Report on the 16th Conference and Labs of the Evaluation Forum (CLEF 2025): Experimental IR Meets Multilinguality, Multimodality, and Interaction

Jorge Carrillo-de-Albornoz
UNED
Spain
jcalbornoz@lsi.uned.es

Guglielmo Faggioli
University of Padua
Italy
faggioli@dei.unipd.it

Nicola Ferro
University of Padua
Italy
nicola.ferro@unipd.it

Alba García Seco de Herrera
UNED
Spain
alba.garcia@lsi.uned.es

Julio Gonzalo
UNED
Spain
julio@lsi.uned.es

Laura Plaza, Josiane Mothe,Florina Piroi, Paolo Rosso, Damiano Spina[†]

Abstract

This report presents the sixteenth edition of the Conference and Labs of the Evaluation Forum (CLEF 2025), held from September 9 to 12, 2025, in Madrid, Spain, and hosted by the National Distance Education University (UNED). CLEF 2025 continued its hybrid format, combining a scientific conference with an evaluation forum. The conference programme featured keynote talks by Sameer Antani, José Hernández-Orallo, and Joanna Bryson, alongside peer-reviewed research papers, poster sessions, and lab presentations. The evaluation forum comprised 14 labs: BioASQ, CheckThat!, ELOQUENT, eRisk, EXIST, ImageCLEF, JOKER, LifeCLEF, LongEval, PAN, QuantumCLEF, SimpleText, TalentCLEF, and Touché. Each lab addresses diverse challenges in multilingual, multimodal, and interactive information access across a variety of domains and media.

Date: 9–12 September, 2025.

Website: https://clef2025.clef-initiative.eu/.

[†]Affiliation not shown for all authors due to space limitations (see Appendix A for details).

1 Introduction

The 2025 edition of the *Conference and Labs of the Evaluation Forum* (CLEF) was hosted by the National Distance Education University of Spain (UNED), and took place in Madrid from 9 to 12 September 2025. This marked the sixteenth year of the CLEF Conference and the twenty-sixth year of the CLEF initiative as a leading forum for evaluation in multilingual and multimodal information access.

CLEF 2025 retained the well-established format of previous editions, comprising keynote talks, contributed papers, lab sessions, and poster presentations. All sessions were conducted in hybrid mode, enabling both in-person and remote participation. However, CLEF 2025 was primarily designed as an in-person event, with online registration permitted only under exceptional circumstances.

CLEF was established in 2000 as a spin-off of the TREC Cross-Language Track, with a focus on stimulating research and innovation in multimodal and multilingual information access and retrieval [Ferro, 2019; Ferro and Peters, 2019]. Over the years, CLEF has fostered the creation of language resources in many European and non-European languages, promoted the growth of a vibrant and multidisciplinary research community, provided sizable improvements in the performance of monolingual, bilingual, and multilingual information access systems [Ferro and Silvello, 2017], and achieved a substantial scholarly impact [Larsen, 2019; Tsikrika et al., 2011, 2013].

During its first decade, CLEF hosted a series of experimental labs whose results were presented at annual workshops co-located with the European Conference on Digital Libraries (ECDL, now TPDL). In 2010, CLEF evolved into a mature evaluation forum by introducing a peer-reviewed conference, broadening its scope to include evaluation methodologies and systems across various data types, formats, and languages. The evaluation labs also expanded to address multimodality, understood as the capacity to process information conveyed through diverse media and modalities, such as the Web, social media, news streams, and domain-specific sources.

Since 2010, CLEF has maintained a consistent structure, integrating keynotes, peer-reviewed papers, lab sessions, and poster presentations, including contributions from other benchmarking initiatives worldwide. In 2013, CLEF established a non-profit association to ensure sustainable coordination and long-term continuity, supported by the financial contributions of its community [Ferro, 2019, 2024].

CLEF 2025 continued the collaborative initiative launched in 2019 with the European Conference on Information Retrieval (ECIR). ECIR 2025 featured a dedicated session for CLEF Labs, where lab organisers presented key outcomes and future plans, followed by a poster session to encourage discussion. These activities were reflected in the ECIR 2025 proceedings [Hauff et al., 2025], helping to disseminate CLEF results and foster engagement with the broader IR community.

The CLEF 2025 edition attracted a record-breaking total of 254 participants, including 214 in-person attendees and 40 remote participants. This turnout marks the highest attendance in CLEF's history, underscoring the continued relevance and vitality of the CLEF community. Participants represented a diverse mix of academic institutions and industry organisations from around the world.

CLEF 2025 welcomed a highly diverse group of participants from all continents. Europe stood out as the most represented continent, accounting for 71% of attendees, reflecting CLEF's

historical roots and continued relevance within the European research community. Asia followed with 14%, and North America with 9%. South America, Oceania, and Africa were represented by four, three, and one participant(s), respectively. This international composition not only enriches the CLEF community but also reinforces its mission to foster collaboration and innovation across borders and disciplines.

2 The CLEF Conference

CLEF 2025 has continued to centre its conference programme around experimental information retrieval (IR), as practised in evaluation forums such as CLEF Labs, TREC, NTCIR, FIRE, and MediaEval. This year's contributions placed particular emphasis on the challenges of multimodality, multilinguality, and interactive search, with a strong presence of work involving large language models (LLMs) and their application to real-world tasks.

The accepted papers reflected a broad spectrum of research directions, including humour classification, sexism detection in social media, and robustness of misinformation classifiers. Collectively, these works offered novel insights into IR evaluation, proposed methodological innovations, and pushed the boundaries of the Cranfield/TREC/CLEF paradigm through both theoretical and applied lenses [Carrillo-de Albornoz et al., 2025; Faggioli et al., 2025].

Keynotes

The CLEF 2025 conference featured keynote addresses by three distinguished scholars, each offering a unique perspective on the future of artificial intelligence, evaluation, and its societal implications:

Sameer Antani (National Institutes of Health, USA)

Title: *Cross-Modal Data Synthesis for AI-driven Biomedical Applications*
Abstract: Biomedical data are inherently multimodal, encompassing structured and unstructured text, images, videos, and other signals. These diverse data types offer complementary insights into a patient's condition and support clinical decision-making. In this talk, Dr Antani presented novel approaches in cross-modal data synthesis, a transformative method that integrates heterogeneous biomedical data to support generative modelling and robust predictive systems. He showcased research on generating synthetic imaging data to enrich training datasets, particularly for rare conditions, and on producing textual descriptions of images to support foundation models. The talk highlighted how integrated multimodal frameworks can make biomedical discoveries more accessible, interpretable, and actionable.

José Hernández-Orallo (Universitat Politècnica de València, Spain & University of Cambridge, UK)

Title: *AI Evaluation Should Make AI Predictable*
Abstract: Evaluation in AI goes far beyond benchmarks and metrics. Dr Hernández-Orallo explored the fragmented landscape of AI evaluation, identifying six distinct paradigms: TEVV (test-

ing, evaluation, verification, validation), benchmark-driven, 'evals', construct-oriented, real-world impact, and exploratory approaches. He argued that the field lacks a coherent understanding of capability versus performance, and that predictability should be the central goal of evaluation—whether assessing general safety or specific operational reliability. By reframing evaluation as a pursuit of explanatory and predictive power, the talk illuminated new research challenges and opportunities for building trustworthy AI systems.

