SIGIR '25 notes and interesting posters

Benedikt Kantz

July 13-18 2025

1 Sunday

1.1 RAG vs. Longcontext LM Retrieval

`https://sigir2025.dei.unipd.it/detailed-program/paper?paper=f6185f0ef02dcaec414a3171cd01c697`

- LLamaIndex - framework for RAG Retrieval
- Summarization based Retrieval
  - Tree/Summarization based retrieval - construct trees and traverse by encoding tree structure
  - "RAPTOR" - flatten tree, search all nodes
- Long Contex LLMs
  - Main challenge: training time $O(n^2), n$ tokens
  - Extend also frequency encoding!
- focuses on LM bascics with long contexts in first half

1.2 Efficient In-Memory Inverted Indexes: Theory and Practice

`https://sigir2025.dei.unipd.it/detailed-program/paper?paper=b38e5ff5f816ac6e4169bce9314b2996`

- Scoring/retrieval based on scores and terms
- Priming index (TODO- slides available)
- Quantization
- Reordering
- stoping (stopwords)
- Efficient Index search - dynamic pruning: WAND (weak AND) good writeup: `https://www.elastic.co/blog/faster-retrieval-of-top-hits-in-elasticsearch-with-block-max-wand`
  - Need to sort cursor when iterating (TODO: investigate further)
  - Or sort once: MaxScore, only keep relevant docs for score by precomputed upper bound for term - all combinations have combined upper bounds!
  - Can even be broken down to blocks - at the cost of storage
  - Ranker is dependent on search algorithm (e.g. BM25, LM-based, ...)
  - Threshold priming - guess initial threshold, skip docs that might not be relevant prior
- Sparse retrieval
  - Documents are naturally sparse - exploited through inverted indices, terms are related to only a few documents (except for stopwords - just remove them) – same for queries! ("Zipf's Law")

  - Exploited through TF-IDF scoring (classics), not perfect - scoring inbalanced, lexical mismatch (semantic mismatch - e.g. F1 vs Formula One)

$\Longrightarrow$ learn term weighting, sparsity, ...

  - e.g. BERT for token re-weighting – learn important tokens
  - Expansion – learn new tokens to add to query, or Document Queries (i.e. Doc2Query); find similar words by using the masked prediction capabilities of BERT
  - Then: mask again, as LMs usually very dense!
  - In Practice: Query Weighting + Document Masked Expansion + Regularisation/top-k
  - But: this learned indices: distributions changes, algorithms become slower – slower retrieval, sparsity reduces significantly

$\Longrightarrow$ improve method by:

    * Combine traditonal BM25 retrieval for index + scoring from learned models (can even be interpolated!)

- boolean retrieval (with inverted indexes) vs. kNN/ANN (image, multimodal, semantic, ...) – future quite open!

1.3 Neural Lexical Search with Learned Sparse Retrieval

`https://sigir2025.dei.unipd.it/detailed-program/paper?paper=3bd4017318837e92a66298c7855f4427`
$\Longrightarrow$ seems interesting, no time (all tutorials are interesting this afternoon!) :/

1.4 Psychological Aspects in Retrieval and Recommendation

`https://sigir2025.dei.unipd.it/detailed-program/paper?paper=110eec23201d80e40d0c4a48954e2ff5`
Slides `https://github.com/aisocietylab/Psy-IR-RecSys-SIGIR25/`

- IR systems integrated in life - play role in life - psychological impact (cognition, decision, ...)!

$\Longrightarrow$ model human using emotions/personality/overload/mood (term: psychological states!)

- "Cognitive architecture": first IR paper in cognitive science journal (by CS researchers!) - go back to roots
- higher level/lower level processes (e.g. attention)
- Key processes
  - Attention: pull focus on key elements
  - Memory: memorize focused elements $\Longrightarrow$ pull focus
  - Decision-making: (ir-)rationality, heuristic-based
- Map processes between IR/RS scenarios

IR/RS Task	Cognitive Process
Query (Re-)formulation	Working memory, attention
Ranking, Interface Design	Attention, perception
Session continuation or abandonment	Episodic memory, fatigue
Item selection	Decision-making, heuristics
Personalization (over time)	Long-term memory, learning

- Cognitive architectures: are "computation frameworks" – helps us to apply models & test them, align models in IR with users!
- Symbolic architecture
  - Discrete symbols, with rules operating on knowledge

⟹ explainable, explicitly encoded; hard to expand due to handcrafted rules

- Emergent/Connectionist architectures
  - across network/graph, parallel processing, nodes represent subsymbolic elements

