Overview of the CLEF-2026 CheckThat! Lab Task 1 on Source Retrieval for Scientific Web Claims

CheckThat! Lab at CLEF 2026

Sebastian Schellhammer[1,2,*], Salim Hafid[4,*], Yavuz Selim Kartal[1], Katarina Boland[2], Dimitar Dimitrov[1], Konstantin Todorov[3] and Stefan Dietze[1,2]

[1]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
[2]Heinrich-Heine-University, Düsseldorf, Germany
[3]LIRMM, CNRS, University of Montpellier, Montpellier, France
[4]médialab Sciences Po, Paris, France

Abstract

We present an overview of Task 1 of the CheckThat! Lab at the 2026 edition of the Conference and Labs of the Evaluation Forum (CLEF). The task focuses on source retrieval for scientific web claims, which given a social media post that contains a scientific claim and an implicit reference to a scientific paper, aims to retrieve the mentioned paper from a larger pool of candidates. The task is a follow-up from Task 4b of the previous lab edition and extends it to a multilingual setting, providing more than 30k posts in English, German, and French along with a collection set of 10k publications. In the context of the CheckThat! Lab that aims to improve automated fact-checking, this task contributes to the retrieval of scientific evidence for claim verification. The task attracted considerable attention with 37 participating teams and 21 submitted system description papers. Most teams employed two-stage retrieval and reranking pipelines, with the best team achieving an average MRR@5 score across all three languages of 0.763. This paper presents a detailed overview of the task, including the datasets, evaluation setting, and the participants' approaches.

Keywords

scientific web discourse, scientific claims, evidence retrieval

1. Introduction

Scientific studies are increasingly mentioned in social media discussions, reflecting the rise of scientific web discourse, which has been studied across multiple disciplines [1, 2, 3]. As such discourse is often informal, scientific resources are typically referenced implicitly rather than explicitly using URLs or DOIs [4]. For example, posts often refer to them through mentions such as *"A #Stanford University #study asserts [...]"* (ID 219 in the English test set), where the actual study is never cited using direct links or other identifiers [4]. The lack of explicit citations poses challenges for the computational analysis of scientific web discourse, including the mining of social media data and the computation of altmetrics, and also contributes to poorly informed online debates. For automated fact-checking, identifying the source of a claim is a crucial step, as it enables verification against the underlying evidence and the assessment of its validity [5]. This is particularly important for scientific web claims, where research suggests that automated verification is more difficult than for general web claims [6]. The difficulty for this task in particular is driven by the linguistic gap between informal social media posts and the formal language of scientific publications, the presence of multiple candidate papers that could plausibly match a vague claim, and the selective or inaccurate representation of scientific findings in posts.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
sebastian.schellhammer@gesis.org (S. Schellhammer); salim.hafid@sciencespo.fr (S. Hafid); yavuzselim.kartal@gesis.org (Y. S. Kartal); katarina.boland@hhu.de (K. Boland); dimitar.dimitrov@gesis.org (D. Dimitrov); todorov@lirmm.fr (K. Todorov); stefan.dietze@hhu.de (S. Dietze)
0009-0001-6413-5823 (S. Schellhammer); 0000-0002-1775-8542 (S. Hafid); 0000-0002-2146-2680 (Y. S. Kartal); 0000-0003-2958-9712 (K. Boland); 0000-0002-4504-5144 (D. Dimitrov); 0000-0002-9116-6692 (K. Todorov); 0009-0001-4364-9243 (S. Dietze)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

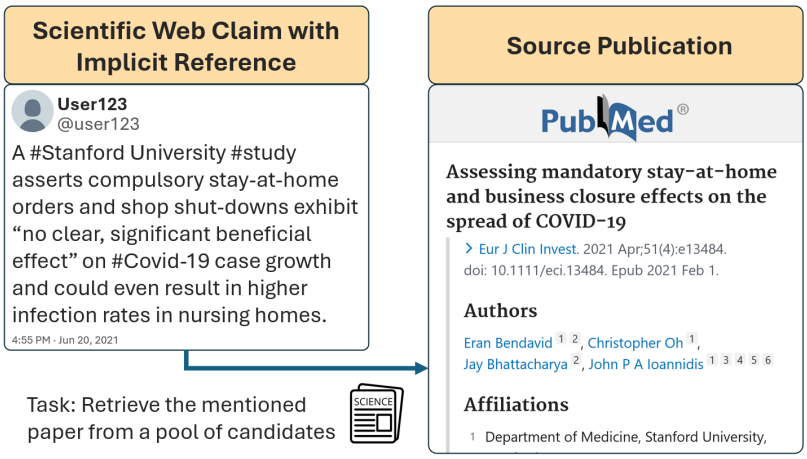

Figure 1: Task overview: the left box shows a scientific web claim with an implicit reference for which the source publication (right box) needs to be retrieved from a pool of candidate papers.

