

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Struß, J. M., Schellhammer, S., Dietze, S., Venkatesh, V., Setty, V., Chakraborty, T., Nakov, P., Anand, A., Chungkham, P., & More Authors (2026). The CLEF-2026 CheckThat! Lab: Advancing Multilingual Fact-Checking. In R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, N. Bai, M. Mansoury, & More Editors (Eds.), *Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Proceedings* (pp. 325-335). (Lecture Notes in Computer Science; Vol. 16486 LNCS). Springer. https://doi.org/10.1007/978-3-032-21321-1_43

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



The CLEF-2026 CheckThat! Lab: Advancing Multilingual Fact-Checking

Julia Maria Struß¹ , Sebastian Schellhammer^{2,3} , Stefan Dietze^{2,3} ,
Venkatesh V.⁴ , Vinay Setty⁵ , Tanmoy Chakraborty⁶ , Preslav Nakov⁷ ,
Avishek Anand⁸ , Primakov Chungkham⁹ , Salim Hafid¹⁰ ,
Dhruv Sahnan⁷ , and Konstantin Todorov¹¹

¹ University of Applied Sciences Potsdam, Potsdam, Germany
julia.struss@fh-potsdam.de

² GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

³ Heinrich-Heine-University, Düsseldorf, Germany

⁴ Stockholm University, Stockholm, Sweden

⁵ University of Stavanger, Stavanger, Norway

⁶ Indian Institute of Technology Delhi, Delhi, India

⁷ Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

⁸ Delft University of Technology, Delft, The Netherlands

⁹ Delft, The Netherlands

¹⁰ Médialab, Sciences Po, Paris, France

¹¹ University of Montpellier, LIRMM, CNRS, Montpellier, France
<https://checkthat.gitlab.io>

Abstract. The CheckThat! lab aims to advance the development of innovative technologies combating disinformation and manipulation efforts in online communication across a multitude of languages and platforms. While in early editions the focus has been on core tasks of the verification pipeline (check-worthiness, evidence retrieval, and verification), in the past three editions, the lab added additional tasks linked to the verification process. In this year's edition, the verification pipeline is at the center again with the following tasks: Task 1 on source retrieval for scientific web claims (a follow-up of the 2025 edition), Task 2 on fact-checking numerical and temporal claims, which adds a reasoning component to the 2025 edition, and Task 3, which expands the verification pipeline with generation of full-fact-checking articles. These tasks represent challenging classification and retrieval problems as well as generation challenges at the document and span level, including multilingual settings.

Keywords: disinformation · fact-checking · claim source retrieval · generating fact-checking articles

Primakov Chungkham—Independent Researcher.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
R. Campos et al. (Eds.): ECIR 2026, LNCS 16486, pp. 325–335, 2026.
https://doi.org/10.1007/978-3-032-21321-1_43

1 Introduction

The **CheckThat!** lab aims at fostering the development of technology to combat the spreading of mis- and disinformation across multiple platforms and languages in online discourses. Over the past eight iterations (e.g., see [1,4]), a multitude of challenges and tasks along the main steps of the verification pipeline used by *journalists* and *fact-checkers* (see Fig. 1) have been offered: Starting with a given document or a claim, the first step is to evaluate its check-worthiness, i.e., whether human effort should be allocated for checking its veracity, which is a task that experts have repeatedly emphasized as being of particular importance to them in the past. As claims that have been proven false before are repeatedly reformulated and circulated, retrieving closely related, previously fact-checked claims is the next step in the verification process. Retrieving additional evidence from diverse sources to verify the claim and finally decide whether the claim is factually true or not, with various degrees in between, follows. Ultimately, the verification process is completed by writing a detailed article that provides the evidence and the final decision, a step which has been added to the **CheckThat!** verification pipeline for the first time in the 2026 edition of the lab.

The three tasks proposed for the 2026 edition cover three of the main tasks in the pipeline, highlighted in Fig. 1, and span a total of five different languages (see in Table 1):

Task 1 Source Retrieval for Scientific Web Claims, is related to the second and third task of the verification pipeline and aims to detect different types of references or mentions to scientific work as well as identifying the original study a social media post refers to (cf. Sect. 2).

Task 2 Fact-Checking Numerical and Temporal Claims addresses the claim verification task of the pipeline, focusing on numerical claims (cf. Sect. 3).