  ⟹ not (or limited) explainabilty!, a lot of data needed
- Hybrid: more flexbility
- Survey: `https://link.springer.com/article/10.1007/s10462-018-9646-y#Fig3`
- ACT-R: Adaptive Control of Thought – Rational (ACT-R): memory types (a la von Neumann), with rules indside
- Application: recency bias in retrieval, e.g. music retrieval for non-mainstream music
- SOAR: goal oriented, episodic memory – agentic framework?
- CLARION: hybrid
- LIDA: Global Workspace Theory, consciousness vs unconscious focus
- Problems: difficult to integrate in fuzzy pipelines like DL models, quite interdisciplinary ...
- Cognitive Biases
  - biases play against rational (logical and optimal decisions)
  - influences in algorithmic decision makings; encoded in various tasks, LM models (generative AND embedding!)
  - Feature-Positive Effect: not aware of *missing* things: which information is ignored; e.g. which aspects of a job RecSys are ignored?
  - Social/Cultural Homophily: users with similar traits like similar things: experiments show users like music from their own country.
  - Conformity bias: conform to opinion of others; e.g. show users ratings of other users first biases – could be used to lead towards higher/lower ratings?
  - Declinism: belief, world was a better place – found in increase in negative terms in music lyrics
  - Primacy/Recency Bias
- Personality and affect
  - States & Traits
  - "Affective Computing": detect emotions from metadata/interactions

2 Monday

2.1 BM25 and All That

- History of IR, back to Cranfield experiments, first probablistic theories and how they theorized/implemented BM1-¿ BM25

2.2 Multimodal Retrieval

2.2.1 DePro: Domain Ensemble using Decoupled Prompts for Universal Cross-Domain Retrieval

- Prompt tuning/embedding optimisation for concatinated prompts

2.3 Reproducibility in Domain-Specific, Multimodal, ... Retrieval

(the critic's section)

2.3.1 Benchmark Granularity

- Image-Text: bidirectional problem
- Problem: benchmarks very curated/clean, real-world: noisy, ...
- Eval: shuffle words, different granularities of descriptions
- Results:
  - performance increases if queries are more fine-grained $\implies$ miss out on performance of model, use multiple granularities in benchmarks!
  - perturbations usually hurt
- Very inconsistent accross models/datasets!

2.3.2 Visual Document Retrieval + Late Interactions

- ViT: patch-based encoding of document (ColPali/ColQwen2 `https://huggingface.co/vidore/colqwen2-v1.0`) using ColBERT paradigm (offline search using MaxSim `https://github.com/stanford-futuredata/ColBERT`)
- Image-Documents has less loss in performance even in big index sizes compared to OCR
- Reason: number of patches/text tokens factor is relevant
- Matching to visual tokens to query is high – seems effective

2.3.3 Assessing effective token length of multimodal retrieval

- Information Alignment: CLIP has a modality gap between image and text!
- Comes from short/descriptive captions
- Long-Text performance/beyond token budget using pooling?
- Domain-Specific Datasets

$\implies$ no model could use full token budget!, also variance between datasets

- How was the token length varied? removing adjectives, just cut? – was just cut... (maybe information afterwards not relevant/good?)

2.3.4 Refined Medical Search via Dense Retrieval and User Interaction

Interesting for A+CHIS/HEREDITARY search engine?

- Enhance User Experience with Interaction integration in Dense (embedding) Retrieval
- TripClick IR Dataset -¿ A+CHIS?
- Fine tune PubMedBERT on TripClick Triplets (Document, D+, D-) – what are D+/D-
- Retrieval by embedding example queries from TripClick dataset + adding similar documents/queries from dataset of session to query

2.3.5 Reassessing LLM for Boolean Queries for Systematic Reviews

- In medical domains: a lot of very complex boolean queries for literature reviews
- Prior paper: boolean queries for systematic reviews, other reproducibility paper found: their results do not generalize – but did not use syntax checker, produces wrong results!
- Also: did not use guide-documents for prompt engineering
- Syntax check - regenerate
- Q: Did they use grammar for queries (constrained generation?) json output can reduce performance of LLMs `https://arxiv.org/abs/2408.02442`

2.3.6 Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings

- TikTok audits are diffictult to reproduce (change of platform/bot banning)
- Personalitation stronger for longer videos
- Country ID by Proxies

$\Longrightarrow$ audits only short-term valid

2.3.7 Beyond Reproducibility: Advancing Zero-shot LLM Reranking Efficiency with Setwise Insertion

- Previous: Comparison by “Bubble-Sort-by-LLM”
- Do chunked reordering using LMs

2.4 Knowledge & KGs

2.4.1 Segmentation Similarity Enhanced Semantic Related Entity Fusion for Multi-modal Knowledge Graph Completion

- Using ViT+Semantic Segmentation to complete KGs with more multimodal knowledge (location based on images, ...)
- Tranformers over Transformers...