Building on this motivation, the task of **source retrieval for scientific web claims** was introduced in the 2025 edition of the lab [7]. In this iteration, the task is extended to a multilingual setting, providing scientific web claims in English, German, and French [8]. An overview of the task is shown in Figure 1.

Task definition Given a social media post that contains a scientific claim and an implicit reference to a scientific paper (mentions it without a URL), retrieve the mentioned paper from a pool of candidate papers.

2. Dataset

The datasets cover three languages: English, German, and French, with language-specific query sets and a single collection set shared across all languages. The query sets consist of social media posts from X (formerly Twitter) that contain scientific claims and implicit references to scientific publications. For each language, the query set is divided into training, development, and test splits. Table 1 summarizes the query set statistics for each language and split. The collection set contains metadata for 10k scientific publications referenced by the posts, including titles, abstracts, author names, and venues. For English, parts of the training and development data splits are reused from last year's Task 4b [7], while for German and French, entirely new data were collected. The datasets were collected by retrieving social media posts from X that contain URLs to scientific publications and filtering for posts where the URL appears at the end. Additional filtering removed posts with insufficient textual content or those that merely restate the publication title, retaining those likely to include claims. Finally, the URLs were removed to obtain posts with implicit references to scientific studies. While this is only an approximation of social media posts with actual implicit references, the approach allows us to obtain large ground-truth datasets without the extensive human annotation effort required to compare posts to all papers in our collection. Tables 2 and 3 show query and collection set examples.

Table 1
Query dataset statistics: number of posts per language and data split.

Split	English	German	French
Train	14,977	1,460	2,807
Dev	3,905	386	702
Test	6,076	876	1,220

Table 2
X posts in English (EN), German (DE), and French (FR) from the query sets along with the publication IDs of the implicitly referenced publications

Language	X post	Publication ID
EN	Peer-reviewed in the New England Journal of Medicine regarding Delta (B.1.617.2): • Pfizer is around 90% effective • AstraZeneca is around 70% effective. This falls in line with vaccine efficacy of other variants. Yes, the vaccines ARE indeed effective against Delta.	123456
EN	Published in the journal Antiviral Research, the study from Monash University showed that a single dose of Ivermectin could stop the coronavirus growing in cell culture – effectively eradicating all genetic material of the virus within two days.	234567
DE	Die in der Fachzeitschrift Antiviral Research veröffentlichte Studie der Monash University zeigte, dass eine einzige Dosis Ivermectin das Wachstum des Coronavirus in Zellkulturen stoppen kann und das gesamte genetische Material des Virus innerhalb von zwei Tagen effektiv zerstört.	234567
FR	L'outil génétique CRISPR-Cas9 a été utilisé pour permettre à la provitamine D3 de s'accumuler dans les feuilles & fruits de la tomate. Une fois exposée aux UV, elle s'est bien convertie en vitamine D (dont 1 milliard de personnes est carencé ds le monde)	345678

Table 3
Publications and their metadata from the collection set

Publication ID	Title	Authors	Venue	Selected Abstract Text
123456	Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant	Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, [...]	New England Journal of Medicine	With the BNT162b2 vaccine, the effectiveness of two doses was 93.7% (95% CI, 91.6 to 95.3) [...]. With the ChAdOx1 nCoV-19 vaccine, the effectiveness of two doses was 74.5% (95% CI, 68.4 to 79.4) [...]
234567	The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro	Caly, Leon; Druce, Julian D.; Catton, Mike G.; Jans, David A.; Wagstaff, Kylie M.	Antiviral Research	We report here that Ivermectin [...] is an inhibitor of the causative virus (SARS-CoV-2), with a single addition to Vero-hSLAM cells 2 h post infection [...]
345678	Biofortified tomatoes provide a new route to vitamin D sufficiency	Jie Li, John Innes Centre, Norwich Research Park, Norwich, UK; Aurelia Scarano, Institute of [...]	Nature Plants	We have engineered the accumulation of provitamin D3 in tomato by genome editing, modifying a duplicated section of phytosterol biosynthesis in Solanaceous plants [...]

