Bayer Oy

On-demand visualization of HSE data

Hackathon Challenge SinceAI, Turku

Martina Fabiani, Alessio Franchini, Christian Petruzzella, Niccolò Settimelli

Bayer Oy

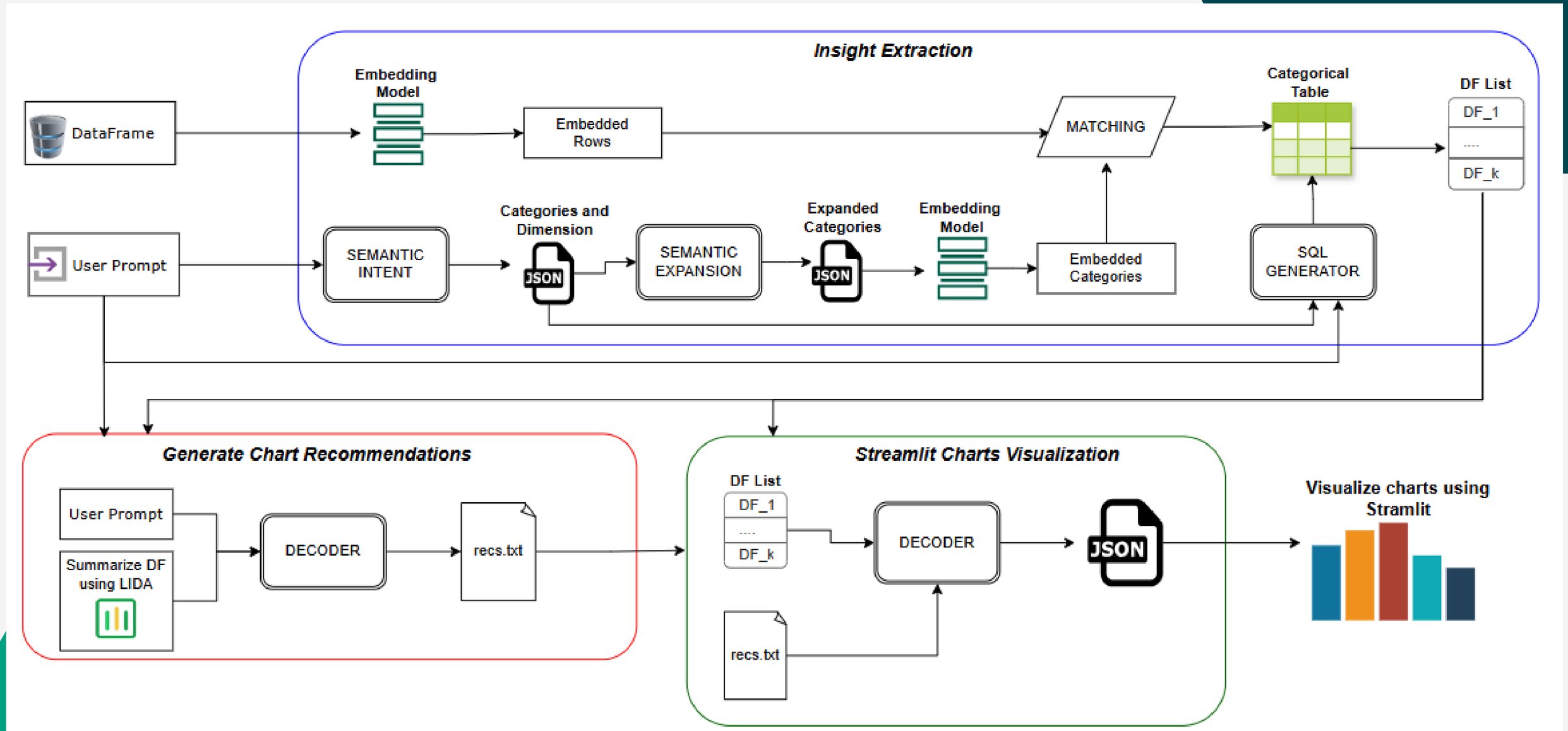
Project Overview

- AI-driven framework for automated insight extraction and chart recommendation from HSE datasets.
- Pipeline integrates:
  - Semantic Intent Parsing
  - Embedding-based Category Expansion & Matching
  - Dataset Profiling (LIDA)
  - LLM-powered Visualization Recommendations
  - Streamlit-based Chart Rendering
- Supports business users in generating meaningful safety analytics with minimal manual effort

