RETECO: A SemEval-2027 Shared Task on Reasoning-Oriented Retrieval

Abdelrahman Abdallah[1], **Mohammed Ali**[1], **Muhammad Abdul-Mageed**[2], **Kevin Duh**[3], **Adam Jatowt**[1]
[1]University of Innsbruck [2]University of British Columbia [3]Johns Hopkins University
{abdelrahman.abdallah,mohammed.ali,adam.jatowt}@uibk.ac.at
muhammad.mageed@ubc.ca, kevinduh@cs.jhu.edu

1 Overview

Standard retrieval benchmarks such as BEIR (Thakur et al., 2021) assume relevance follows from topical similarity. Yet many real-world queries are not fact lookups: they require multi-step inference, contextual interpretation, or reasoning over relationships never stated explicitly, making surface matching insufficient (Su et al., 2024; Abdallah et al., 2026c). **RETECO** (Reasoning-Oriented Retrieval with Temporal & Conversational Context) is a new SemEval shared task on passage retrieval where relevance depends on **semantic interpretation** of either temporal constraints or multi-turn conversational context (Campos et al., 2014a; Abdallah et al., 2026b; Joho et al., 2014; Anantha et al., 2021; Aliannejadi et al., 2024; Katsis et al., 2025; Su et al., 2024; Abdallah et al., 2026d,e; Gruber et al., 2025). Existing benchmarks typically isolate these phenomena—temporal datasets often reduce the problem to date lookup, while conversational datasets reward shallow similarity. RETECO instead evaluates **semantic grounding for retrieval** in two settings central to retrieval-augmented generation (RAG), conversational search, and evidence-based QA. We propose two complementary tracks:

- **Track 1: Temporal Grounded Retrieval.** Given a temporally grounded query (e.g., "How has privacy law evolved *after* GDPR?") and a domain corpus, systems retrieve passages that are both topically relevant *and* temporally aligned with the information need, spanning the required time periods.
- **Track 2: Reasoning-Intensive Conversational Retrieval.** Given a multi-turn conversation, systems retrieve passages for the current turn under two simultaneous challenges: resolving discourse dependencies (anaphora, ellipsis, references to prior turns) and identifying passages whose relevance requires multi-step inference rather than surface overlap.

The task targets the IR, RAG, conversational search, and temporal NLP communities, providing reusable data, scorers, baselines, and a public leaderboard. Figure 1 illustrates one example per track.

2 Task Definition and Data

RETECO is grounded in two existing pilot benchmarks already created and publicly released by our team, which substantially reduces task risk. The Track 1 pilot is **TEMPO** (Abdallah et al., 2026a), a temporal reasoning-intensive retrieval benchmark with 1,730 queries over a corpus of 1,654,055 documents across 13 domains. The Track 2 pilot is **RECOR** (Ali et al., 2026), a reasoning-focused multi-turn conversational retrieval benchmark with 707 conversations and 2,971 turns across 11 domains. Both pilots are in English and already provide public data releases (used here as training/development data), baseline systems, and evaluation scripts.

Track 1: Temporal Grounded Retrieval. Track 1 is grounded in the TEMPO benchmark (Abdallah et al., 2026a) and comprises two subtasks:

Sub-Track 1a – Temporal Retrieval (Query → Documents): Given a temporally grounded query and a domain corpus, systems retrieve and rank documents that are both topically relevant *and* temporally aligned with the required time periods. Evaluation uses nDCG@10 as the primary metric alongside four temporal-specific metrics: Temporal Precision@10 (temporally relevant documents ranked early), Temporal Relevance@10 (fraction of the top-10 judged temporally relevant), Temporal Coverage@10 (fraction of required time periods covered), and NDCG|FC@10 (NDCG restricted to queries achieving full temporal coverage).

Sub-Track 1b – Step-wise Temporal Retrieval (Query → Step → Documents): Given a query

and its decomposed step-wise retrieval plan, systems retrieve documents for each individual step (e.g., "Retrieve baseline statistics from 2013–2017"; "Retrieve current statistics from 2020–present"). Evaluation is performed at the step level using step-specific gold documents. Three query strategies are permitted: *Step-Only* (each step as the sole query), *Query+Step* (the original query concatenated with each step), and *Query+All* (the original query with all steps combined).

Track 2: Reasoning-Intensive Conversational Retrieval. Track 2 is grounded in the RECOR benchmark (Ali et al., 2026) and comprises three subtasks:

Sub-Track 2a – Conversational Retrieval: Given all prior turns plus the current question and a domain corpus, systems retrieve passages for the current turn, resolving coreferences and context dependencies while handling queries whose relevance requires multi-step reasoning. Results are reported by turn position (T1–T5+) to capture how retrieval degrades as conversations deepen.