3. Results and Overview of the Systems

In total, 37 teams participated in Task 1. Table 4 provides an overview of the different approaches and their performance for teams that submitted a system description paper. The teams are ranked by their average MRR@5 performance, while teams that did not participate in all languages or submitted after the competition ended are unranked.

Most teams relied on, or extended, two-stage retrieval-reranking pipelines using retrievers for candidate generation and rerankers to rank the top-k candidate publications. Retrieval methods predominantly include dense models rather than sparse approaches such as BM25. To optimize performance,

teams experimented with translating German and French posts into English, augmenting posts and publications, and fine-tuning retriever and reranker models. Overall, the performance between English and French posts was comparable, while the average performance for German posts was slightly lower. The following gives an overview of the approaches for which a system description paper was submitted.

Among all participating teams, team **Claim2Source** [9] achieved the highest performance on the English and French subsets, ranking first with an average MRR@5 of 0.763. Their approach translated German and French posts into English to create bilingual claim representations and relied on source representations that included all available publication metadata. The post and publication representations were used in a three-stage retrieval process: a `GritLM-7B` dense retrieval model for candidate generation, a similarity-based reranking using an instruction-tuned `Qwen3-8B` model, followed by a verification-based reranking with `Kimi-K2.6`.

Team **Outsider** [10] achieved the best performance on the German subset and ranked second overall. They employed a two-stage pipeline with a fine-tuned `Harrier-OSS-v1-27B` dense retriever and a fine-tuned `Qwen3-VL-Reranker-8B` for reranking. For fine-tuning the retrieval and reranking model, only English data was used.

Team **Boyes Boys** [11], ranked third, combined a `Harrier-OSS-v1-27B` zero-shot retriever with a sparse BM25 retriever, merging the two candidate lists with a fine-tuned Random Forest fusion model. The top-30 candidate publications were reranked using a `Qwen3-Reranker-8B`.

Team **CheckMate** [12] ranked fourth with a fine-tuned Qwen3-Embedding-8B retriever and two different fine-tuned reranker models: `Llama-nemotron-rerank-1b-v2` and `Qwen3-Reranker-8B`. The final ranking of scientific publications was derived using multi-model fusion, combining and normalizing the relevance scores from the three components.

Team **RETRIX** [13], which ranked fifth, first translated the social media posts to English and expanded them with additional context from semantically similar publications. The expanded queries were then used in a two-stage retriever-reranker pipeline with a `BGE-M3` retrieval model and a fine-tuned `Llama-nemotron-rerank-1b-v2` for reranking.

Team **MeVer** [14] ranked sixth, combining a `bge-large-en-v1.5` retrieval model with a `jina-reranker-v2-base-multilingual` reranker, both fine-tuned with a focus on different hard negative mining strategies. Disagreements between the retriever and reranker predictions were resolved by a `GPT-5.5` LLM-as-a-judge.

Team **PEI** [15] achieved seventh place by translating German and French posts into English and normalizing them. They used a zero-shot two-stage approach with a `Harrier-OSS-v1-27b` retriever and a `jina-reranker-v3`.

Team **Awakened** [16], which ranked tenth, augmented the representation of publications by prompting `GPT-4o-mini` to create tweet-like paper summaries. The enriched paper representations were used with the social media posts to fine-tune a `multilingual-e5-large` retrieval model. The top-10 candidates, together with posts translated into English, were passed to the `bge-reranker-v2-m3` reranker. The final ranking was based on the scores from both the retrieval and reranking models.

Team **MG CheckThat2026** [17] ranked eleventh with a pipeline combining hybrid candidate generation with language-specific initial ranking and a final LLM-based listwise reranking. For hybrid candidate generation, the top-30 papers from the sparse BM25 algorithm and the top-100 papers from the dense retrieval model (`bge-m3`) were concatenated. In the second stage, language-specific ranking models (`bge-reranker-v2-m3`) were fine-tuned using hard negative mining based on the output of the candidate generation step. Ultimately, for the LLM reranking step, prompts containing the top-10 paper candidates were sent to the DeepSeek API.