Motivation & Business Value

- Safety teams need fast, accurate analytics from large multi-source datasets.
- Manual interpretation of categories, filters, and visualizations is slow and error-prone.
- The **pipeline enables**:
  - Consistent visual analytics across teams
  - Faster understanding of safety trends and anomalies
  - Reduced bottlenecks in reporting
  - Increased reliability of category matching and filtering

Workflow Pipeline



Insight Extraction Module

- A user prompt rarely maps directly to dataset columns.
- Natural language contains synonyms, implicit filters, incomplete references.
- **Semantic Intent Mapping** ensures:
  - Correct identification of dimensions (e.g., “events”, “locations”)
  - Extraction of time windows (e.g., “in 2024”)
  - Recognition of metrics (e.g., “proportion”, “trend”, “comparison”)
- **Result**: dramatically fewer interpretation errors.

Insight Extraction Module

- HSE datasets contain hundreds of categories, often inconsistent or domain-specific.
- Users describe categories with different words than those in the data.
- We embed every dataset row across all dimensions and match it with the embedding of the user's category
- **Embeddings allow:**
  - Semantic matching of user terms to dataset labels
  - Expansion to related terms or misspellings
  - Capturing hidden relationships not present in the raw text
- This increases **robustness** and makes the system work on real, messy data.

Dataset preview

	Title	Observation
0	Water puddle in corridor	There was a small puddle of water on the floor of the packaging departme
1	Loose handrail on stairs	The handrail on the metal stairs between production areas was dangerous
2	Inadequate protective equipment	An employee was handling a powdery substance in the cleanroom without
3	Items left on stairs	Empty boxes had been left on the top landing of the stairs leading to the w
4	Chemical spill on floor	A small amount of unknown chemical was on the laboratory floor with no

+ PROMPT =

Row_id	Observation_date	Processed_date	Time	Department	Observation_type	Processing_time_days	Event_year	Event_month
0	2024-01-10	2024-01-11		warehouse	hazard_report	1	2024	1
1	2024-01-18	2024-01-19		logistics	near_miss	1	2024	1
2	2024-01-25	2024-01-26		production	safety_observation	1	2024	1
3	2024-02-02	2024-02-03		warehouse	near_miss	1	2024	2
4	2024-02-09	2024-02-10		production	safety_observation	1	2024	2
5	2024-02-16	2024-02-17		production	safety_observation	1	2024	2
6	2024-02-23	2024-02-24		maintenance	unsafe_act	1	2024	2
7	2024-03-01	2024-03-02		warehouse	unsafe_act	1	2024	3
8	2024-03-08	2024-03-09		logistics	near_miss	1	2024	3
9	2024-03-15	2024-03-16		engineering	incident	1	2024	3
10	2024-03-22	2024-03-23		facility_manager	incident	1	2024	3
11	2024-03-29	2024-03-30		production	quality_issue	1	2024	3
12	2024-04-05	2024-04-06		maintenance	maintenance_request	1	2024	4
13	2024-04-12	2024-04-13		production	incident	1	2024	4
14	2024-04-19	2024-04-20		production	incident	1	2024	4
15	2024-04-26	2024-04-27		production	environmental_observatic	1	2024	4
16	2024-05-03	2024-05-04		administration	unsafe_act	1	2024	5

CATEGORICAL TABLE

SQL Aggregations for the Chart Recommendations

Observation_type	Avg_processing_time_days
first_aid_case	3
	2.43
unsafe_act	2.14
safety_observation	2.14
incident	2.11
near_miss	2.09
unsafe_condition	2
quality_issue	2
hazard_report	2
maintenance_request	1
environmental_observation	1

