# An (LMM free[1]) Benchmark Collection for Assessing Scholarly Search by Non-Educated Users[2]

**Stéphane Huet** **Éric SanJuan**




CLEF 2025, Madrid, Spain

[1] <5%

[2] Computer, NLP, and data scientists

# Motivation


- Improving Access to Scientific Texts for Everyone
  - Everyone agrees on the importance of objective scientific information
  - But scientific documents are inherently complex...
- Can we improve accessibility for everyone?
  - Experts
  - Students
  - Lay persons
  - **Ourselves**
- Useful for:
  - Scientific communication
  - Science journalism
  - Political communication
  - Education

# The SimpleText-1 Test Collection


- A resource for scientific information access, shared in the context of the **CLEF SimpleText Track 1 (2022-2024)**.
- Designed to help users retrieve scientific abstracts in response to popular science queries.
- The collection includes:
  - A large corpus of scientific abstracts and more.
  - A set of relevance labels (**qrels**).
  - Additional automatic judgments based on **dense vector representations** ~~and **LLMs**~~.
  - A relational framework for efficient storage and retrieval of multiple embeddings.
  - Multiple baseline systems for meta-evaluation of qrels.

# The Corpus: the Citation Network Dataset (12th version)[3]


- Scientific publications in computer science and related fields from MS Open Academic Graph (OAG) / OpenAlex: DBLP + ACM citation network + MS FoS + *author's abstracts*
- Provides a large source of scientific documents for our test collection.
- Key statistics (publications before 2020):
  - 4,894,083 bibliographic references
  - 4,232,520 abstracts in English
  - 3,058,315 authors with their affiliations
  - 45,565,790 ACM citations

[3] J. Tang, A. C. Fong, B. Wang, J. Zhang, A Unified Probabilistic Framework for Name Disambiguation in Digital Library, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 975–987.

# Topics and Queries


- **Topics:** 40 press articles from *The Guardian* and *Tech Xplore*, written for a general audience.
- These articles cover various domains of computer science and electronics (AI, cybersecurity, bioinformatics, etc.).
- **Queries:**
  - Short keyword queries (1-4 per topic) crafted by computer scientists to pinpoint technical concepts.
  - Manually verified to ensure at least 5 relevant abstracts exist in the corpus.
  - Long queries (62 total for 2024), generated by GPT4 and manually reviewed, to explore subtopics.

# Topics and Queries


- Guardian
  - Topics: G01-G20
  - Queries: short and long
    - 42 short queries G*.[1-4], e.g. gene editing, drug discovery, crispr, forensics, advertising, Snowden
    - 63 long queries: G*.C[1-5], e.g. how algorithms are designed with human interaction in mind
- Tech Xplore
  - Topics: T01-T20
  - 67 Queries: T*.[1-4], e.g. phototransistor, 3G, energy efficiency, empathy, Bayesian approach

# Relevance Judgments (Qrels)


- **Iterative Construction:** The Qrels were built over three editions (2022-2024) using a cumulative process.
- **Methodology:**
  - A pooling method was used, where a sample of documents submitted by participants was manually judged.
  - New judgments were added to the training set for the following edition.
  - The official test sets were always on new, unassessed queries to prevent overfitting.
- **Judgment Criteria:**
  - Assess how well the title and abstract addressed the query.
  - Ensure correspondence with the key themes of the original news article.
- **Scoring Scales:**
  - 2022: 0-5 scale.
  - 2023-2024: Streamlined to a 0-2 scale to improve efficiency and consistency.

# CLEF SimpleText Task 1 Qrels Collection Statistics.


- **Total Judgments:** Across the 3 editions, **16,011** relevance judgments were assessed.
- This resource is split into two sets:

| Qrels | Topics | #Queries | #Assessed abstracts | | |
|---|---|---|---|---|---|
| | | | 0 | 1 | 2 |
| 2023 train | G01–G15 | 29 | 672 | 271 | 356 |
| 2023 test | G16–G20, T01-T05 | 34 | 2174 | 345 | 1207 |
| 2024 train | G01–G15 | 30 | 790 | 130 | 83 |
| 2024 extended test | G01–G20, T01-T05, T12-T20 | 66 | 3,681 | 991 | 457 |
| 2024 test | G01.C1–G10.C1, T06–T11 | 30 | 2,775 | 1,500 | 579 |

# Qrels Expansion


- The training set can be expanded using our database framework.
- An SQL procedure based on title embedding similarity finds documents with similar titles to those already judged relevant.
- This method can add over 2,000 new relevant documents, bringing the training set to **13,407** total judgments.

# Comparison of LLM Models to generate q-rels (Promps ...)


- Prompt:

**prefix** Here is a societal question and a scientific paper in computer science. Please, do not recommend papers that are off topic. I do not have time to read them all.