Sub-Track 2b – Generation with Gold Passages: Given a conversation task and gold reference passages, systems generate a grounded answer evaluated against the reference. This subtask isolates generation quality from retrieval quality, establishing an upper bound.

Sub-Track 2c – Full Conversational RAG: Given a conversation task, systems first retrieve the top-5 passages, then generate a grounded answer. Both retrieval (nDCG@10) and generation (GPT-4o LLM-judge, 1–5 per dimension across Correctness, Completeness, Relevance, Coherence, Faithfulness) are evaluated end-to-end.

Submission format. For each test instance, participants submit a ranked list of passage identifiers; participants may submit to either track or both. **Allowed resources:** official-leaderboard retrieval must use the *provided* corpus; open-source and proprietary models (embedders, LLMs, judging APIs) are permitted but must be disclosed for fair open-vs.-closed comparison.

The public TEMPO and RECOR releases serve as **training/development data**. For SemEval-2027, evaluation uses a **separate held-out test set** that was already annotated by our human annotators during dataset construction and has *never been publicly released*, eliminating contamination. The test is **stratified across all domains** (13 for Track 1, 11 for Track 2) rather than a single subset, with approximately **350 Track 1 test queries** and **200 Track 2 test turns**. It was constructed with the same validated pipeline as the pilots: for Track 1, temporally grounded questions were collected from unseen source material, normalized into passage-retrieval instances, and source documents segmented into passages; for Track 2, multi-turn grounded conversations were constructed from unseen source material and supporting passages annotated for each target turn. **Annotation protocol.** For Track 1, a passage is labelled gold only when it is both topically relevant *and* temporally aligned—i.e., its time scope (start/end dates) matches a required anchor period; an unrelated date does not count. For Track 2, each instance preserves the full conversation history for the target turn, and a passage is labelled gold only when it supports the answer through multi-step inference (reasoning-dependent relevance) rather than surface overlap, with discourse dependencies (anaphora, ellipsis) resolved against prior turns.

Data quality. The pilot resources already provide strong evidence that the data and annotations are reliable. In TEMPO, expert verification achieved Cohen's $\kappa = 0.82$ for temporal reasoning class assignment and $\kappa = 0.85$ for gold document relevance. In RECOR, human evaluation reported strong quality scores for naturalness (4.2/5), turn coherence (4.6/5), and groundedness (4.4/5), with $\kappa \geq 0.73$ across criteria. The benchmarks also span diverse reasoning: TEMPO covers 10 temporal reasoning classes (led by event analysis/localization 36%, time-period contextualization 21%) and RECOR six discourse-dependency types (continuation 35%, coreference 29%, self-contained 24%, topic-shift 11%). Every held-out test item has already been annotated; as a final quality pass before the evaluation phase, each item undergoes manual re-verification, at least 25% are double-annotated, disagreements are adjudicated by an organizer, and final inter-annotator agreement statistics are reported in the task-description paper.

3 Task Details

Licensing and release. The full task package (splits, relevance annotations, scorers, format checkers, baselines, and leaderboard utilities) will be released under **CC-BY-4.0** and archived on Zenodo.

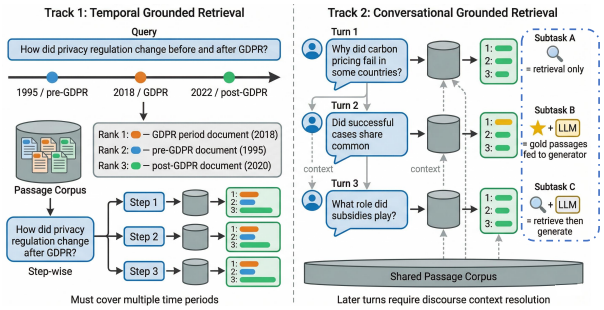

Figure 1: RETECO tracks. **Track 1** (left): a temporal query needs passages spanning multiple time periods; the step-wise sub-track decomposes it into sequential steps. **Track 2** (right): each turn retrieves from a shared corpus, resolving discourse context and reasoning-based relevance (A = retrieval only; B = gold passages + generator; C = retrieve then generate).

Resources required. Because the pilot datasets, the held-out test, baseline code, and scorers already exist, the main remaining work is a final QA pass over the *already-annotated* held-out test, packaging, and competition infrastructure. We estimate roughly 180–220 person-hours (organizing team plus 2 student annotators) for final re-verification, sample double-annotation, and adjudication—not for building the test from scratch. **Anticipated risks and mitigations:** disagreement on temporal alignment or reasoning-dependent relevance is mitigated by explicit guidelines, $\geq$25% double annotation, organizer adjudication, and reported κ; residual LLM-assisted annotation errors are caught by mandatory human verification of every test item plus an independent LLM-judge cross-check; schedule risk is low because the pipelines already exist (refresh, not design). Compute is modest (CPU-based BM25 indexing; dense baselines on standard GPUs). We will provide a task website, mailing list, format checker, scorer, starter baseline code, and a CodaLab/CodaBench competition.