Task 3 Generating Full-Fact-Checking Articles introduces a new, final task into the CLEF CheckThat! lab pipeline, which attempts to automate the fact-checking article writing process (cf. Sect. 4).

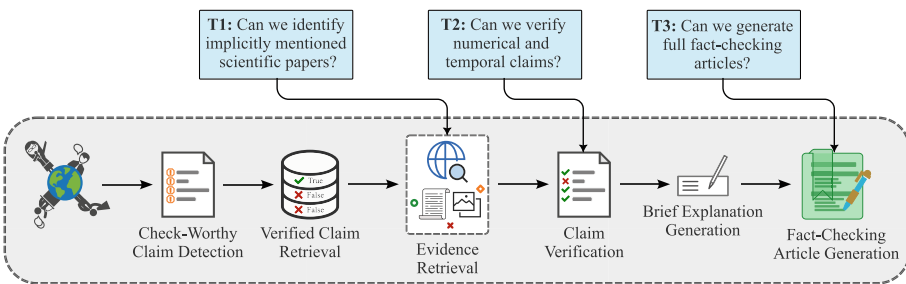


Fig. 1. Overview of the **CheckThat!** verification pipeline, which features the core fact-checking tasks along with the tasks we address in the 2026 edition of the lab.

Table 1. Languages targeted in the three tasks of the CheckThat! 2026 edition.

Task	Arabic	English	German	French	Spanish
Task 1		■	■	■	
Task 2	■	■			■
Task 3		■			

2 Task 1: Source Retrieval for Scientific Web Claims

Motivation. Scientific web discourse, i.e., discourse about scientific claims or resources on the social web, has increased substantially throughout the past years and has been studied across a wide range of disciplines [5, 12]. However, the scientific web discourse is usually informal, including posts such as “*covid vaccines just don’t work on children*”. The scientific studies at the center of debates from which claims originate are usually not cited explicitly but only mentioned informally, such as “*Stanford study shows that vaccines don’t work*”. This poses challenges both from a computational perspective when mining social media or computing altmetrics, but also from a societal perspective, leading to poorly informed online debates [32]. For fact-checking, identifying the source of a claim is important, especially for false claims where counter-evidence may not exist, because human fact-checkers use the source to verify a claim by assessing the quality of the source publication and checking whether the claim accurately reflects its findings [13]. Based on this motivation, we propose the task of Source Retrieval for Scientific Web Claims:

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. (Languages: English, German, and French.)*

Data. Ground-truth datasets for training and evaluation will be provided in three languages: English, German, and French. Each dataset consists of pairs of social media posts from X and the scientific studies they implicitly reference. For English, we reuse the existing 15,699 data pairs from [17] for training, while new data will be annotated for the evaluation set. For German and French, new data will be annotated for both training and evaluation sets, with approximately 1,500 pairs per language. For each language, a query dataset will be provided, containing social media posts, along with a single collection set that includes information about the scientific studies, which will be used across all three languages.

Evaluation. For each language, the participating systems will be evaluated in terms of Mean Reciprocal Rank at 5 ($MRR@5$), which is an established measure for evaluating known-item search tasks. It focuses on ranking the correct study within the first five positions, as users are only rarely examine more results.

3 Task 2: Fact-Checking Numerical and Temporal Claims

Motivation. In recent years, tremendous progress has been made in automated fact-checking approaches [16, 37], along with the development of large-scale benchmarks [3, 36], for tackling misinformation at scale. Automating fact-checking is inherently knowledge-intensive, as it requires reasoning over heterogeneous evidence sources to arrive at a reliable verdict. This challenge is further amplified for claims involving numerical information [41], which require numerical contextualization and precise quantitative reasoning. Moreover, numerical claims are prone to the *Numeric-Truth Effect* [33], whereby the presence of numbers induces a false sense of credibility, thereby accelerating the spread of misinformation. Although Large Language Models (LLMs) have achieved remarkable advances on knowledge-intensive tasks, recent studies demonstrate that verifying numerical claims remains challenging for LLMs than verifying non-numerical ones [2, 41]. Motivated by recent advances in test-time reasoning [7], we propose a test-time scaling framework with the goal of training more effective verification models to enhance the performance of LLMs on fact-checking numerical claims.

