

SafeLens: AI-Augmented Exploratory Analysis over Organizational Safety Records for Actionable Insights Generation

Sanjay Das
Oak Ridge National Laboratory
Oak Ridge, USA
dass3@ornl.gov

Ryan Burchfield
Oak Ridge National Laboratory
Oak Ridge, USA
burchfieldra@ornl.gov

Gavin Wiggins
Oak Ridge National Laboratory
Oak Ridge, USA
wigginsg@ornl.gov

Alan Guy
Oak Ridge National Laboratory
Oak Ridge, USA
guyas@ornl.gov

Tirthankar Ghosal
Oak Ridge National Laboratory
Oak Ridge, USA
tirthankar.slg@gmail.com

Abstract

Safety programs at research institutions, industrial facilities, and public agencies routinely accumulate tens of thousands of free-text observation records per year, documenting unsafe behaviors, near-misses, safety conversations, and best practices, yet these records are seldom analyzed beyond simple counts. We present SAFELENS, an interactive system for AI-augmented exploratory analysis of large-scale organizational safety observation corpora. SAFELENS addresses this gap through a four-stage pipeline: (1) multi-dimensional structured filtering over a Parquet-encoded corpus of 468,397 real safety records; (2) stratified adaptive sampling (SAQS) that guarantees representation of all interaction-type strata; (3) large language model (LLM) inference that returns ranked insight objects in three categories: *riskiest behaviors*, *common patterns*, and *best practices*; and (4) an interactive Streamlit dashboard with statistical charts, scrollable insight panels, a data viewer, and a conversational chat interface with persistent session memory. The demonstration allows users to issue live queries against the full corpus, inspect individual records, rate insight quality, and observe how the system’s sampling strategy affects result coverage. A built-in evaluation dashboard captures all human-AI interaction metrics in real time. A full demonstration of the system is presented in the supplementary video here¹.

CCS Concepts

• **Information systems** → **Data analytics**; *Information extraction*; *Digital libraries and archives*.

Keywords

Safety analytics, LLM-augmented retrieval, organizational knowledge mining, digital libraries, exploratory data analysis, stratified adaptive sampling.

¹<https://tinyurl.com/5n6bys7w>

This paper is authored by an employee(s) of the United States Government and is in the public domain. Non-exclusive copying or redistribution is allowed, provided that the article citation is given and the authors and agency are clearly identified as its source. Request permissions from owner/author(s).

JCDL '26, Frisco, TX, USA

2026. ACM ISBN 979-8-4007-2597-5/2026/10
<https://doi.org/10.1145/3805696.3846509>

ACM Reference Format:

Sanjay Das, Ryan Burchfield, Gavin Wiggins, Alan Guy, and Tirthankar Ghosal. 2026. SafeLens: AI-Augmented Exploratory Analysis over Organizational Safety Records for Actionable Insights Generation. In *The 2026 ACM/IEEE Joint Conference on Digital Libraries (JCDL '26)*, October 13–16, 2026, Frisco, TX, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3805696.3846509>

1 Introduction and Motivation

Workplace safety observation programs generate continuous streams of human-authored records that document operational hazards, behavioral deviations, and procedural compliance in field environments. At scale, these programs produce repositories of considerable depth: a single large institution may accumulate hundreds of thousands of free-text observations spanning multiple years, buildings, divisions, and organizational levels. Yet despite this accumulation, these corpora remain largely unmined. Periodic reports count observations by building or week; rare manual reviews sample a handful of records; and the semantic content of the majority such as what workers actually said, what behaviors were observed, what patterns recur, goes unexamined.

The gap is not a data problem. The institution whose corpus underpins SAFELENS has collected **468,397 safety observations** over 11 years, spanning several buildings, divisions and directorates. The gap is an **analysis infrastructure** problem: existing tools for digital collections such as keyword search, document management systems, are not designed to surface semantic insights from mixed structured-unstructured corpora at organizational scale.