Timeline and readiness. We will align with the SemEval-2027 schedule: sample data by 15 July 2026, training/development data by 1 September 2026, and the already-annotated hidden evaluation data finalized and packaged internally by 1 December 2026. Because both pilot datasets, baseline pipelines, and scorers already exist, the remaining work is primarily final validation and infrastructure packaging rather than task design or new data collection.

Ethics, privacy, and security. The task uses public technical, scientific, historical, legal, and policy-oriented text. We remove or avoid personally identifiable information and do not include tasks aimed at identifying individuals. Although one RECOR domain is medical science, the task evaluates retrieval from public informational texts only and does not support diagnosis or medical decision-making. We do not anticipate major security risks, and all released resources are screened for privacy and licensing compliance.

Model	Track 1 (TEMPO pilot)	Track 2 (RECOR pilot)
BM25 (Robertson et al., 2009)	10.8	44.6
BGE (Xiao et al., 2023)	22.0	41.1
ReasonIR (Shao et al., 2025)	27.3	49.6
DiVeR (Long et al., 2025)	**32.0**	**54.5**

Table 1: Representative pilot baselines (nDCG@10) on the public pilot resources underlying the two RETECO tracks, showing substantial room for improvement.

4 Pilot Studies and Lessons Learned

A major strength of this proposal is that it is grounded in pilot studies with existing data, baselines, and evaluation scripts. Table 1 summarizes representative pilot results underlying the two tracks; both settings show clear headroom and remain challenging for strong modern baselines. On Track 1, even the best system (DiVeR) reaches only 32.0 nDCG@10 and 71.4% Temporal Coverage@10—i.e., it fails to retrieve temporally complete evidence for many queries. A controlled ablation confirms that temporal grounding is genuinely required: stripping temporal signals from Track 1 queries lowers nDCG@10 for nearly all models (e.g., E5 30.4$\rightarrow$27.0, SFR 30.0$\rightarrow$27.8, DiVeR 32.0$\rightarrow$29.9), so relevance cannot be recovered by topical matching alone. On Track 2, augmenting the query with conversation history and explicit reasoning roughly doubles retrieval quality on the RECOR pilot, underscoring that reasoning over dialogue history—not surface matching—drives relevance.

These pilots motivate three design choices: (1) RETECO is primarily **retrieval**-focused, with optional generation sub-tracks, since strong retrieval is a prerequisite for reliable RAG and ranked passage retrieval permits transparent, reproducible evaluation; (2) each track ranks on a **single official metric** (nDCG@10) for leaderboard clarity; and (3) diagnostic analyses (temporal coverage, turn depth, per-domain) are reported separately to characterize where systems fail.

5 Evaluation

We will release format checkers, a local scorer for the public splits, and a CodaLab/CodaBench leaderboard for the hidden-test phase. The official leaderboard ranks retrieval by **nDCG@10** for both tracks (the generation sub-tracks 2b/2c are scored by the LLM-judge and reported alongside, not used for ranking); gold labels remain hidden during the competition and final scores are computed by the organizers using the released official scorer. **Metric definitions.** At cutoff k=10: *nDCG@10* is standard normalized discounted cumulative gain; *Temporal Precision@10* is a position-weighted precision rewarding temporally relevant documents ranked earlier; *Temporal Relevance@10* is the fraction of the top-10 judged temporally relevant; *Temporal Coverage@10* is the fraction of the query's required time periods covered by at least one retrieved document; and *NDCG|FC@10* is nDCG computed only over queries achieving full temporal coverage. Temporal relevance and period coverage are assigned by an LLM-as-judge with documented prompts and "golden-case" meta-evaluation. For **Track 1**, we will additionally report temporal diagnostic metrics including **Temporal Coverage@10** and **Temporal Precision@10**. For **Track 2**, we report results by turn position (T1–T5+) and, diagnostically, by domain (per-domain figures are reported for analysis, not for ranking). For **Sub-Track 2c**, the GPT-4o judge scores each of the five dimensions with a separate focused prompt on a fixed 1–5 Likert rubric (normalized to 0–1 and macro-averaged); the rubric is calibrated via "golden-case" meta-evaluation, and judge outputs are spot-validated against human ratings (in the RECOR pilot, three PhD annotators rated 200 conversations and automatic scores tracked humans within $+0.14$ to $+0.35$), which we will repeat on the SemEval test set. **Competition phases.** A development phase releases the public train/dev splits, the official scorer, and starter baselines; an evaluation phase then opens the hidden-test leaderboard with a capped number of daily submissions to discourage tuning on the leaderboard, and final rankings are computed by the organizers on withheld gold labels. All official evaluation scripts, baseline runs, and data readers will be released and archived after the shared task to support reproducible follow-up research.