Team **varytrieval** [18], which ranked 12th, combined rule-based string matching with dense retrieval and reranking. For the string matching, the system tried to match publications based on explicit signals in the posts, such as fragments of the publication's title, author names, journals, and conferences, or quotations from the paper's abstract. In case no matching publication was found, a fine-tuned `Qwen3-Embedding-0.6B` model was fine-tuned for dense retrieval using English translations of the posts as input. Ultimately, a `bge-reranker-v2-gemma` reranker was applied for German and French posts. Overall, the rules-based string matching improved the MRR@5 performance slightly by about

0.01.

Team **FELLA** [19], ranked thirteenth, developed a multilingual multi-stage pipeline that translated German and French posts into English using LibreTranslate and explored BM25, SPECTER2, Reciprocal Rank Fusion, and several bi-encoder retrievers. Their strongest approach used `Stella-en-1.5B-v5` to retrieve the top-50 candidate publications, followed by cross-encoder reranking with `bge-reranker-v2-m3`.

Team **BoPC** [20], ranked fifteenth, used a two-stage retrieval architecture with a multilingual fine-tuned `multilingual-e5-large-instruct` encoder and a multilingual `bge-reranker-v2-m3` cross-encoder reranker via an existing document retrieval framework. They employed this approach for all languages.

Team **SEUPD Cobalt** [21], ranked sixteenth, explored sparse and neural retrieval pipelines. Their sparse pipeline, thought for *"realistic commodity hardware"* such as laptops, combined multilingual translation, synonym-based normalization, multi-field BM25 retrieval, and `ms-marco- TinyBERT-L-2-v2` cross-encoder reranking. Their neural approach used contrastive domain fine-tuning of a multilingual bi-encoder and `doc2query` expansion of publication abstracts. The neural configuration was selected for the official submission.

Team **KhyontekAI** [22], which ranked 18th, utilized a multilingual two-stage retrieval framework with a fine-tuned `bge-base-en-v1.5` retrieval model, and a fine-tuned `ms-marco-MiniLM-L6-v2` cross-encoder for reranking. To improve retrieval quality, they experimented with joint multilingual training and oversampling German posts.

Team **NEXA** [23], ranked twentieth, developed a system that incorporates sparse retrieval based on BM25 and dense retrieval using a `bge-m3` model. The scores from both approaches were added with a weight tuned on the development sets.

Team **SAGASU** [24], ranked twenty-fourth, combined BM25-based sparse retrieval with a fine-tuned `ms-marco-MiniLM-L6-v2` reranker, translating German and French posts into English with `opus-mt` models. They experimented with stemming, stop-word removal, and weighting of different publication metadata, such as the title or the authors, for the first-stage retrieval. Their strongest configuration used four publication metadata fields followed by neural reranking.

Team **Jesters** [25] applied an approach that combines BM25 sparse retrieval and dense retrieval using the `multilingual-e5-large embedding model`. They fine-tuned the multilingual E5 model with contrastive learning and hard negatives mined from combined retrievers. They also translated English posts into French and German to improve the performance in these languages.

Team **Infraglyph** [26], which ranked thirty-third, implemented a hybrid two-stage retrieval approach, where the initial ranking stage included both a BM25 lexical ranker and a semantic-based dense retriever (`multilingual-e5-base`), which were then fused into a reciprocal rank before the final reranking stage.

Team **SINAI-UGPLN** [27] ranked thirty-fourth by employing a pretrained retrieval model (`multilingual-e5-large`) that ranked publications based on their semantic similarity to the target post.

Team **AKSH** [28] developed a hybrid retrieval system that fine-tuned `multilingual-e5-large` on pooled English, French, and German tweet–paper pairs and combined its dense similarity scores with BM25 scores through per-query min–max normalization and linear fusion.

Team **Sciclaim Seekers** [29], ranked 11th on the English subset, used a three-stage approach. They first retrieve candidate documents using a sparse retriever, BM25, and a dense retriever, `E5-large-instruct`. Both results are combined by reciprocal rank fusion ($k = 60$). Finally, they used `Qwen2.5-14B-Instruct` to rerank the final results.