Three failure modes characterize current approaches. *Business intelligence dashboards* count and chart structured metadata but cannot interpret free-text content; they answer “how many” but not “what” or “why.” *Manual expert review* is precise but does not scale: an analyst reviewing records individually cannot cover 468K entries on any useful cadence. *Naïve LLM application* fails because few current models accept a corpus of this size in a single context window, and random sampling systematically under-represents rare but high-consequence interaction types. Furthermore, even if a model possesses a sufficiently large context window to accommodate the entire dataset, ingesting the whole corpus simultaneously introduces “lost in the middle” or “needle in a haystack” problems, preventing the model from reliably identifying and analyzing critical patterns buried deep within the text.

SAFELENS addresses these failure modes through a pipeline that combines principled data management (structured filtering, stratified sampling) with LLM semantic mining. The result is a system that delivers organizational safety insights, ranked by risk, frequency, or exemplary practice, in under 60 seconds, at any granularity the user specifies.

The demonstration is practically motivated: Safety managers at an institution require a scalable tool for periodic review. The proposed approach is also scientifically grounded: the system instantiates a query model, a sampling algorithm, and a consistency metric for validating insight hierarchies. This paper describes the system as a demonstration vehicle for the Joint Conference on Digital Libraries (JCDL) community, which has a long-standing interest in tools for large-scale document collection analysis, knowledge extraction from institutional archives, and human-AI collaboration over digital corpora [4, 5, 9].

2 Related Work

Digital library analytics. The challenge of extracting actionable knowledge from large institutional document collections is central to the digital libraries community [2]. Prior work has addressed automatic metadata enrichment [10], topic modeling over scholarly corpora [3], and faceted search over heterogeneous collections. **SAFELENS** extends this tradition to operational free-text records, where the extraction target is not topical classification but *organizational actionability* with pattern identification, and best-practice recognition.

LLM-augmented information access. Large language models have been applied to question answering over document collections [11], structured data generation from unstructured text [13], and exploratory analysis assistance [8]. A key challenge unaddressed by these systems is the token budget problem: how to select a representative, semantically-diverse sample from a corpus larger than any model’s context window. **SAFELENS**’s Stratified Adaptive Query Sampling algorithm (SAQS) addresses this directly by ensuring all interaction types appear in the sample sent to the LLM, unlike uniform random sampling which can omit rare but high-consequence categories entirely.

Safety analytics. Behavior-based safety (BBS) research [12] focuses predominantly on data collection protocols rather than computational analysis at scale. **REFSafe**[6] has demonstrated that AI augmented vulnerability analysis can enhance organizational safety by incorporating historical incident corpora. Similarly, **ORCHID** [7] demonstrated capability of automated agentic systems in handling complex policy documents and performing high-risk property identification. However, No prior system, to our knowledge, applies LLM-based semantic mining to large institutional safety observation corpora.

Human-AI interaction over collections. Systems such as **LIDA** [8] and conversational search interfaces demonstrate the value of iterative, natural language-driven exploration of data collections. **SAFELENS** contributes a session-persistent chat interface grounded in live filter context and statistical summaries, enabling follow-up questions without re-running the full analysis.

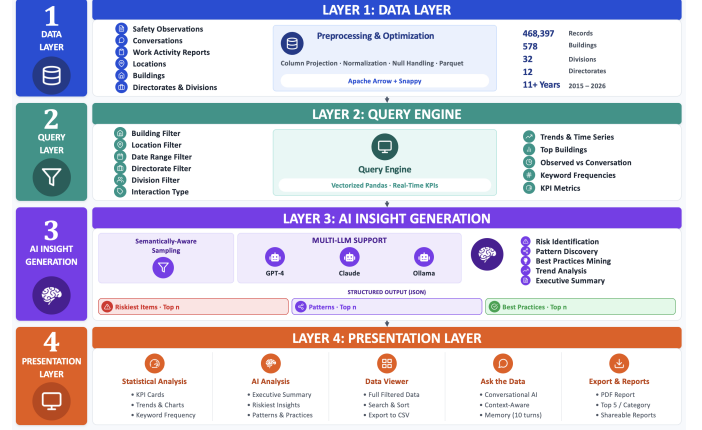


Figure 1: **SAFELENS four-layer architecture. The deterministic pipeline (Layers 1–2) executes instantly; LLM inference (Layer 3) completes in 15–40 seconds depending on provider.**

3 Problem Statement

In this section, we define the problem statement formally and explain the technical challenges that we face in our endeavour.

