W. Chang, Y. Yang, Large dual encoders are generalizable retrievers, in: EMNLP'22, ACL, Abu Dhabi, United Arab Emirates, 2022, pp. 9844–9855. doi:10.18653/v1/2022.emnlp-main.669.

[44] Y. Luan, J. Eisenstein, K. Toutanova, M. Collins, Sparse, dense, and attentional representations for text retrieval, TACL 9 (2021) 329–345. doi:10.1162/tacl_a_00369.

[45] P. Mandikal, R. Mooney, Sparse meets dense: A hybrid approach to enhance scientific document retrieval, CoRR (2024). arXiv:2401.04055.

[46] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, et al., The llama 3 herd of models, 2024. arXiv:2407.21783.

A. Style Transfer Prompting Template

Context: You are a language model specialized in text optimization for scientific claim-source retrieval. In this task informal social media posts like tweets (claim) have to be matched with the most relevant scientific abstract (source). To increase the retrieval performance, a style transfer of the given tweet must be performed.

Task: Pre-process the given tweet with minimal changes, according to the following instructions. C1
Instructions:
- Remove non-essential elements like hashtags, emojis, and informal symbols.
- Retain all key details, and all factual information.
- Do not alter the terminology or the core meaning of the tweet.
- Focus on improving readability without over-processing, ensuring the tweet remains suitable for matching with corresponding scientific papers.

Task: Rewrite each tweet into a more formal scientific-claim format, according to the following instructions. C2
Instructions:
- Transform the tweet into a clear, standalone scientific claim or set of claims.
- Remove non-essential elements like hashtags, emojis, and informal symbols.
- Do not alter the meaning of the tweet.
- Use precise academic language and domain-specific terminology. Do not alter the terminology of the tweet, if it is already scientific.

Task: For each tweet write a scientific abstract, according to the following instructions. C3
Instructions:
- Write a concise, formal scientific abstract (~150 words).
- Use precise academic language and domain-specific terminology. Do not alter the terminology of the tweet, if this is already scientific.
- Ensure coherence and logical flow, but do not include section headers.
- Present key aspects of the study, including context, methodology (if available), key findings, and significance, in a continuous and structured manner. Do not introduce information which is not present in the tweet.
- Maintain neutrality unless the tweet explicitly expresses a stance, in which case incorporate the sentiment appropriately.

Task: For each tweet write one scientific question, according to the following instructions. C4
Instructions:
- The question should only be answerable with the paper the tweet refers to.
- The question should be formulated precisely without being divided into several sub-questions.
- Do not alter the terminology used in the tweet.
- The question should be optimized for the best possible retrieval performance.

Output: Only return the pre-processed question without any explanations or step-by-step reasoning!

Figure 2: Prompt template for style transfer of claims (tweets) into four distinct styles: Formal (C1), Scientific (C2), Abstract (C3), and Question (C4). The modular prompt includes four components: Context, Task, Instructions, and Output Specification. Each component guides the LLM in transforming tweets with stylistic and structural consistency.

Context: You are a language model specialized in text optimization for scientific claim-source retrieval. In this task informal social media posts like tweets (claim) have to be matched with the most relevant scientific abstract (source). To increase the retrieval performance, a style transfer of the given scientific abstract must be performed.

Task: Pre-process the given abstract with minimal changes, according to the following instructions. S1
Instructions:
- Retain all key details, and all factual information.
- Do not alter the terminology or the core meaning of the paper abstract.
- Remove unnecessary filler or verbose academic phrasing.
- Fix unclear grammar or overly long sentences for readability.
- Omit references to the paper itself, if not essential to meaning.

Task: For each abstract write a summary, according to the following instructions. S2
Instructions:
- Do not alter the terminology or the core meaning of the paper abstract.
- Summarize the core findings, relevant methodology and key contextual details.
- Focus on content that supports scientific claims and aids in source retrieval.
- Exclude general background, broad motivation, or unrelated information.

Task: For each abstract write a short tweet, according to the following instructions. S3
Instructions:
- Write a concise tweet that draws attention to the key research findings.
- The tweet should be written in such a way that either the author of the paper or interested third parties can post the tweet on social media.
- A tweet cannot be longer than 280 characters.
- Use relevant hashtags to increase discoverability.

Output: Only return the pre-processed text without any explanations or step-by-step reasoning!