4. Related Work

Scientific claim source retrieval is closely related to the problem of evidence retrieval in automated fact-checking [30], as explored by previous research [31, 32, 33, 34]. For instance, the FEVER shared

Table 4
Leaderboard and overview of the approaches. Teams that did not submit a system description paper are uncategorized. **Bold scores** represent the best performance on average and for each language.

Rank	Team	Approach							MRR@5 Performance			
		Dense Retrieval	Sparse Retrieval	Reranking	Fine-tune Retriever	Fine-tune Ranker	Translations	Other Optimizations	Avg.	EN	DE	FR
1	Claim2Source [9]	■		■	■	■	■	■	**0.763**	**0.782**	0.715	**0.791**
2	Outsider [10]	■		■	■	■			0.758	0.780	**0.723**	0.771
3	Boyes Boys [11]	■	■	■			■	■	0.746	0.757	0.707	0.775
4	CheckMate [12]	■		■	■	■		■	0.719	0.734	0.686	0.738
5	RETRIX [13]	■	■	■		■	■	■	0.711	0.712	0.681	0.740
6	MeVer [14]	■		■	■	■		■	0.684	0.683	0.661	0.707
7	PEI [15]	■		■			■	■	0.682	0.698	0.645	0.702
8	AlexCheck								0.660	0.665	0.623	0.691
9	LeaderboardAddicts								0.649	0.653	0.632	0.662
10	Awakened [16]	■		■			■	■	0.640	0.644	0.613	0.664
11	MG_CheckThat2026 [17]	■	■	■		■	■	■	0.639	0.638	0.597	0.683
12	varytrieval [18]	■	■	■	■		■	■	0.639	0.691	0.582	0.644
13	FELLA [19]	■		■			■		0.596	0.606	0.564	0.618
14	Galaxy								0.596	0.602	0.560	0.625
15	BoPC [20]	■		■	■	■			0.594	0.593	0.562	0.627
16	SEUPD Cobalt [21]	■		■	■	■		■	0.575	0.578	0.541	0.606
17	FactX								0.568	0.597	0.521	0.585
18	KhyontekAI [22]	■		■	■	■			0.556	0.596	0.501	0.572
19	Os reis da bola								0.550	0.574	0.490	0.586
20	NEXA [23]	■	■	■			■		0.543	0.570	0.487	0.574
21	meth3group2								0.543	0.565	0.499	0.565
22	Performers								0.533	0.542	0.491	0.566
23	CT-Task1-Methodology1-Group1								0.514	0.547	0.457	0.537
24	SAGASU [24]		■	■		■			0.510	0.561	0.438	0.531
25	AIR-M3-G1								0.506	0.525	0.449	0.544
26	DELL								0.500	0.530	0.446	0.524
27	DSPY								0.497	0.486	0.466	0.540
28	mailuefterl								0.493	0.557	0.416	0.505
29	FMI-BioNLP-AAC								0.490	0.488	0.435	0.547
30	CheckMates								0.489	0.538	0.412	0.517
31	Jesters [25]	■	■		■		■		0.488	0.480	0.455	0.530
32	Unibuc-NLP								0.487	0.555	0.407	0.499
33	Infraglyph [26]	■	■	■			■	■	0.482	0.506	0.432	0.510
34	SINAI-UGPLN [27]	■	■						0.433	0.494	0.373	0.431
35	CLaC@CT								0.240	0.454	0.129	0.138
36	Group1								0.235	0.535	0.085	0.084
-	AKSH [28]	■	■		■				0.520	0.453	0.572	0.535
-	Sciclaim Seekers [29]	■	■	■					-	0.644	-	-

task focused on evidence retrieval from Wikipedia for human-authored claims [31]. In a previous iteration of the CheckThat! Lab, a similar task was introduced in which given a rumour, evidence tweets from authority Twitter accounts should be retrieved [33]. However, these tasks differ from scientific claim source retrieval in several aspects. They often rely on synthetic claims [31], claims from different sources such as scientific publications [32], or on evidence from sources other than scientific

publications. In addition, previous evidence retrieval tasks aim to retrieve any relevant evidence useful to verify a claim, rather than identifying the specific source that is most likely the basis for the stated claim. Targeting the specific source of a claim is particularly important as prior work shows that the source is among the primary pieces of evidence used by fact-checkers to verify claims [5]. Another related line of work is the retrieval of publications referred to by news articles [35, 36, 37]. While similar to the implicit references in our dataset, we assume the references in news articles are more formal and have a larger context.