Let R be a relation of safety observation records with schema $\langle id, conv, bldg, loc, date, itype, obs \rangle$, where $conv \in \Sigma^*$ is a free-text conversation record, $bldg, loc, itype$ are categorical dimensions, $date$ is a timestamp, and $obs \in \{0, 1\}$ flags whether the record reflects a direct field observation or a safety conversation.

A **Safety Insight Query** (SIQ) is a tuple $Q = \langle \sigma, gran, k, \tau \rangle$ where σ is a structured filter predicate over dimensions (building, date range, interaction type), $gran$ specifies an aggregation granularity (Building | Division | Directorate | All), k is the desired result count per category, and $\tau \in \{RISK, PATTERN, BESTPRACTICE, ALL\}$ is the insight type.

The answer to Q over R is a ranked list of k typed *insight objects* $I_Q = [i_1, \dots, i_k]$ grounded in the filtered slice $R_\sigma = \sigma(R)$ and aggregated at granularity $gran$. Each insight i carries a title, description, representative location, estimated frequency, and a supporting verbatim excerpt.

The core technical challenge is that $|R_\sigma|$ may be large (10^3 – 10^5) while LLM context budgets admit at most $N \ll |R_\sigma|$ records. Selecting the N -record sample S that maximizes semantic *coverage* (i.e., ensures all interaction-type strata are represented) is a data management problem that we solve with our SAQS algorithm (Section 4).

4 System Architecture

Figure 1 illustrates the four-layer **SAFELENS** architecture. We describe each layer in turn.

4.1 Layer 1: Data Layer

The full 468,397-record corpus is preprocessed once by a developer-executed script that projects the seven semantically meaningful columns, normalizes date fields, fills null conversation entries, and serializes the result as a Parquet file using Apache Arrow columnar encoding with Snappy compression. This reduces the dataset

from the original XLSX format to a fraction of its size and enables full-corpus load in under two seconds at serving time. At runtime the application reads only the Parquet file; the original spreadsheet is not required. Streamlit’s `@cache_resource` decorator pins the loaded DataFrame in server memory for the duration of a deployment session.

4.2 Layer 2: Query Layer

Given a user-specified predicate σ with building selection, date range, and optionally interaction type, the query layer evaluates σ against the in-memory DataFrame using vectorized pandas operations, returning R_σ in under 5 ms for all tested filter combinations on the full 468K corpus. Streamlit’s `@cache_data` decorator memoizes results for identical predicates, making repeat queries instantaneous. Alongside the filtered slice, the query layer computes a suite of deterministic statistics: observation counts, observed-versus-conversation splits, top buildings and locations by frequency, keyword distributions, and temporal histograms.

4.3 Layer 3: Inference Layer — SAQS and LLM Mining

Given R_σ and a token budget N (default 300 records), SafeLens applies the **Stratified Adaptive Query Sampling** (SAQS) algorithm. SAQS uses the existing Interaction Type field as a proxy for semantic category: records sharing the same type tend to describe related safety phenomena, so stratifying by this field maximizes thematic diversity in the sample without requiring a separate embedding step. SAQS partitions R_σ into strata $\{C_1, \dots, C_m\}$, allocates sample slots proportionally to stratum size while enforcing a minimum floor $n_{\min} = 5$ per stratum, scales down if the total exceeds N , and samples uniformly within each stratum. In contrast, uniform random sampling of N records allocates expected count $N \cdot |C_i|/|R_\sigma|$ to stratum C_i , which approaches zero for rare types and produces insight lists that systematically omit low-frequency but operationally critical categories.

