DEMONSTRATION

Result Assessment Tool (RAT): An Open-Source Toolkit for Conducting Studies based on Search Results

SEBASTIAN SÜNKLER, Hamburg - University of Applied Sciences, Hamburg, Hamburg, Germany

KARDELEN BILIR, Hamburg - University of Applied Sciences, Hamburg, Hamburg, Germany

TUHINA KUMAR, Hamburg - University of Applied Sciences, Hamburg, Hamburg, Germany

OLIVER KOOP, Hamburg - University of Applied Sciences, Hamburg, Hamburg, Germany

SEBASTIAN SCHULTHEIß, Hamburg - University of Applied Sciences, Hamburg, Hamburg, Germany

DIRK LEWANDOWSKI, Hamburg - University of Applied Sciences, Hamburg, Hamburg, Germany

Open Access Support provided by:

Hamburg - University of Applied Sciences

Published: 22 March 2026

Citation in BibTeX format

CHIIR '26: 2026 ACM SIGIR Conference on Human Information Interaction and Retrieval
March 22 - 26, 2026
Seattle, USA

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3786304.3787925
ISBN: 9798400724145

Result Assessment Tool (RAT): An Open-Source Toolkit for Conducting Studies based on Search Results

Sebastian Sünkler
HAW Hamburg
Hamburg, Germany
sebastian.suenkler@haw-hamburg.de

Kardelen Bilir
HAW Hamburg
Hamburg, Germany
kardelen.bilir@haw-hamburg.de

Tuhina Kumar
HAW Hamburg
Hamburg, Germany
tuhina.kumar@haw-hamburg.de

Oliver Koop
HAW Hamburg
Hamburg, Germany
Oliver.Koop@haw-hamburg.de

Sebastian Schultheiß
HAW Hamburg
Hamburg, Germany
sebastian.schultheiss@haw-hamburg.de

Dirk Lewandowski
HAW Hamburg
Hamburg, Germany
dirk.lewandowski@haw-hamburg.de

Abstract

The Result Assessment Tool (RAT) is an open-source software toolkit for conducting research with results from commercial search engines and other web-based information retrieval systems. Conducting such research is challenging due to the "black-box" nature of these systems and limited data access. RAT addresses this by providing an integrated software environment that unifies modules for study design, automated data collection, manual assessment in a dedicated interface, and automated analysis via an extensible classifier framework. The software is designed to assist with various research tasks, including comparative assessments of result quality, investigations of source variety, and content analysis. It emphasizes transparency in methodologies, reproducibility of outcomes, and responsible data collection.

CCS Concepts

• **Information systems** → **Information retrieval**.

Keywords

search engines, web scraping, retrieval tests, research software

ACM Reference Format:
Sebastian Sünkler, Kardelen Bilir, Tuhina Kumar, Oliver Koop, Sebastian Schultheiß, and Dirk Lewandowski. 2026. Result Assessment Tool (RAT): An Open-Source Toolkit for Conducting Studies based on Search Results. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), March 22–26, 2026, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3786304.3787925

1 Introduction

Auditing commercial web search engines poses an ongoing and fundamental challenge for the information retrieval (IR) community and beyond. Traditional evaluation paradigms, such as those used by the Text REtrieval Conference (TREC) [17], are not applicable because they require access to system internals, test collections, and ranking functions. In contrast, commercial search engines operate like "black boxes". They treat their ranking algorithms and result data as crucial intellectual property, offering minimal access, if any, to their data. This requires the development of specialized methods and tools for conducting external observational studies. These studies require not only the collection of search engine results pages (SERPs), but also the archiving of the full content of the linked web pages — a key capability that RAT provides.

In recent years, the IR community has advanced evaluation methods through large-scale initiatives like MS MARCO [9] and the TREC Deep Learning Track [4]. However, initiatives like these do not address the problem of auditing external, closed systems. In addition, access to APIs provided by search engine vendors is limited and often insufficient for rigorous academic research. Large providers such as Google offer only limited access, while others, such as Microsoft Bing, have completely discontinued their public search APIs [8]. Even if some smaller providers, such as Brave Search, offer APIs, the overall landscape indicates that relying solely on APIs is not a sufficient solution. These APIs often come at a significant cost, are subject to strict limitations, and usually do not deliver the same results or comprehensive SERP features (e.g., universal search, featured snippets, or AI-generated overviews) that are presented to end users. This discrepancy makes reliable comparative studies, such as comparing the retrieval performance and output quality of two commercial search engines, exceptionally difficult without considerable, non-standardized effort.

