Report on the 18th Round of NII Testbeds and Community for Information Access Research (NTCIR-18)

Chung-Chi Chen
National Institute of Advanced Industrial Science and Technology
Japan
c.c.chen@acm.org

Qingyao Ai
Tsinghua University
China
aiqy@tsinghua.edu.cn

Shoko Wakamiya
Nara Institute of Science and Technology
Japan
wakamiya@is.naist.jp

Makoto P. Kato
University of Tsukuba
Japan
mpkato@acm.org

Yiqun Liu
Tsinghua University
China
yiqunliu@tsinghua.edu.cn

Charles L.A. Clarke
University of Waterloo
Canada
claclark@gmail.com

Noriko Kando
National Institute of Informatics
Japan
kando@nii.ac.jp

Abstract

This event report summarizes the eighteenth round of the NII Testbeds and Community for Information Access Research (NTCIR-18), held on June 10–13, 2025 in Tokyo, Japan. NTCIR-18 organized seven core tasks (AEOLLM, FairWeb-2, FinArg-2, Lifelog-6, MedNLP-CHAT, RadNLP, Transfer-2) and three pilot tasks (HIDDEN-RAD, SUSHI, U4), spanning evaluation of generative LLMs, fair ranking, temporal reasoning in finance, multimodal lifelog retrieval, safety assessment for medical dialogue, bilingual radiology staging, resource transfer for dense retrieval, causal explanation in radiology, search over archival metadata, and table-centric QA over annual reports. Across 178 registrations from 113 teams worldwide, participants submitted runs and analyses that combined traditional IR pipelines with LLM-centric methods. This report outlines each task's motivation, data, and methodology, and summarize key findings, including the complementary roles of LLM-based and feature-based evaluators, trade-offs and mitigations in fairness-aware ranking, the importance of structure-aware approaches for tables, and the persistent challenges of sparse metadata and clinical reasoning.

Date: 10–13 June 2025.

Website: https://research.nii.ac.jp/ntcir/ntcir-18/.

1 Introduction to NTCIR

The NTCIR (NII Testbeds and Community for Information Access Research) is a long-running series of open evaluation campaigns for information access technologies, organized by Japan's National Institute of Informatics (NII) since 1999. It was initiated as an Asian counterpart to the U.S. Text REtrieval Conference (TREC), with a focus on East Asian languages and cross-lingual information retrieval (IR) tasks [Kando, 2006]. Over the past two decades, NTCIR has grown into a major international forum for developing and benchmarking a broad range of information access techniques (including IR, question answering, text summarization, etc.) using shared test collections and standardized evaluation methodologies. A core goal of NTCIR has been to create reusable benchmark datasets and evaluation frameworks beyond English, enabling fair comparison of approaches across different languages and domains [Sakai et al., 2021].

According to recent statistics, more than 5,000 research groups worldwide have utilized NTCIR test collections [Sakai et al., 2021], illustrating its wide impact. By providing common datasets, well-defined tasks, and relevance judgments, NTCIR enables researchers to develop and compare systems on an equal footing, driving progress especially in multilingual and cross-lingual settings. The series has also pioneered methodologies such as cross-language evaluation and advanced pooling techniques for collecting human relevance judgments. Over its history, NTCIR has closely mirrored evolving information access challenges and languages: the very first NTCIR conference in 1999 included tasks for Japanese monolingual retrieval and Japanese–English cross-lingual retrieval, attracting participants from multiple countries [Kando, 2006]. Since then, the scope has continuously broadened. By the late 2010s and 2020s, NTCIR tasks encompass emerging areas like fair ranking, large language model evaluation, and complex information extraction. In summary, NTCIR serves as a platform that not only produces valuable resources (datasets and evaluation tools), but also fosters a community dedicated to rigorously assessing new information access techniques, particularly emphasizing languages and challenges under-represented in other initiatives (such as those beyond English-speaking contexts).

2 NTCIR-18 Overview

NTCIR-18 is the eighteenth round of the NTCIR conference series, running from January 2024 to June 2025 and culminating in a conference held on June 10–13, 2025 in Tokyo, Japan [Chen et al., 2025a]. In this cycle, seven *core tasks* and three *pilot tasks* were organized, each targeting a specific problem domain or emerging research question. Table 1 summarizes the ten tasks of NTCIR-18.

NTCIR-18's tasks cover a wide spectrum of modern information access challenges, from the evaluation of generative large language models to domain-specific information extraction. Notably, several tasks align with current trends in AI: for instance, one task focuses on automatic evaluation methods for large language models, and another addresses fairness in search engine results. Other tasks continue NTCIR's tradition of domain-focused evaluations (financial text analysis, medical and radiological text processing, personal lifelog data retrieval) and push into new areas such as searching in undigitized archives and structured data question answering.

The NTCIR-18 process began with task proposal and selection in early 2024, followed by a kick-off event in March 2024 [Chen et al., 2025a]. Task organizers prepared datasets (released around

Task (Type)	Brief Description
AEOLLM (Core)	Automatic Evaluation of LLMs on generative tasks; encourages reference-free metrics for summaries, QA, expansions, dialogues.
FairWeb-2 (Core)	Fair Web Search; return relevant yet demographically group-fair search results or responses for topics (researchers, movies, videos).
FinArg-2 (Core)	Financial Argument Mining (2nd round); focus on temporal inference in financial opinions (analyst reports, earnings calls, social media).
Lifelog-6 (Core)	Personal Lifelog Retrieval; multimodal (photo, sensor) lifelog data organization, search across multiple years, and lifelog-based Q&A.
MedNLP-CHAT (Core)	Medical Chatbot Risk Assessment; evaluate chatbot responses in patient-doctor conversations for medical, legal, and ethical risks.
RadNLP (Core)	Radiology NLP; automated lung cancer staging from radiology reports (extended to bilingual English/Japanese data in this round).
Transfer-2 (Core)	Resource Transfer for Dense Retrieval; techniques to transfer models/resources to new dense retrieval tasks (cross-lingual, multimodal, generative retrieval).
HIDDEN-RAD (Pilot)	Hidden Causality in Radiology Reports; generate or identify causal explanations for radiology diagnoses (making AI reasoning explicit).
SUSHI (Pilot)	Searching Unseen Sources for Historical Info; retrieval methods for undigitized archive documents using sparse metadata (folder search, reference detection).
U4 (Pilot)	Unifying and Utilizing Unstructured Data in Financial Reports; information extraction from annual reports, specifically table retrieval and table Q&A.