The resulting sample S is serialized into a structured prompt containing: (a) statistical context derived from Layer 2; (b) one line per sampled record encoding its metadata and conversation text; and (c) a JSON output schema specifying the required structure of each insight object. The prompt is dispatched to a configurable LLM provider (OpenAI GPT-4, or a locally-hosted Ollama model). The response is parsed into three typed arrays: `riskiest`, `common_patterns`, `best_practices`.

4.4 Layer 4: Presentation Layer

The Streamlit dashboard exposes four tabs. *Statistical Analysis* presents the deterministic Layer 2 outputs: KPI cards, an observations-over-time area chart, bar charts for top buildings and locations, a donut chart for observed-versus-conversation splits, and a keyword frequency chart. This tab requires no LLM call and returns instantly. *AI Analysis* displays the LLM-generated executive summary, followed by three side-by-side scrollable panels for the ranked insight categories. Each insight card shows title, description, location, frequency, and a representative verbatim excerpt from the corpus. A PDF report download bundles statistics and top-5 insights per category into a formatted document. *Data Viewer* exposes the full

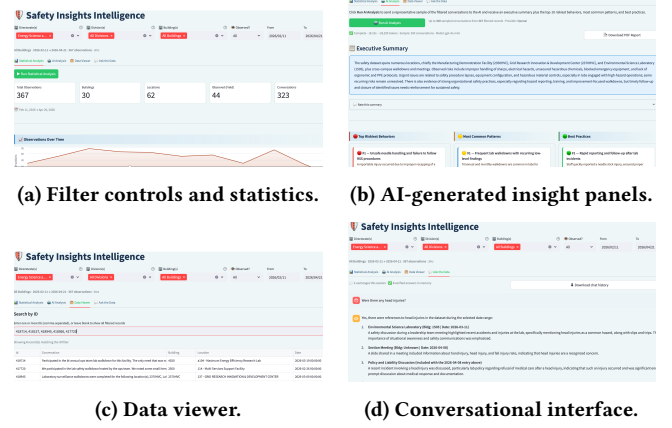


Figure 2: SAFELENS interface. (a) Users set filters at the top of the page; the filter bar is always visible. (b) AI insight panels appear after clicking “Run AI Analysis.” (c) Statistical analysis is available instantly without any LLM call. (d) The chat interface maintains session history and injects live filter context into every response.

R_σ slice in a sortable, scrollable table with an ID-based search filter. *Ask the Data* tab provides a session-persistent conversational interface: the last ten turns of conversation history, the current filter context, Layer 2 statistics, and any AI Analysis summary are all injected as system context for each LLM call. Every AI interaction including run metadata, response latency, sample size, provider, and user thumbs-up/down feedback, is logged to a local SQLite database and visualized in a separate *Evaluation* dashboard page.

5 Demonstration

The demonstration executes on a standard, internet-enabled client system (e.g. Laptop) utilizing a locally stored corpus of 468,397 records. To ensure strict data privacy, no sensitive, record-level information leaves the host machine. Artificial intelligence capabilities are facilitated either on-premises using Ollama or externally via a secure Azure OpenAI API endpoint, ensuring full compliance with data protection standards. Figure 2 shows the system interface. We walk attendees through four scripted scenarios and invite open exploration.

Scenario 1 — Granularity Drill-Down. The demonstrator begins with a directorate-wide query (all 468K records, unrestricted dates) to establish the global risk profile. The attendee then progressively narrows the filter to a single building and a single quarter, observing how the ranked risk list updates. This illustrates the SIQ query model in practice and motivates the need for multi-granularity analysis.

Scenario 2 — Sampling Strategy Comparison. Running the same query with SAQS versus uniform random sampling at $N = 100$, the demonstrator shows side-by-side how rare interaction types vanish from the random sample but remain represented in the SAQS result.

