Preface

Eva Sánchez Salido[1,*], Alberto Barrón-Cedeño[2], Alba García Seco de Herrera[1,*], Sean MacAvaney[3] and Julia Maria Struß[4]

[1]National University of Distance Education - UNED, Calle Juan del Rosal 16, 28040, Madrid, Spain
[2]DIT, Università di Bologna, Viale Filippo Corridoni 20, Forlì, Italy
[3]University of Glasgow, Lilybank Gardens, G12 8QQ Glasgow, United Kingdom
[4]University of Applied Sciences Potsdam, Kiepenheuerallee 5, 14469 Potsdam, Germany

CLEF 2026[1] marks the seventeenth edition of the CLEF conference series, and the twenty-seventh year of the CLEF initiative[2]. Since its launch in 2000, CLEF has provided a leading forum for the systematic evaluation of multilingual and multimodal information access systems, fostering scientific progress through shared tasks, benchmark datasets, and collaborative experimentation. Since its transition to a conference format in 2010, CLEF has combined a scientific conference programme with evaluation-driven research activities. The event brings together keynote lectures, contributed papers, poster sessions and a broad range of laboratory-style evaluation campaigns. These labs are proposed and organised by teams of researchers who voluntarily contribute their expertise to design, coordinate and promote evaluation activities addressing current and emerging research challenges.

CLEF 2026 was hosted by Friedrich-Schiller-Universität in Jena, Germany, from 21 to 24 September 2026. The programme followed the established CLEF format and took place primarily in person while also providing opportunities for remote participation.

The call for lab proposals attracted 18 submissions, which were assessed through a peer review process considering their innovation potential and the quality of the resources created. Following this evaluation, 16 labs were selected. Together they address a diverse range of challenges in multimodal and multilingual information access, spanning both established research areas and emerging topics of growing interest to the community. The selected labs provide collections, benchmarks, and evaluation frameworks that enable researchers to compare approaches under common conditions and advance the state of the art. These resources offer opportunities to develop and evaluate novel solutions, obtain comparative performance results, and engage with fellow researchers in discussions of shared challenges and future directions.

The labs participating in CLEF 2026 further demonstrate the vitality and maturity of the CLEF evaluation ecosystem. They introduce new tasks, new datasets, and new evaluation methodologies, consolidating successful activities that have become well-established reference points within their respective research communities.

More broadly, the continued success of CLEF relies on the combined efforts of researchers, organisers, participants, and supporting institutions. Throughout its history, CLEF has benefited from the support of European-funded projects, which provided the framework and resources necessary to complement the extensive voluntary effort contributed by lab organisers and the wider CLEF community. This support played a key role in enabling the coordination and long-term sustainability of CLEF as a major international evaluation initiative, alongside other well-established benchmarking forums such as TREC, NTCIR, FIRE, and MediaEval. Since 2014, CLEF has operated independently of direct European project funding. Its long-term sustainability has been ensured through the CLEF Association, an independent

The 17th Conference and Labs of the Evaluation Forum (CLEF) 2026. September 21 – 24, Jena, Germany.
*Corresponding author.
evasan@lsi.uned.es (E. Sánchez Salido); a.barron@unibo.it (A. Barrón-Cedeño); alba.garcia@lsi.uned.es (A. García Seco de Herrera); Sean.MacAvaney@glasgow.ac.uk (S. MacAvaney); julia.struss@fh-potsdam.de (J. M. Struß)
0000-0001-8665-3018 (E. Sánchez Salido); 0000-0003-4719-3420 (A. Barrón-Cedeño); 0000-0002-6509-5325 (A. García Seco de Herrera); 0000-0002-8914-2659 (S. MacAvaney); 0000-0001-9133-4978 (J. M. Struß)
[1]https://clef2026.clef-initiative.eu/
[2]https://www.clef-initiative.eu/


© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

non-for-profit legal entity established in October 2013 as a result of the activities of the PROMISE Network of Excellence, which supported CLEF between 2010 and 2013. Through the commitment of its members, the Association provides the organisational and financial foundations required to coordinate and maintain CLEF's activities. In 2026, the CLEF Association entered a new phase of its development through a structural reorganisation and the adoption of an updated legal framework, strengthening its capacity to support the continued growth and long-term evolution of the CLEF initiative.