2.4.2 Mitigating Modality Bias in Multi-modal Entity Alignment from a Causal Perspective

- finding equality/alignment between KGs using semantic information

2.5 Human and Interfaces

2.5.1 The In-Situ Effect of Offensive Ads on Search Engine Users

- people see avg. 5000 ads/day
- offensive ads more effective, but at a risk – evoke negative emotions
- complaints: usually only very few offensive ads (Bing: a single offensive ad spreadsheet!)
- then: user study to count (offensive) ads
- Result: users can barely distinguish ads, irrelevance and offensiveness is very correlated; irrelevance is considered more offensive than offensive itself!
- People do not recall offensive ads on bing/google

2.5.2 Simulating Before Planning: Constructing Intrinsic User World Model for User-Tailored Dialogue Policy Planning

- Agentic framework ...

3 Tuesday

3.1 Keynote: Digital Health

- Vaguely related to IR, advertisement for the industrial lab
- Prediction of treatment/preselection of people based on disease history
- Kenneth Church: Chatbots are not the solution for Callcenters - could be just an app...

3.2 Search and Ranking

(Peter's favorite session!)

3.2.1 CoDIME: A Counterfactual Approach for Dimension Importance Estimation through Click Logs

- Dimension Importance Model Estimation (DIME): only keep good dimensions for dense retrieval
- Estimate through LLM Gen docs (inconsistent) OR relevant docs from collection (need ground truth)
- Solution: use implicit feedback (click data)
- Estimate Correlation between rankings/clicks; fit linear model between this correlation (for dimension) as estimator
- Improves retrieval significantly - interesting for dense semantic retrieval!

3.2.2 Stitching Inner Product and Euclidean Metrics for Topology-aware Maximum Inner Product Search

`StitchingInnerProductandEuclideanMetricsforTopology-awareMaximumInnerProductSearch`

- Nearest neighbour search using euclidian distance is not great, as inner poroduct not a natural norm (no triangle ineq); need it for speedup of MIPS
- This seems like a convex hull over neighbourhood?

3.2.3 On the Scaling of Robustness and Effectiveness in Dense Retrieval

- Robustness of models largely ignored, lower limit of system relevant (in OOD settings)!
- Incorporate Robustness into loss
- Danger zone: did they use the test data to compute the robustness in the loss? – not a great answer in session
- Also no comparison to BM25 baseline...

3.2.4 Optimizing Compound Retrieval Systems

- Cascading retrieval systems (rerankers...)
- LLM Ranking does not really fit into reranker approach
- New paradigm: compound retrieval: *any combination* – cascading just a specific instance
- Propose fully differentiable rerank pipeline - can just learn to design the compound system
- Results: dynamic trade-off between LM calls and
  - BM25 ranker for pointwise comparison
  - Cascading system seems to be good choice
  - sparse selection of top seems also nice (or even other patterns of selection)
- Conclusion: could be interesting for novel compund systems

3.2.5 Hypencoder: Hypernetworks for Information Retrieval

- Problem: the linearity of dot product could result in non-linearly-sperarable subspaces for retrieval
- Solution: non-linear query-specific NN using Hypernetworks - learned NN that produces weights for other NNs

3.2.6 Classifying Term Variants in Query Formulation

- Variety of query for the same goal – 95/100 queries unique even for simple query!
- over 10k query formulations for 100 query back stories
- Analysis: remove back story; analyse added termns – users added 70% terms by themselves
- User modify semantics, misspell...
- Or enrich terms with information types or sources

3.2.7 Locality-Sensitive Indexing for Graph-Based Approximate Nearest Neighbor Search

- kNN search: graph based, often fsated by searching topologies
- Update patterns: reconnection best, deletion/hiding hurt performance (recall/time)
- LIGS: bucketize/hash vectors together – simulated graph by overlaying buckets – connections only implicit!
- good for high-churn applications (social media/networks)

3.2.8 Constrained Auto-Regressive Decoding Constrains Generative Retrieval

- Generative retrieval methods use constrained generative retrieval
- Either: step-wise elimination or beam-search over k sampling steps
- Result: erroneous retrieval, and will give high precision and low recall (just as we observed in the GutBrainIE evaluation!!)