Scenario 3 — Temporal Risk Trend. Four consecutive quarterly queries on a high-activity building generate top-3 risk lists

per quarter. Displayed as a temporal sequence, this reveals whether identified risks from an earlier period were subsequently addressed, which are making it the most operationally actionable scenario for safety managers.

Scenario 4 — Conversational Follow-Up. After generating an AI insight set, attendees pose natural language follow-up questions through the chat interface: “Are any of these risks seasonal?”; “Which building shows the most improvement?”; “Compare the risk profiles of two buildings.” Session history is maintained across turns, and the AI has access to the current filter context and statistical summary without the user needing to re-specify them.

Open exploration. Attendees are invited to formulate their own queries using any combination of the building and date filters, trigger a live AI analysis, and rate insight quality using the thumbs-up/down buttons. All interactions are logged to the evaluation dashboard, which is displayed on a second screen throughout the demonstration. Table 1 lists the six pre-scripted queries available at the demonstration station.

Table 1: Pre-scripted queries at the SAFELENS demo station.

ID	Scope	Description
D1	All, all time	Directorate-wide risk profile
D2	Top building, 2023	Single-site annual analysis
D3	Top bldg, Q1 vs Q3	Quarterly risk comparison
D4	All, observed only	Field observation risks only
D5	All, conversations	Conversation-reported patterns
D6	Two buildings, 2024	Cross-site best-practice sharing

6 Discussion: Usage and Feedback

SAFELENS has been used by safety managers at the host institution across multiple review cycles. Three patterns of use have emerged.

Insight as a starting point, not an endpoint. Managers consistently use the AI-generated risk list not as a final answer but as a ranked agenda for deeper investigation. They accept the top-3 risks as plausible, investigate the supporting record excerpts, and sometimes promote a lower-ranked item based on domain knowledge. This collaborative modality where, AI narrows the search space and human expert decides, aligning with best practices in human-AI decision support [1].

Statistical analysis satisfies many queries without LLM. In a substantial fraction of sessions, users found the keyword frequency chart and temporal histogram sufficient to answer their question without invoking the LLM. The two-layer design (instant deterministic analysis always available, LLM on demand) proved well-matched to real usage patterns. Users routinely begin with the Statistical Analysis tab, form a hypothesis, and only then run the AI Analysis for confirmation or elaboration.

The observed/unobserved distinction matters. Early deployments mixed field observations and safety conversations without stratification. Users reported that the resulting insight lists conflated two fundamentally different safety data types. The current system’s

interaction-type stratification (via SAQS) and the observed-versus-conversation filter address this directly and were specifically cited in post-session feedback as essential for interpretability.

“This is the best tool I have used in the last year and a half. While future upgrades will come, gathering feedback from a broader user base right now would be invaluable.” — Expert,

Integrated Research Operations

Feedback metrics. The evaluation dashboard captures granular, multi-dimensional telemetry across all system interactions. By monitoring critical metrics, including latency, sample size, model provider, and user feedback, the platform ensures continuous operational visibility and comprehensive auditing capabilities.

An analysis of 136 total runs demonstrates strong user approval alongside robust operational stability. The evaluation spans two distinct operational phases. The initial phase comprised 69 runs leveraging the GPT-4o-mini model, followed by a subsequent phase of 67 runs using the GPT-5.6 Luna engine. While the initial phase recorded a faster average response time of 17.13 seconds, the subsequent phase observed an increased latency profile of 21.92 seconds.

However, this increase in processing time resulted in a substantial gain in architectural reliability. System errors, which is defined strictly as any non-standard behavior including blank outputs, workflow disruptions, or system freezes, dropped significantly from 8.7% in the initial phase down to 4.5% in the subsequent phase. A complete breakdown of these comparative performance characteristics is detailed in Table 2.