The task of source retrieval for scientific web claims was first introduced by Hafid et al. [4] and later included in the 2025 edition of the CheckThat! Lab [7]. The participating teams explored two-stage pipelines combining sparse and/or dense retrieval with reranking [38, 39, 40]. To overcome the linguistic differences between web claims and scientific publications, Schreieder and Färber [41] investigate the impact of style transfer and find that formalizing claims generally improves retrieval performance, while changing the style of publications tends to decrease it. Similarly, Schofield et al. [42] find that concatenating web claims with their formalized version slightly improves results. Staudinger et al. [43] examine the influence of enriching the scientific publications with full-texts and summaries but do not find that this improves performance. Building on the previous edition, this iteration extends the task to a multilingual setting incorporating English, French, and German.

5. Conclusion and Future Work

We presented an overview of Task 1 of the CheckThat! Lab at CLEF 2026, which aims to retrieve the scientific source publication that is implicitly mentioned in a given scientific web claim. In total, 37 teams submitted their predictions, and 21 submitted system description papers, attracting considerable interest. Most teams relied on two-stage retrieval pipelines combining initial sparse retrieval, dense retrieval, or both with a subsequent reranking step. Many of the high-performing approaches included large dense retrieval models combined with neural rerankers, often supplemented by model fine-tuning and the translation of German and French posts into English.

The winning team **Claim2Source** [9] achieved an average MRR@5 of 0.763. Their approach constructs bilingual claim representations and applies a three-stage pipeline consisting of candidate generation using a `GritLM` model, similarity-based reranking with an instruction-tuned `Qwen3-8B` model, and a final verification-based reranking step using a `Kimi-2.6` model.

Compared to the previous task iteration [7], which included only English claims, the performance increased significantly from the highest MRR@5 score in 2025 of 0.68 to 0.78 on the English subset in this iteration. One possible explanation is the use of larger models. For instance, the second-best team in 2025 [38] used a fine-tuned `e5-large-v2` retrieval model (0.3B parameters) with a `scibert_scivocab_uncased` (110M parameters) model for reranking. In contrast, team **Claim2Source** [9] used three different models with the number of parameters ranging from 7B to 32B.

While the performance of the high-ranked teams increased compared to the previous iteration, MRR@5 scores close to 0.8 suggest that there is still room for improvement. This is especially true for German posts, where performance is generally weaker.

6. Acknowledgements

This work has been funded by the AI4Sci grant (co-funded by MESRI [France, grant UM-211745], BMBF [Germany, grant 01IS21086], and the French National Research Agency ANR), the Leibniz Association as part of the Leibniz Collaborative Excellence funding programme (grant no K490/2022), and the EMO-SCI grant (funded by ANR [grant ANR-25-CE23-2066]).

7. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly to perform the following tasks: "Grammar and spelling check", "Paraphrase and reword". After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Brüggemann, I. Lörcher, S. Walter, Post-normal science communication: exploring the blurring boundaries of science and journalism, Journal of Science Communication 19 (2020) A02.

[2] S. Dunwoody, Science journalism: Prospects in the digital age, in: Routledge handbook of public communication of science and technology, Routledge, 2021, pp. 14–32.

[3] S. Hafid, S. Schellhammer, S. Bringay, K. Todorov, S. Dietze, Scitweets-a dataset and annotation framework for detecting scientific online discourse, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 3988–3992.

[4] S. Hafid, Y. S. Kartal, S. Schellhammer, V. Jacot, S. Bringay, S. Dietze, K. Todorov, Disambiguation of implicit scientific references on x, in: Proceedings of the 36th ACM Conference on Hypertext and Social Media, HT '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 165–170. URL: https://doi.org/10.1145/3720553.3746676. doi:10.1145/3720553.3746676.

[5] M. Glockner, Y. Hou, I. Gurevych, Missing counter-evidence renders NLP fact-checking unrealistic for misinformation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5916–5936.

