NIST Special Publications
SP 1348

The 34th Text REtrieval Conference

TREC 2025

Ian Soboroff
George Awad

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1348

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publications
SP 1348

The 34th Text REtrieval Conference

TREC 2025

Ian Soboroff
George Awad
Technology Test and Evaluation Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1348

March 2026

U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2026-03-19

How to cite this NIST Technical Series Publication:
Soboroff IM, Awad GM (2026) The 34th Text REtrieval Conference. (National Institute of Standards and Technology, Gaithersburg, MD), SP 1348. https://doi.org/10.6028/NIST.SP.1348

Author ORCID iDs
Ian Soboroff: 0000-0003-2363-3014
George Awad: 0000-0003-2058-8422

Contact Information
trec@nist.gov

SP 1348
March2026

Abstract

TREC 2025 is the thirty-fourth edition of the Text REtrieval Conference (TREC). The main goal of TREC is to create the evaluation infrastructure required for large-scale testing of information retrieval (IR) technology. This includes research on best methods for evaluation as well as development of the evaluation materials themselves. "Retrieval technology" is broadly interpreted to include a variety of techniques that enable and/or facilitate access to information that is not specifically structured for machine use. The TREC 2025 meeting was held online on December 11–12, 2025.

Keywords

information retrieval; natural language processing; search; evaluation

Table of Contents

List of Tables

List of Figures

1. Introduction

TREC 2025 is the thirty-fourth edition of the Text REtrieval Conference (TREC). The main goal of TREC is to create the evaluation infrastructure required for large-scale testing of information retrieval (IR) technology. This includes research on best methods for evaluation as well as development of the evaluation materials themselves. "Retrieval technology" is broadly interpreted to include a variety of techniques that enable and/or facilitate access to information that is not specifically structured for machine use. The TREC 2025 meeting was held online on December 11–12, 2025.

Each TREC is organized around a set of focus areas called "tracks". A track has a motivating use case, which is generally an abstraction of a user task. TREC 2025 contained ten tracks:

Adhoc Video Search (AVS): content-based search in internet video

BioGen: retrieval-augmented generation for bioinformatics.

Detection, Retrieval, and Augmented Generation for Understanding News (DRAGUN): generating questions about a news article, and searching for answers.

Interactive Knowledge Aquisition (IKAT): conversational search

Million LLM: selecting among many LLMs in order to answer hard queries.

Product Search and Recommnendation: searching for products in an e-commerce setting.

Retrieval-Augmented Generation (RAG): RAG for web search.

RAG TREC Instrument for Multilingual Evaluation (RAGTIME): multilingual search and RAG for reports.

Tip-of-the-Tongue Search (ToT): searching for very ill-defined known item search needs.

Video Question Answering (VQA): Asking multimodal AI systems questions about short videos.

In addition, a Change Detection track was planned but has been postponed to TREC 2026.

Seventy-three groups from 23 different countries participated in TREC 2025. Table 1 lists the participating organizations.

This paper serves as an introduction to the research described in detail in the remainder of the proceedings. The next section provides a summary of the retrieval background knowledge that is assumed in the other papers. Section 3 presents a short description of each track—a more complete description of a track can be found in that track's overview paper in the proceedings.