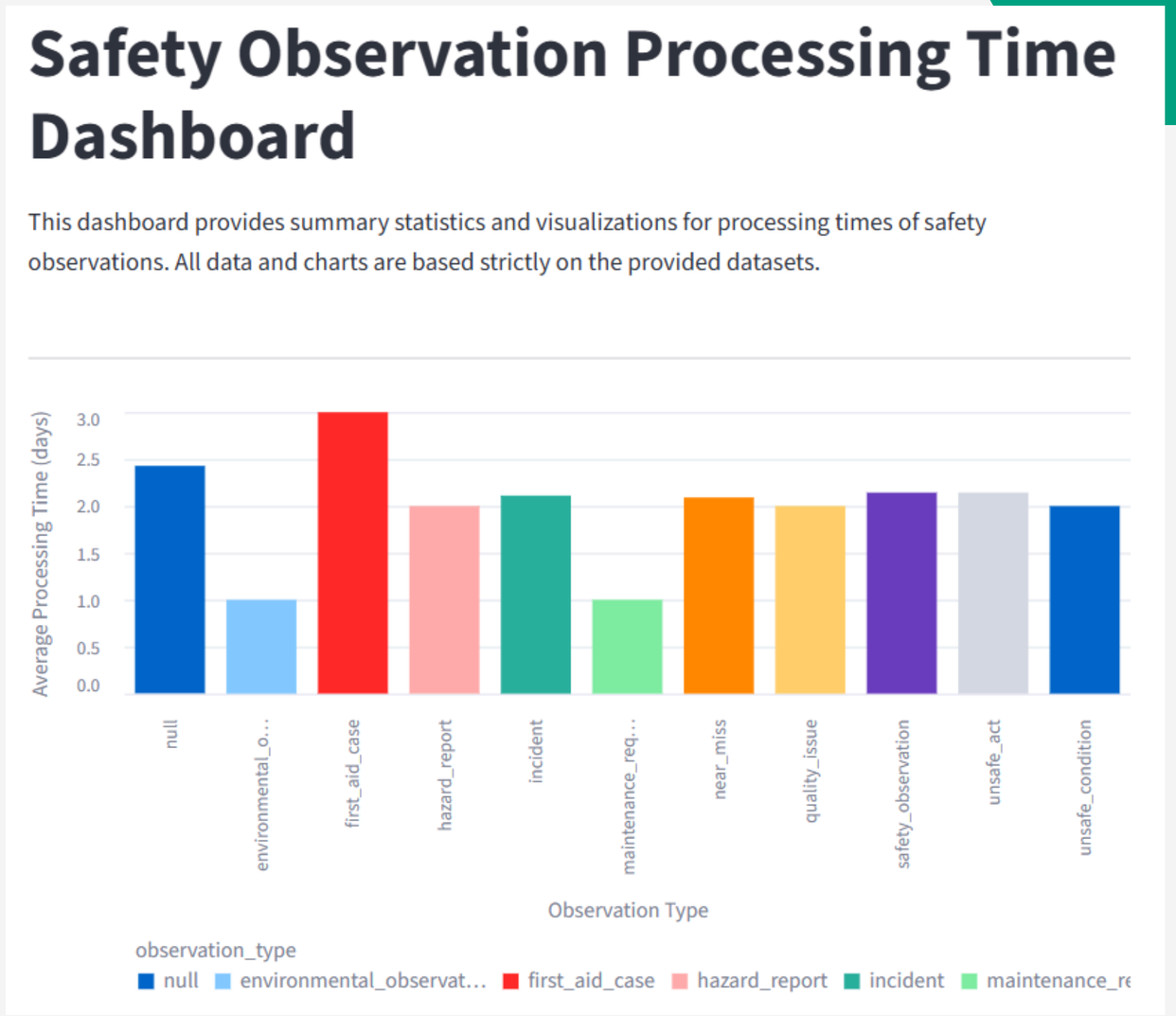


Chart Recommendation Engine

- **LIDA** extracts:
  - Schema
  - Value ranges
  - Category frequencies
- These insights **help the LLM**:
  - Understand what columns actually represent
  - Avoid making up fields
  - Suggest realistic visualizations
- Increases accuracy and reduces hallucinations.

Why Separate Recommendation from Visualization?

- Decoupling the two steps provides:
  - **Flexibility** (any frontend can consume the recommendations)
  - **Consistency** (same logic for all dashboards)
  - **Reusability** (the recommendation engine can be used outside Streamlit)
- The system produces a neutral **recs.txt** specification that any renderer can interpret.

Future Improvements

- Enhanced chart rendering with automatic layout optimization and color logic (severity-based).
- Feedback loop to let users rate chart quality and improve recommendations over time.
- Domain-tuned embeddings for stronger matching of HSE-specific categories and terminology.
- Expanded visualization types (risk heatmaps, anomaly timelines, incident funnels).
- Full Streamlit integration for dynamic dashboards and multi-chart reports.
- Multi-language support for prompts and dataset labels.

Business Impact

- **Faster insights:** charts and analysis generated in seconds instead of hours
- **Higher accuracy:** consistent, error-free interpretation of HSE data
- **Lower workload:** automation reduces manual reporting and repetitive tasks
- **Accessible analytics:** anyone can request charts without technical skills
- **Better decisions:** clearer visibility on risks and trends for safer operations