Joanna Bryson (Hertie School, Germany)

Title: *Do We Co-evolve with What We Design? DevOps, AGI, and Human Frailties*
Abstract: Prof Bryson examined the complex relationship between humanity and artificial intelligence, asking whether we are co-evolving with the technologies we create. She challenged the notion of AI systems as collaborators, warning against anthropomorphising designed systems. The talk addressed the biological and cultural evolution of humans, contrasted with the engineered nature of digital systems, and emphasised the political dimensions of accountability—whether AI itself or the corporations behind it should be held responsible. Prof Bryson also discussed four possible futures for AI and humanity, and provided insights into EU and global regulatory efforts aimed at governing essential digital infrastructure.

Technical Programme

CLEF 2025 received a total of 14 scientific submissions, of which 6 papers were accepted for presentation (5 talks and 1 poster). Each submission underwent peer review by at least two members of the programme committee, with oversight and coordination provided by the programme chairs. Several of the accepted papers were the result of international collaborations, reflecting the global reach and interdisciplinary nature of the CLEF community.

The selected contributions addressed a diverse set of challenges at the forefront of information retrieval and evaluation. Topics included selective search as a first-stage retrieval strategy, detection of AI-generated biomedical texts, and the creation of multilingual datasets for authorship attribution, with a particular focus on medical disinformation detection. Other papers explored authorship analysis through writing style change detection, the use of LLMs for authorship attribution, taxonomy generation, and personalised education in K–12 settings. The programme also featured work on document visual question answering using vision–language models, highlighting the growing importance of multimodal approaches in IR research.

CLEF 2025 has continued the *methodology-focused review process* introduced in two previous CLEF editions, inspired by the "Dagstuhl Seminar 23031 on Frontiers of Information Access Experimentation for Research and Education" [Bauer et al., 2023a,b]. This innovative approach divides the reviewing process into two distinct phases. In the first phase, submissions are evaluated based on their *methodological contribution*, the clarity and relevance of their *research questions*, and the soundness of their *experimental design*. Importantly, papers submitted at this stage do not include any experimental results. Only those papers that pass this initial assessment proceed to the second phase, where reviewers examine the *experimentation*, *analysis*, and *insights* derived from the results. The primary objectives of this review model are threefold: (i) to avoid accepting papers solely on the basis of performance improvements over a baseline; (ii) to ensure that accepted

papers are grounded in robust methodology; and (iii) to confirm that research questions are driven by methodological rigour rather than post-hoc interpretation of results.

Since 2015, CLEF 2025 has continued the tradition of inviting CLEF lab organisers to nominate a "best of the labs" paper. These papers, originally submitted to the CLEF 2024 labs, were reviewed as full paper submissions to the CLEF 2025 conference, following the same review criteria and under the supervision of the programme committee. A total of **five full papers** were accepted in this special section, showcasing high-quality research outcomes from the CLEF 2024 evaluation activities and reinforcing the strong link between lab experimentation and scholarly dissemination.

3 The CLEF Lab Sessions

CLEF 2025 received a total of 20 lab proposals, which were evaluated through a peer-review process based on their innovation potential, the quality of the resources provided, and their relevance to real-world challenges. From these, 14 labs were selected, each addressing scientific problems grounded in datasets and practical applications in multimodal and multilingual information access.

These labs offer researchers unique opportunities to explore diverse collections, develop and test solutions, receive feedback on system performance, and engage in in-depth discussions with peers during the conference.

The selected labs for CLEF 2025 are: **BioASQ**, **CheckThat!**, **ELOQUENT**, **eRisk**, **EXIST**, **ImageCLEF**, **JOKER**, **LifeCLEF**, **LongEval**, **PAN**, **QuantumCLEF**, **SimpleText**, **TalentCLEF**, and **Touché**.

Details of each lab are provided by the organisers in the Conference proceedings [Carrillo-de Albornoz et al., 2025] and CLEF Working Notes [Faggioli et al., 2025]. A brief overview of the labs is presented below, in alphabetical order.

BioASQ: Large-scale biomedical semantic indexing and question answering[1] [Nentidis et al., 2025] aims to push the research frontier towards systems that use the diverse and voluminous information available online to respond directly to the information needs of biomedical scientists. This edition of BioASQ offered the following tasks: *Task 1 (13b) – Biomedical Semantic Question Answering:* Benchmark datasets of biomedical questions, in English, along with gold standard (reference) answers constructed by a team of biomedical experts. The participants have to respond with relevant articles, and snippets from designated resources, as well as exact and "ideal" answers. *Task 2 – Synergy: Question Answering for developing problems:* Biomedical experts pose unanswered questions for developing problems, such as COVID-19, receive the responses provided by the participating systems, and provide feedback, together with updated questions in an iterative procedure that aims to facilitate the incremental understanding of developing problems in biomedicine and public health. *Task 3 – MultiClinSum: Multilingual Clinical Summarization:* a shared task on the automatic summarization of lengthy clinical case reports written in different languages. The organisers distribute lengthy clinical case reports written in English, Spanish, French, and Portuguese. The participants generate summaries of the clinical case reports. The evaluation is based on a comparison with manual summaries of the clinical case reports.

[1] https://www.bioasq.org/workshop2025

Task 4 – BioNNE-L: Nested Named Entity Linking in Russian and English: A shared task on NLP challenges in entity linking, also known as medical concept normalization (MCN), for English and Russian languages. The train/dev datasets include annotated mentions of disorders, anatomical structures, and chemicals. The participants normalize the entity mentions to concept names and unique UMLS identifiers. The evaluation is based on a comparison with manual nested named entity linking annotations. *Task 5 – ElCardioCC: Clinical Coding in Cardiology:* The ELCardioCC on the automated clinical coding concerns i) the assignment of cardiology-related ICD-10 codes to discharge letters from Greek hospitals, ii) the extraction of the specific mentions of ICD-10 codes from the discharge letters. The evaluation is based on metrics, such as micro and macro F-measure for Subtask (i) and token F-measure for Subtask (ii). *Task 6 – GutBrainIE: Gut-Brain Interplay Information Extraction:* The GutBrainIE task aims to foster the development of Information Extraction (IE) systems that support experts by automatically extracting and linking knowledge from scientific literature, facilitating the understanding of gut-brain interplay and its role in neurological disease. The task is divided into two subtasks: i) extraction of named entities and linking them to concepts in a reference ontology, and ii) identifying binary relations between entity pairs.