Table 1: Organizations participating in TREC 2025

ANR Guidance	Meisei University
Adam Mickiewicz University	Missouri University of Science and Technology
Air Force Research Laboratory	NCSU LAS
CFDA Lab, Academia Sinica	NIT Agartala TREC RAG Team
Centre for Research and Technology Hellas	Nagaoka University of Technology
Centrl South University	National Chengchi University
Chittagong University of Engineering & Technology	National Institute of Informatics (NII), HCM-UIT
Co-creation Informatics Laboratory, Doshisha Univ.	National Institute of Technology, Agartala
Computational Finance and Data Analytics Lab	National University of Singapore
Computational Linguistics at Concordia	Paul Hildebrandt
CyCraft	Politecnico di Torino
DFKI	RMIT University
Dalhousie University	Ricoh Software Research Center (Beijing) Co., Ltd.
Dario Giovannini Thesis at TU Wien	Shenzhen University Artificial Intelligence
Data Science @ Georgia Tech	Siena College Institute of Artificial Intelligence
Democritus University of Thrace	Swiss Institute of Bioinformatics
Deutsche Forschungszentrum für Künstliche Intellig	TREMA lab at the University of New Hampshire
Digital Science	The University of Tokyo
GE Healthcare	Tianjin University
GRILL Lab	Tokyo University of Science
GenAIus Technologies	Toyota Connected North America Inc
Human Lang Tech COE at Johns Hopkins University	Trec Product Search and Recommendations Organizers
Human Language Technology Center of Excellence	Universidade Federal de Minas Gerais
Human Language Technology Center of Excellence - Evaluation Team	University of Amsterdam
IDA Center for Computing Sciences	University of Amsterdam, Information Retrieval Lab
II-Lab, University of Tsukuba	University of Delaware
IRKM Lab at University of California Santa Cruz	University of Glasgow Terrier Team
Indian Institute of Technology Jodhpur	University of Illinois Urbana Champaign
Infolab at University of Delaware	University of Regensburg
Institut de Recherche en Informatique de Toulouse	University of Udine
Intelligent Data Engineering Lab, UvA	University of Waterloo
JHU HLTCOE SCALE25 gen cascade	University of Waterloo (Clarke)
JHU HLTCOE SCALE25 gen multiagent	Università della Svizzera Italiana
JHU HLTCOE SCALE25 rerank	WHU-NERCMS, Wuhan, China
JHU HLTCOE SCALE25 retrieve	Webis @ Jena, Leipzig, Weimar
Jeonbuk National University	WueRAG
MIT Lincoln Laboratory	

2. Information Retrieval

Information retrieval is concerned with locating information that will help satisfy a user's information need. Traditionally, the emphasis has been on text retrieval: providing access to natural language texts where the set of documents to be searched is large and topically diverse. There is increasing interest, however, in finding appropriate information regardless of the medium that happens to contain that information. Thus "document" can be interpreted as any unit of information such as a tweet, an email message, a medical record, a web page, a video clip, or an academic paper.

The prototypical retrieval task is a researcher doing a literature search in a library. In this environment the retrieval system knows the set of documents to be searched (the library's holdings), but cannot anticipate the particular topic that will be investigated. We call this

an *ad hoc* retrieval task, reflecting the arbitrary subject matter that is the focus of the search and its short duration. A retrieval system's response to an ad hoc search is generally an ordered list of documents sorted such that documents the system believes are more likely to help satisfy the information need are ranked before documents it believes are less likely to satisfy the need.

Information retrieval tasks have proliferated beyond the ad hoc task, to encompass the type of desired response, the objective of the user, the frequency of the search, and the organization of the results. LLMs have given birth to "Retrieval-Augmented Generation", a combination of retrieval, question answering, and summarization. Here, the systems response is generated natural language text, with citations into the documents that are presumably the sources for the output. In the following section, test collections are discussed with respect to the ad hoc task, but the principles and many of the techniques are identical in other tasks.

2.1. Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [1, 2], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test collections generally consist of three parts: a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics (or more generally responses judged to be correct). We call the result of a retrieval system executing a task on a test collection a run.

2.1.1. Documents

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. It is important that the document set reflect the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task. The initial TREC test collections contained 2 to 3 gigabytes of text and 500,000 to 1,000,000 documents. The document sets used in various tracks throughout the years have been smaller and larger than these initial sets depending on the needs of the track and the availability of data, but the general trend has been toward ever-larger document sets to enhance the realism of the evaluation tasks. Similarly, the initial TREC document sets consisted mostly of newspaper or newswire articles, but later document sets have included a much broader spectrum of document types (such as recordings of speech, web pages, scientific documents, blog posts, email messages, video clips, and business documents). Each document is assigned an unique identifier called the DOCNO. For most document sets, high-level structures within a document are tagged using a mark-up

language such as SGML or HTML, or broken into fields of a JSON object. In keeping with the spirit of realism, the text is kept as close to the original as possible.