3.3 Biomedical and Health

3.3.1 Cooking with Conversation: Enhancing User Engagement and Learning with a Knowledge-Enhancing Assistant

- Allow users either an active or passive guidance for Cooking
- include annotated recipes (KG/facts)

3.3.2 General Neural Embedding for Sequence Distance Approximation

- Usually: Distance metrics like edit distance hard to compute
- distance approximated using NN, existing approachs lack generality
- Their Approach: Small domain head, shared CNN+Transformer block for all domains!
- Show good performance on both DNA and GPS data (??), no significance test

3.3.3 ProtChatGPT: Towards Understanding Proteins with Hybrid Representation and Large Language Models

- PLP-Former for special protein tokens

3.3.4 Combining Evidence and Reasoning for Biomedical Fact-Checking

- 3-stage framework for verification of facts, including justification aoutput of LMs

4 Wednesday

4.1 Keynote: Please meet AI, our dear new colleague. In other words: can scientists and machines truly cooperate?

(Iryna Gurevych)

- Peer review assistance through various tools: ACL and AAAI already incorporate this!
- Specific Review Datasets
- Workflow for Review Process, AI cannot fully arrive at review but help prefill fields
- Related works: difficult to write a good argumentative section, not just summarize!
- Connection between reviews and revisions – hard, as the task is quite undefined!
- Aspects in review: Detect aspects in fields semi-autonomously (ontology!); LLM generated reviews are very similar w.r.t. aspects! – use for detection of LLM detection;
- Lazy Review heuristic: i.e. did not do ...
- Reviews and promised changes: relate them automatically, verifiy completion!

4.2 Efficiency

4.2.1 WebANNS: Fast and Efficient Approximate Nearest Neighbor Search in Web Browsers

(Peter: nice WASM usage); very good presentation!

- Previously: USed IndexedDB within browser / prefetch data points, us HNSW + JS (?)
- Proposed: WASM, problem WASM $\rightleftharpoons$ IndexedDB, IndexedDB slow
- Tree In-Memory, rest in IndexedDB+JS (three-tier data storage: WASM-JS-DB), Lazyload per tree level
- Benchmarks show sub-second performance!

4.2.2 TITE: Token-Independent Text Encoder for Information Retrieval

- Reduce computational cost by aggregrating tree-wise, but the model can increase dimensionality
- Related to Funnel Transformers: pool (e.g. 2×) over one of the attention dimensions
- without loss of performance!

4.2.3 WARP: An Efficient Engine for Multi-Vector Retrieval

(Best Paper Award)

- Combine XTR+ColBERTv2/PLAID to do token-wise retrieval

4.2.4 An Analysis on Matching Mechanisms and Token Pruning for Late-interaction Models

- Prune tokens based on attention/...
- Little effiency loss

4.2.5 Data Augmentation for Sample Efficient and Robust Document Ranking

- Augment data by splitting docs & rerank passages of docs based on simple methods
- Solid eval on various loss/augmentation combinations
- Augmentation is useful, especially in low-resource
- Acts as regularizer, allows transfer from e.g. MSMARCO to other domains

4.3 Reranking

4.3.1 Bridging Personalization and Control in Scientific Personalized Search

- Focus: Reranker with personalized search
- Cross-Encoder for reranking, their contribution: personalization
- Approach: represent history as context to the reranker; add personalization by mixing concepts using 'routing'
- Could be interesting to visulize within OnSET to improve user guidance?

4.3.2 Reason-to-Rank: Distilling Direct and Comparative Reasoning from Large Language Models for Document Reranking

- Add reasoning model for reranking model

4.4 Low-Resouce

4.4.1 When Less is Enough: Optimizations for Low-Cost Recommendation Systems

- Low-ARPU regions: profitability @ $4/user

4.4.2 Efficient Approximate Nearest Neighbor Search on a Raspberry Pi

4.4.3 IR for AAC Users: A Hyperdimensional Computing (Vector Symbolic Architectures) Approach

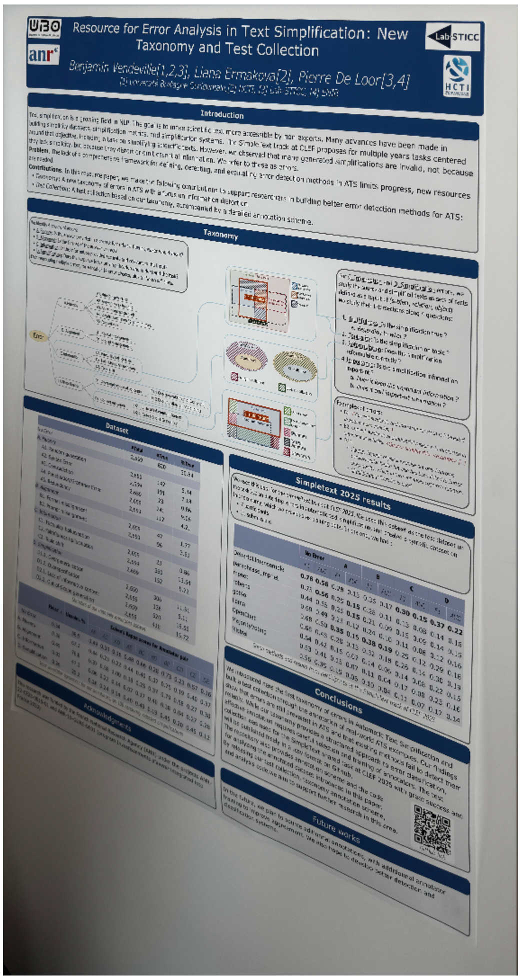
4.4.4 Some Things Never Change: Overcoming Persistent Challenges in Children IR