[6] S. Hafid, S. Schellhammer, Y. S. Kartal, T. Papastergiou, S. Dietze, S. Bringay, K. Todorov, An in-depth analysis of the linguistic characteristics of science claims on the web and their impact on fact-checking, ACM Trans. Web 19 (2025). URL: https://doi.org/10.1145/3746170. doi:10.1145/3746170.

[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: [44], 2025.

[8] J. M. Struß, S. Schellhammer, S. Dietze, V. V., V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, D. Sahnan, K. Todorov, Overview of the CLEF-2026 CheckThat! Lab: Advancing multilingual fact-checking, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, Alba García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[9] T. Schreieder, H. Khandelwal, Y.-L. Zhong, M. Färber, Claim2Source at CheckThat! 2026: Improving Multilingual Scientific Claim–Source Retrieval with Verification-based Re-Ranking, in: [45], 2026.

[10] Q. T. Le, I. Badache, M. E. A. Hamri, Outsider at CheckThat! 2026: Retrieval, Verification, and Generation for Multilingual Fact-Checking, in: [45], 2026.

[11] O. Svenaeus, E. Nazário, M. Kvist, D. Suzener, Boyes Boys at CheckThat! 2026: Go big or go home, in: [45], 2026.

[12] T. Munawar, A. Amir, F. Alvi, A. Samad, Checkmate at checkthat! 2026: Fusion-based retrieval, reasoning-trace ranking, and citation-controlled article generation, in: [45], 2026.

[13] F. Baldan, G. Padoan, D. Donati, A. Garberino, M. Tessari, RETRIX at CheckThat! 2026: Source Retrieval for Scientific Web Claims, in: [45], 2026.

[14] J. Bakagianni, S. Papadopoulos, MeVer at CheckThat! 2026: Cluster-Aware Hard-Negative Mining for Multilingual Scientific-Source Retrieval, in: [45], 2026.

[15] F. Zeidi, F. Zeidi, R. Christof, R. König, L. Childs, PEI at CheckThat! 2026: Investigating Query Normalization and Retrieve-and-Rerank Strategies for Linking Scientific Web Claims to Publications, in: [45], 2026.

[16] M.-D. Cotelin, C.-O. Truică, E.-S. Apostol, Awakened at CheckThat! 2026: X-IBIS: an Iterative

Bi-Encoder Fine-Tuning with Hard-Negative Mining for Cross-Lingual Scientific Source Retrieval, in: [45], 2026.
[17] L. Junyi, P. Zhenyu, MG_CheckThat2026 at CheckThat! 2026: A Multilingual Cascade Pipeline for Scientific Claim Source Retrieval, in: [45], 2026.
[18] A. Hövelmeyer, Varytrieval at CheckThat! 2026: Combining Rule-Based String Matching and Dense Retrieval for Scientific Claim Source Retrieval, in: [45], 2026.
[19] F. Corradi, E. De Lorenzo Buratta, L. Fantin, A. Sartor, L. Tamburi, N. Ferro, FELLA at CheckThat! 2026: From BM25 to Stella, a Multilingual Pipeline for Scientific Source Retrieval, in: [45], 2026.
[20] N.-P. Mac, M.-P. Mac, D. X. Tran, V. V.-A. Nguyen, BoPC at CheckThat! 2026: Efficient Source Retrieval for Social Media Claims Using Two-Stage Dense and Cross-Attention Networks, in: [45], 2026.
[21] F. Battisti, I. Da Silva, M. Trevisan, M. Trifunovic, T. Zogno, SEUPD Cobalt at CheckThat! 2026: Exploring Sparse and Neural Retrieval from Social Media Claims to Scientific Sources, in: [45], 2026.
[22] B. J. Bhuyan, P. Deka, KhyontekAI at CheckThat! 2026: Multilingual Dense Retrieval and Cross-Encoder Reranking for Scientific Web Claims, in: [45], 2026.
[23] P. Arlot, A. Di Tillo, G. G. Flagstad, B. Khashechian, D. Smirnov, M. Tomaiuoli, N. Ferro, NEXA at CheckThat! 2026: Hybrid Source Retrieval for Cross-Lingual Scientific Web Claims, in: [45], 2026.
[24] G. Brocco, D. Daccordo, V. S. Ho, D. Pinzan, A. Sancetta, N. Ferro, SAGASU at CheckThat! 2026: Combining Sparse and Dense Approaches for Implicit Source Retrieval, in: [45], 2026.
[25] D. Vaghasiya, V. Goti, Jesters at CheckThat! 2026: Multilingual Scientific Source Retrieval with Hybrid Dense-Sparse Ranking, in: [45], 2026.
[26] H. Patel, U. Kava, T. Mandl, S. Modha, Infraglyph at CheckThat! 2026: Deep Hybrid Multi-Stage Scientific Source Retrieval from Implicit Social Media Claims, in: [45], 2026.
[27] M. Toapanta, M. Á. Garcia-Cumbreras, L. A. Ureña-López, SINAI-UGPLN at CheckThat! 2026: Dense Retrieval and Neural Reranking for Scientific Source Retrieval from Social Media Claims, in: [45], 2026.
[28] S. Siddiqui, F. Alvi, A. Samad, Team AKSH at CheckThat! 2026: Resource-Efficient Hybrid Retrieval for Multilingual Scientific Source Retrieval, in: [45], 2026.
[29] M. Rashid, N. Baniya, A. S. Anik, X. Song, L. Hong, SciClaimSeekers at CheckThat! 2026: Retrieving Scientific Sources for Social Media Claims with LLM Reranking, in: [45], 2026.
[30] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barron-Cedeno, P. Papotti, S. Shaar, G. Da San Martino, et al., Automated fact-checking for assisting human fact-checkers, in: IJCAI, International Joint Conferences on Artificial Intelligence, 2021, pp. 4551–4558.
[31] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal (Eds.), Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–9. URL: https://aclanthology.org/W18-5501/. doi:10.18653/v1/W18-5501.
[32] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609/. doi:10.18653/v1/2020.emnlp-main.609.
[33] F. Haouari, T. Elsayed, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab task 5 on rumor verification using evidence from authorities, in: Conference and Labs of the Evaluation Forum, 2024.
[34] J. Chen, G. Kim, A. Sriram, G. Durrett, E. Choi, Complex claim verification with evidence retrieved in the wild, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3569–3587. URL: https://aclanthology.org/2024.naacl-long.196/. doi:10.18653/