The labs running as part of CLEF 2026 comprise a mixture of established labs, namely BioASQ, CheckThat!, ELOQUENT, eRisk, EXIST, ImageCLEF, JOKER, LifeCLEF, LongEval, PAN, QuantumCLEF, SimpleText, TalentCLEF, and Touché. CLEF also marks the return of the HIPE lab after several years, while welcoming FinMMEval as a new lab.

Details of the individual labs are provided in these proceedings by the lab organisers. A brief description of each lab is provided below, in alphabetical order:

BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering[1]

Since 2012, BioASQ has pushed the research frontier towards systems that exploit the diverse and voluminous information available online to respond directly to the information needs of biomedical scientists. In its fourteenth edition, the challenge comprises *Task 14b: Biomedical Semantic QA*, which asks participants to return relevant articles, snippets, exact answers, and ideal answers for English biomedical questions; *Task Synergy 14: Biomedical Semantic QA for developing topics*, which follows an iterative setting for evolving questions on emerging biomedical issues; *Task MultiClinSum-2: Multilingual Clinical Summarisation*, which evaluates the automatic summarisation of clinical case reports across fifteen languages; *Task BioNNE-R: Nested Relation Extraction in Russian and English*, focused on extracting relations between nested named entities in biomedical texts; *Task ELCardioCC: Clinical Coding in Cardiology*, which addresses ICD-10 coding of Greek cardiology discharge letters; and *Task GutBrainIE: Gut-Brain Interplay Information Extraction*, which targets entity mention classification and mention-level relation extraction from publications on the gut–brain axis. The resulting shared tasks continue to ground progress in expert-created gold-standard annotations and carefully designed evaluation protocols.

CheckThat!: Advancing Multilingual Fact-Checking[2]

Misinformation travels fast across languages and platforms, and CheckThat! addresses this challenge by fostering technology and resources along the full fact-checking verification pipeline. The ninth edition of the lab offers three tasks in mono-, multi-, and cross-lingual settings covering five languages: *Task 1: Source Retrieval for Scientific Web Claims* asks participants to identify the source publication of informal scientific references appearing in social media posts; *Task 2: Fact-Checking Numerical and Temporal Claims* extends claim verification with a reasoning component focused on numerical and temporal statements; and *Task 3: Generating Full Fact-Checking Articles*, which requires systems to produce complete fact-checking articles with inline citations given a claim, its verdict, and supporting evidence. Classification, retrieval, and generation are thus evaluated jointly as complementary steps towards more reliable automated and assisted fact-checking.

ELOQUENT: Evaluating Generative Language Model Quality[3]

Standard test suites often miss the quality criteria that matter once generative language models are embedded in real applications spanning languages, cultures, and sectors of society. ELOQUENT's third year continues to explore alternative evaluation regimes through Task *Voight-Kampff*, which challenges

[1] https://www.bioasq.org/workshop2026
[2] https://checkthat.gitlab.io/clef2026/
[3] https://eloquent-lab.github.io

classifiers to distinguish machine-generated from human-authored text in a builder–breaker setting with the PAN lab; Task *Cultural Robustness and Diversity*, which studies how models' responses to value-oriented questions vary across languages, cultural backgrounds, and system settings; Task *PISA Quiz Item Generation*, organised in collaboration with the OECD, which investigates how well models can formulate useful topical quizzes fitted to given target texts; and Task *PISA Quiz Item Scoring* (also known as Sensemaking), which evaluates how reliably models can score topical quiz responses. The lab thus promotes multilingual, culturally aware assessment of generative AI beyond conventional accuracy-oriented benchmarks.

eRisk: Early Risk Prediction on the Internet[4]

When personal risk situations leave traces online, even small gains in anticipation can support earlier interventions —especially in mental health and safety-related domains. Marking its tenth edition, eRisk consolidates a shift towards contextual, conversational, and symptom-aware evaluation with *Task 1: Conversational Depression Detection via LLMs*, where participants interact with fine-tuned LLM personas to infer depression severity and active depressive symptoms; *Task 2: Contextualised Early Detection of Depression*, where systems process full conversational threads rather than isolated user posts; and *Task 3: ADHD Symptom Sentence Ranking*, a new symptom-oriented retrieval task covering the eighteen symptoms of the Adult ADHD Self-Report Scale. Methods are encouraged not only to predict risk early, but also to surface clinically meaningful and interpretable evidence.