**prompt** Answer only returning a relevance score 0, 1 or 2. 0: Not really relevant, 1: relevant, 2: very relevant.

**system** You are a journalist writing about a tech topic that raises societal questions. You are looking for scientific publications that could feed your paper for a large audience.

# Comparison of LLM Models to generate q-rels (... results.)


| Model | tau | P-value | Accuracy (3) | Accuracy (2) |
|---|---|---|---|---|
| Qwen | -1.25 % | 63.29 % | 23.24 % | 48.44 % |
| Qwq | 30.00 % | *** | 32.17 % | 52.73 % |
| Gemma3 :small | 33.26 % | *** | 37.00 % | 58.45 % |
| Gemma3:12b | 31.95 % | *** | 38.29 % | 59.59 % |
| Phi4 | 41.54 % | *** | 45.16 % | 62.79 % |
| Llama 4 | 40.04 % | *** | 52.11 % | 69.58 % |

# Credibility and Simplicity


- **Goal:** Beyond relevance, retrieved information must be *trustworthy* and *understandable*.
- **Credibility:**
  - Scientific publications are considered a reliable source.
  - We provide metadata like **number of citations** and **bibliographic references** to assess peer recognition and scientific rigour.
  - A small subset of documents was human-assessed for credibility on a 0-2 scale.
- **Simplicity (Complexity):**
  - **Automatic Metrics:**
    - We computed readability scores, such as the **Flesch-Kincaid Grade Level (FKGL)**.
    - Provided complementary indices, including average characters per word, syllables, and vocabulary size.
  - **Human Assessments:**
    - A subset of documents was assessed by students on a 0-2 scale for complexity.

# Human assessment


- Number of assessed abstracts:

| **Annotators** | **Credibilty** | | | **Complexity** | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 |
| B. stud. in Humanity | 402 | 907 | 549 | 542 | 815 | 515 |
| B. stud. in Comp. Sci. | | | | 72 | 114 | 109 |
| M. Sc. stud. in Comp. Sci. | 1172 | 361 | 712 | 1274 | 548 | 424 |