5 Thursday (Workshops)

5.1 ReNeuIR at SIGIR 2025: The Fourth Workshop on Reaching Efficiency in Neural Information Retrieval

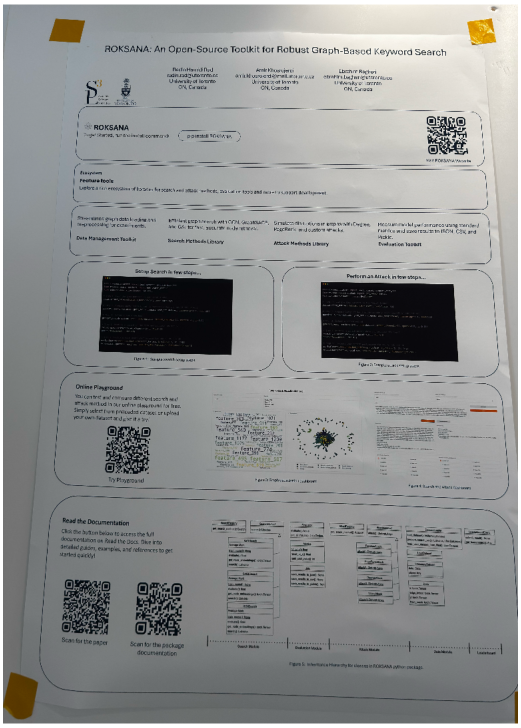
6 Posters

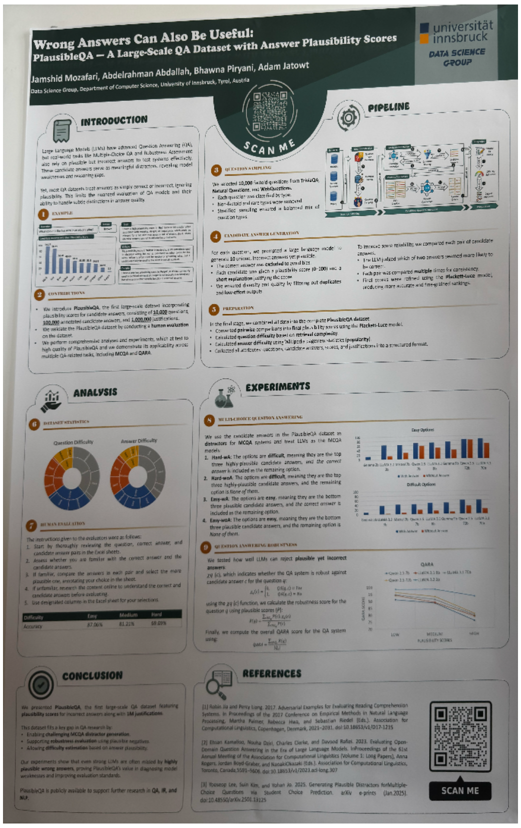
Table 1:

Poster	Information
	**The Agent Perspective In LLM-Based Strategic Information Retrieval Ecosystems** *0.56 0.35 0.15 0.06 0.10 0.04 0.11 0.07 0.17 0.14* `https://doi.org/10.1145/3726302.3730125`

Continued on next page

Table 1: (Continued)