6 Expected Impact

RETECO fills a gap between temporal understanding, discourse-level conversational interpretation, and reasoning-oriented retrieval. Because grounded RAG fails when the temporally or conversationally appropriate evidence is not retrieved (e.g., medical, legal, and policy QA), the task targets a concrete bottleneck for applied RAG. It provides the community with a reusable benchmark for semantic grounding in retrieval and is expected to stimulate work on temporally grounded retrieval, context-aware multi-turn retrieval, reasoning-guided query reformulation, and retrieval modules for grounded RAG systems. All released artifacts remain publicly available after the workshop via GitHub and Zenodo.

7 Task Organizers

Abdelrahman Abdallah (University of Innsbruck; lead organizer) is a PhD candidate working on LLM-based reranking, temporal reasoning, and conversational retrieval. He led the development of TEMPO and contributes to RECOR, and will coordinate data release, evaluation infrastructure, leaderboard maintenance, and participant support.

Mohammed Ali (University of Innsbruck; co-organizer) is a PhD candidate working on multi-turn conversational retrieval; he is first author of RECOR and will lead data preparation, evaluation design, and participant support. **Muhammad Abdul-Mageed** (University of British Columbia; advisory organizer) is Canada Research Chair in NLP and Machine Learning and Associate Professor at UBC; he will advise on benchmark design, NLP evaluation, and dissemination.

Kevin Duh (Johns Hopkins University; advisory organizer) is Associate Research Professor of Computer Science and Senior Research Scientist at the Human Language Technology Center of Excellence at Johns Hopkins University. He will contribute guidance on shared-task design, evaluation methodology, and community outreach.

Adam Jatowt is a Professor at Dept. of Computer Science at University of Innsbruck. His research interests span NLP and IR with particular focus on temporal information retrieval (T-IR). He has co-organized 7 shared tasks at CLEF, NTCIR and ICDAR including two related to T-IR, and co-authored surveys on T-IR (Campos et al., 2014b; Piryani et al., 2025) and temporal commonsense reasoning (Wenzel and Jatowt, 2023). He will de-

liver a tutorial on T-IR at WebConf 2026.

References

Abdelrahman Abdallah, Mohammed Ali, Muhammad Abdul-Mageed, and Adam Jatowt. 2026a. Tempo: A realistic multi-domain benchmark for temporal reasoning-intensive retrieval. *arXiv preprint arXiv:2601.09523*.

Abdelrahman Abdallah, Jamie Holdcroft, Mohammed Ali, and Adam Jatowt. 2026b. Are llm-based retrievers worth their cost? an empirical study of efficiency, robustness, and reasoning overhead. *arXiv preprint arXiv:2604.03676*.

Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and Hyun-Soo Kang. 2026c. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2601.09562*.

Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Andreas Herzinger, Jamie Holdcroft, and Adam Jatowt. 2026d. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 208–219.

Abdelrahman Abdallah, Bhawna Piryani, Jonas Wallat, Avishek Anand, and Adam Jatowt. 2026e. Tempretriever: Fusion-based temporal dense passage retrieval for time-sensitive questions. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 5–15.

Mohammed Ali, Abdelrahman Abdallah, Amit Agarwal, Hitesh Laxmichand Patel, and Adam Jatowt. 2026. Recor: Reasoning-focused multi-turn conversational retrieval benchmark. *arXiv preprint arXiv:2601.05461*.

Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. Trec ikat 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 819–829.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. 2014a. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):1–41.

Ricardo Campos, Gaël Dias, Alípio Mário Jorge, and Adam Jatowt. 2014b. Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15:1–15:41.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2025. Complextempqa: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9111–9123.

Hideo Joho, Adam Jatowt, and Roi Blanco. 2014. Ntcir temporalia: a test collection for temporal information access research. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 845–850, New York, NY, USA. Association for Computing Machinery.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yecheng Luo, Yue Shen, Jian Wang, Hualei Zhou, Chunxiao Guo, Peng Wei, and 1 others. 2025. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's high time: A survey of temporal question answering. *arXiv preprint arXiv:2505.20243*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and 1 others. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, and 1 others. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *arXiv preprint arXiv:2308.00002*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.