CheckThat! Lab on Checkworthiness, Subjectivity, Persuasion, Roles, Authorities and Adversarial Robustness[2] [Alam et al., 2025] The eighth edition of the CheckThat! lab at CLEF presented a diverse set of challenges aimed at advancing technology to support and enhance the journalistic verification process. This edition revisited core tasks in the verification pipeline while also introducing auxiliary tasks such as subjectivity identification, claim normalization, and fact-checking numerical claims, with a particular emphasis on scientific web discourse. These tasks pose complex classification and retrieval problems at both the document level, including in multilingual contexts. The lab was organised into the following tasks: *Task 1 – Subjectivity:* Given a sentence from a news article, determine whether it is subjective or objective. This is a binary classification task and is offered in Arabic, English, Bulgarian, German, and Italian for mono- and multi-lingual settings. Additionally, unseen languages like French and Spanish are considered for zero-shot settings. *Task 2 – Claim Normalization:* Given a noisy, unstructured social media post, the task is to simplify it into a concise form. This is a generation task, offered in 20 languages: English, Arabic, Bengali, Czech, German, Greek, French, Hindi, Korean, Marathi, Indonesian, Dutch, Punjabi, Polish, Portuguese, Romanian, Spanish, Tamil, Telugu, Thai. *Task 3 – Fact-Checking Numerical Claims* This task focuses on verifying claims with numerical quantities and temporal expressions. Numerical claims are defined as those requiring validation of explicit or implicit quantitative or temporal details. Participants must classify each claim as True, False, or Conflicting based on a short list of evidence. *Task 4 – Scientific Web Discourse Processing (SciWeb)* which was further divided into two subtasks. *Subtask 4.1 – SciWeb Discourse Detection:* This task aims at classifying the different forms of science-related online discourse. Namely, given a tweet, this multilabel task aims at detecting if a tweet contains a scientific claim or scientific reference or is referring to science contexts or entities. *Subtask 4.2 – SciWeb Claim-Source Retrieval:* Given a tweet containing a scientific

[2] https://checkthat.gitlab.io/clef2025/

claim and an informal reference to a scientific paper, this task aims at retrieving the scientific paper that serves as the source for the claim from a given pool of candidate scientific papers.

ELOQUENT lab for evaluation of generative language model quality[3] [Karlgren et al., 2025] addresses high-level quality criteria through a set of open-ended shared tasks implemented to require minimal human assessment effort. It offered the following tasks: *Task 1 – Voight-Kampff:* Generate text samples for a classifier to distinguish between human-authored and machine-generated text. *Task 2 – Robustness and Consistency:* Explore how much a generative language model's output is affected by stylistic, dialectal, or other non-topical variation in the input. *Task 3 – Preference Score Prediction:* Predict human preferences between sets of LLM-generated responses collected from human assessors, and generate judgments to explain the choice made. *Task 4 – Sensemaking:* Given a set of possibly noisy texts, generate questions and answers about the topic.

eRisk: Early Risk Prediction on the Internet[4] [Parapar et al., 2025] explores the evaluation methodology, effectiveness metrics and practical applications (particularly those related to health and safety) of early risk detection on the Internet. This year's edition of eRisk included the following tasks: *Task 1 – Search for Symptoms of Depression:* Rank sentences from users according to their relevance to each of the 21 symptoms of the BDI-II questionnaire. Training data consists of sentence-tagged datasets from 2023 and 2024, with new test data including contextual information (previous and next sentences). *Task 2 – Contextualized Early Detection of Depression:* Participants analyze full conversational interactions to classify users with signs of depression, considering the conversational context beyond isolated user writings. The test phase includes writings with full conversational dynamics, while the training phase uses isolated user submissions. *Pilot Task – Conversational Depression Detection via LLMs:* Participants interact with a persona powered by a large language model (LLM) that is fine-tuned using types of depressive and non depressive users. The objective is to detect signs of depression, with participants limited to a specified number of messages to engage with the LLM.

EXIST: sEXism Identification in Social neTworks[5] [Plaza et al., 2025] aims to capture and categorize sexism, from explicit misogyny to other subtle behaviors, in social networks. In 2024 the EXIST campaign included multimedia content in the format of memes, stepping forward research on more robust techniques to identify sexism in social networks. Following this line, in 2025 we will focus on TikTok videos in the challenge, thus including in the dataset the three most important sources of sexism spreading: text, images and videos. Consequently, it is essential to develop automated multimodal tools capable of detecting sexism in text, images, and videos, to raise alarms or automatically remove such content from social network because platforms' algorithms often amplify content that perpetuates gender stereotypes and internalized misogyny. This lab will contribute to the creation of applications that identify sexist content in social media across all three formats. This task was divided into three tasks, each split into three subtasks. *Task 1 – Sexism Identification*

[3] https://eloquent-lab.github.io/
[4] https://erisk.irlab.org/
[5] https://nlp.uned.es/exist2025/

and Characterization in Tweets Subtask 1.1 – Sexism Identification in Tweets: The first subtask is a binary classification. The systems have to decide whether or not a given tweet contains or describes sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior). *Subtask 1.2 – Source Intention in Tweets* This subtask aims to categorize the sexist messages according to the intention of the author in one of the following categories: (i) direct sexist message, (ii) reported sexist message and (iii) judgmental message. *Subtask 1.3 – Sexism Categorization in Tweets* The third subtask is a multiclass task that aims to categorize the sexist messages according to the type or types of sexism they contain (according to the categorization proposed by experts and that takes into account the different facets of women that are undermined): (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence. *Task 2 – Sexism Identification and Characterization in Memes Subtask 2.1 – Sexism Identification in Memes:* Similar to Subtask 1.1, Subtask 2.1 is a binary classification task where participants must determine when a meme contains or describes sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior). *Subtask 2.2 – Source Intention in Memes:* This subtask aims to categorize the sexist messages according to the intention of the author in one of the following categories: (i) direct sexist message, (ii) judgmental message. *Subtask 2.3 – Sexism Categorization in Memes:* Finally, this subtask addresses the problem of categorizing a sexist meme according to the type of sexism that it encloses: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence. *Task 3 - Sexism Identification and Characterization in TikTok Videos Subtask 3.1 – Sexism Identification in Videos:* Similar to Subtasks 1.1 and 2.1, this subtask is a binary classification task where participants must determine when a video contains or describes sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior). *Subtask 3.2 – Source Intention in Videos:* This subtask aims to categorize the sexist messages according to the intention of the author in one of the following categories: (i) direct sexist message, (ii) judgmental message. *Subtask 3.3 – Sexism Categorization in Videos:* Finally, this subtask addresses the problem of categorizing a sexist meme according to the type of sexism that it encloses: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence.

ImageCLEF: Multimodal Challenge in CLEF[6] [Ionescu et al., 2025] focuses on evaluating technologies for annotating, indexing, classifying, retrieving and generating multimodal data, providing access to large datasets across a variety of scenarios, including medical, social media, and internet-based applications. Building on the success of recent editions, it encourages interdisciplinary methods by engaging participants in diverse domains, providing large amounts of challenging multimodal data and providing an evaluation platform for a large number of use cases. This year's edition of ImageCLEF involved the following tasks: *Task 1 – ImageCLEFmedical:* In its 21st edition, the task will continue all the medical subtasks from the last 2 years, namely: (i) the Caption task with medical concept detection and caption prediction, (ii) the GAN task focused on synthetic medical images, (iii) MEDVQA

[6] https://www.imageclef.org/2025

regarding Visual Question Answering for gastrointestinal data, and (iv) MEDIQA-MAGIC, introducing a new use-case on multimodal dermatology response generation. *Task 2 – Image Retrieval/Generation for Arguments:* As a joint task between Touché and ImageCLEF since 2022, the task aims to show the impact of images in arguments, making them more compelling. In this year's task, participants shall find suitable images that convey a given argument. Two submission styles are possible, either as a retrieval task or as prompt generation for an image generator. *Task 3 – ImageCLEFtoPicto:* The aim of this task to convert either speech or text into a meaningful sequence of pictograms, aiding communication for people with language impairments, enhancing user understanding or helping with translation. Therefore, 2 sub-tasks are derived from this: (i) Text-to-Picto, involving generating pictograms starting from a French text and (ii) Speech-to-Picto, which focuses on translating speech to pictograms directly. *Task 4 – MultimodalReason:* is a new task, focusing on Multilingual Visual Question Answering. Participants are given multiple-choice questions and corresponding images and are asked to identify the correct answer, in multiple languages, disciplines and difficulty levels. The task aims to assess the reasoning abilities of modern LLMs across a wide range of real-world situations.