2.1.2. Topics

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of the criteria that make a document relevant. What is now considered the "standard" format of a TREC topic statement—a topic id, a title, a description, and a narrative—was established in TREC-5 (1996). But topic formats vary in support of the task, and few current TREC tasks use topics in this traditional format.

Participants are (usually) free to use any method they wish to create queries from the topic statements. TREC distinguishes among two major categories of query construction techniques, automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention whatsoever; a manual method is anything else. The definition of manual query construction methods is very broad, ranging from simple tweaks to an automatically derived query, through manual construction of an initial query, to multiple query reformulations based on the document sets retrieved. Since these methods require radically different amounts of (human) effort, care must be taken when comparing manual results to ensure that the runs are truly comparable.

TREC topics are generally constructed specifically for the task they are to be used in. When outside resources such as search engine logs are used as a source of topics the sample selected for inclusion in the test set is vetted to insure there is a reasonable match with the document set (i.e., neither too many nor too few relevant documents). Topics developed at NIST are created by the NIST *assessors*, the set of people hired to both create topics and make relevance judgments. Most of the NIST assessors are retired intelligence analysts. The assessors receive track-specific training by NIST staff for both topic development and relevance assessment.

2.1.3. Relevance judgments

Relevance judgments turn a set of documents and topics into a test collection. Given a set of relevance judgments, the ad hoc retrieval task is then to retrieve all of the relevant documents and none of the irrelevant documents. Most of the traditional measures of retrieval effectiveness treat relevance judgments as binary indicators—either a document is relevant to the topic or it is not—and the judgments themselves are binary in the original TREC collections. Use of evaluation measures that incorporate different levels (or grades) of relevance has become much more prevalent in recent TRECs, and today relevance judgments are generally made on a graded scale to support the use of these measures.

Relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [3]. Furthermore, a set of static relevance judgments makes no provision for the fact that a real user's perception of relevance changes as he or she interacts with the retrieved documents. Despite the idiosyncratic nature of relevance, test collections are useful abstractions because the *comparative* effectiveness of different retrieval methods is stable in the face of changes to the relevance judgments [4].

The relevance judgments in the first retrieval test collections were complete. That is, a relevance decision was made for every document in the collection for every topic. The size of the TREC document sets makes complete judgments infeasible, so by necessity TREC collections are created by judging only a subset of the document collection for each topic and then estimating the effectiveness of retrieval results from the judged sample.

"Pooling" is the technique used in many TREC tracks for selecting the sample of documents for the human assessor to judge [5]. In pooling, the top results from a set of runs are combined to form the pool and only those documents in the pool are judged. Runs are subsequently evaluated assuming that all unpooled (and hence unjudged) documents are not relevant. In more detail, the TREC pooling process proceeds as follows. When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST chooses a number of runs to be merged into the pools, and selects that many runs from each participant respecting the preferred ordering. For each selected run, the top X documents per topic are added to the topics' pools.

The critical factor in pooling is that unjudged documents are assumed to be not relevant when computing traditional evaluation scores such as mean average precision (MAP). This treatment is a direct result of the original premise of pooling: that by taking top-ranked documents from sufficiently many, diverse retrieval runs, the pool will contain the vast majority of the relevant documents in the document set. If this is true, then the resulting relevance judgment sets will be "essentially complete", and the evaluation scores computed using the judgments will be very close to the scores that would have been computed had complete judgments been available.