# Automatic complexity measures toward human assessments

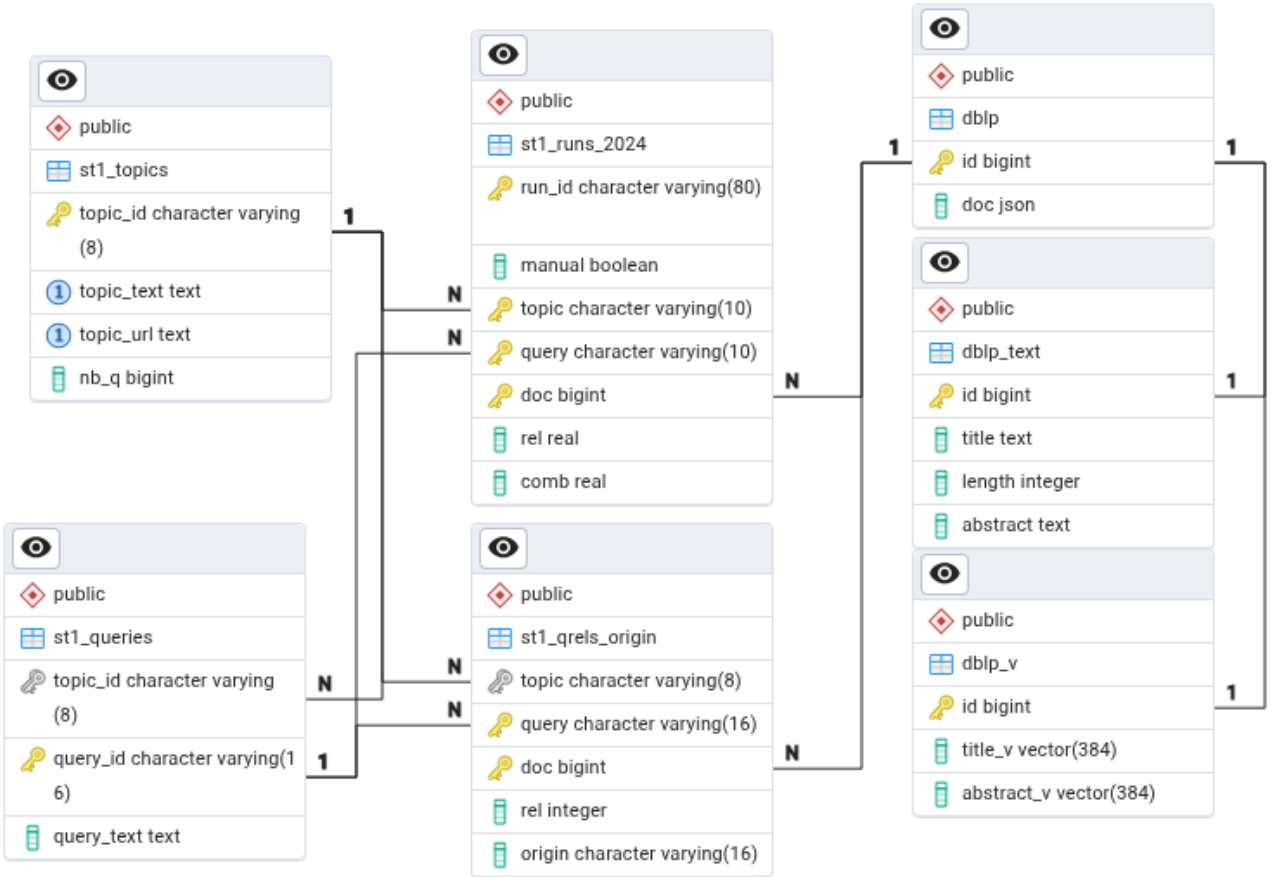

- Evolution of 4 automatic metrics w.r.t. to human scores:

| **Metrics** | **Bachelor stud.** | | | **Master stud.** | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 |
| FKGL | 15.02 | 15.16 | 15.25 | 15.1 | 15.38 | 14.72 |
| #words | 116.36 | 138.24 | 146.13 | 128.48 | 149.49 | 154.06 |
| #complex words | 29.93 | 36.61 | 38.11 | 34.01 | 39.5 | 39.98 |
| vocabulary size | 76.91 | 88.76 | 92.26 | 83.15 | 95.02 | 96.21 |

# Relational Vector Database


- We use **PostgreSQL** as our relational database management system.
- It leverages key features for our system:
  - **JSON type** for flexible document storage and data extraction.
  - **Generalized Inverted Indexes (GIN)** for efficient full-text search.
  - A simple **ivvflat index** for fast k-nearest neighbors search on vectors.
- The complete database is available in SQL code or a Docker image for participants at: https://inex:qatc2011@wayback.simpletext-project.com/

# Relational schema including JSON corpora and q-rels



# Decoupled Architecture


- Documents are stored in a central dblp table.
- Two derived relations: `dblp_text` (textual content) and `dblp_v` (embeddings).
- This separation allows for:
  - Adding new dense representations without altering document management.
  - Considwering multiple vector dimensions and models in separate relations.
- Topics and queries are stored on different relations allowing to build agregated scores.
- Participant runs per year and qrels are stored on distinct relations with references to queries and documents.

# Integrated Baseline System


- We provide a complete online baseline system to demonstrate the framework's capabilities.
- This system is based on the **lightweight cross-encoder MS MARCO Mini LM**.
- It uses **pgcurl** to integrate an online embedding service, allowing for real-time, CPU-based query processing.
- **Retrieval Methods:**
  - **Dense Retrieval:**
    - Using the MS MARCO Mini LM model.
    - We compare results based on embeddings of document titles and full abstracts.
  - **Sparse Retrieval:**
    - Using PostgreSQL GIN indexes for conjunctive queries.
    - BM25 results from an external ElasticSearch instance (not in the docker image) are included for comparison.

# Evaluation Results

| Run | MRR | Precision | | NDCG | | MAP |
|---|---|---|---|---|---|---|
| | | 10 | 20 | 10 | 20 | |
| title | **0.8454** | **0.6933** | **0.4383** | **0.5090** | **0.4010** | 0.1534 |
| abstract | 0.7683 | 0.6000 | 0.4067 | 0.4269 | 0.3539 | **0.1603** |
| bool | 0.7242 | 0.5233 | 0.3633 | 0.3409 | 0.2906 | 0.1199 |
| BM25 | 0.6173 | 0.3733 | 0.2900 | 0.2818 | 0.2442 | 0.1325 |

- **Key Finding:** Mini LM models are effective for integrating efficient IR into relational databases. Combining different embeddings could yield further improvements.
- **Titles vs. Abstracts:** Title-based embeddings achieve higher NDCG10, the official measure. Abstract-based embeddings perform better on MAP and Bpref.

# Future Work


- Develop and evaluate **user-centric information retrieval systems** that consider not only relevance but also comprehensibility.
- Further explore and refine the **hybrid retrieval approach** to better understand the thematic scope of the corpus.
- Investigate the effectiveness of **different neural models** for dense retrieval on our collection.
- **Extend the PostgreSQL framework** to support new retrieval models and passage aggregation methods.
- Utilize the collection to **authentically evaluate neural approaches** on textual content without the influence of modern LLMs.
- Apply the methodology to other domains to expand the scope of our work.


cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

# *SimpleText project*
# *Way Back Data website*

*Non contaminated, LLM free (<5%)*

Website : https://wayback.simpletext-project.com
Login : inex ; Password : qatc2011
E-mail : wayback@simpletext-project.com