JOKER: Automatic Humour Analysis[7] [Ermakova et al., 2025b] aims to foster interdisciplinary approaches to the (semi-)automatic analysis and processing of humor and wordplay. *Task 1 – Humor-aware Information Retrieval:* For Task 1, the aim is to retrieve short humorous texts from a document collection based on a given query. The languages are English and Portuguese. *Task 2 – Wordplay Translation:* For Task 2, the goal is to translate English punning jokes into French. *Task 3 – Onomastic Wordplay:* For Task 3, the goal is to classify proper names according to whether they are humorous, and to translate them from English into French. *Task 4 – Controlled Creativity:* For Task 4, the goal is to identify the introduction of distorted or spurious content ("hallucinations") in generated creative texts.

LifeCLEF: Challenges on Species Presence Prediction and Identification, and Individual Animal Identification[8] [Picek et al., 2025] focuses on advancing AI-driven solutions for biodiversity monitoring through challenges on species and individuals recognition and prediction. *Task 1 – AnimalCLEF:* Multi-species individual animal identification. *Task 2 – BirdCLEF:* Bird species identification in soundscape recordings. *Task 3 – FungiCLEF:* Few-shot classification with rare fungi species. *Task 4 – GeoLifeCLEF:* Multi-modal species prediction using remote sensing and large-scale biodiversity data. *Task 5 – PlantCLEF:* Multi-species plant identification in vegetation plot images.

LongEval: Longitudinal Evaluation of Model Performance[9] [Cancellieri et al., 2025] aims to ignite the development of Information Retrieval systems that can handle temporal data evolution. The retrieval systems evaluated in this task are expected to be persistent in their retrieval efficiency, as collection documents and queries change over time. To evaluate such features of systems, we rely on collections of documents and queries, corresponding to data acquired from two Web search engines. The time distance between the training and test

[7] http://joker-project.com/
[8] https://www.imageclef.org/LifeCLEF2025
[9] https://clef-longeval.github.io/

data for the LongEval tasks ranges from one to 6 months. LongEval 2025 included two tasks, *WebRetrieval* and *SciRetrieval*, each tailored to the two types of data available in Longeval: Web document entries, collected from the Qwant search engine, and scientific document entries, collected from the CORE (COnnecting REpositories) service.

PAN: Lab on Stylometry and Digital Text Forensics[10] [Bevendorff et al., 2025] aims to advance the state of the art and provide for an objective evaluation on newly developed benchmark datasets in those areas. The tasks proposed by PAN Lab this year included: *Task 1 – Generated Content Analysis:* Given a document, decide if it was written by a human, an AI, or both. *Task 2 – Multilingual Text Detoxification:* Given a toxic piece of text, re-write it in a non-toxic way while saving the main content as much as possible. *Task 3 – Multi-author Writing Style Analysis:* Given a document, determine at which positions the author changes. *Task 4 – Generated Plagiarism Detection:* Given a generated and a human-written source document, identify the passages of reused text between them.

QuantumCLEF: Quantum Computing at CLEF[11] [Pasin et al., 2025] The second edition of the QuantumCLEF lab is composed of three tasks and aims at: Discovering and evaluating Quantum Annealing approaches compared to their traditional counterpart; Identifying new ways of formulating Information Retrieval and Recommender Systems algorithms and methods, so that they can be solved with Quantum Annealing; Establishing collaborations among researchers from different fields to harness their knowledge and skills to solve the considered challenges and promote the usage of Quantum Annealing. This lab allows participants to use real quantum computers provided by CINECA, one of the most important computing centers worldwide. *Task 1 – Feature Selection:* focuses on formulating the well-known NP-Hard Feature Selection problem and solving it with quantum annealers. Feature Selection is a widespread problem for both Information Retrieval and Recommender systems which requires to identify a subset of the available features (e.g., the most informative, less noisy, etc.) to train a learning model. This problem is very impacting since many of these systems involve the optimization of learning models, and reducing the dimensionality and noise of the input data can improve their performance. *Task 2 – Instance Selection:* focuses on formulating the Instance Selection problem to solve it through Quantum Annealing. Currently, transformer-based architectures, including 1st and 2nd generation transformers (e.g., RoBERTa) as well as current large language models (e.g., Llama3), are used and considered state-of-the-art in several fields. Given the LLMs high-cost application, one of the big challenges is to fine-tune these models efficiently. Instance Selection focuses on selecting a representative subset of instances from a dataset to make the training of these models faster while maintaining a high level of effectiveness of the trained model. *Task 3 – Clustering:* focuses on the formulation of the clustering problem to solve it with a quantum annealer. Clustering is a relevant problem for Information Retrieval and Recommender systems which involves grouping items together according to their characteristics. Clustering can be helpful for organising large collections, helping users to explore a collection and providing similar results to a query. It can also be used to divide users according to their interests or build user

[10] http://pan.webis.de/
[11] https://qclef.dei.unipd.it/

models with the cluster centroids boosting efficiency or effectiveness for users with limited data.

SimpleText: Simplify Scientific Text (and Nothing More)[12] [Ermakova et al., 2025a] aims at improving accessibility to scientific information for everyone, developing corpora, evaluation measures, and new IR/NL models able to reduce scientific text complexity with strict faithfulness to the original text. *Task 1 – Text Simplification: simplify scientific text:* aims to simplify scientific text, using aligned biomedical abstracts and lay summaries for sentence-level, paragraph-level, and document-level text simplification. *Task 2 – Controlled Creativity: identify and avoid hallucination:* aims to identify and avoid hallucination, by either post-hoc detection on CLEF submissions with over-generation, or by avoiding creative license of models by design. *Task 3 – SimpleText 2024 Revisited: selected tasks by popular request:* aims to rerun selected tasks by popular request, on scientific passage retrieval and complex terminology detection, and on tracking the state-of-the-art in scholarly papers.