As research questions expand beyond relevance to include issues such as source diversity, content classification, and fairness [15], the need for a sustainable, open-source, and integrated platform becomes critical. This paper demonstrates the Result Assessment Tool (RAT), an open-source toolkit designed to fill this gap. RAT provides a flexible, end-to-end framework that unifies the entire research workflow: from study design and robust data collection (including both SERP metadata and full-page archival of result documents, and also AI-generated content like AI Overviews and answers provided by Chatbots) to a dedicated human assessment interface and an extensible automated analysis framework [15]. The software is designed to support a wide range of empirical studies, from retrieval performance tests and source diversity audits

CHIIR '26, Seattle, WA, USA

ACM ISBN 979-8-4007-2414-5/26/03
https://doi.org/10.1145/3786304.3787925

to qualitative content analysis and investigations of algorithmic fairness. Our key contribution is this unified, sustainable software solution, which we validate through its successful application in multiple empirical studies. The software can be downloaded and installed as a self-hostable package and is also available as a ready-to-use web application.

2 RAT Research Workflow and System Architecture

RAT is a customizable, web-based application developed in Python, intended for researchers and study participants. Its modular architecture consists of two main applications: RAT Frontend and RAT Backend, which interact via a shared PostgreSQL database. This separation allows applications to be installed on different computers, enabling researchers to utilize resources for computationally intensive tasks (such as scraping dynamic pages or running classifiers) without affecting the frontend interface. The entire system was designed according to user-centered design (UCD) principles and includes automated testing and heuristic evaluations to ensure adaptability for both conventional information retrieval studies and qualitative content analysis, as well as other study types. Figure 1 shows RAT's software architecture, including its applications and components.

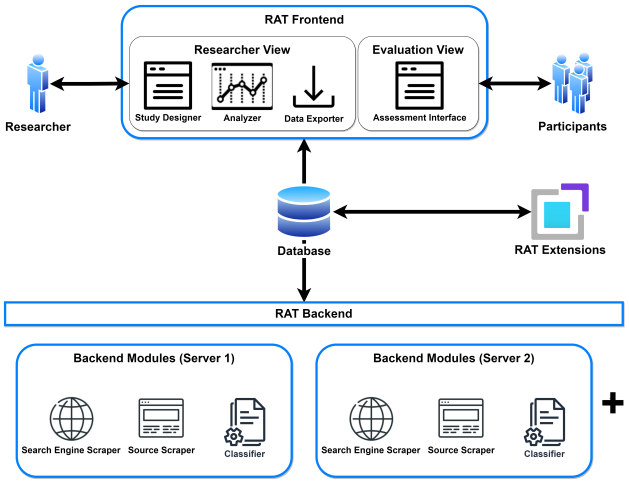

Figure 1: Software architecture of RAT [15]

2.1 Frontend Application

The RAT Frontend is a Flask-based GUI that provides two distinct views for the system's different user groups. The Researcher View serves as the central control panel for managing a study. It includes the Study Designer, a module where researchers define all study parameters, such as result types (e.g., snippets only, full-page archival, AI-generated content) and query lists. This module also integrates the Query Sampler, a tool to systematically generate query sets using the Google Ads API [12]. Within the Researcher View, researchers can create questionnaires for jurors that support a wide range of question types, including Likert scales, open-ended questions, and sliders. The Analyzer module computes and displays real-time statistics on collection progress, participant answers, and classifier results. Finally, the Data Exporter allows downloading all collected data in XLSX format.

The Evaluation View is the assessment interface for study participants (jurors). Participants can register via a unique URL, enabling anonymous data collection. Jurors are presented with an archived screenshot of a search result (i.e., a web page) or an AI-generated answer, along with the corresponding questionnaire, ensuring that all participants evaluate the exact same content.

2.2 Backend Application and Data Collection

All modules in the backend, including the Search Engine Scraper, Source Scraper, and Classifier, are controlled by a job management system (see Figure 2), which ensures that all necessary processes are executed in the background using the Advanced Python Scheduler (APScheduler). This system sets up jobs for scraping and classifying data in database tables, tracks their status (running, failed, completed), and automatically repeats failed jobs up to a specified maximum number of attempts to ensure that they are executed correctly [15].