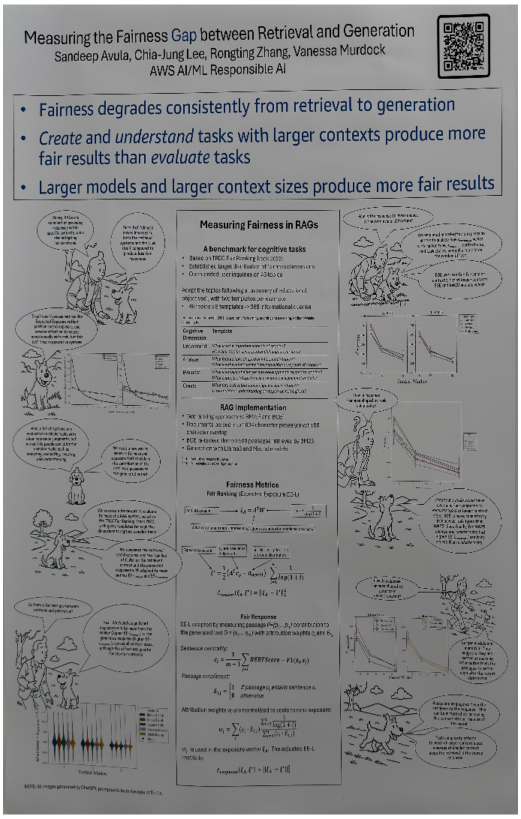
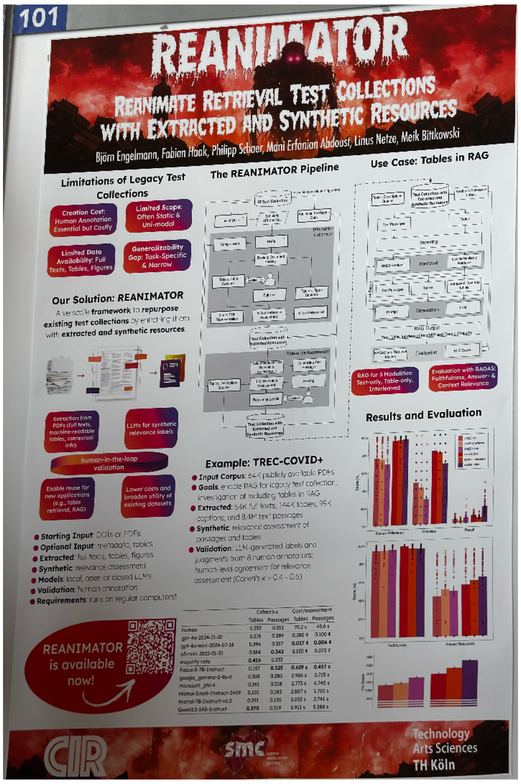
<table>
<tr>
<td>


</td>
<td>ROKSANA: An Open-Source Toolkit for Robust Graph-Based Keyword Search equeries, gold set, resus before, k. values-15, 10, 201) https://doi.org/10.1145/3726302.3730154</td>
</tr>
<tr>
<td>


</td>
<td>Wrong Answers Can Also Be Useful: PlausibleQA — A Large-Scale QA Dataset with Answer Plausibility Scores 100,000 annotated candidate answers, and 1,000,000 justifications. https://doi.org/10.1145/3726302.3730299</td>
</tr>
</table>

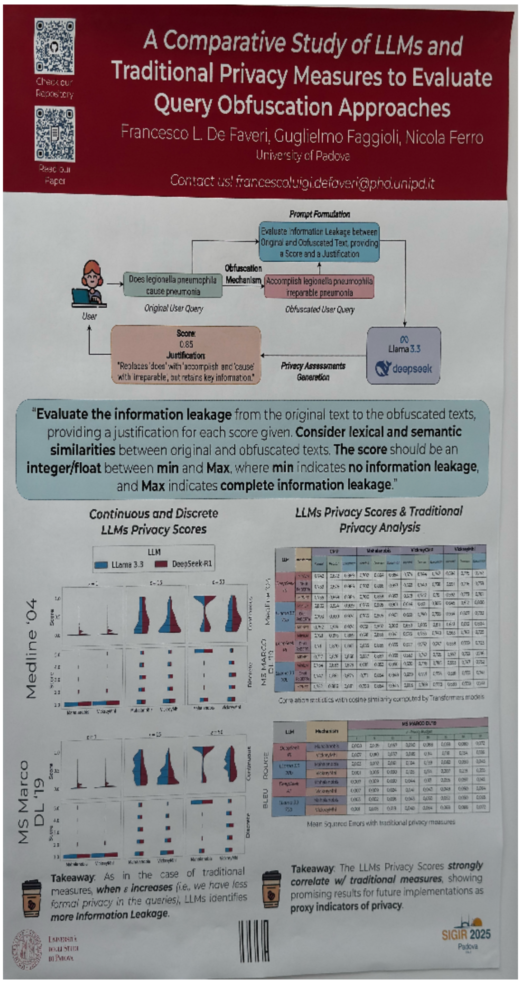
Continued on next page

Table 1: (Continued)

<table>
<tr>
<td>
</td>
<td>Conversational Search: Towards Personalization and Evaluation the generated text G = (5,... Sm with attribution weights c; and E,
https://doi.org/10.1145/3726302.3730126</td>
</tr>
<tr>
<td>
</td>
<td>REANIMATOR: Reanimate Retrieval Test Collections with Extracted and Synthetic Resources Björn Engelmann, Fabian Haak, Philipp Scher, Mani Erfanian Abdoust, Linus Netze, Meik Bikowski
https://doi.org/10.1145/3726302.3730342</td>
</tr>
</table>