EXIST: sEXism Identification in Social neTworks[5]

Online sexism is rarely purely textual: in memes and short videos, meaning often emerges from the interplay of visual, linguistic, temporal, and contextual cues, and human judgements themselves remain subjective. EXIST 2026, the sixth edition of the lab, focuses exclusively on multimedia content and adopts a Human-Centred AI perspective by enriching the data with eye-tracking, heart-rate, and EEG signals collected from subjects exposed to potentially sexist material. Six subtasks in English and Spanish organise three core objectives over memes and TikToks: *Sexism Identification*, deciding whether content is sexist; *Source Intention Detection*, distinguishing direct sexism from judgemental messaging; and *Sexism Categorisation*, multi-label classification into types such as ideological inequality, stereotyping and dominance, sexual violence, and misogyny. Following the Learning With Disagreement paradigm, conscious labels and unconscious physiological responses are combined to better capture diverse perceptions of sexism.

FinMMEval: Multilingual and Multimodal Evaluation of Financial AI Systems[6]

Financial applications demand more than fluent generation: systems must reason over numbers, regulations, filings, and multilingual reports, and increasingly act as agents over live or semi-live market data. Debuting at CLEF 2026, FinMMEval studies these requirements through *Task 1: Financial Exam Question Answering*, a multilingual multiple-choice setting covering English, Chinese, Arabic, and Hindi; *Task 2: Multilingual Financial Question Answering*, which requires short-answer generation over heterogeneous financial evidence such as narrative reports, statements, and filing-derived fields; and *Task 3: Financial Decision Making*, which evaluates live trading-agent endpoints that issue daily BUY, HOLD, or SELL decisions for selected assets. Hidden-test protocols for the QA tasks and endpoint-based live evaluation for trading jointly expose both model quality and operational constraints in finance.

[4] https://erisk.irlab.org
[5] https://nlp.uned.es/exist2026/
[6] https://mbzuai-nlp.github.io/CLEF-2026-FinMMEval-Lab/

HIPE: Person–Place Relation Extraction from Multilingual Historical Texts[7]

Was this person ever at that place and, if so, when? Answering such questions from noisy, multilingual historical documents is the central challenge of HIPE-2026. the third edition of the HIPE series, which moves from named entity recognition and linking towards temporally grounded relation extraction. The lab targets two relation types —*at*, indicating that a person was present at a location at some point prior to a document's publication date, and *isAt*, indicating presence contemporaneous with that date—over nineteenth- and twentieth-century newspapers in French, German, and English, with a surprise-domain generalisation set drawn from early modern French literary texts. Participants must cope with OCR noise, historical language variation, and indirect contextual cues, while a three-fold evaluation framework jointly assesses predictive accuracy, computational efficiency, and cross-domain generalisation.

ImageCLEF: Multimodal Challenges in Medicine, Science, Agritech, and Security[8]

Now in its twenty-fourth edition, ImageCLEF remains one of CLEF's broadest multimodal laboratories, spanning medicine, assistive communication, scientific reasoning, agriculture, and media authenticity. This year features *ImageCLEFmedical*, with subtasks on concept detection and caption prediction, veracity and privacy of GAN-generated CT scans, visual question answering and explainability in gastrointestinal endoscopy, and MEDIQA-CORE challenges on brain tumour subtype classification and radiology report discrepancy summarisation; *ToPicto*, focused on text-to-pictogram translation and pictogram sequence prediction; *Multimodal Reasoning*, multilingual multi-discipline question answering over multiple modalities; *Deepfake Detection and Generation*, covering images and audio in a cyclical generation–detection setup; and *AI4Agriculture*, focused on viticulture potential prediction and crop identification. The diversity of tasks mirrors the expanding role of multimodal AI across scientific and societal applications.