Task Definition. *This task performs test-time scaling for fact-verification. Given a claim, relevant evidences and reasoning traces with corresponding verdicts, the task is to rank the reasoning traces based on their utility in leading to the correct verdict and also output the final verdict from the top-ranked reasoning traces. (Languages: English, Spanish and Arabic.)*

The task focuses on verifying naturally occurring claims that contain quantitative quantities and/or temporal expressions by improving the reasoning capabilities of LLMs through test-time scaling. Unlike previous editions, which primarily focused only on fact-checking accuracy, this year’s task explicitly incorporates rationale generation into the evaluation. This is achieved in the form of re-ranking reasoning traces from the LLM, assessing both the correctness of the predicted veracity and the quality of the underlying reasoning traces.

While the claims are drawn from the previous iteration of the task, the overall setup differs substantially. Specifically, we introduce a test-time scaling framework designed to enhance LLM reasoning for claim verification. Given a claim and its associated evidence, multiple reasoning traces are generated using an LLM with varying temperature values to induce diversity. Redundant traces are subsequently removed through a de-duplication step. Based on this data, participants are required to train a verifier model that ranks the reasoning traces for each test claim and derives a final verdict from the top-ranked traces. To avoid leakage and ensure rigorous evaluation, we will release new test sets.

Data. We collect the dataset from various fact-checking websites, completed with detailed metadata and an evidence corpus sourced from the web. We use the English dataset released in [41], as well as the 3,260 Arabic and 2,808 Spanish claims from CLEF 2025 CheckThat! [1] with corresponding evidence. Table 2 summarizes the statistics of the dataset.

Each claim in the training and the validation datasets comprises 20 reasoning traces generated by **gpt-4o-mini** that can be used by participants to train a verification model, for selecting the best reasoning traces that lead to the

Table 2. Dataset statistics for task 2.

Language	# of claims
English	8,000
Spanish	2,808
Arabic	3,260

correct verdict. Apart from top-10 evidences, the entire evidence corpus will also be provided to the participants so they can run their own retrieve and re-rank pipelines for providing additional context to the verification model. Each claim is annotated with relevant reasoning traces that lead to the correct verdict (ground truth). The traces re-ranked by the verifier can be evaluated based on their ability to identify such relevant traces.

Evaluation. We use Recall@ k and MRR@ k to measure the quality of the reasoning traces ranked by the verifier model, with $k = 5$. For claim verification, we use macro-averaged F1 and classwise F1 scores. Participants will be judged based on high performance on both macro F1 and Recall@ k . The final score is obtained by averaging ranking across these measures (macro F1 and Recall@ k , with $k = 5$).

4 Task 3: Generating Full-Fact-Checking Articles

Motivation. Previous work on automated fact-checking has focused on five main tasks: (i) check-worthy claim detection, (ii) previously verified claim retrieval, (iii) evidence retrieval, (iv) claim verification, and (v) justification generation [16, 26, 42]. However, professional fact-checkers also undertake the critical task of *publishing detailed fact-checking articles*, which guide readers toward a clear understanding of the claim by presenting factual arguments and explaining how they lead to the final verdict [15]. Writing fact-checking articles is time-consuming, as it involves synthesizing information from various sources, penning it down in an easy to understand draft, and refining it multiple times to ensure quality and correctness. Recently, we proposed the generation of fact-checking articles automatically using large language models [34]. While LLMs can generate fluent, long-form text [43–45], our evaluations show that LLM-generated fact-checking articles considerably lag behind expert-written articles, uncovering several issues, but especially their inability to construct coherent, evidence-based arguments that properly explain the claim’s veracity.

Task Definition. *Given a claim, its veracity, and a set of evidence documents consulted for fact-checking the claim, generate a full fact-checking article with inline citations. (Language: English.)*

Note that, we do not address automatic evidence retrieval or claim verification here: the claim, its veracity, and the evidence set to use are all explicitly provided.

Data. We reuse the WatClaimCheck dataset [21] for training and validation, and the union of the already publicly available ExClaim [46] dataset and currently private AmbiguousSnopes [34] dataset for testing. Overall, the training set comprises 26k data examples, and the validation set comprises 3.3k data examples. The test set contains 1.2k data examples in total, comprising 987 from ExClaim and 220 from AmbiguousSnopes. Each example corresponds to a claim and comprises the claim’s veracity label, a set of evidence documents as listed by the fact-checking expert on the source website, and the expert-written fact-checking article, which we treat as the ground truth.