Various studies have examined the validity of pooling's premise in practice. Harman [6] and Zobel [7] independently showed that early TREC collections in fact had unjudged documents that would have been judged relevant had they been in the pools. But, importantly, the distribution of those "missing" relevant documents was highly skewed by topic (a topic that had lots of known relevant documents had more missing relevant), and uniform across runs. Zobel demonstrated that these "approximately complete" judgments produced by pooling were sufficient to fairly compare retrieval runs. Using the leave-out-uniques (LOU) test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 0.5 %. The maximum increase for

any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

As document sets continue to grow, the proportion of documents contained in standard-sized pools shrinks. At some point, pooling's premise must become invalid. The test collection created in the Robust and HARD tracks in TREC 2005 showed that this point is not at some absolute pool size, but rather when pools are shallow relative to the number of documents in the collection [8]. With shallow pools, the sheer number of documents of a certain type fill up the pools to the exclusion of other types of documents. This produces judgments sets that are biased against runs that retrieve the less popular document type, resulting in an invalid evaluation.

Several TREC tracks have investigated new ways of sampling from very large documents sets to obtain judgment sets that support fair evaluations. The primary goal of the Terabyte track that was part of TRECs 2004–2006 was to investigate new pooling strategies to build reusable, fair collections at a reasonable cost despite collection size. The Million Query track (TRECs 2007–2009) was a successor to the Terabyte track in that it had the same goal, but a different approach. The Common Core track of TREC 2017 and 2018 used multi-arm bandit optimization techniques to select documents to be judged. TRECs since 2019 have experimented with a different approach based on the University of Waterloo's HiCAL [9] system to select the judgment set in the Deep Learning track. Each of these methods reduces the number of relevance judgments made, but can bias the test collection more towards submitted systems, introduce logistical challenges, or both.

2.2. Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, ad hoc tasks that use pooling are evaluated using the `trec_eval` package [10]. This package reports about 85 different numbers for a run, including *recall* and *precision* at various cut-off levels plus single-valued summary measures that are derived from recall and precision. Precision is the proportion of retrieved documents that are relevant (number-retrieved-and-relevant/number-retrieved), while recall is the proportion of relevant documents that are retrieved (number-retrieved-and-relevant/number-relevant). A cut-off level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list. The `trec_eval` program reports the scores as averages over the set of topics where each topic is equally weighted. (An alternative is to weight each relevant document equally and thus give more weight to topics with more relevant documents. Evaluation of retrieval effectiveness historically weights topics equally since all users are assumed to be equally important.)

Precision reaches its maximal value of 1.0 when only relevant documents are retrieved, and recall reaches its maximal value (also 1.0) when all the relevant documents are retrieved. Note, however, that these theoretical maximum values are not obtainable as an average over a set of topics at a single cut-off level because different topics have different

numbers of relevant documents. For example, a topic that has fewer than ten relevant documents will have a precision score at ten documents retrieved less than 1.0 regardless of how the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score at ten documents retrieved less than 1.0. For a single topic, recall and precision at a common cut-off level reflect the same information, namely the number of relevant documents retrieved. At varying cut-off levels, recall and precision tend to be inversely related since retrieving more documents will usually increase recall while degrading precision and vice versa.

Recently the emergence of large language models and highly tuned applications based on them has created a lot of interest in generative information seeking. This could be retrieval-augmented generation, where results from a search are summarized (instead of snippets on search results); question answering in situations where answers could be long or detailed (instead of passage retrieval); iterative conversational search (instead of query reformulation and suggestion); and more.

The current practice in evaluating generative tasks comes from methods used for evaluating complex question answering and summarization [11]. To use summarization as an example, a task situation is established and a set of documents to be summarized forms the topic. The assessor composes a model summary that can be used for automatic metrics like ROUGE [12]. The assessor then proceeds to compose a set of "nuggets". The definition of a nugget varies from task to task but essentially articulates something that the generated summary must contain in order to be a good summary. A nugget could be a fact that needs to be mentioned, or a question that needs to be answered, or entities or relations that need to be identified, etc. Usually nuggets are defined such that the generated text could cover them in a single sentence. The assessor then reviews each sentence in the generated text to identify which if any nuggets are addressed by that sentence. Precision and recall can then be used to score the generated summary.