TalentCLEF: Skill and Job Title Intelligence for Human Capital Management[13] [Gasco et al., 2025] aims to drive technological advancement in Human Capital Management by establishing a public benchmark for NLP models that facilitates their application in real-world Human Resources (HR) scenarios, incorporating evaluation criteria including multilingualism, fairness, and cross-industry adaptability. The lab also seeks to build a community for researchers and practitioners to generate, evaluate, and discuss ideas on the use of AI in Human Resources, pushing the state-of-the-art of NLP applications for Human Resources. *Task 1 – Multilingual Job Title Matching:* involves the development of systems that can identify and rank job titles most similar to a given one. For each job title in a provided test set, participants must generate a ranked list of similar job titles from a specified knowledge base. The task includes multilingual and cross-lingual tracks, requiring participants to develop systems adapted to English, Spanish, German, and optionally Chinese. *Task 2 – Job Title-Based Skill Prediction:* involves developing systems capable of retrieving relevant skills associated with a given job title. Participants must train models that can retrieve a list of relevant skills from a provided knowledge base, ranking them according to their relevance to the job title. This task is in English.

Touché: Argumentation Systems[14] [Kiesel et al., 2025] focuses on computational argumentation and causality. Touché 2025 included 4 tasks. *Task 1 – Retrieval-Augmented Debating:* it served to develop generative retrieval systems that argue against their users to support users in forming or confirming opinions or to train their debating skills. *Task 2 – Ideology and Power Identification in Parliamentary Debates:* it concerned with predicting ideology and power in the parliamentary debates on a multi-lingual, multi-country dataset. *Task 3 – Image Retrieval/Generation for Arguments (Joint task with ImageCLEF):* aimed to find images that support a particular point of view. *Task 4 – Advertisement in Retrieval-Augmented Generation:* analyzed possibilities and counter-measures for advertisements in retrieval-augmented search results.

[12] http://simpletext-project.com/
[13] https://talentclef.github.io/talentclef/
[14] https://touche.webis.de/

4 Overall Trends for CLEF

Figure 1 illustrates how participation at CLEF has evolved since the event's first edition. The attendance evolution from 2000 to 2025 reflects a steady increase in participation, particularly following the format change in 2010 that introduced a peer-reviewed conference. The 2020 and 2021 editions, held entirely online due to the COVID-19 pandemic and featuring minimal registration fees, saw a notable surge with exclusively remote attendance. In the post-pandemic years, CLEF embraced a hybrid format, successfully balancing both in-person and remote participation. The 2025 edition in Madrid marked a record in both overall (excluding the exceptional virtual years) and in-person attendance, attracting 254 participants—214 on site and 40 remotely. This milestone not only highlights the renewed enthusiasm for face-to-face scientific exchange, but also celebrates CLEF's vibrant reinstatement as a dynamic and engaging community event, reaffirming the value of in-person collaboration in the research landscape, while continuing to offer remote access for exceptional circumstances.

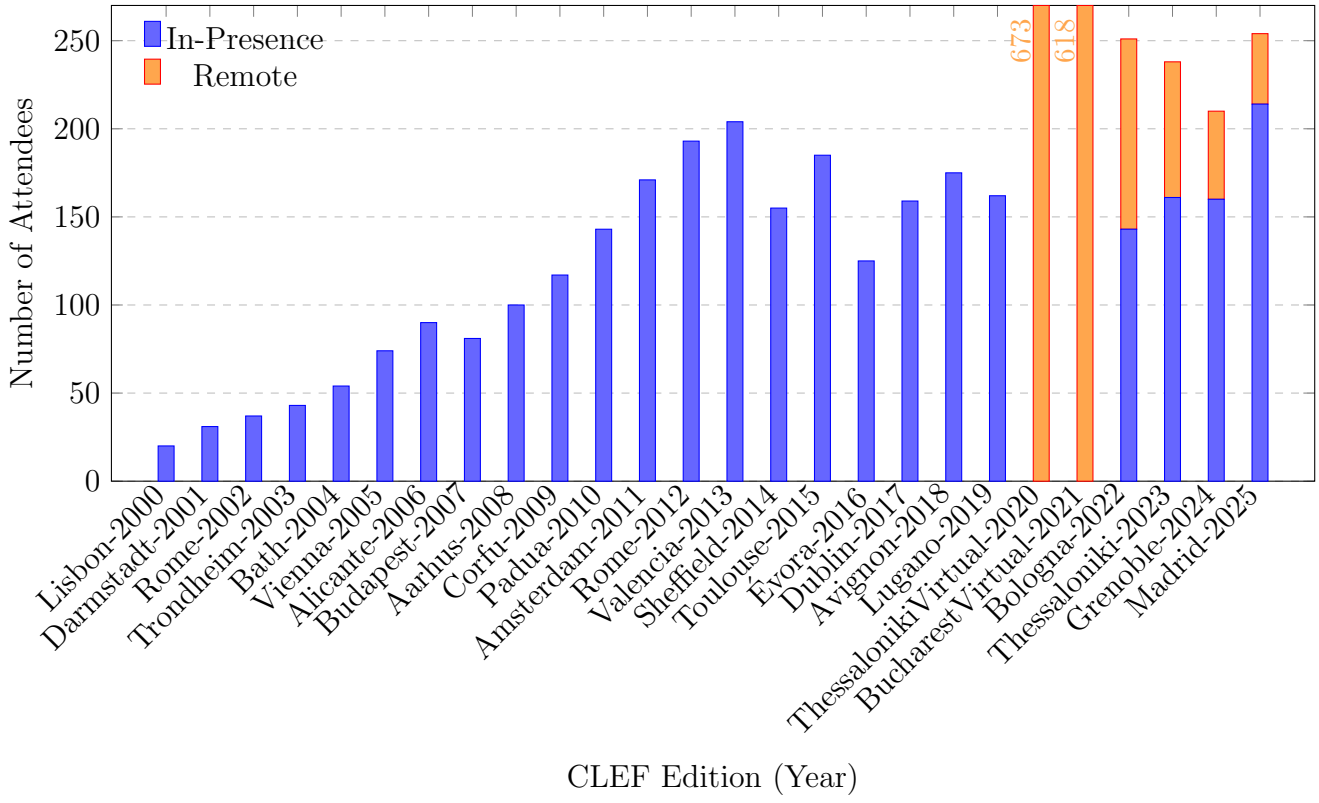

Figure 1. Evolution of In-Person and Remote Attendees at CLEF 2000–2025 (bars capped at 270).

Figure 2 illustrates the number of papers published in the CLEF Working Notes from 2000 to 2025. We report the Working Notes because they include both the lab overview papers and all participant contributions, offering a comprehensive view of the scientific output associated with CLEF each year. The data reveals a steady increase over time, with a particularly sharp rise in recent years, culminating in a record 392 papers in 2025.