User sentiment remains exceptional across both evaluation periods. Across the 20 total explicit feedback submissions gathered via the dashboard’s inline thumbs-up and thumbs-down utility, the system achieved a 100.0% user satisfaction rate. Every recorded user submission validated the accuracy, relevance, and presentation of the generated insights, indicating that the system’s outputs consistently align with user expectations despite varying latency profiles.

Table 2: Evaluation Dashboard Performance Metrics.

Metric	Performance
Total Evaluation Runs	136
Average Latency (seconds)	19.49
System Error Rate (%)	6.6%
User Feedback Submissions	20
User Satisfaction Rate (%)	100.0%

7 Conclusion

We present SAFELENS, a deployed, interactive system for AI-augmented exploratory analysis over a 468,397-record organizational safety observation corpus. The system demonstrates that principled data management (structured filtering, stratified adaptive sampling) rather than LLM capability alone, determines the quality and reliability of insights extracted from large free-text institutional collections. For the JCDL community, SAFELENS offers a concrete instantiation of how digital library infrastructure principles including

efficient retrieval, faceted access, provenance-aware result presentation, can be extended to AI-augmented semantic mining over operational document corpora. The demonstration invites the community to explore a live system on real institutional data and to reflect on the broader implications for knowledge management in organizations.

8 Future Work

Future work proceeds along four directions. First, we plan to extend SAQS to budget-optimal allocation under variable document length, replacing the current record-count budget with a token-count budget for tighter integration with model context limits. Second, we are exploring fine-tuning a domain-adapted language model on the safety corpus to improve sensitivity to institution-specific terminology and reduce hallucination in rare-event descriptions. Third, we will investigate differential privacy mechanisms to enable broader data sharing for research purposes while protecting individual reporter identity. Finally, we aim to generalize the SIQ query model beyond safety observation records to other operational institutional corpora including maintenance logs, audit findings, HR observations, where the same mixed structured-unstructured analysis challenge applies.

9 Use of Large Language Models

LLM were used for editing purposes. All LLM-generated materials and outputs were reviewed, filtered, and analyzed by the authors.

Acknowledgments

This research is sponsored by the Office of the Laboratory Director, Oak Ridge National Laboratory's Operational Excellence Initiatives,

which is supported by the United States Department of Energy (DOE)'s Office of Science under Contract No. DE-AC05-00OR22725.

References

- [1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *Proceedings of the International Conference on Software Engineering (ICSE): Software Engineering in Practice*. 291–300.
- [2] William Y. Arms. 2000. *Digital Libraries*. MIT Press.
- [3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.
- [4] Christine L Borgman and Heather Hessel. 2001. Report on the First Joint Conference on Digital Libraries. *D-Lib Magazine* 7, 10 (2001), 1082–9873.
- [5] Gregory Crane. 1996. Building a digital library: The Perseus Project as a case study in the humanities. In *Proceedings of the first ACM international conference on Digital libraries*. 3–10.
- [6] Sanjay Das, Ran Elgedawy, Ethan Seefried, et al. 2026. REFSafe: A RAG-Enabled Framework for Predictive Risk Analysis and Automated Safety Report Generation in Mission-Critical Environments. In *Proceedings of the 1st Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports 2026)*. 47–56.
- [7] Sanjay Das, Maria Mahbub, Vanessa Lama, et al. 2026. ORCHID: Orchestrated Retrieval-Augmented Classification of High-Risk Property with Intelligent Decision-Making. In *Proceedings of the 1st Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports 2026)*. 57–64.
- [8] Victor Dibia. 2023. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics Using Large Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*. 113–126.
- [9] Edward A. Fox. 1993. Digital Libraries. *Commun. ACM* 36, 11 (1993), 22–28.
- [10] Jane Greenberg. 2003. Metadata and the World Wide Web. *Encyclopedia of Library and Information Science* 3 (2003), 1876–1888.
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 9459–9474.
- [12] Richard C. McNamee and Sarah A. Corbet. 2019. Behavior-Based Safety: A Practical Approach. *Professional Safety* 64, 3 (2019), 28–34.
- [13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 27730–27744.