In a RAG scenario with a modern LLM, there are some novel twists that need to be addressed. Since the documents are not provided up front, but are discovered by the system, we need to determine how the document relates to the generated output. Further, LLMs may generate false or misleading information ("hallucinations"), and these should be identified and impact the score appropriately. Currently, we require systems to cite documents for each generated sentence. The assessor can then determine if the cited document is relevant and if it supports the information in the sentence. Only relevant and supported sentences are aligned to nuggets. Hallucinations would either cite irrelevant documents or documents that do not support the sentence, and so can be marked wrong. We do not currently try to establish whether some other document supports the sentence, or if the information in the sentence is true, so this is a limited level of hallucination capture. The current practice does not measure the quality of the presentation, for example if the information is provided in a usefully structured way; a generated output and an random permutation of the sentences of that output would both currently score the same.

3. TREC 2025 Tracks

TREC's track structure began in TREC-3 (1994). The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups. Table 2 lists the different tracks that were in each TREC, the number of groups that submitted runs to that track, and the total number of groups that participated in each TREC.

This section describes the tasks performed in the TREC 2025 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1. Adhoc Video Search (AVS)

AVS is traditional ad hoc search, but the "documents" are internet videos. The topics are a sentence-length query like "A man is putting on a jacket or a t-shirt". The videos come from the Vimeo Creative Commons dataset and the track provides a master shot boundary reference. Systems return at most 1000 shots for each query, and those shots are pooled and assessed following the traditional TREC protocol.

Eight participating groups submitted 29 runs to the AVS track.

3.2. BioGen

BioGen is retrieval-augmented generation in the biomedical context. Inaccurate generations in the biomedical context could be particularly harmful. The track this year had two tasks. Task A asked systems to ground a provided generated sentence in a PubMed document. Task B ("reference attribution") was to generate answers with citations into PubMed.

Ten participating groups submitted 22 runs to the grounding task and 49 runs to the reference attribution task.

3.3. DRAGUN

DRAGUN stands for Detection, Retrieval, and Augmented Generation for Understanding News. The track started in 2024 as "Lateral Reading". Someone reading a web page or an opinion piece in a newspaper, instead of seeking further information on the topic might engage in what researchers have called "lateral reading": examining source credibility, finding supporting evidence, and cross-referencing other sources through web searches

Table 2: Number of participants per track and total number of distinct participants in each TREC. [1]The DRAGUN track was originally called “Lateral Reading” in 2024. [2] The VTT and AVS tasks ran for many years in TRECVID. [3]PLABA ran in 2023 in TAC.