Continued on next page

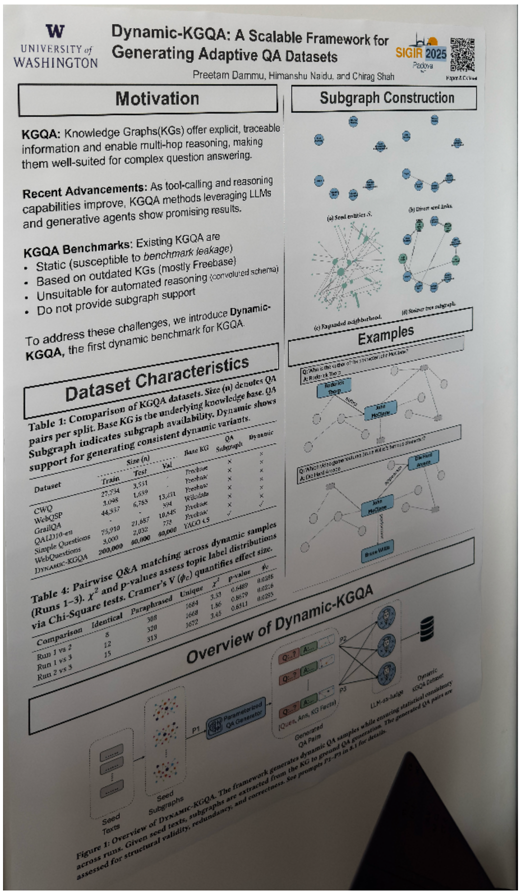
Table 1: (Continued)

<table>
<tr>
<td>


</td>
<td>
A Comparative Study of Large Language Models and Traditional Privacy Measures to Evaluate Query Obfuscation Approaches measures, when e increases (i.e., we have less
<code>https://doi.org/10.1145/3726302.3730158</code>
</td>
</tr>
</table>

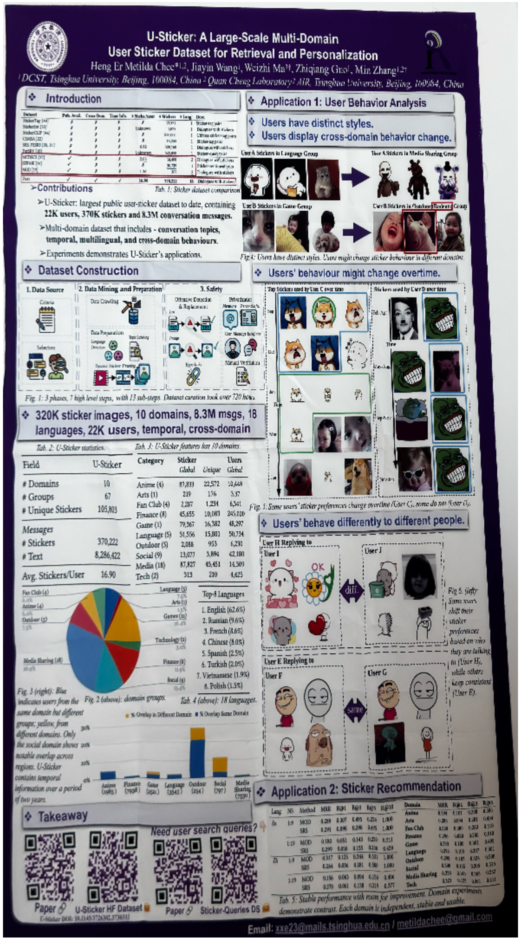
Continued on next page

Table 1: (Continued)

	**Dynamic-KGQA: A Scalable Framework for Generating Adaptive Question Answering Datasets** *assessed for structural validity, redundancy, and correctness. See prompts P1-P3 in 8.1 for details.* https://doi.org/10.1145/3726302.3730324