JOKER: Humour Detection, Search, and Translation[9]

Humour and other non-literal uses of language remain difficult for AI systems because they often depend on implicit cultural references and on unorthodox interactions between form and meaning. In its fifth year, the JOKER lab brings together researchers in NLP, IR, translation studies, and the humanities to build reusable test collections and to advance the automatic analysis, retrieval, translation, and generation of humour and wordplay. The 2026 edition comprises *Task 1: Humour-Aware Information Retrieval*, which asks systems to retrieve short humorous texts that are both topically relevant and instances of wordplay, humour, or punning, in English and Hinglish; *Task 2: Pun Translation*, which focuses on translating English puns into French while preserving both meaning and wordplay; *Task 3: Onomastic Wordplay Translation*, which targets the translation of name-based wordplay from English into French across literature, games, comics, and related sources; and *Task 4: Humour Generation*, a new constrained generation task in which systems produce short humorous texts in English, French, and Spanish from structured briefs specifying a target pun word and the two senses it must activate. Creative language understanding and generation thus remain the lab's unifying evaluation agenda.

LifeCLEF: AI Challenges for Biodiversity Understanding and Ecosystem Management[10]

Biodiversity monitoring increasingly relies on AI to turn large volumes of citizen-science and field observations into actionable indicators of species occurrence, abundance, and range change. LifeCLEF has supported this agenda since 2011. The 2026 edition broadens its scope with five complementary

[7] https://hipe-eval.github.io/HIPE-2026/
[8] https://www.imageclef.org/2026
[9] https://www.joker-project.com/2026/
[10] https://www.imageclef.org/LifeCLEF2026

challenges spanning visual, acoustic, and textual data: *AnimalCLEF*, focused on discovery and re-identification of individual animals; *BirdCLEF+*, focused on multi-taxonomic species recognition in complex soundscapes; *FathomNetCLEF*, focused on detecting marine species in underwater imagery under positive-unlabelled constraints; *PestCLEF*, focused on extracting information on plant pests from heterogeneous textual sources; and *PlantCLEF*, focused on multi-species plant identification in quadrat images. The lab connects AI researchers with ecologists and practitioners working under realistic ecological and data conditions.

LongEval: Longitudinal Evaluation of IR Model Performance[11]

Retrieval effectiveness is rarely static: as document collections grow and user behaviour shifts, systems that look strong on a fixed snapshot may degrade over time. LongEval investigates this temporal dimension in cooperation with the CORE academic search platform, providing participants with successive SciRetrieval snapshots derived from real queries, documents, and interactions. Four tasks structure the 2026 edition: *LongEval-Sci: Ad-Hoc Scientific Retrieval*, measuring ranking effectiveness across corpus updates; *LongEval-TopEx: Topic Extraction from Query Logs*, formalising information needs from production logs to enable Cranfield-style evaluation; *LongEval-USim: User Simulations*, simulating user interactions via query (re-)formulations; and *LongEval-RAG: Retrieval-Augmented Generation*, assessing RAG over an evolving scientific collection. Beyond standard metrics, robustness measures quantify how performance changes as the corpus expands.

PAN: Digital Text Forensics and Stylometry[12]

For almost two decades, PAN has advanced text forensics and stylometry through objective evaluation on curated benchmarks covering authorship, originality, and more recently generative AI. The 2026 edition organises five shared tasks: *Voight-Kampff Generative AI Detection*, distinguishing human- from machine-written texts with particular attention to robustness under LLM style obfuscation, in collaboration with ELOQUENT; *Text Watermarking*, evaluating the inconspicuousness and robustness of post-hoc watermarking under unknown obfuscations; *Multi-Author Writing Style Analysis*, identifying positions in a document where the author changes; *Generative Plagiarism Detection*, focused this year on generated-text source retrieval and paraphrastic reuse; and *Reasoning Trajectory Detection*, addressing source identification and safety assessment of generative reasoning chains. Authorship analysis, provenance, and the integrity of AI-generated text remain at the core of the lab's evaluation agenda.

QuantumCLEF: Quantum Computing, Information Retrieval and Recommender Systems[13]

Quantum Computing promises new ways to tackle combinatorial optimisation problems that arise naturally in Information Retrieval and Recommender Systems, yet access to real quantum hardware remains limited. QuantumCLEF's third edition focuses on Quantum Annealing and gives participants hands-on experience with cutting-edge annealers through three tasks: *Task 1: Feature Selection for IR and RS*, selecting informative feature subsets for ranking and recommendation; *Task 2: Instance Selection for IR*, targeting noise removal and instance selection in retrieval collections; and *Task 3: Clustering for IR*, grouping semantically related documents. Tutorials and supporting materials help teams compare quantum annealing with simulated annealing and classical baselines, fostering both algorithmic innovation and practical literacy in this emerging paradigm.