Figure 3: Prompt template for style transfer of source documents (abstracts) into three distinct styles: Formal (S1), Summary (S2), Tweet (S3). The modular prompt includes four components: Context, Task, Instructions, and Output Specification. Each component guides the LLM in transforming tweets with stylistic and structural consistency.

B. Qualitative Analysis of Style Transfer

Table 4
Qualitative analysis of claim and source style transfer. The table presents one representative claim example in its original form (C0) from the development dataset, alongside its transformed versions using four distinct claim style transfer methods: Formal (C1), Scientific (C2), Abstract (C3), and Question (C4). Below the midrule, a corresponding source example (S0) matching the claim is shown in its original form, along with its transformed versions across three source style transfer methods: Formal (S1), Summary (S2), and Tweet (S3).

Style	Text
C0	Myth: I sometimes see the downplaying of COVID deaths because those dying "are old" on average and would have "died soon anyway". This study looked at the first 1,279,866 reported #covid19 deaths. The estimated average years of life lost was 16 years.
C1	Myth: I sometimes see the downplaying of COVID deaths because those dying are old on average and would have died soon anyway. This study looked at the first 1,279,866 reported covid19 deaths. The estimated average years of life lost was 16 years.
C2	A common misconception surrounding COVID-19 is that the deaths are often downplayed due to the advanced age of the individuals, with the implication that they would have died soon regardless. However, a study examining the first 1,279,866 reported COVID-19 deaths found that the estimated average years of life lost was 16 years.
C3	This study examined the first 1,279,866 reported COVID-19 deaths, aiming to contextualize the impact of the pandemic on mortality. A common misconception downplays the significance of COVID-19-related deaths, attributing them to the advanced age of the deceased and implying that these individuals would have succumbed to other causes shortly thereafter. However, the findings of this investigation reveal a substantial loss of life expectancy, with an estimated average of 16 years of life lost per deceased individual. This metric underscores the considerable premature mortality associated with COVID-19, highlighting the need to reevaluate the perceived insignificance of these deaths. The results emphasize the profound consequences of the pandemic, particularly in terms of the years of life forfeited, and underscore the importance of acknowledging the true extent of COVID-19's impact on human lifespan.
C4	What is the estimated average years of life lost among the first 1,279,866 reported COVID-19 deaths?
S0	Years of life lost to COVID-19 in 81 countries. Understanding the mortality impact of COVID-19 requires not only counting the dead, but analyzing how premature the deaths are. We calculate years of life lost (YLL) across 81 countries due to COVID-19 attributable deaths, and also conduct an analysis based on estimated excess deaths. We find that over 20.5 million years of life have been lost to COVID-19 globally. As of January 6, 2021, YLL in heavily affected countries are 2–9 times the average seasonal influenza; three quarters of the YLL result from deaths in ages below 75 and almost a third from deaths below 55; and men have lost 45% more life years than women. The results confirm the large mortality impact of COVID-19 among the elderly. They also call for heightened awareness in devising policies that protect vulnerable demographics losing the largest number of life-years.
S1	To understand the mortality impact of COVID-19, it's essential to analyze the prematurity of deaths, not just the number of deaths. We calculated years of life lost (YLL) due to COVID-19 in 81 countries and found that over 20.5 million years of life have been lost globally. As of January 6, 2021, heavily affected countries experienced 2-9 times more YLL than the average seasonal influenza. Notably, three-quarters of YLL resulted from deaths in people under 75, and almost a third from deaths in people under 55. Men lost 45% more life years than women. The results highlight the significant mortality impact of COVID-19, particularly among the elderly and other vulnerable demographics, emphasizing the need for policies that protect these groups.
S2	We calculated years of life lost (YLL) due to COVID-19 in 81 countries, finding over 20.5 million years lost globally. Analysis revealed that YLL in heavily affected countries are 2-9 times higher than average seasonal influenza, with 75% of YLL resulting from deaths under 75 years and 30% from deaths under 55. Men lost 45% more life years than women, highlighting the significant mortality impact of COVID-19, particularly among the elderly and younger demographics.
S3	"COVID-19 has claimed 20.5M+ years of life globally! 75% of years lost are from people under 75. Men lost 45% more years than women. #COVID19 #GlobalHealth #MortalityRate"