Table 1. Summary of the NTCIR-18 evaluation tasks (core and pilot).

May 2024) and defined the evaluation methodology for their tasks. Participant teams from around the world then conducted a dry run (mid-2024, optional for practice) and a formal run (late 2024 through early 2025) for each task, submitting system outputs for evaluation. In total, NTCIR-18 attracted 113 participating teams, with 178 total task registrations (many teams participated in multiple tasks). After systems were evaluated by the organizers (using relevance judgments, answer keys, or other criteria as appropriate per task), results were returned to participants by February 2025. The participating teams and organizers then prepared papers documenting their approaches and findings. Finally, at the June 2025 conference, organizers presented overall results and analyses, and participants shared their system descriptions. The official proceedings[1] include an overview paper by the program chairs and individual task overview papers detailing the test collections, evaluation metrics, and summary of results for each task. In the following section, we provide concise summaries of each NTCIR-18 task, including their motivations, data, evaluation methods, and key outcomes. Beyond the task-specific design, NTCIR-18 was also notable as a community event. A total of 113 teams from over twenty countries and regions registered

[1] https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings18/NTCIR/toc_ntcir.html

Figure 1. Scenes from the plenary sessions and panel discussion during the NTCIR-18 conference program.

Figure 2. Workshop and invited-talk scenes at NTCIR-18.

for the ten tasks, covering topics that range from automatic evaluation of large language models and fairness-aware web search to financial and medical NLP, personal lifelog analysis, and search over historical archival materials. This diversity of tasks and participants reflects the continued broadening of the NTCIR community beyond traditional ad-hoc retrieval.

3 NTCIR-18 Conference Program

The NTCIR-18 conference itself was held as a four-day, in-person event at the National Institute of Informatics (NII) in Tokyo, Japan, from June 10 to 13, 2025. Figure 1 illustrates the atmosphere of the plenary sessions and the panel discussion. The workshops and invited talks are shown in Figure 2. Figure 3 shows the lively poster and task sessions.

Workshops on Day 1

The first day featured two co-located workshops. The *FinTech in AI CUP special session* reported on a series of AI competitions held in Taiwan, with a particular emphasis on information access problems such as financial retrieval and analysis. In parallel, the *EVIA 2025* workshop—the 11th in a long-running series on Evaluation of Information Access—brought together researchers to discuss evaluation measures, test collections, and experimental design for information access

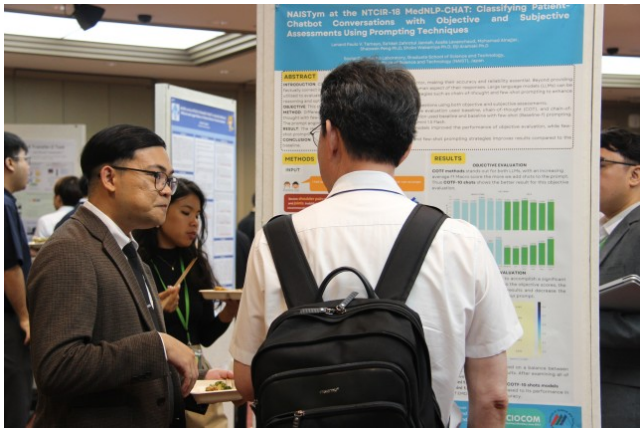

Figure 3. Poster session and task-specific technical sessions at NTCIR-18.

systems. A central theme in EVIA this year was whether and how LLMs should be used in evaluation, for instance as automatic judges or as components in new metrics.

Keynote and Panel on Day 2

The second day began with an overview of NTCIR-18 from the program chairs, followed by a keynote talk by Maarten de Rijke (University of Amsterdam) titled "Measuring the Generative Information Retrieval Universe." His talk surveyed recent work on generative information retrieval and highlighted open challenges in evaluating systems that produce answers or summaries rather than ranked lists of documents. After the keynote, each task organizer delivered a brief overview presentation, summarizing the task design, datasets, evaluation methodology, and headline results for their respective tasks.

In the afternoon, Mark Sanderson (RMIT University) moderated a panel discussion on "LLMs and offline test collections: a dangerous distraction or a vital new tool?" Panelists and audience members debated the opportunities and risks of using LLMs in the evaluation pipeline—for instance, as surrogates for human relevance assessors, as generators of synthetic queries or documents, or as tools for analyzing and explaining system behavior.

Task Sessions and Poster Presentations (Days 3–4)

The third and fourth mornings were devoted to task-specific technical sessions. For each task, selected participating teams gave oral presentations describing their systems, experimental findings, and error analyses, enabling in-depth discussion among groups who had tackled the same problem from different angles. Around lunchtime, a joint poster session was held in which all participating teams were invited to present their work. The poster format facilitated active interaction among participants across different tasks, and made it easy to compare approaches and results informally.

Invited Talks and Closing Session

In the final afternoon, Douglas W. Oard (University of Maryland) delivered a keynote talk titled "Things We Know That Aren't (Always) True," reflecting on common assumptions in information retrieval and cases where those assumptions break down, drawing on his work on searching public-record archives and email collections. The conference also featured invited talks by Ian Soboroff (NIST), who reported on recent developments in the TREC evaluation workshop series, and by Gareth Jones (Dublin City University), who discussed the activities of the MediaEval multimedia evaluation campaign. The closing session concluded with an introduction to the tasks planned for NTCIR-19, providing participants with a preview of the next round in the NTCIR series.