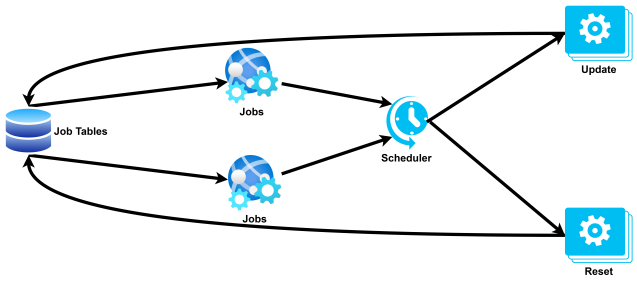

Figure 2: Job management system in RAT [15]

Data collection is a two-step process: First, the Search Engine Scraper collects metadata (title, URL, snippet) from SERPs or AI-generated content. Maintaining scrapers is challenging, as search engines frequently change their layouts. To address this, RAT employs multiple technical collection methods while respecting website policies and without attempting to bypass protective measures like CAPTCHAs. This includes a traditional backend collector using Selenium WebDriver[1] in headless mode to scrape without rendering web pages, enabling parallel scraping on web servers in the background, as well as a new, in-development browser extension. This extension integrates with the RAT platform but runs in a standard web browser to more closely mimic human interaction.

Second, the Source Scraper retrieves the full content of each collected URL. We also use Selenium WebDriver for this task. It can interact with dynamic web pages (e.g., by executing JavaScript or scrolling), which is essential for capturing the full content of modern websites. Simpler tools, such as BeautifulSoup[2] or cURL[3], are inadequate because they do not support server-side rendering. While scraping search results is the primary use case, RAT also

[1] https://www.selenium.dev/
[2] https://pypi.org/project/beautifulsoup4/
[3] https://curl.se/

includes scrapers to capture AI-generated content, including Google AI Overviews and responses from chatbots such as ChatGPT.

2.3 Classification and System Extensibility

RAT is designed as an extensible platform. The classification module allows researchers to add their own automated analysis scripts. These classifiers can use the data collected for each result, including the URL, domain, position, snippet, complete source code, and original search query. Classification results are stored as key-value pairs in the database. A pre-implemented SEO classifier [14] that estimates the probability of search engine optimization for a web page is already included. Furthermore, RAT supports community contributions via a dedicated GitHub repository for RAT extensions[4]. This allows researchers to develop and share new modules that integrate with the RAT database.

3 Application Scenarios and Impact

The broad applicability of RAT is demonstrated by its successful application across various scientific disciplines, indicating that its usefulness extends beyond information retrieval (IR). For traditional information retrieval, RAT's original use case was large-scale comparative retrieval audits, such as a comparison of Google and Bing that utilized crowdsourcing to evaluate results for a set of 1,000 queries [6]. RAT has been used to conduct large-scale studies of source diversity. For instance, it was used to collect and analyze 141,480 results from Google, Bing, DuckDuckGo, and MetaGer to compare the overlap of their results and the diversity of domains presented to users [18]. Following this, it was applied to analyze the visibility of extreme-right content in Google's results for sensitive queries in Germany and Sweden [10].

Other examples of research include the study by Haider et al. [7], which used the Query Sampler [12] to generate search queries related to the keyword "vindkraft", the Swedish word for wind power. In another study by Rödl and Haider [11], RAT was used to scrape keyphrases from the Swedish climate obstruction network on Google and DuckDuckGo. In political science, the automated classifier was used to analyze the use of search engine optimization (SEO) by political candidates in the 2021 German federal election. A total of 1,372 websites were classified to measure the use of optimization techniques [5].

RAT is not commercial, but rather methodological and technical. Its primary technical impact is providing a single, sustainable, and extensible framework. This framework is what enables scalable, replicable audits of live information systems and search engines. Its research impact is demonstrably interdisciplinary, as evidenced by the studies mentioned. The tool is already in active use, as multiple studies have been published.

4 Significance and Comparison to Related Work

The novelty of RAT lies in its sustainable integration of the entire research workflow, closing a gap in the research software ecosystem [15]. The tool positions itself between three categories of existing software, each with significant limitations for scientific research.

First, there are raw extraction libraries such as Selenium, Scrapy[5], and Playwright[6]. While these provide the technical capability to render and interact with web pages, they are not complete research instruments. They lack the necessary infrastructure for scientific accuracy, such as reproducible job scheduling, error recovery, data management, and separation of user roles, forcing researchers to rebuild these complex systems from scratch for every study.