Track	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19	'20	'21	'22	'23	'24	'25
Ad Hoc	18	24	26	23	28	31	42	41																										
Routing	16	25	25	15	16	21																												
Interactive			3	11	2	9	8	7	6	6	6																							
Spanish			4	10	7																													
Confusion				4	5																													
Merging				3	3																													
Filtering				4	7	10	12	14	15	19	21																							
Chinese					9	12																												
NLP					4	2																												
Speech						13	10	10	3																									
Xlingual						13	9	13	16	10	9																							
High Prec						5	4																											
VLC							7	6																										
Query							2	5	6																									
QA								20	28	36	34	33	28	33	31	28								14	16	5								
Web								17	23	30	23	27	18					26	24	16	12	15	10											
Video										12	19																							
Novelty											13	14	14																					
Genomics												29	33	41	30	25																		
HARD												14	16	16																				
Robust												16	14	17																				
Terabyte													17	19	21																			
Enterprise														23	25	20	16																	
Spam														13	9	12																		
Legal															6	14	15	14	17	11														
Blog															16	24	25	11	16															
Mlln Query																11	7	8																
Feedback																	15	20	7															
Chemical																		8	4	9														
Session																			10	13	10	6	11											
Crowd																				11	8	4												
Medical																				29	24													
Microblog																				58	36	20	23	16										
Contextual																					13	19	17	12	13									
KBA																					11	13	11											
Temporal Summ																						7	6	12										
Federated																						11	12											
Clinical																							26	36	26									
Dynamic Domain																								7	6	3								
Tasks																								5	4	2								
Recall																								10	5									
RTS																									19	18	6							
OpenSearch																									9	3								
CAR																										7	9	7						
Core																										14	11							
Precision Medicine																										32	27	15	16					
CENTRE																											1							
Incident Streams																											11	15	6	5				
News																											7	13	10	8				
CAsT																												21	15	15	10			
Misinfo																												4	9	7	4			
Deep Learning																												16	25	20	14	14		
Fair Ranking																												5	6	4	5			
Podcast																													14	11				
Clinical Trials																														26	11	11		
NeuCLIR																															12	6	6	
CrisisFACTs																															7	10		
AToMiC																																4		
IKAT																																8	8	8
Product																																4	5	4
BioGen																																	5	10
DRAGUN[1]																																	5	
VTT[2]																																	5	
AVS[2]																																	7	8
MedVidQA																																	7	
PLABA[3]																																	10	
RAG																																	25	23
ToT																																	6	9
MLLM																																	25	6
RAGTIME																																	25	20
VQA																																		7
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117	107	95	56	67	75	121	83	60	75	87	74	67	57	66	84	74	48	46	68	73

beyond the web page being read. DRAGUN/Lateral Reading topics are news articles in the ClueWeb22-B-English collection. The first task for systems was to generate questions that readers might ask when evaluating the article's trustworthiness. For the second task, systems searched ClueWeb22-B-English for answers to the generated questions.[13]

Ten participating groups submitted 37 runs to question generation task and 28 runs to the generation task.

3.4. Interactive Knowledge Acquisition (IKAT)

The Interactive Knowledge Assistance Track (IKAT) is the successor to the Conversational Assistance track (CAsT), which ran from 2019 to 2022. The focus of these tracks is on systems that support conversational information seeking. One inspiration for this is personal assistant "smart speakers", but imagine them able to help you with a complex, multistep search. Such systems need to be able to maintain information about the state of the dialogue ("context") to properly interpret the current information need.

Conversations were composed by the coordinators. The conversations could branch and merge at several points. Users were modeled with a "personal text knowledge base", a series of facts about the user that may have been pertinent to the conversation. Different users could follow through similar conversations but have different outcomes. A new task this year allowed participant systems to interact with a simulated user.

Eight groups submitted 35 conversation generation runs (with 9 runs submitting generations without passage ranking) and 23 interactive simulation runs.

3.5. Million LLM

The new Million LLM track is based on the notion that in the future, large language models will be more task- and domain-specific, and consequently there will be a need to select the appropriate model to answer a question.[14]

NIST was asked to create queries that current LLMs did not seem to answer well. We did this using an iterative process, where the assessor composed a question, which was sent to three current-generation models. The assessor then modified their question until they arrived at one that none of the models could generate a satisfactory answer for.

Six groups participated in the MLLM track, submitting 19 runs.

3.6. Product Search

Product Search is TREC's first track in the e-commerce domain. This year there was a search task and a new recommendation task. The documents and queries came from the ESCI Challenge for Improving Product Search [1], a KDD Cup challenge in 2022 administered by

[1] https://amazonkddcup.github.io/

Amazon.[2] The search task was essentially identical to 2024 except that the queries this year were intended to be harder and more ambiguous.

In the recommendation task, systems produced three rankings given a query product: a list of products that complement (go along with) the product, a list of products that could substitute for the product, and a combined list of related products. Products were pooled from all three rankings and judged to be complementary, substitutes, related, or neither. In addition to familiar nDCG scoring, rankings were scored as to the diversity of their products, based on a tree of product categories.