4 Task Summaries

4.1 Automatic Evaluation of LLMs (AEOLLM)

The AEOLLM task was motivated by the need for better automated methods to evaluate large language models. Traditionally, new language models are often assessed either by costly human judgment or by automatic metrics that rely on reference answers (e.g., BLEU for translation, which compares against a given reference text). However, many generative tasks do not have a single correct answer, and most existing automatic evaluation benchmarks (e.g. in QA) focus on multiple-choice formats. AEOLLM specifically tackled open-ended generative tasks and encouraged *reference-free* evaluation metrics [Chen et al., 2025c]. That is, participants were asked to develop evaluation methods that can judge the quality of an LLM's output without necessarily comparing it to a predefined reference. Four diverse subtasks were included to cover different generation scenarios: summarization, non-factoid question answering, text expansion (e.g. elaborating a given prompt), and dialogue generation. The organizers provided datasets and prompts for each subtask (covering various domains and query types), along with baseline outputs from LLMs. Participant systems had to output quality scores or rankings for the LLM-generated answers.

The evaluation methodology for AEOLLM centered on measuring how well each automatic metric or evaluator agreed with human assessments of the LLM outputs. After participants submitted their evaluations (scores/rankings) for the test outputs, the organizers compared these against human judgments to compute correlation and accuracy measures [Chen et al., 2025c]. In total, 4 teams submitted 48 runs (evaluation systems) [Chen et al., 2025c]. The results showed that approaches using LLMs themselves as judges (through prompt-based scoring) could achieve the highest agreement with human preferences, with one submission using a GPT-4-family model for scoring attaining the top performance. At the same time, feature-based methods (e.g., combining transformer-based semantic similarity metrics and classifiers) also performed competitively and had the advantage of interpretability. A notable observation was the trade-off between using powerful LLMs as black-box evaluators versus more transparent metrics: the former tend to align well with human judgment but offer little insight into their reasoning, whereas the latter can provide explainable factors (like content overlap or keyword coverage) but may require careful tuning to reach comparable accuracy [Kim et al., 2025].

4.2 FairWeb-2 (Fair Web Task)

The FairWeb-2 task addressed the growing concern of bias and fairness in information retrieval, particularly in web search results and conversational search responses. Modern search engines aim to retrieve documents relevant to a user's query, but they may inadvertently amplify popularity biases or under-represent certain groups (e.g., minority groups or less mainstream content). FairWeb-2 built upon an initial FairWeb task in the previous NTCIR round, continuing the effort to evaluate systems not just on relevance but also on *group fairness*. The task defined three types of search topics: (R) academic researchers, (M) movies, and (Y) YouTube video content creators [Tao et al., 2025]. For each topic type, a set of sensitive attributes was identified (for example, a researcher's gender or a movie's genre could be considered for fairness grouping). FairWeb-2 was divided into two subtasks: a Web Search subtask, where systems returned a ranked list of results (simulating a search engine results page, SERP) for a given topic query, and a Conversational Search subtask, where systems returned a single synthesized answer or description as if responding in a dialogue. In both cases, the objective was to produce outputs that are both highly relevant to the query and balanced with respect to the protected attributes (ensuring no group is systematically favored or sidelined in the results) [Tao et al., 2025].

The organizers created a test collection with topics in each category (researcher, movie, YouTube) and ground-truth judgments for relevance, along with definitions of fairness criteria for each topic type. Evaluation of submitted runs was thus two-dimensional: using traditional IR measures for relevance (like NDCG or precision) and fairness metrics assessing the distribution of results across groups. One such metric introduced is the GFRC (Group Fairness and Relevance Combined) framework, which combines these aspects in a single evaluation [Tao et al., 2025]. Five teams participated in the Web Search subtask (submitting 23 runs in total), and two teams participated in the conversational subtask (4 runs) alongside several baseline runs by organizers. The official results and analysis indicated that achieving fairness often entails a trade-off with relevance, but some systems managed to improve the exposure of under-represented groups with only a minor drop in relevance scores. For example, certain runs re-ranked results to increase visibility of female researchers or non-blockbuster movies while maintaining acceptable relevance. A key finding was that improving the underlying relevance (by using strong rankers or neural re-ranking) could indirectly improve fairness as well, since more diverse relevant content gets retrieved initially [Tao et al., 2025]. Among techniques employed, participants explored both pre- and post-processing approaches: some methods explicitly adjusted rankings to satisfy demographic proportions, while others modified queries or used fairness-aware loss functions during learning to balance the results.

4.3 FinArg-2 (Temporal Financial Arguments)

FinArg-2 is the second installment of a shared task series on Financial Argument Mining. The initial FinArg-1 task (in a previous NTCIR) [Chen et al., 2023] introduced fundamental tasks of identifying argumentative statements and sentiment in financial texts. FinArg-2 built upon this by focusing on *temporal inference* within financial arguments [Chen et al., 2025b]. In finance-related documents (e.g., stock analyst reports, earnings call transcripts, social media posts about companies), forward-looking statements are common—these are claims or predictions about a company's future performance or conditions. The FinArg-2 task aimed to capture three aspects

of temporal reasoning in such arguments: (1) determining the duration of impact of a premise (e.g., does a cited factor affect the company in the short-term or long-term?), (2) identifying the temporal reference of an argument (e.g., does an opinion refer to next quarter, next year, or some timeframe), and (3) assessing the validity period of a claim (how long the claim holds true). By addressing these, the task moves beyond simple sentiment or stance detection and into understanding when and for how long a financial argument applies.