Evaluation. As a primary evaluation measure, we use the mean of the following evaluation measures to assess the quality of the generated fact-checking article: (i) *entailment score*, which is a reference-based metric that measures if the generated text is entailed by the reference [34], (ii) *citation correctness*, which verifies if a text attributed to a citation can be entailed by the corresponding evidence, and (iii) *citation completeness*, which computes the proportion of input evidence that is correctly cited in the generated text. As a secondary measure, we also assess writing quality by computing Elo ratings for the participants at the end of the testing phase, through a setup similar to Chatbot Arena [6], but using LLM-as-a-judge [22, 47] in our case. We choose to use reference-free evaluation measures to assess the utility of the participating systems for professional fact-checkers, rather than how similar the generated articles are to the expert-written articles. This design choice also aims to prevent cheating on the publicly-available subset of our test dataset.

5 Related Work

The tasks offered in the different editions of the **CheckThat!** lab are closely related to several tasks at other evaluation initiatives such as SemEval, e.g. determining rumor veracity [9, 14], stance detection [25], fact-checking in community question-answering forums [24], propaganda detection in various contexts such as English news articles [8], English memes [10, 11], and Arabic news paragraphs, social media posts, and memes [19, 20], as well as multi- and cross-lingual fact-checked claim retrieval [28], emphasizing the focus on multi-linguality, also including low-resourced languages. The tasks offered in the **CheckThat!** lab are further related to the FEVER tasks [35, 40] on fact extraction and verification, the Fake News Challenge [18, 23, 31], and the detection of online hostile posts [27]. Furthermore, it aligns with the FakeNews task at MediaEval [29, 30]. Other recent shared tasks include claim detection and span identification in social media posts [38] as well as multi-modal fake news detection [39], further emphasizing the growing focus on analyzing and verifying information across diverse modalities and platforms.

6 Conclusion and Future Work

We presented the tasks of the 2026 edition of the **CheckThat!** lab, featuring complementary tasks to assist journalists and fact-checkers along the fact-checking

pipeline: from retrieving informally mentioned sources of scientific web claims to verifying numerical and temporal claims and ultimately generating full-fact-checking articles - a task that has been newly added to the **CheckThat!** verification pipeline. In line with the CLEF mission, two of the three tasks are multilingual by offering tasks in five languages, namely Arabic, English, German, French, and Spanish.

In future work, we plan to expand the multilingual coverage, particularly for low-resource languages. We also plan to explore stronger cross-document and numerical reasoning, improved handling of implicit evidence, and closer alignment of task design and evaluation with real-world fact-checking workflows to maximize practical impact.

Acknowledgments. The work of Stefan Dietze, Konstantin Todorov, Salim Hafid, and Sebastian Schellhammer is partially funded under the AI4Sci grant, co-funded by MESRI (France, grant UM-211745), BMBF (Germany, grant 01IS21086), and the French National Research Agency (ANR). The work of Dhruv Sahnan, Tanmoy Chakraborty, and Preslav Nakov is supported by the Multi-Institutional Faculty Interdisciplinary Research Project (MFIRP) between IIT Delhi and MBZUAI. The responsibility for the contents of this publication lies with the authors.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alam, F., et al.: Overview of the CLEF-2025 CheckThat! lab: subjectivity, fact-checking, claim normalization, and retrieval. In: Carrillo-de Albornoz, J., Gonzalo, J., Plaza, L., García Seco de Herrera, A., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)* (2025)
2. Aly, R., et al.: FEVEROUS: fact extraction and verification over unstructured and structured information. In: Vanschoren, J., Yeung, S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual* (2021)
3. Augenstein, I., et al.: MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4685–4697. Association for Computational Linguistics, Hong Kong, China (2019)
4. Barrón-Cedeño, A., et al.: Overview of the CLEF-2024 CheckThat! Lab: Checkworthiness, subjectivity, persuasion, roles, authorities and adversarial robustness. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Soulier, L., Di Nunzio, G.M., Galuščáková, P., García Seco de Herrera, A., Faggioli, G., Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)* (2024)