Four groups submitted 21 search runs and 5 recommendation runs.

3.7. RAG

Retrieval-augmented generation continues to be a compelling and popular research topic. The second year of the track continued its three tasks: retrieval, augmented-generation (generation using a provided baseline ranking) and full retrieval-augmented generation. The track also had an experimental relevance judgments task, essentially a baseline reranking task. The document collection is the MS MARCO 2.1 passage collection.

Assessment for RAG was done similarly to RAGTIME: the assessors first reviewed passages for relevance. When the assessor found a relevant documents, they supplied a short sentence to seed the nugget creation process. The assessors then refined the sentences down to a set of nuggets, which were then matched to full generated outputs, yielding a set of nuggets covered by the response.

The RAG track was again the most popular track, with 23 groups submitting 46 retrieval runs, 25 augmented-generation runs, and 51 generation runs.

3.8. RAGTIME

RAGTIME stands for RAG TREC Instrument for Multilingual Evaluation, is the successor to the NeuCLIR track, and focuses on multilingual search and generation. The task is multilingual – systems operate over multiple languages and return a single result based on documents from any or all languages. The search task is traditional ad hoc search. The generation task requires systems to produce "reports" that summarizes critical information, grounded by citations back to the source documents.

The document collection for this task is a combination of Russian, Chinese, Arabic, and English news text from CommonCrawl News. The relevance assessors are all bilingual in English and one of the target languages, and many of them are native speakers in the target language.

[2]The data is Apache-2.0 licensed and so was available to reuse in TREC.

Documents are pooled from both retrieval and generation runs and assessed for each topic in all three languages. This essentially means judging three times the number of topics, and so the assessment process takes quite a while. It also introduces a level of assessor disagreement noise into the relevance judgments, since those three assessors can't match the thinking of each other exactly.[3]

The scale for relevance is not relevant, topical (being in the ballpark), valuable and very valuable. The distinction between the top two levels is stated in terms of the user task: the user is writing a report on the topic, and very valuable items are the citations of the highest importance and usefulness. They would be cited early in the report, perhaps even on the first page. Documents that were merely "valuable" were valid citations, but they were useful more in a supporting role. When documents are labeled valuable or very valuable, the assessor records an English sentence they will later use to create "nuggets".

Following relevance assessment, citations are assessed in their context to determine if they support the sentence they are cited for. This is still a bilingual task. Next, assessors integrate their recorded English sentences into set of "nuggets", which for RAGTIME take the form of short, factual questions that are grouped into higher-level questions the report seeks to answer. Lastly, assessors match sentences in the generated report against the nugget set. These last two tasks are in English and so any assessor on the team can do them.

Twenty groups submitted 55 runs to a dry run generation task, 32 multilingual retrieval runs, and 61 report generation runs.

3.9. Tip-of-the-Tongue Search (ToT)

Tip-of-the-Tongue search is inspired by sites such as irememberthismovie.com and the Reddit board `r/tipofmytongue`. On these sites, people ask about a movie or book or song or place that they don't remember the title of, and don't even completely remember anything about it. Rather, they have an incomplete memory that may even mix up two separate things. People on the site then respond with the book or movie that the poster was trying to remember.

This is a special type of known-item search where the information need is vague, incomplete, and possibly erroneous. Such searches happen in many domains (we've all struggled to find that email…) but as a search query type they are not well studied in the information retrieval community.[15]

The track concept originated from a paper [16] and the accompanying MS-TOT dataset. The NIST assessors created topics using a special elicitation process to yield these special type of queries.

Nine groups submitted 32 runs.

[3] I do not expect to find assessors who individually are fluent in all four languages.

3.10. Video Question Answering

The Video Question Answering track, new for 2025, challenges systems to answer questions about video content. This requires systems that can work with multimodal data. The video collection is a set of short videos (around 30 seconds on average) from YouTube that are Creative-Commons licensed.