Second are commercial marketing suites and proprietary data-as-a-service platforms like Ahrefs [20], ScraperAPI [13], and Screaming Frog [19]. These services are not built for academic research but for marketing and SEO. More critically, their data collection methods are opaque, and their terms of service often explicitly prohibit data storage, sharing, and deep analysis.

Third are previous academic tools (e.g., [1, 3, 16]), which provided valuable contributions but often suffer from a lack of sustainability. Typically designed for a single study, they are rarely maintained or extensible. RAT addresses this by offering a modular, community-extensible framework with a long-term organizational structure, ensuring it adapts as web technologies evolve.

RAT, by contrast, is designed to solve the limitations of all three categories. It is an open-source, integrated environment built by researchers for researchers. It provides the complete, sustainable infrastructure that standalone libraries lack, offers the methodological transparency that commercial tools hide, and is built for long-term maintenance and extensibility in a way single-study academic tools are not. Its ongoing developments, such as the new browser extension that captures localized and personalized results more accurately, demonstrate an evolving platform that adapts to new research challenges.

5 Ethical Considerations and Methodological Integrity

A fundamental design concept of RAT is to enable responsible and systematic research. This is important in a context where there is no consistent method for accessing publicly available search engine data. Recent regulations, such as the EU's Digital Services Act (DSA) [21], are intended to give researchers access to platform data. However, these regulations are often cumbersome, often restrict data access to specific topics, and do not guarantee comparable data across different platforms. The DSA is also not suitable for live observation audits. The lack of data, along with the inadequacy of public APIs, indicates that independent external research is essential.

RAT is explicitly designed to be a tool for targeted academic studies (e.g., hundreds or thousands of queries), not for mass-scale commercial data harvesting [15]. We employ several technical and procedural measures to ensure methodological and ethical integrity. The system's job manager distributes requests over time and can be configured with delays to prevent server overload. With respect to anti-bot measures like CAPTCHA, RAT does not attempt to break them; the job simply fails and is rescheduled for later. This approach respects the protective measures of search systems and websites while enabling necessary academic research.

[4] https://github.com/rat-software/rat-extensions

[5] https://scrapy.org/

[6] https://playwright.dev/

Beyond ethical data collection, RAT is designed to support core scientific quality criteria in an environment where traditional validity checks are difficult. To ensure reliability, the automated Job Manager eliminates human error in data collection timing and execution, ensuring uniform conditions across all queries. Validity is prioritized by capturing the full DOM and visual rendering of results (via screenshots and source code) exactly as presented to users, rather than relying on simplified text data from APIs. Finally, regarding sampling adequacy, RAT integrates the Query Sampler. This allows researchers to construct systematic, representative query sets based on real-world search volume data, ensuring that audits are grounded in actual user behavior rather than arbitrary keyword selection. We argue that the ethical imperative for transparency and independent auditing of public-facing information access systems, especially when no sufficient, official data access exists, necessitates a responsible, methodological tool like RAT.

6 Demonstration Plan

Our demonstration is designed to be modular, allowing attendees to focus on the workflow stages most relevant to them on a live instance accessible online[7]. To demonstrate the Researcher View and analysis capabilities, we will showcase the study configuration using the Query Sampler. To ensure a smooth experience without waiting for live scraping times, we will utilize a pre-computed dataset. This allows us to immediately demonstrate the Analyzer View, highlighting visualizations of source domain overlap and results from the automated SEO classifier without delay. Furthermore, we offer a streamlined hands-on experience for the Evaluation View. Attendees can use a guest login to assess 1–2 search results, which provides a quick, tactile understanding of the participant's perspective — displaying archived screenshots and questionnaires — without requiring a full study run. Finally, for interested developers, we will briefly outline the system's architecture and extensibility via our GitHub repository [7], explaining how custom scrapers or classifiers can be integrated into the RAT ecosystem.

7 Conclusion

The Result Assessment Tool (RAT) is an integrated, open-source toolkit for reproducible science. It is available both as a self-hostable package and as a ready-to-use web application, lowering the technical barrier for researchers. It fills a gap between simple scripts and commercial tools. By providing a common, extensible platform for study design, collection, assessment, and analysis, RAT empowers the broader research community to conduct the essential work of understanding the complex information systems that influence our world and their users. RAT is an actively evolving project. Future work is focused on expanding our collection capabilities (such as the new browser extension) and ensuring the long-term sustainability of the software. To this end, we are establishing a RAT Association to provide an organizational home for the tool beyond current funding, ensuring its continued development and support for the research community.