The data for FinArg-2 comprised the same collection of documents as FinArg-1: a multilingual set of analyst reports, transcripts of earnings calls, and financial microblogs or social media content, all annotated with argument structures. Additional annotation was performed to mark temporal expressions and link arguments with temporal attributes (such as explicit dates or relative time indicators). A total of 20 teams registered interest in FinArg-2, and ultimately 7 teams actively submitted system outputs [Chen et al., 2025b]. The evaluation involved standard text classification and information extraction metrics (like precision, recall, F1) for tasks such as identifying temporal spans or classifying an argument's time horizon. Key challenges observed were the sparsity of explicit temporal cues in some documents and the need for external knowledge (for instance, understanding that "next quarter" refers to a specific timeframe depending on context). The best-performing systems often combined traditional NLP methods (for extracting and normalizing temporal expressions) with transformer-based language models fine-tuned on the argument mining tasks. One observation was that domain-specific pre-training (e.g., financial news corpora) helped in improving accuracy. The results suggested that systems did relatively well on detecting explicit temporal references, but struggled on implicit ones like "in the long run" or when inferring how long an effect might last. FinArg-2's introduction of temporal evaluation is an important step in creating more realistic financial opinion mining systems. By requiring systems to say not just "what is the argument and sentiment" but also "when does this apply," it paves the way for time-aware financial analytics.

4.4 Lifelog-6 (Personal Lifelog Organization and Retrieval)

The Lifelog-6 task continues a series of NTCIR tasks (running since NTCIR-12) dedicated to lifelog data, which are rich multimedia recordings of an individual's daily activities (typically including wearable camera photos, videos, location and biometric sensor data, etc.). The motivation behind lifelog tasks is to develop methods for organizing and retrieving personal data in ways that help users browse or search their digital memories. As people accumulate years of lifelog data, tools are needed to index and find specific events or answer questions about one's past experiences. Lifelog-6 aimed to push the state of the art in *multimodal lifelog retrieval and analytics* [Zhou and Gurrin, 2025]. It provided a testbed with multi-year lifelog archives from multiple users, larger and more diverse than earlier editions. The task had multiple components: participants were asked to perform asynchronous retrieval (find moments matching a query across a long timeline), a question-answering task based on lifelog content (e.g., "Find all days when the user visited a beach and took photos of food"), and to explore novel analytics such as summarizing patterns in the lifelog.

The dataset for Lifelog-6 included annotated lifelog data spanning at least two years per individual, with annotations for events, locations, and semantic content of images. Because privacy is a concern in lifelogs, the data was likely partially obfuscated or carefully shared with participants

under agreements. Systems were evaluated on their ability to retrieve relevant moments or images given textual queries (using metrics like mean average precision or recall@N, adapted to this domain). For the QA-style tasks, accuracy of the answer or overlap with ground-truth sets was measured. Over 100 teams have participated in lifelog challenges since 2015. Many participants leveraged existing multimedia retrieval frameworks and adapted it to index lifelog images with metadata [Rossetto, 2025]. Some teams focused on interactive retrieval interfaces to refine queries. Generally, successful approaches combined visual analysis with temporal context and metadata to improve search accuracy [Tran et al., 2025]. A common pattern was the use of ensemble strategies: combining text-based search on annotations with content-based image retrieval (CBIR) on visual features. The tasks also revealed challenges: purely text-based approaches might miss relevant images if events are not annotated, whereas purely image-based approaches can be misled by visual similarity that is irrelevant to the user's query intent.

4.5 MedNLP-CHAT (Medical Chatbot Dialogue Risk Assessment)

The MedNLP-CHAT task was a timely pilot in evaluating AI chatbots in the medical domain, specifically focusing on the detection of risks in automated patient-doctor conversations. With the rise of medical advisory chatbots, it is crucial to ensure that their responses are not only factually correct but also safe from a medical, legal, and ethical standpoint. The task introduced by NTCIR-18 aimed to create a testbed where chatbot responses to patient queries could be analyzed for potential *medical inaccuracies, legal liabilities, or ethical issues* [Aramaki et al., 2025]. Each question-response pair (where the question simulates a patient inquiry and the response is generated by a chatbot) was annotated for whether it contains any of these three types of risk. For example, giving harmful medical advice would be a medical risk; violating privacy or disclaimers could be a legal risk; and an insensitive or biased remark might be an ethical risk. Importantly, multiple risk categories could apply to a single response.

The data set spanned multiple languages or regions (subtasks were provided for at least Japanese and German dialogues, with translations to English for common evaluation) [Aramaki et al., 2025]. Each team's goal was to develop a system that can automatically classify a chatbot response as containing a medical, legal, ethical risk or none. This was essentially a multi-label classification problem, complicated by class imbalance (genuinely problematic responses are rarer than benign ones). Nine teams participated, applying methods ranging from traditional machine learning to prompt-based large language model techniques. Some systems trained separate classifiers (or separate output layers) for each risk type, while others attempted a unified model. A few teams did not use fine-tuning and instead used prompting strategies with large pre-trained models to see if the model itself can judge the response (e.g., asking GPT-4 "Is there any medical risk in this answer?"). Evaluation metrics included accuracy and F1-score for each risk category, and some continuous scoring measures (like an Earth Mover's Distance to evaluate how well systems predicted the distribution of risk severity when provided) [Aramaki et al., 2025].

Key findings from the results were that medical misinformation was relatively easier to catch (some systems achieved reasonably high precision on medical risk detection), whereas ethical and legal risks proved more challenging. This is likely because medical facts can be verified against known knowledge (especially with external medical databases or by using medically fine-tuned models), but ethical judgment is more subjective and context-dependent, and legal acceptability

varies by jurisdiction. One interesting approach combined the strengths of different models: for instance, one team found that using a knowledge-enhanced ChatGPT-based model excelled at identifying medical issues, while a BERT-based classifier was better for legal/ethical cues, leading them to build a hybrid ensemble that improved recall [Ohara et al., 2025]. Another observation was the benefit of *few-shot prompting*: techniques like chain-of-thought prompts and providing a few examples improved an LLM's consistency in flagging risks [Tamayo et al., 2025; Van Supranes et al., 2025].