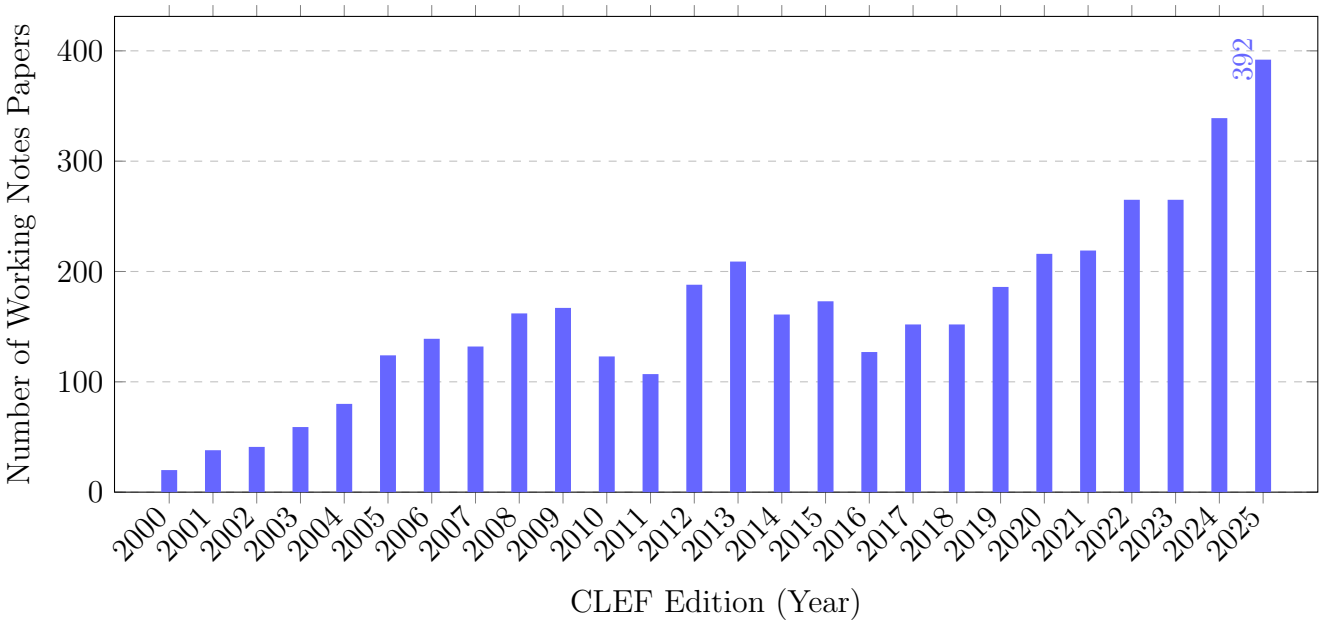

Figure 2. Number of papers published in the CLEF Working Notes from 2000 to 2025.

Figure 3 shows the Google Scholar metrics for CLEF.[15] The h5-index and h5-median values have shown a consistent upward trend from 2016 to 2024, reaching peaks of 47 and 65 respectively. These metrics reflect CLEF's increasing scholarly impact and visibility. A slight dip in 2025 may be attributed to indexing delays or natural variation, but overall, the data confirms CLEF's strong and sustained presence in the research landscape.

According to Google Scholar Metrics, when searching for venues related to *Information Retrieval*, *Information Access*, *Information and Knowledge Extraction*, and *Evaluation*, CLEF ranks among the top 10 international venues, positioned 7th, highlighting its recognised influence and impact in the field of information access evaluation.

5 CLEF 2026

CLEF 2026 will take place at the Friedrich-Schiller-Universität Jena, in the heart of Jena, Germany, on 21–24 September 2026.

More information on CLEF 2026, the call for papers and the ongoing labs is available at:

- https://clef2026.clef-initiative.eu

As far as labs are concerned, CLEF 2026 will run 16 evaluation activities out of 18 proposals received: fourteen will be a continuation of the labs running during CLEF 2025:

- BioASQ – A challenge in large-scale biomedical semantic indexing and question answering;[16]

[15] Note that Google Scholar still indexes CLEF as "Cross-Language Evaluation Forum", even if the name has changed to "Conference and Labs of the Evaluation Forum" since 2010.

[16] https://www.bioasq.org/workshop2026

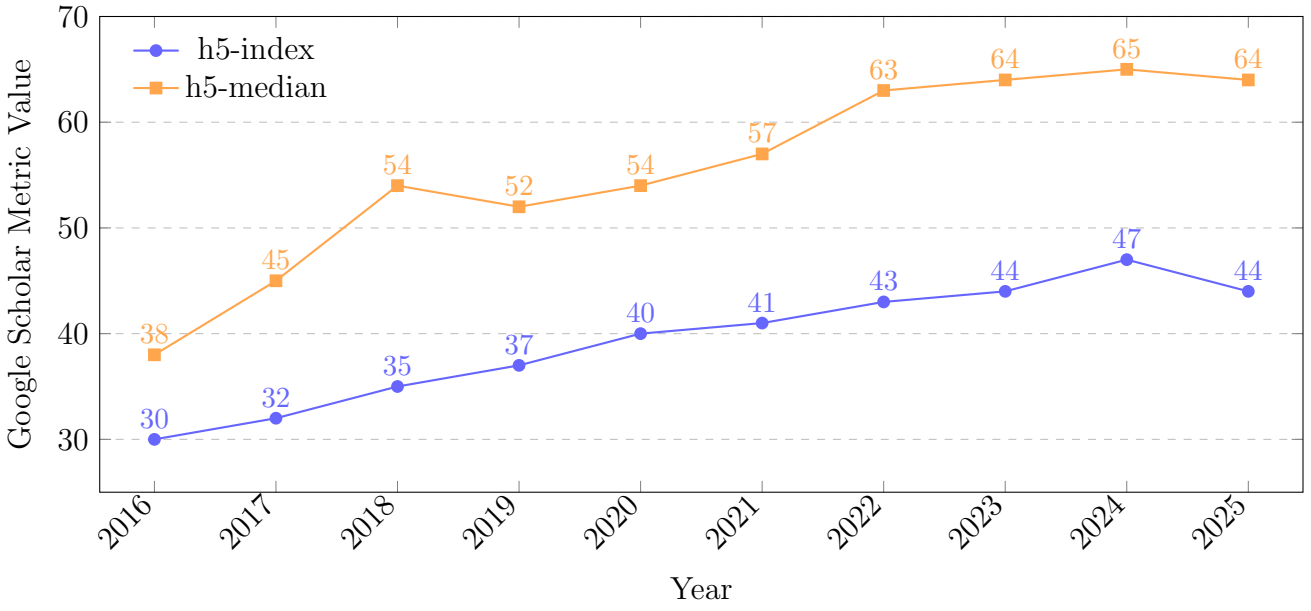

Figure 3. Google Scholar metrics for "Cross-Language Evaluation Forum" since 2016: h5-index and h5-median values.

- CheckThat! – Lab on Subjectivity, Fact-Checking, Claim Extraction & Normalization, and Retrieval;[17]
- ELOQUENT – Lab for evaluation of generative language model quality;[18]
- eRisk – Early risk prediction on the Internet;[19]
- EXIST – sEXism Identification in Social neTworks;[20]
- ImageCLEF – Multimodal Challenge in CLEF;[21]
- JOKER – Humour in the Machine;[22]
- LifeCLEF – Challenges on Species Presence Prediction and Identification, and Individual Animal Identification;[23]
- LongEval – Longitudinal Evaluation of Model Performance;[24]
- PAN – Lab on Stylometry and Digital Text Forensics;[25]
- QuantumCLEF – Quantum Computing at CLEF;[26]
- SimpleText – Simplify Scientific Text (and Nothing More);[27]
- TalentCLEF – Skill and Job Title Intelligence for Human Capital Management;[28]

[17] http://checkthat.gitlab.io/clef2026/
[18] https://eloquent-lab.github.io
[19] https://erisk.irlab.org
[20] https://nlp.uned.es/exist2026/
[21] https://www.imageclef.org/2026/
[22] https://www.joker-project.com
[23] http://www.lifeclef.org
[24] https://clef-longeval.github.io
[25] https://pan.webis.de
[26] https://qclef.dei.unipd.it
[27] https://simpletext-project.com
[28] https://talentclef.github.io/talentclef

- Touché – Argumentation Systems.[29]

One new pilot lab:

- FinMMEval – Multilingual and multimodal evaluation of financial AI systems. The aim of FinMMEval is to advance the multilingual and multimodal evaluation of financial AI systems. Recognising that real-world financial decision-making depends on diverse data types—ranging from textual and visual documents to time-series signals across multiple languages—the lab seeks to foster robust, interpretable, and auditable AI solutions. By bringing together expertise from natural language processing, computer vision, time-series modelling, and financial analysis, FinMMEval aspires to establish a benchmark for next-generation financial AI and promote cross-disciplinary collaboration in this critical domain.