[7] https://rat-software.org/

Availability of the Software and Research Data

In compliance with the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles [2], we make the research data from our RAT studies accessible through the Open Science Framework (OSF), provide a software demonstration, offer a demo video, and make the source code available [7].

Acknowledgments

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft (DFG)); Grant No. 557366357.

References

[1] Judit Bar-Ilan and Mark Levene. 2011. A method to assess search engine results. Online Information Review 35, 6 (2011), 854–868. https://doi.org/10.1108/14684521111193166
[2] Michelle Barker et al. 2022. Introducing the FAIR Principles for research software. Sci Data 9, 1 (October 2022), 622. https://doi.org/10.1038/s41597-022-01710-x
[3] Peter Carragher, Evan M. Williams, and Kathleen M. Carley. 2024. Detection and Discovery of Misinformation Sources Using Attributed Webgraphs. ICWSM 18, (May 2024), 214–226. https://doi.org/10.1609/icwsm.v18i1.31309
[4] Nick Craswell et al. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In Proceedings of the 44th International ACM SIGIR Conference. ACM, 2369–2375. https://doi.org/10.1145/3404835.3463249
[5] Kay Hinz, Sebastian Sünkler, and Dirk Lewandowski. 2023. SEO im Wahlkampf. In Die Bundestagswahl 2021. Springer Fachmedien Wiesbaden, 1–28.
[6] Dirk Lewandowski. 2015. Evaluating the retrieval effectiveness of web search engines using a representative query sample. Journal of the Association for Information Science and Technology 66, 9 (2015), 1763–1775. https://doi.org/10.1002/asi.23304
[7] Dirk Lewandowski, Sebastian Sünkler, Nurce Yagci, Tuhina Kumar, and Kardelen Bilir. 2025. Result Assessment Tool (RAT). https://doi.org/10.5281/ZENODO.14049810
[8] Microsoft. 2025. Bing Search APIs Retiring on August 11, 2025. https://learn.microsoft.com/en-us/lifecycle/announcements/bing-search-api-retirement
[9] Tri Nguyen et al. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. CoRR abs/1611.09268. http://arxiv.org/abs/1611.09268
[10] Ov Cristian Norocel and Dirk Lewandowski. 2023. Google, data voids, and the dynamics of the politics of exclusion. Big Data & Society 10, 1 (2023). https://doi.org/10.1177/20539517221149099
[11] Malte Rödl and Jutta Haider. 2025. 'Google this!' How performative links and search engines organise information disorders in a climate obstruction network. Information, Communication & Society (2025), 1–21. https://doi.org/10.1080/1369118X.2024.2445640
[12] Sebastian Schultheiß et al. 2023. Query sampler: generating query sets for analyzing search engines using keyword research tools. PeerJ Computer Science 9, e1421. https://doi.org/10.7717/peerj-cs.1421
[13] ScraperAPI. 2025. ScraperAPI - The Proxy API For Web Scraping. https://www.scraperapi.com/
[14] Sebastian Sünkler. 2025. SEO-Effekt 2: SEO Analysis and Scoring Tool. https://doi.org/10.5281/ZENODO.17224644
[15] Sebastian Sünkler, Dirk Lewandowski, Sebastian Schultheiß, and Nurce Yagci. 2025. Result Assessment Tool (RAT): empowering search engine data analysis. PeerJ Computer Science 11, e2962. https://doi.org/10.7717/peerj-cs.2962
[16] Daniel Trielli and Nicholas Diakopoulos. 2020. Partisan search behavior and Google results in the 2018 U.S. midterm elections. Information, Communication & Society 25, 1 (2020), 145–161. https://doi.org/10.1080/1369118X.2020.1764605
[17] E. Voorhees and D. K. Harman (Eds.). 2005. TREC: experiment and evaluation in information retrieval. MIT Press.
[18] Nurce Yagci et al. 2022. A Comparison of Source Distribution and Result Overlap in Web Search Engines. Proc. of the ASIST 59, 1 (2022), 346–357. https://doi.org/10.1002/pra2.758
[19] Screaming Frog. 2025. Screaming Frog SEO Spider Website Crawler. https://www.screamingfrog.co.uk/seo-spider/
[20] Ahrefs. 2025. Ahrefs—Marketing Intelligence Tools. https://ahrefs.com/
[21] The EU's Digital Services Act. 2025. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en