4.6 RadNLP (Radiology Report NLP for Lung Cancer Staging)

The RadNLP task focused on a very specific yet clinically significant challenge: automatically determining the stage of lung cancer from free-text radiology reports. In oncology, "staging" refers to assessing the extent of cancer (often denoted by the TNM system: T for tumor size/local extent, N for lymph node involvement, M for metastasis). Radiologists often describe findings in narrative form in their reports, and clinicians must interpret these to assign a TNM stage (e.g., T2 N1 M0). Automating this process can assist in decision support and save time. In NTCIR-17, a task called RR-TNM addressed this for English radiology reports; RadNLP at NTCIR-18 expanded it to a bilingual setting (English and Japanese reports) [Nakamura et al., 2025]. This added complexity, as systems had to potentially handle two languages and possibly align staging criteria across them.

The dataset consisted of de-identified radiology reports of lung cancer patients, labeled with the "gold standard" TNM stage for each case, presumably provided by experts or derived from structured records. For NTCIR-18, an equivalent set of Japanese radiology reports with staging labels was also included. Participants could focus on one language or both. The task had subtasks for classification: predicting the full stage (combination of T, N, M categories), and possibly separate tasks for each component (some teams reported separate accuracy for T, N, and M). A total of around 15 teams participated, making RadNLP one of the more popular tasks in this round. Various approaches were tried.

The evaluation metric was primarily accuracy (exact match of predicted stage to the true stage) and possibly some softer measures (partial credit if T, N, or M is correct individually). Achieving high accuracy is challenging because an error in any component leads to an incorrect overall stage. The top systems in RadNLP attained around 70–80% accuracy on test data for English or Japanese, with performance often better for certain components (for instance, many systems found N and M easier to classify than the precise T category) [Yamagishi et al., 2025]. One notable trend was that combining strategies improved results: for example, one team used an ensemble of an LLM with a rule-based system, thereby capturing both the nuanced context and specific key words. Another team applied a two-stage reasoning approach: first generating intermediate conclusions (like "is there metastasis? yes/no") and then combining them to produce the stage, which helped break down the complex task [Nakamura et al., 2025]. Also, domain adaptation efforts, such as customizing tokenizers for medical vocabulary or pre-training on radiology text, yielded gains in performance for some participants. Comparing the English and Japanese tracks, it was observed that language differences (such as how lung anatomy and staging are described) required careful handling; direct translation approaches did not perform as well as having dedicated models for each language.

4.7 Transfer-2 (Resource Transfer for Dense Retrieval)

The Transfer-2 task was a novel addition in NTCIR-18, aiming to explore how resources (data or models) developed for one scenario can be *transferred* to improve performance in another scenario, within the context of *dense retrieval*. Dense retrieval refers to using vector embeddings (often from neural networks) to retrieve documents, as opposed to traditional sparse keyword matching. The idea of Transfer-2 was to leverage existing resources—such as trained models, labeled data, or knowledge from one domain—to benefit another task or domain, especially when the target has limited data. In particular, Transfer-2 targeted Japanese information access problems and was provisionally structured into three subtasks: Dense Cross-Language Retrieval (DCLR), Dense Multimodal Retrieval (DMR), and Retrieval-augmented Generation (RAG) [Joho et al., 2025]. DCLR would involve using resources from one language (like English) to aid Japanese search or vice versa; DMR focused on combining text with other modalities (e.g., sensor data or images in a search scenario, implying possibly a lifelog or IoT context); and RAG examined using retrieval to aid answer generation (for example, using a Japanese QA system boosted by external documents retrieved).

In practice, participants gravitated to the DMR and RAG subtasks. The DMR subtask provided a dataset of queries with both textual and numeric sensor-related context and asked systems to retrieve relevant items considering both types of data. The RAG subtask gave questions where the answer needed to be generated from a knowledge base, requiring systems to first retrieve relevant passages (using dense retrieval) and then produce a written answer. Only a small number of teams (two or three) participated, making this a smaller pilot effort [Joho et al., 2025]. One team (ditlab) [Tachioka and Terao, 2025] tackled RAG by proposing a late fusion approach: they retrieved multiple relevant passages and then combined them in the answer generation process, rather than feeding them serially to the LLM, which improved answer quality. The same team in DMR used a specialized encoder for sensor data alongside a vision encoder for images to handle queries that involve, say, finding moments with certain physical readings and visual content. Another team (YMX2L) in DMR applied data augmentation and object detection on images to better integrate the visual information with textual queries, and reported improvements with these techniques. [Mizuguchi et al., 2025]

Evaluation in Transfer-2 depended on each subtask: for RAG, measures included answer accuracy or ROUGE (since answers were textual, some had reference answers), while for DMR, ranking metrics like precision at rank cutoff or recall were used. Because the tasks were pilots, the results were analyzed qualitatively as much as quantitatively. The outcomes indicated that transferring knowledge across modalities or languages is promising but requires careful alignment. For cross-language retrieval, while not many submissions were made, it was noted that multilingual embeddings can serve as a starting point. For multimodal retrieval, simple concatenation of modalities was less effective than architectures explicitly handling each modality's features. For retrieval-augmented generation, using multiple retrieved sources and fusing their information led to more accurate answers than relying on a single best document.