And one returning lab:

- HIPE – Evaluating accurate and efficient person-place relation extraction from multilingual historical texts.[30] The aim of HIPE-2026 is to advance the extraction and interpretation of person–place relations in multilingual historical documents, building on the foundations of previous CLEF Evaluation Labs HIPE-2020 and HIPE-2022. By focusing on a single but fundamental relation type and its temporal scope, the lab seeks to support the development of methods essential for constructing historical knowledge graphs, enabling spatial analysis, and reconstructing biographies. HIPE-2026 also introduces efficiency as a key evaluation criterion, recognising the computational demands of processing large-scale cultural heritage collections and encouraging approaches that balance accuracy with resource awareness.

6 CLEF 2027

CLEF 2027 will be hosted in Bucharest, Romania, and organised once again by the University "Politehnica" of Bucharest. This edition marks a symbolic return to the city, as the 2021 conference —also organised by the University "Politehnica" of Bucharest— was held entirely online due to the COVID-19 pandemic [Candan et al., 2021; Faggioli et al., 2021]. CLEF 2027 will offer the opportunity to reconnect in person with the CLEF community and continue fostering multilingual and multimodal evaluation research in a city that previously welcomed CLEF under exceptional circumstances.

7 Bids for CLEF 2028

The call for proposals is currently open and will remain so until December 2025. Institutions interested in hosting CLEF 2028 are invited to submit their bids. Submissions should be sent to the CLEF Steering Committee Chairs at chair@clef-initiative.eu. A template to guide the

[29] https://touche.webis.de
[30] https://hipe-eval.github.io/HIPE-2026/

preparation of bids is available at: https://www.clef-initiative.eu/assets/CLEF-Template_for_bids.docx.

Acknowledgments

The success of CLEF 2025 would not have been possible without the dedicated efforts of many individuals and organisations, including the CLEF Association, the Programme Committee, the Lab Organising Committee, the reviewers, and the numerous students and volunteers whose contributions were essential to the event.

We gratefully acknowledge the Friends of SIGIR programme for covering the registration fees of several students. UNED has generously supported CLEF 2025 by providing funding for coffee breaks, institutional backing, and access to the venues at the Faculties of Education and Psychology. We also extend our sincere thanks to the HiTZ Chair of Artificial Intelligence and Language Technology at the University of the Basque Country for their generous sponsorship.

Finally, CLEF owes its existence to the enthusiasm, creativity, and hard work of the authors, the organisers of the selected labs, the colleagues and friends involved in running them, and all the participants who contribute their time and commitment to the labs and the conference, as well as their financial support through the CLEF Association.

Thank you all very much!

A Authors and Affiliations

Workshop organizers:

- Jorge Carrillo-de-Albornoz, UNED, Spain, *jcalbornoz@lsi.uned.es*
- Guglielmo Faggioli, University of Padua, Italy, *faggioli@dei.unipd.it*
- Nicola Ferro, University of Padua, Italy, *nicola.ferro@unipd.it*
- Alba García Seco de Herrera, UNED, Spain, *alba.garcia@lsi.uned.es*
- Julio Gonzalo, UNED, Spain, *julio@lsi.uned.es*
- Laura Plaza, UNED, Spain, *lplaza@lsi.uned.es*
- Josiane Mothe, Université de Toulouse, UT2J, IRIT, France, *josiane.mothe@irit.fr*
- Florina Piroi, TU Wien, Austria, *florina.piroi@tuwien.ac.at*
- Paolo Rosso, Universitat Politècnica de València, Spain, *prosso@dsic.upv.es*
- Damiano Spina, RMIT University, Australia,*damiano.spina@rmit.edu.au*

References

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and V. Venktesh. Overview of the CLEF-2025 CheckThat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In Carrillo-de Albornoz et al. [2025], pages 199–223. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_13. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Christine Bauer, Ben Carterette, Nicola Ferro, Norbert Fuhr, Joeran Beel, Timo Breuer, Charles L. A. Clarke, Anita Crescenzi, Gianluca Demartini, Giorgio Maria Di Nunzio, Laura Dietz, Guglielmo Faggioli, Bruce Ferwerda, Maik Fröbe, Matthias Hagen, Allan Hanbury, Claudia Hauff, Dietmar Jannach, Noriko Kando, Evangelos Kanoulas, Bart P. Knijnenburg, Udo Kruschwitz, Meijie Li, Maria Maistro, Lien Michiels, Andrea Papenmeier, Martin Potthast, Paolo Rosso, Alan Said, Philipp Schaer, Christin Seifert, Damiano Spina, Benno Stein, Nava Tintarev, Julián Urbano, Henning Wachsmuth, Martijn C. Willemsen, and Justin Zobel. Report on the Dagstuhl seminar on frontiers of information access experimentation for research and education. 57(1), December 2023a. ISSN 0163-5840. doi: 10.1145/3636341.3636351. URL https://doi.org/10.1145/3636341.3636351.

Christine Bauer, Ben A. Carterette, Nicola Ferro, Norbert Fuhr, and Guglielmo Faggioli, editors. *Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education*, Dagstuhl Reports, Volume 13, Number 1, 2023b. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.

Janek Bevendorff, Daryna Dementieva, Maik Fröbe, Bela Gipp, André Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, Artem Shelmanov, Efstathios Stamatatos, Benno Stein, Yuxia Wang, Matti Wiegmann, and Eva Zangerle. Overview of PAN 2025: Voight-kampff generative ai detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection. In Carrillo-de Albornoz et al. [2025], pages 388–411. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_21. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Matteo Cancellieri, Alaa El-Ebshihy, Tobias Fink, Maik Fröbe, Petra Galuščáková, Gabriela Gonzalez-Saez, Lorraine Goeuriot, David Iommi, Jüri Keller, Petr Knoth, Philippe Mulhem, Florina Piroi, David Pride, and Philipp Schaer. LongEval at CLEF 2025: Longitudinal evaluation of ir systems on web and scientific data. In Carrillo-de Albornoz et al. [2025], pages 363–387. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_20. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

K.S̃elçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021)*, 2021. Lecture Notes in Computer Science (LNCS) 12880, Springer, Heidelberg, Germany.