5. Brüggemann, M., Lörcher, I., Walter, S.: Post-normal science communication: exploring the blurring boundaries of science and journalism. *J. Sci. Commun.* **19**(3), A02 (2020)
6. Chiang, W.L., et al.: Chatbot arena: an open platform for evaluating LLMs by human preference. In: *Proceedings of the 41st International Conference on Machine Learning. ICML 2024, JMLR.org* (2024)
7. Chungkham, P., V, V., Setty, V., Anand, A.: Think right, not more: test-time scaling for numerical claim verification. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24345–24363. Association for Computational Linguistics, Suzhou, China (2025)
8. Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., Nakov, P.: SemEval-2020 task 11: detection of propaganda techniques in news articles. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1377–1414. International Committee for Computational Linguistics, Barcelona (online) (2020)
9. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., Zubiaga, A.: SemEval-2017 task 8: RumourEval: determining rumour veracity and support for rumours. In: *Proceedings of the 11th International Workshop on Semantic Evaluation*, pp. 69–76. SemEval 2017 (2017)
10. Dimitrov, D., et al.: SemEval-2024 task 4: multilingual detection of persuasion techniques in memes. In: Ojha, A.K., Doğruöz, A.S., Tayyar Madabushi, H., Da San Martino, G., Rosenthal, S., Rosá, A. (eds.) *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 2009–2026. Association for Computational Linguistics, Mexico City, Mexico (2024)
11. Dimitrov, D., et al.: SemEval-2021 task 6: detection of persuasion techniques in texts and images. In: *Proceedings of the International Workshop on Semantic Evaluation*, pp. 70–98. SemEval 2021 (2021)
12. Dunwoody, S.: Science journalism: prospects in the digital age. In: *Routledge Handbook of Public Communication of Science and Technology*, pp. 14–32. Routledge (2021)
13. Glockner, M., Hou, Y., Gurevych, I.: Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5916–5936 (2022)
14. Gorrell, G., et al.: SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 845–854. SemEval 2019 (2019)
15. Graves, L.: Anatomy of a fact check: objective practice and the contested epistemology of fact checking. *Commun. Cult. Critique* **10**(3), 518–537 (2017)
16. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. *Trans. Assoc. Comput. Linguist.* **10**, 178–206 (2022)
17. Hafid, S., et al.: Overview of the CLEF-2025 CheckThat! lab task 4 on scientific web discourse. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum. CLEF 2025*, Madrid, Spain (2025)
18. Hanselowski, A., et al.: A retrospective analysis of the fake news challenge stance-detection task. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1859–1874. COLING 2018 (2018)

19. Hasanain, M., et al.: ArAIEval shared task: persuasion techniques and disinformation detection in Arabic text. In: Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., Haddad, H., Zitouni, I., Mrini, K., Almatham, R. (eds.) *Proceedings of ArabicNLP 2023*, pp. 483–493. Association for Computational Linguistics, Singapore (Hybrid) (2023)
20. Hasanain, M., et al.: ArAIEval shared task: propagandistic techniques detection in unimodal and multimodal Arabic content. In: *Proceedings of The Second Arabic Natural Language Processing Conference*, pp. 456–466 (2024)
21. Khan, K., Wang, R., Poupart, P.: WatClaimCheck: a new dataset for claim entailment and inference. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1293–1304. Association for Computational Linguistics, Dublin, Ireland (2022)
22. Kim, S., et al.: Prometheus 2: an open source language model specialized in evaluating other language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353. Association for Computational Linguistics, Miami, Florida, USA (2024)
23. Malliga, S., et al.: Overview of the shared task on fake news detection from social media text. In: *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pp. 59–63 (2023)
24. Mihaylova, T., Karadzhov, G., Atanasova, P., Baly, R., Mohtarami, M., Nakov, P.: SemEval-2019 task 8: fact checking in community question answering forums. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 860–869. SemEval 2019 (2019)
25. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 task 6: Detecting stance in tweets. In: *Proceedings of the 10th International Workshop on Semantic Evaluation*, pp. 31–41. SemEval 2016 (2016)
26. Nakov, P., et al.: Automated fact-checking for assisting human fact-checkers. In: Zhou, Z.H. (ed.) *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4551–4558. International Joint Conferences on Artificial Intelligence Organization (2021), survey Track
27. Chakraborty, T., Shu, K., Bernard, H.R., Liu, H., Akhtar, M.S. (eds.): CON-*STRAINT* 2021. CCIS, vol. 1402. Springer, Cham (2021). <https://doi.org/10.1007/978-3-030-73696-5>
28. Peng, Q., et al.: SemEval-2025 task 7: multilingual and crosslingual fact-checked claim retrieval. In: Rosenthal, S., Rosá, A., Ghosh, D., Zampieri, M. (eds.) *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 2498–2511. Association for Computational Linguistics, Vienna, Austria (2025)
29. Pogorelov, K., Schroeder, D.T., Brenner, S., Maulana, A., Langguth, J.: Combining tweets and connections graph for FakeNews detection at MediaEval 2022. In: *Proceedings of the MediaEval 2022 Workshop*. MediaEval 2022 (2022)
30. Pogorelov, K., et al.: FakeNews: corona virus and 5G conspiracy task at MediaEval 2020. In: *Proceedings of the MediaEval 2020 Workshop*. MediaEval 2020 (2020)
31. Pomerleau, D., Rao, D.: The fake news challenge: exploring how artificial intelligence technologies could be leveraged to combat fake news (2017). <http://www.fakenewschallenge>