4.8 HIDDEN-RAD (Hidden Causality in Radiology Reports)

HIDDEN-RAD was a pilot task introduced to encourage more explainable and reasoning-oriented outputs from medical AI systems. In conventional radiology reports, radiologists typically describe

findings and give a conclusion (diagnosis) but do not explicitly spell out every causal reasoning step that led from finding to conclusion. For instance, a report might conclude "likely pneumonia" after describing "infiltrates in the right lower lobe," but it may not explicitly say "because infiltrates in the lung suggest pneumonia." The goal of HIDDEN-RAD was to have systems generate those missing causal explanations or to otherwise include the causal reasoning that is usually implicit. Two subtasks were defined [Choi and Cho, 2025]: Task 1 was *diagnostic explanation generation* given a radiology report (and optionally the associated chest X-ray image), where the system should produce a textual explanation linking findings to diagnosis. Task 2 was to evaluate a system's capability to interpret diagnostic reasoning from structured inputs, possibly a questionnaire or form that captures clinical facts, and output a reasoning chain or verification.

The dataset for HIDDEN-RAD was built on an enriched subset of the MIMIC-CXR corpus, a public dataset of chest X-ray images and reports, augmented with annotations for causal links or reasoning steps. For Task 1, systems could use the report text and even image data to produce explanations; for Task 2, a structured representation (like key findings in a template) was provided. Evaluation was challenging because it involved free-text outputs. Organizers used both automatic metrics (like BLEU/ROUGE comparing to reference explanations) and expert judgment to assess the quality of the causal explanations [Choi and Cho, 2025]. Quantitatively, they also measured to what extent important causal factors (e.g., mentioning that a certain symptom leads to a conclusion) were included. Three teams participated in Task 1 (submitting a total of 40 runs) and two teams in Task 2 (16 runs) [Choi and Cho, 2025]. The top systems achieved about 69% and 79% of the maximum achievable scores in Task 1 and Task 2 respectively, indicating room for improvement but also a strong baseline for this novel challenge.

Participants in HIDDEN-RAD explored diverse approaches combining recent advances in prompting and retrieval-augmented generation. Teddysum applied a chain-of-thought prompting strategy using their fine-tuned LLM Blossom, enriched with a medical knowledge graph (RadGraph) to capture typical finding–diagnosis relations, which helped them achieve top performance in the structured explanation task [Won et al., 2025]. RADPHI3 fine-tuned a smaller, radiology-specialized model (Rad-Phi-3.5-Vision-CXR) and compared it against GPT-4, finding that domain-specific tuning improved terminology precision and conciseness while maintaining competitive overall quality [Ranjit et al., 2025]. Nash combined retrieval of similar past cases with a reasoning-oriented prompting framework, using analogical cues (e.g., "in that case X was the reason, so here Y") to guide causal inference; this approach ranked first in one subtask and second in the other [Cho et al., 2025]. Collectively, the results demonstrate that structured reasoning (e.g., chain-of-thought) and integration of domain knowledge (through knowledge graphs or retrieval) substantially enhance the quality of causal explanations in medical report generation, advancing practical XAI for healthcare.

4.9 SUSHI (Searching Unseen Sources for Historical Information)

The SUSHI pilot task tackled an often overlooked problem in information access: how to search for information that exists in nondigital or not-directly-searchable forms, such as physical archives or collections with only minimal metadata. In many historical archives, the documents (letters, records, etc.) may not be fully digitized or OCRed, but there might be catalog entries or folder-level descriptions. SUSHI set out to develop and evaluate methods to find relevant information

from such "unseen" sources by leveraging whatever metadata or partial information is available [Suzuki et al., 2025]. The task had two subtasks. The first was *Folder Search*: given a textual query about a historical fact or event, retrieve the archive folder(s) that are likely to contain relevant documents (assuming each folder groups documents by topic or origin, and has a brief description). The second was *Archival Reference Detection*: essentially a subtask to identify specific references or pointers within archival metadata that pertain to the query (for example, identifying which collection or index entry could lead a historian to the desired info).

To support this, the organizers constructed a test collection using real archival metadata. This included hierarchical data: collections divided into boxes, boxes into folders, and possibly folder into items, each with titles or short notes. Only the metadata (titles, descriptions) was searchable text, while the actual content was assumed inaccessible (thus "unseen"). Relevance judgments were made by domain experts mapping queries to the correct folders or reference entries. Three teams participated. The evaluation metrics were similar to standard ad-hoc retrieval (precision, recall, MAP) but applied to the ranked lists of folders or references. One particular difficulty was the small amount of text in each folder description, which makes it hard for retrieval algorithms to discriminate.

Participants explored techniques to address the data sparsity. One team (KASYS) aggregated metadata across levels: for example, they concatenated item-level titles up to the folder level, effectively enriching the folder representation with all the terms from its contents [Fujimaki and Kato, 2025]. They compared this multi-level metadata aggregation using both BM25 (a classical IR model) and modern dense retrieval embeddings. Interestingly, their results showed that the hierarchical approach did not significantly outperform a simpler baseline of just using the folder descriptions with BM25 [Suzuki et al., 2025]. This implies that the noise from concatenating many items might have hurt precision. Another team (University of Maryland) provided analysis on both subtasks and even generated additional data for training, indicating that carefully engineered heuristics can slightly improve recall but systematic gains were elusive with the current approaches. [Oard et al., 2025] also analyzed results and found that current IR techniques struggle when faced with extremely sparse and specialized metadata; for some queries, no team could retrieve the correct folder in the top ranks without some form of manual tuning.

4.10 U4 (Unstructured Data Utilization in Financial Reports)

The U4 task dealt with information extraction and question answering from a very common but complex source: annual financial reports of companies. Annual reports contain a wealth of quantitative and qualitative information, often presented in tables embedded in text. For an AI system, extracting specific facts or answering questions from these reports requires handling both natural language and structured tables, which can be difficult because tables are not easily parsed by standard text models. The U4 task aimed to improve methods for reading and using such semi-structured documents by dividing the problem into two subtasks [Kimura et al., 2025]: (1) *Table Retrieval*, where given a query (for example, "What is the total revenue of Company X in 2021?"), the system must find the most relevant table from a collection of reports that contains the answer; and (2) *Table Question Answering*, where after identifying the relevant table, the system must extract or generate the precise answer (in the example, the revenue figure).