Jorge Carrillo-de Albornoz, Alba Garcá Seco de Herrera, Julio Gonzalo, Luara Plaza, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025)*, volume 16089 of *Lecture Notes in Computer Science (LNCS)*. Springer, Cham, 2025. ISBN 978-3-032-04354-2. doi: https://doi.org/10.1007/978-3-032-04354-2. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Liana Ermakova, Hosein Azarbonyad, Jan Bakker, Benjamin Vendeville, and Jaap Kamps. Overview of the CLEF 2025 SimpleText track. In Carrillo-de Albornoz et al. [2025], pages

436–463. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_23. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Liana Ermakova, Ricardo Campos, Anne-Gwenn Bosser, and Tristan Miller. Overview of the CLEF 2025 JOKER lab: Humour in machine. In Carrillo-de Albornoz et al. [2025], pages 315–337. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_18. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi, editors. *CLEF 2021 Working Notes*, 2021. CEUR Workshop Proceedings (CEUR-WS.org). URL http://ceur-ws.org/Vol-2936/.

Guglielmo Faggioli, Nicola Ferro, Paolo Rosso, and Damiano Spina, editors. *CLEF 2025 Working Notes*, volume 4038 of *CEUR Workshop Proceedings*. CEUR Workshop Proceedings (CEUR-WS.org), Madrid, Spain, September 2025. URL https://ceur-ws.org/Vol-4038/.

Nicola Ferro. What Happened in CLEF... For a While? In Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*, pages 3–45. Lecture Notes in Computer Science (LNCS) 11696, Springer, Heidelberg, Germany, 2019.

Nicola Ferro. What Happened in CLEF... For Another While? In L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) – Part I*, pages 3–57. Lecture Notes in Computer Science (LNCS) 14958, Springer, Heidelberg, Germany, 2024.

Nicola Ferro and Carol Peters, editors. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, 2019. Springer International Publishing, Germany.

Nicola Ferro and Gianmaria Silvello. 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc-*ish* CLEF? *Information Processing & Management*, 53(1): 175–202, January 2017.

Luis Gasco, Hermenegildo Fabregat, Laura García-Sardiña, Paula Estrella, Daniel Deniz, Alvaro Rodrigo, and Rabih Zbib. Overview of the TalentCLEF 2025: Skill and job title intelligence for human capital management. In Carrillo-de Albornoz et al. [2025], pages 464–485. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_24. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors. *Advances in Information Retrieval: 47th European Conference on Information Retrieval (ECIR 2025) – Part V*, volume 15576 of

Lecture Notes in Computer Science (LNCS), 2025. Springer. ISBN 978-3-031-88719-2. doi: 10.1007/978-3-031-88720-8. URL https://link.springer.com/book/10.1007/978-3-031-88720-8.

Bogdan Ionescu, Henning Müller, Dan-Cristian Stanciu, Alexandra-Georgiana Andrei, Ahmedkhan Radzhabov, Yuri Prokopchuk, Liviu-Daniel Ştefan, Mihai Gabriel Constantin, Mihai Dogariu, Vassili Kovalev, Hendrik Damm, Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia Sabrina Schmidt, Tabea M. G. Pakull, Benjamin Bracke, Obioma Pelka, Bahadır Eryılmaz, Helmut Becker, Wen-Wai Yim, Noel Codella, Roberto Andres Novoa, Josep Malvehy, Dimitar Dimitrov, Rocktim Jyoti Das, Zhuohan Xie, Ming Shan Hee, Preslav Nakov, Ivan Koychev, Steven A. Hicks, Sushant Gautam, Michael A. Riegler, Vajira Thambawita, Pål Halvorsen, Diandra Fabre, Cécile Macaire, Benjamin Lecouteux, Didier Schwab, Martin Potthast, Maximilian Heinrich, Johannes Kiesel, Moritz Wolter, Sharat Anand, and Benno Stein. Overview of ImageCLEF 2025: Multimedia retrieval in medical, social media and content recommendation applications. In Carrillo-de Albornoz et al. [2025], pages 290–314. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_17. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Jussi Karlgren, Ekaterina Artemova, Ondřej Bojar, Marie Isabel Engels, Vladislav Mikhailov, Pavel Šindelář, Erik Velldal, and Lilja Øvrelid. Overview of ELOQUENT 2025: Shared tasks for evaluating generative language model quality. In Carrillo-de Albornoz et al. [2025], pages 224–241. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_14. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Johannes Kiesel, Çağrı Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaž Erjavec, Matthias Hagen, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Harrisen Scells, Moritz Wolter, Ines Zelch, Martin Potthast, and Benno Stein. Overview of Touché 2025: Argumentation systems. In Carrillo-de Albornoz et al. [2025], pages 486–508. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_25. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Birger Larsen. The Scholarly Impact of CLEF 2010-2017. In N. Ferro and C. Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 547–554. Springer International Publishing, Germany, 2019.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, Grigorios Tsoumakas, George Giannakoulas, Alexandra Bekiaridou, Athanasios Samaras, Giorgio Maria Di Nunzio, Nicola Ferro, Stefano Marchesin, Marco Martinelli, Gianmaria Silvello, and Georgios Paliouras. Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In Carrillo-de Albornoz et al. [2025], pages 173–198. ISBN 978-3-032-04354-2. doi:

10.1007/978-3-032-04354-2_12. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. Overview of eRisk 2025: Early risk prediction on the internet. In Carrillo-de Albornoz et al. [2025], pages 242–265. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_15. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Andrea Pasin, Maurizio Ferrari Dacrema, Washington Cunha, Marcos André Gonçalves, Paolo Cremonesi, and Nicola Ferro. Overview of QuantumCLEF 2025: The second quantum computing challenge for information retrieval and recommender systems at CLEF. In Carrillo-de Albornoz et al. [2025], pages 412–435. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_22. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Lukáš Picek, Stefan Kahl, Hervé Goëau, Lukáš Adam, Théo Larcher, Cesar Leblanc, Maximilien Servajean, Klára Janoušková, Jiří Matas, Vojtěch Čermák, Kostas Papafitsoros, Robert Planqué, Willem-Pier Vellinga, Holger Klinck, Tom Denton, Juan Sebastián Cañas, Giulio Martellucci, Fabrice Vinatier, Pierre Bonnet, and Alexis Joly. Overview of LifeCLEF 2025: Challenges on species presence prediction and identification, and individual animal identification. In Carrillo-de Albornoz et al. [2025], pages 338–362. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_19. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Laura Plaza, Jorge Carrillo-de Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, and Roser Morante. Overview of EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos. In Carrillo-de Albornoz et al. [2025], pages 266–289. ISBN 978-3-032-04354-2. doi: 10.1007/978-3-032-04354-2_16. URL https://link.springer.com/book/10.1007/978-3-032-04354-2.

Theodora Tsikrika, Alba Garcia Seco de Herrera, and Henning Müller. Assessing the Scholarly Impact of ImageCLEF. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*, pages 95–106. Lecture Notes in Computer Science (LNCS) 6941, Springer, Heidelberg, Germany, 2011.

Theodora Tsikrika, Birger Larsen, Henning Müller, Stefan Endrullis, and Erhard Rahm. The Scholarly Impact of CLEF (2000–2009). In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013)*, pages 1–12. Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany, 2013.