v1/2024.naacl-long.196.
[35] J. Wang, B. Yu, News2pubmed: A browser extension for linking health news to medical literature, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2605–2609.
[36] K. Kousha, M. Thelwall, An automatic method to identify citations to journals in news stories: A case study of uk newspapers citing web of science journals, Journal of Data and Information Science 4 (2019) 73–95.
[37] J. Ravenscroft, A. Clare, M. Liakata, HarriGT: A tool for linking news to science, in: F. Liu, T. Solorio (Eds.), Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 19–24. URL: https://aclanthology.org/P18-4004/. doi:10.18653/v1/P18-4004.
[38] C. Ashbaugh, L. Baumgärtner, T. Greß, N. Sidorov, D. Werner, AIRwaves at CheckThat! 2025: Retrieving Scientific Sources for Implicit Claims on Social Media with Dual Encoders and Neural Re-Ranking, in: [44], 2025.
[39] P. J. Sager, A. Kamaraj, B. F. Grewe, T. Stadelmann, Deep Retrieval at CheckThat! 2025: Identifying Scientific Papers from Implicit Social Media Mentions via Hybrid Retrieval and Re-Ranking, in: [44], 2025.
[40] G. Marchetti, G. Rocha, H. L. Cardoso, Team SeRRa at CheckThat! 2025: Sequential Re-Ranking in a Scientific Claim Source Retrieval Pipeline, in: [44], 2025.
[41] T. Schreieder, M. Färber, Claim2Source at CheckThat! 2025: Zero-Shot Style Transfer for Scientific Claim-Source Retrieval, in: [44], 2025.
[42] J. Schofield, S. Tian, H. Truong, M. Heil, DS@GT at CheckThat! 2025: Exploring Retrieval and Reranking Pipelines for Scientific Claim Source Retrieval on Social Media Discourse, in: [44], 2025.
[43] M. Staudinger, A. El-Ebshihy, W. Kusa, F. Piroi, A. Hanbury, ATOM at CheckThat! 2025: Retrieve the Implicit - Scientific Evidence Retrieval, in: [44], 2025.
[44] 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.
[45] E. S. Salido, A. Barrón-Cedeño, Alba García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CLEF 2026, Jena, Germany, 2026.