The dataset for U4 consisted of annual securities reports from TOPIX 100 companies (major Japanese companies), which are lengthy documents with many financial tables. The organizers provided a set of questions along with ground truth answers and the corresponding table that contains each answer [Kimura et al., 2025]. For the Table Retrieval subtask, systems would output a ranked list of tables (each table having an identifier); success was measured by metrics like precision at 1 (did you get the right table at the top) and MRR. For the Table QA subtask, systems could either output a text answer or a cell from a table. The outputs were evaluated for exact match accuracy or numeric error tolerance as appropriate. A leaderboard-style evaluation was used, and participants uploaded JSON-formatted results for automatic scoring, ensuring consistency and reproducibility [Kimura et al., 2025]. Ten teams actively participated in U4, reflecting significant interest, and together they submitted 210 runs (including development phase submissions).

The approaches varied widely. Some teams treated it as an end-to-end QA problem by fine-tuning large language models with the reports: essentially feeding the model a serialized form of the table plus surrounding text. However, it turned out that off-the-shelf LLMs often struggle to understand table structure when given as plain text, leading to errors. More successful were pipeline approaches that explicitly handle table structure. For example, the WhiteME team implemented a hybrid retrieval mechanism: they first used a combination of term matching and language model embeddings to retrieve candidate tables, and then for QA, they identified which row and column in the selected table likely held the answer, using similarities between the question and table headers [Tanaka et al., 2025]. Their method achieved a high accuracy (around 74.6% correct answers), outperforming a GPT-4 based approach that achieved about 64% on the same data [Kimura et al., 2025]. This demonstrated that a tailored solution using table structure could beat a generic LLM on this task. Other teams similarly reported that leveraging the table's format (e.g., identifying row/column by keywords) plus using BERT-like models for cell selection worked well. In the retrieval subtask, fine-tuning a model specifically for table relevance (some used a model pre-trained on web tables) improved the chances of picking the correct table. [Takasago and Akiba, 2025] used a straightforward approach of splitting each report's tables into separate documents and indexing them with Lucene; interestingly, this classical IR approach was a strong baseline, showing that if the query and table share obvious terms, traditional retrieval suffices. The harder questions required some semantic understanding (e.g., knowing that "revenue" might be labeled as "Net Sales" in the table).

5 Conclusion

Across all the tasks in NTCIR-18, a common thread is the emphasis on evaluating systems in realistic, complex scenarios that go beyond traditional ad-hoc document retrieval. First, the influence of large language models (LLMs) is evident throughout: the AEOLLM task directly evaluated methods for scoring LLM outputs, while many participant systems in other tasks (FairWeb-2, MedNLP-CHAT, RadNLP, etc.) made use of LLMs either as components (for reranking, prompting, or generating answers) or as baselines to beat. This reflects a broader shift in the information access field where learning-based and generative models are central. NTCIR-18 provided a controlled way to measure how well these models perform and can be evaluated, highlighting both their potential and their limitations (for example, LLMs can judge content fairly well, but need careful prompting and lack transparency, as seen in AEOLLM).

Second, the tasks underscored the importance of domain-specific evaluation. By setting up dedicated challenges for financial text, medical dialogues, radiology reports, personal lifelogs, and so on, NTCIR-18 helped ensure that advances in these niches are systematically measured. In FinArg-2 and U4, for instance, unique metrics and test collections were developed that account for temporal reasoning and table-structured data respectively, which would not be captured by generic IR evaluations. The outcome is a richer understanding of how to handle these formats: e.g., that temporal cues require combining NLP with time normalization, or that table QA is best served by a hybrid of IR and structured parsing techniques. Additionally, tasks like MedNLP-CHAT and FairWeb-2 introduced multi-faceted evaluation criteria (accuracy along with fairness or risk metrics), moving evaluation beyond single-number metrics. This multifactor evaluation is crucial as AI systems are deployed in sensitive contexts; it demonstrates that we can, and should, quantify aspects like fairness and safety in tandem with raw performance.

Another notable pattern is the continuous interplay between human expertise and automation in these tasks. Many tasks still rely on human annotations and judgments (for relevance, for risk labeling, for ground-truth answers), and the quality of those directly impacts the evaluation. Through the conference, best practices in pooling, judging, and inter-annotator agreement were likely discussed. The shared experience of NTCIR-18 reinforced known challenges—for example, SUSHI revealed that without extensive human metadata, searching physical archives remains extremely difficult for machines. Meanwhile, HIDDEN-RAD's results suggested that even when humans provide some causal annotations, generating full explanations is complex and requires new strategies.

NTCIR-18's contributions also include the tangible resources created: new datasets for each task have been released to the community, from multilingual financial argument corpora to annotated medical conversations, bilingual radiology report corpora, and QA pairs on financial tables. These will enable further research beyond the conference itself. Moreover, the methodologies that emerged—such as fairness-aware reranking algorithms, prompt engineering for risk assessment, retrieval-augmented generation pipelines—serve as reference points for practitioners tackling similar problems. The conference facilitated comparisons of these approaches under fair conditions, so that clear progress (or sometimes lack thereof) was identified. For example, the top systems in RadNLP showed only incremental gains over earlier results, suggesting that lung cancer staging by text remains challenging and might need new ideas (like joint vision-text models or larger specialized pretraining), whereas in U4, a substantial jump was achieved by a novel hybrid approach.

Looking ahead, the NTCIR series will continue with NTCIR-19,[2] which is scheduled to run until December 2026. As in previous rounds, NTCIR-19 offers a broad set of tasks spanning information retrieval, natural language processing, and multimedia applications. Researchers and practitioners interested in these areas are warmly encouraged to consult the NTCIR web site, consider participating in existing tasks, and even propose new tasks in future rounds. In this way, the community can continue to use NTCIR as a shared platform for building reusable test collections, evaluating emerging techniques such as LLM-based systems, and fostering collaboration across domains and regions.