[11] https://clef-longeval.github.io
[12] https://pan.webis.de
[13] https://qclef.dei.unipd.it

SimpleText: Simplify Scientific Text (and Nothing More)[14]

Primary scientific sources are among the most trustworthy references available to the public, yet their specialised language often places them out of reach for non-experts. SimpleText tackles this accessibility gap —especially in biomedicine, where aligned Cochrane abstracts and plain-language summaries provide a concrete use case— while also confronting the hallucinations and information distortions that generative models may introduce. The 2026 edition comprises *Task 1: Text Simplification*, making complex scientific texts easier to understand at sentence and document level; *Task 2: Controlled Creativity*, detecting, classifying, and avoiding over-generation and other distortions in generated content; and *Task 3: Research Area Classification*, a pilot on classifying scientific articles by research field and discipline. The shared goal is clearer public access to scientific knowledge without sacrificing factual fidelity.

TalentCLEF: Skill and Job Title Intelligence for Human Capital Management[15]

As labour markets transform, organisations increasingly need NLP systems that can connect jobs, people, and skills across languages and textual genres —from vacancies and CVs to skill taxonomies. TalentCLEF addresses the fragmentation of prior Human Capital Management research by offering shared multilingual benchmarks; in its second edition, the lab featured *Task A: Contextualised Job–Person Matching*, identifying and ranking the most suitable candidates (represented by their CVs) for a given job vacancy in English and Spanish and *Task B: Job–Skill Matching with Skill Type Classification*, retrieving the most relevant skills for a given job title in English while distinguishing core from contextual skills. Common datasets and evaluation protocols enable comparable progress on recruitment, career guidance, skill gap analysis, and related workforce applications.

Touché: Argumentation Systems[16]

Decision-making and opinion-forming are everyday tasks, yet today's search engines and language models rarely provide transparent quality assurance of the arguments they surface —and may inject commercial bias. Motivated by these threats to the information ecosystem, the seventh edition of Touché investigates supportive technologies for argumentation through four shared tasks, three of them new: *Fallacy Detection*, detecting fallacious arguments, classifying fallacy types, and identifying argument schemes in short texts; *Causality Extraction*, detecting causal information, marking related entities, and classifying relations as procausal, concausal, or uncausal; *Generalisability of Argument Identification in Context*, classifying whether a sentence constitutes an argument given provenance and context; and *Advertisement in Retrieval-Augmented Generation*, detecting, locating, and removing ads from search responses without harming the underlying answer. Computational argumentation and the evaluation of mediated argumentative content remain the lab's unifying focus.

Acknowledgments

We would like to thank the members of the CLEF-LOC (the CLEF Lab Organisation Committee) for their thoughtful and elaborate contributions to assessing the proposals during the selection process: (...).

We thank sponsors, institutions... for their support, and all the enthusiastic and creative proposal authors, the organisers of the selected labs and workshops, the colleagues and friends involved in running them, and the participants who contribute their time to making the labs and workshops a success. Without their effort, the CLEF labs would not be possible.

[14] https://simpletext-project.com
[15] https://talentclef.github.io/talentclef/
[16] https://touche.webis.de

July, 2026

Eva Sánchez Salido
Alberto Barrón-Cedeño
Alba García Seco de Herrera
Sean MacAvaney
Julia Maria Struß

Organisation

General Chairs

Matthias Hagen, Friedrich-Schiller-Universität Jena, Germany

Martin Potthast, University of Kassel, Germany

Benno Stein, Bauhaus-Universität Weimar, Germany

Programme Chairs

Philipp Schaer, Technische Hochschule Köln, Germany

Eva Zangerle, Universität Innsbruck, Austria

Lab Chairs

Sean MacAvaney, University of Glasgow, United Kingdom

Julia Maria Struß, Fachhochschule Potsdam University of Applied Sciences, Germany