32. Rocha, Y.M., de Moura, G.A., Desidério, G.A., de Oliveira, C.H., Lourenço, F.D., de Figueiredo Nicolete, L.D.: The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *J. Public Health*, 1–10 (2021)
33. Sagara, N.: Consumer understanding and use of numeric information in product claims. University of Oregon (2009)
34. Sahnan, D., et al.: Can LLMs automate fact-checking article writing? *Transactions of the Association for Computational Linguistics* (2026)
35. Schlichtkrull, M., et al.: The automated verification of textual claims (AVeriTeC) shared task. In: Schlichtkrull, M., Chen, Y., Whitehouse, C., Deng, Z., Akhtar, M., Aly, R., Guo, Z., Christodoulopoulos, C., Cocarascu, O., Mittal, A., Thorne, J., Vlachos, A. (eds.) *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 1–26. Association for Computational Linguistics, Miami, Florida, USA (2024)
36. Schlichtkrull, M., Guo, Z., Vlachos, A.: AVeriTeC: a dataset for real-world claim verification with evidence from the web. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS 2023*, Curran Associates Inc., Red Hook, NY, USA (2023)
37. Setty, V.: FactCheck editor: multilingual text editor with end-to-end fact-checking. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2744–2748 (2024)
38. Sundriyal, M., Akhtar, M.S., Chakraborty, T.: Overview of the CLAIMSCAN-2023: uncovering truth in social media through claim detection and identification of claim spans. In: *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pp. 7–9 (2023)
39. Suryavardan, S., et al.: Findings of factify 2: multimodal fake news detection. In: *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection*, co-located with AAAI 2023. Washington, DC, USA (2023)
40. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819 (2018)
41. Venkatesh, V., Anand, A., Anand, A., Setty, V.: QuanTemp: a real-world open-domain benchmark for fact-checking numerical claims. In: *47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pp. 650–660. Association for Computing Machinery (ACM) (2024)
42. Vlachos, A., Riedel, S.: Fact Checking: task definition and dataset construction. In: Danescu-Niculescu-Mizil, C., Eisenstein, J., McKeown, K., Smith, N.A. (eds.) *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 18–22. Association for Computational Linguistics, Baltimore, MD, USA (2014)
43. Wang, Q., et al.: Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1352–1391. Association for Computational Linguistics, Albuquerque, New Mexico (2025)
44. Xie, Z., Cohn, T., Lau, J.H.: The next chapter: a study of large language models in storytelling. In: Keet, C.M., Lee, H.Y., Zarrieß, S. (eds.) *Proceedings of the 16th International Natural Language Generation Conference*, pp. 323–351. Association for Computational Linguistics, Prague, Czechia (2023)

45. Yang, K., Klein, D., Peng, N., Tian, Y.: DOC: improving long story coherence with detailed outline control. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3378–3465. Association for Computational Linguistics, Toronto, Canada (2023)
46. Zeng, F., Gao, W.: JustiLM: few-shot justification generation for explainable fact-checking of real-world claims. *Trans. Assoc. Comput. Linguist.* **12**, 334–354 (2024)
47. Zheng, L., et al.: Judging LLM-as-a-judge with MT-Bench and Chatbot arena. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10-16, 2023* (2023)