[2] https://research.nii.ac.jp/ntcir/ntcir-19/

References

Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, Shohei Hisada, Tomohiro Nishiyama, Lenard Paulo Tamayo, Jingnan Xiao, Axalia Levenchaud, Pierre Zweigenbaum, Christoph Otto, Jerycho Pasniczek, Philippe Thomas, Nathan Pohl, Wiebke Duettmann, Lisa Raithel, and Roland Roller. Ntcir-18 mednlp-chat determining medical, ethical and legal risks in patient-doctor conversations: Task overview. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-17 finarg-1 task: Fine-grained argument understanding in financial analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*, pages 12–15, 2023.

Chung-Chi Chen, Qingyao Ai, and Shoko Wakamiya. Overview of ntcir-18. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025a.

Chung-Chi Chen, Chin-Yi Lin, Cheng-Chih Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-18 finarg-2 task: Temporal inference of financial arguments. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025b.

Junjie Chen, Haitao Li, Zhumin Chu, Yiqun Liu, and Qingyao Ai. Overview of the ntcir-18 automatic evaluation of llms (aeollm) task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025c.

Ju-Min Cho, Ho-Jin Yi, Myung-Kyu Kim, Se-Jin Jeong, and Seung-Hoon Na. Optimizing causality-based radiology reporting with retrieval-augmented and structured reasoning approaches for the ntcir-18 hidden-rad task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Key-Sun Choi and You-Sang Cho. Overview of the ntcir-18 hidden-rad task: Hidden causality inclusion in radiology report generation. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Haruki Fujimaki and Makoto P Kato. Kasys at the ntcir-18 sushi task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Hideo Joho, Atsushi Keyaki, Yuuki Tachioka, and Shuhei Yamamoto. Overview of the ntcir-18 transfer-2 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Noriko Kando. Evaluation of information access technologies with asian languages at ntcir workshop. 2006.

Yumi Kim, Meen Chul Kim, and Jongwook Lee. Knuir at the ntcir-18 aeollm: Automatic evaluation of llms. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Yasutomo Kimura, Sato Eisaku, Kazuma Kadowaki, and Hokuto Ototake. Overview of the ntcir-18 u4 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Riku Mizuguchi, Takeshi Yamazaki, and Shuhei Yamamoto. Ymx2l at the ntcir-18 transfer-2 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Yuta Nakamura, Koji Fujimoto, Eiji Aramaki, Shouhei Hanaoka, and Shuntaro Yada. Ntcir-18 radnlp 2024 overview: Dataset and solutions for automated lung cancer staging. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Douglas W Oard, Shashank Bhardwaj, and Emi Ishita. Biting into sushi: The university of maryland at ntcir-18. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Aoi Ohara, Nanami Murata, Ami Yuge, and Rei Noguchi. Tusnlp at the ntcir-18 mednlp-chat task: Utilization of external medical knowledge and hybrid approach of bert and chatgpt. In *Proceedings of the NTCIR-18 Conference*, 2025.

Mercy Ranjit, Rahul Kumar, Shaury Srivastav, Anirban Porya, and Tanuja Ganu. Rad-phi3 at the ntcir-18 hidden-rad: Hidden causality inclusion in radiology reports with multimodal small language models. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Luca Rossetto. vitrivr-engine at the ntcir-18 lifelog-6 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Tetsuya Sakai, Douglas W Oard, and Noriko Kando. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer Nature, 2021.

Tokinori Suzuki, Douglas Oard, Shashank Bhardwaj, Emi Ishita, and Yoichi Tomiura. Ntcir-18 sushi pilot task overview. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Yuuki Tachioka and Yasunori Terao. ditlab at the ntcir-18 transfer-2 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

So Takasago and Tomoyoshi Akiba. Akbl at ntcir-18 u4 tableretrieval and tableqa. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Lenard Paulo V Tamayo, Sa'idah Zahrotul Jannah, Axalia Levenchaud, Mohamad Alnajjar, Shaowen Peng, Shoko Wakamiya, and Eiji Aramaki. Naistym at the ntcir-18 mednlp-chat: Classifying patient-chatbot conversations with objective and subjective assessments using prompting techniques. In *Proceedings of the NTCIR-18 Conference*, 2025.

Koji Tanaka, Daiki Shirafuji, and Tatsuhiko Saito. Whiteme at u4 shared task: Hybrid retrieval with table-structured clues for economic table qa. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Sijie Tao, Tetsuya Sakai, Junjie Wang, Hanpei Fang, Yuxiang Zhang, Haitao Li, Yiteng Tu, Nuo Chen, and Maria Maistro. Overview of the ntcir-18 fairweb-2 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Quang-Linh Tran, Binh T Nguyen, Gareth JF Jones, and Cathal Gurrin. Dcu memoriease at the ntcir-18 lifelog 6 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Michael Van Supranes, Martin Augustine Borlongan, Joseph Ryan Lansangan, Genelyn Ma Sarte, Shaowen Peng, Shoko Wakamiya, and Eiji Aramaki. Upxsocio at ntcir-18 mednlp-chat task: Similarity-based few-shot example selection for prompt-based detection. In *Proceedings of the NTCIR-18 Conference*, 2025.

Youngseob Won, Younggyun Hahm, Chanhyuk Yoon, and Seong Tae Kim. Teddysum at the ntcir-18 hidden-rad task: Using rag and tree-of-thought for causal explanation generation. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Yosuke Yamagishi, Ryosuke Tomiyama, and Yui Ueda. Uty at the ntcir-18 radnlp 2024 task: Possibilities and limitations of a hybrid rule-based and llm approach for lung cancer tnm classification. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.

Liting Zhou and Cathal Gurrin. Overview of the ntcir-18 lifelog-6 task. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*, 2025.