Proceedings Chairs

Alberto Barrón-Cedeño, Università di Bologna, Italy

Alba García Seco de Herrera, National University of Distance Education - UNED, Spain

Eva Sánchez Salido, National University of Distance Education - UNED, Spain

Website Chairs

Lukas Gienapp, University of Kassel, Germany

Community Chairs

Jan Heinrich Merker, Friedrich-Schiller-Universität Jena, Germany

Local Organisation Committee/Support

Janek Bevendorff, Bauhaus-Universität Weimar, Germany

Niklas Deckers, University of Kassel, Germany

Theresa Elstner, University of Kassel, Germany

Ben Gerhards, University of Kassel, Germany

Maik Fröbe, Bauhaus-Universität Weimar, Germany

Marcel Gohsen, Bauhaus-Universität Weimar, Germany

Tim Gollub, Bauhaus-Universität Weimar, Germany

Klara Gutekunst, University of Kassel, Germany

Tim Hagen, University of Kassel, Germany

Sebastian Heineking, University of Kassel, Germany

Midhun Kanadan, Bauhaus-Universität Weimar, Germany

Jan Heinrich Merker, Friedrich-Schiller-Universität Jena, Germany

Wilhelm Pertsch, Friedrich-Schiller-Universität Jena, Germany

Simon Ruth, University of Kassel, Germany
Ines Zelch, Friedrich-Schiller-Universität Jena, Germany

Steering Committee

Steering Committee Chairs

Alba García Seco de Herrera, National University of Distance Education - UNED, Spain

Alberto Barrón-Cedeño, Università di Bologna, Italy

Deputy Steering Committee Chair for the Conference

Josiane Mothe, IRIT, Université de Toulouse, France

Deputy Steering Committee Chair for the Evaluation Lab

Martin Braschler, Zurich University of Applied Sciences (ZHAW), Switzerland

Members

Avi Arampatzis, Democritus University of Thrace, Greece

Khalid Choukri, Evaluations and Language resources Distribution Agency (ELDA), France

Fabio Crestani, Università della Svizzera italiana, Switzerland

Carsten Eickhoff, University of Tübingen, Germany

Nicola Ferro, University of Padua, Italy
(Steering Committee Chair 2010–2025)

Petra Galuščáková, University of Stavanger, Norway

Anastasia Giachanou, Utrecht University, The Netherlands

Lorraine Goeuriot, Université Grenoble Alpes, France

Julio Gonzalo, National Distance Education University (UNED), Spain

Donna Harman, National Institute for Standards and Technology (NIST), USA

Bogdan Ionescu, University "Politehnica" of Bucharest, Romania

Evangelos Kanoulas, University of Amsterdam, The Netherlands

Birger Larsen, University of Aalborg, Denmark

Maria Maistro, University of Copenhagen, Denmark

Henning Müller, University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Jian-Yun Nie, Université de Montréal, Canada

Gabriella Pasi, University of Milano-Bicocca, Italy

Florina Piroi, TU Wien, Austria

Paolo Rosso, Universitat Politècnica de València, Spain

Eric SanJuan, University of Avignon, France

Damiano Spina, RMIT Melbourne, Australia

Theodora Tsikrika, Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Greece

Past Members

Paul Clough, University of Sheffield, United Kingdom

Norbert Fuhr, University of Duisburg-Essen, Germany

Djoerd Hiemstra, Radboud University, The Netherlands

Jaana Kekäläinen, University of Tampere, Finland

Séamus Lawless, Trinity College Dublin, Ireland

David E. Losada, Universidade de Santiago de Compostela, Spain

Mihai Lupu, Field consultant, Vienna, Austria

Carol Peters, ISTI, National Council of Research (CNR), Italy
(Steering Committee Chair 2000–2009)

Emanuele Pianta, Centre for the Evaluation of Language and Communication Technologies (CELCT), Italy

Maarten de Rijke, University of Amsterdam UvA, The Netherlands

Giuseppe Santucci, Sapienza University of Rome, Italy

Jacques Savoy, University of Neuchêtel, Switzerland

Alan Smeaton, Dublin City University, Ireland

Laure Soulier, Sorbonne Université, France

Christa Womser-Hacker, University of Hildesheim, Germany