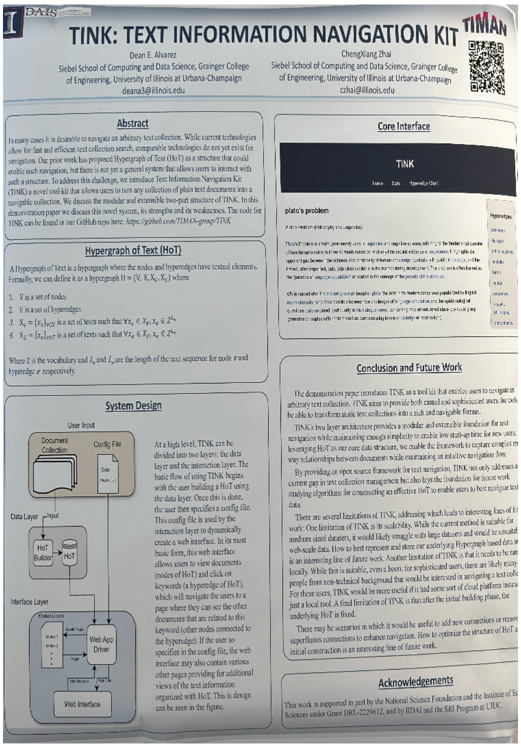
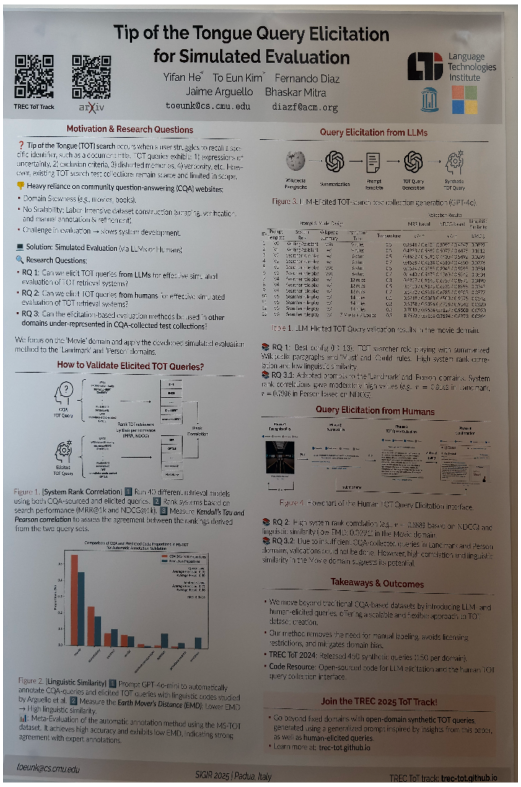
Continued on next page

Table 1: (Continued)

<table>
<tr>
<td>
</td>
<td>The Agent Perspective In LLM-Based Strategic Information Retrieval Ecosystems 'DCST, Tsinghua University, Beijing, 100084, China 2 Quan Cheng Laboratory2 AIR, Tsinghua University, Beijing, 100084, China
https://doi.org/10.1145/3726302.3730125</td>
</tr>
</table>

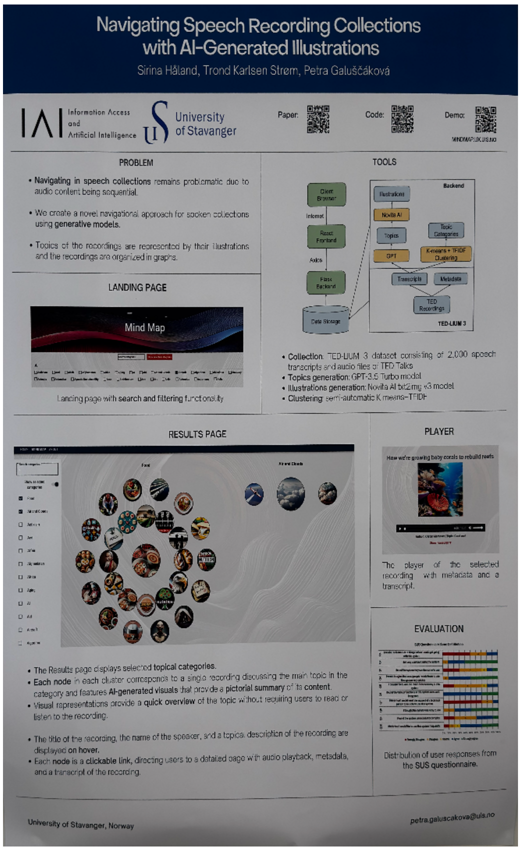
Continued on next page

Table 1: (Continued)

	**TINK: Text Information Navigation Kit** *3. X, = [Xo}vev is a set of texts such that VX, E Xv, X, € Z* https://doi.org/10.1145/3726302.3730141
	**Tip of the Tongue Query Elicitation for Simulated Evaluation** *rank correlations gave moderately high values (e.g., r = 0.9142 in Landmark,* https://doi.org/10.1145/3726302.3730335

Continued on next page

Table 1: (Continued)

<table>
<tr>
<td>


</td>
<td>
Navigating Speech Recording Collections with AI-Generated Illustrations Sirina Häland, Trond Karlsen Strom, Petra Galuscáková
https://doi.org/10.1145/3726302.3730136
</td>
</tr>
</table>