The track had two tasks: answer generation, a familiar RAG-style task, and multiple choice, where systems selected an answer from a set of four options.

Seven groups participated in the VQA track, submitting eight multiple-choice runs and 20 answer generation runs.

4. Future

TREC will continue in 2026. The RAG, RAGTIME, Million LLM, and VQA tracks will continue. The Change Detection track, originally intended to start in 2025, will run in 2026. IKAT is ending, but a new track on User Simulation will start. Lastly, a meta-track, AutoJudge, will offer a space for researchers to test automated and semi-automated methods of assessing generated outputs from RAG, RAGTIME, and VQA.

AI-generated content continues to be a major thrust of current IR research as well as a challenge to evaluation. The current evaluation methods, refined from techniques born from question-answering and summarization, are labor intensive and do not create reusable test collections, since new generations can't be matched against ones that have already been judged, as we can do with documents in ranked retrieval. This is a barrier to data-driven research. Fortunately, the IR community is already working on a number of automated and semi-automated approaches, all grounded in an understanding that evaluation requires human understanding and judgment.

Acknowledgments

TREC could not happen without a program committee, track coordinators, participants, and assessors, and we are grateful to them all for the contributions they make to the community. This year we faced an uncertain and challenging calendar, and could not have made it without the combined efforts of the entire TREC community.

References

[1] Cleverdon CW, Mills J, Keen EM (1968) Factors determining the performance of indexing systems. Two volumes, Cranfield, England.

[2] Spärck Jones K (1981) *Information Retrieval Experiment* (Butterworths, London).

[3] Schamber L (1994) Relevance and information behavior. *Annual Review of Information Science and Technology* 29:3–48.

[4] Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management* 36:697–716.

[5] Spärck Jones K, van Rijsbergen CJ (1975) Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.

[6] Harman D (1996) Overview of the fourth Text REtrieval Conference (TREC-4). *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, ed Harman DK, pp 1–23. NIST Special Publication 500-236.

[7] Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, eds Croft WB, Moffat A, van Rijsbergen C, Wilkinson R, Zobel J (ACM Press, New York, Melbourne, Australia), pp 307–314.

[8] Buckley C, Dimmick D, Soboroff I, Voorhees E (2007) Bias and the limits of pooling for large collections. *Information Retrieval* 10:491–508.

[9] Abualsaud M, Ghelani N, Zhang H, Smucker MD, Cormack GV, Grossman MR (2018) A system for efficient high-recall retrieval. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* SIGIR '18, pp 1317–1320.

[10] Buckley C, et al. trec_eval IR evaluation package.. Available from https://github.com/usnistgov/trec_eval.git.

[11] Dang HT, Lin J (2007) Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, eds Zaenen A, van den Bosch A (Association for Computational Linguistics, Prague, Czech Republic), pp 768–775. Available at https://aclanthology.org/P07-1097/.

[12] Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (Association for Computational Linguistics, Barcelona, Spain), pp 74–81. Available at https://aclanthology.org/W04-1013/.

[13] Zhang D, Smucker MD, Clarke CLA (2025) Overview of the TREC 2025 DRAGUN track: Detection, retrieval, and augmented generation for understanding news (notebook). *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*, eds Soboroff I, Awad G.

[14] Kanoulas E, Eustratiadis P, Callan J, Sanderson M, Li Y, Qiao J, Poerwawinata G, Pal V (2025) Overview of the TREC 2025 million large language models track (notebook).

Proceedings of the 34th Text REtrieval Conference (TREC 2025), eds Soboroff I, Awad G.

[15] Arguello J, Diaz F, Fröebe M, Kim TE, Mitra B (2025) Overview of the trec 2025 tip-of-the-tongue track. *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*, eds Soboroff I, Awad G.

[16] Arguello J, Ferguson A, Fine E, Mitra B, Zamani H, Diaz F (2021) Tip of the tongue known-item retrieval: A case study in movie identification. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp 5–14.