CAAP Quarterly Report

12/28/2025

Project Name: A Framework and Integrated Solution of a Dynamic Pipeline Hazard and Risk Data Repository for All Pipelines

Contract Number: 693JK32450004CAAP

Prime University: University of Dayton

Prepared By: Hui Wang, Ph.D., hwang12@udayton.edu, 937-229-2039 (University of Dayton)

Hao Wang, Ph.D., hw261@soe.rutgers.edu, 848-445-2874 (Rutgers University)

Lei Wang, Ph.D., wang4li@ucmail.uc.edu, 513-556-7649 (University of Cincinnati)

Homero Castaneda, Ph.D. hcastaneda@tamu.edu, 979-458-9844 (Texas A&M University)

Reporting Period: 10/01/2025 – 12/31/2025

1. Project Activities for Reporting Period:

Items Completed During this Quarterly Period:

Per the proposal, progress of Task 4 and Task 5 are associated with the fifth quarterly report. The following activities have been completed.

Item #	***Task #***	***Activity/Deliverable/Title***
1	4,5	4th Quarterly Report (this report)
2	4	Initial testing of database prototype and integrating with risk assessment models
3	5	Database updating, tuning, and refinement based on internal feedback
4	5	Preliminary study on LLM model integration with database

Items in Progress During this Quarterly Period:

Based on the database prototype developed in the past quarter, the research partner teams conducted testing and validation activities under Task 4. More detailed, University of Cincinnati (UC) team, Rutgers University (RU) team, and Texas A&M University (TAMU) team validated the database by integrating geological, hydrological, and electrochemical risk models (as presented in Appendix B). Feedback from the partner teams was collected and incorporated, leading to updates to the database schema and additional data integration for risk assessment. In addition, this feedback helped define future plans to integrate risk assessment results into the database, i.e., a two-way information transferring between database and risk assessment tools.

As part of Task 5, the team updated the database to include private data attributes (i.e., pipeline structure, baseline, location, in-line inspection, and other integrity inspection data) and additional data from soil survey, informed by the industry experts from ROSEN and Integrity Solutions Field Services, collaborative teams, and ongoing validation results. The key outputs for this period include an updated database prototype, a revised XML schema supporting private data, and review of the large language model (LLM) llama3 to integrate to the database. On-going work is focusing on further refinement based on testing feedback, scaling support for additional data types.

In the annual project report meeting, the team proposed to further integrate a large language model to enhance natural-language query interfaces, smart data ingestion, and schema alignment for the pipeline database. During the past quarter, we have explored and generated some preliminary results. To summarize, the following items are on-going with progress to report during this past quarter.

Item #	***Task #***	***Activity/Deliverable/Title***
1	5	LLM model integration with database
2	5	Continue to update and refine the database
3	5	Continue to validate the updated database

Task #4 Objective:

The main objective of task 4 for the current quarter is to test the developed database prototype for downstream risk assessment and analyze the test findings for further development of the database.

Summary of work performed:

Building on the database prototype developed in the previous quarter, each collaborative team developed a qualitative risk assessment model to test and validate the database. Validation was carried out using ground truth data from a natural geohazard case history that impacted a pipeline, with the associated data for the specific event date obtained from the database to ensure accurate comparison. Testing was conducted on the three georeferenced pipeline datasets requested by the university team and provided by PHMSA. These datasets were expanded to include a range of geological, hydrological, and electrochemical factors, using the in-house developed python scripts to pull data from several publica repositories.

The UC team developed a semi-quantitative geohazard risk assessment system, accompanied by an automated Python-based data processing program, to enable scalable, repeatable, and batch-runnable risk assessment in accordance with API RP-1187 and PHMSA guidelines. The system calculates Likelihood (Hazard Susceptibility: HS, Pipeline Vulnerability: PV, Monitoring & Detection: MD) and Consequence (Preventive Measures: PM, Exposure & Consequence: EC) scores on a 0–5 scale, with defined weights (HS 30%, PV 25%, MD 20%, PM 15%, EC 10%) to produce a total risk score and classification. Environmental parameters, such as slope, NDVI, rainfall, flood depth, soil moisture, and particle composition, are structured into standardized XML files and batch-processed, enabling automated HS scoring that aligns with expert judgment. The testing confirmed that the automated scores correctly reflect geological hazard trends and highlighted areas for data extraction refinement, such as soil moisture variability and slope instability.

To account for attributes not directly measurable via remote sensing, an interactive module collects expert inputs for PV, MD, PM, and EC, which are then applied across pipeline sections with similar conditions. The program outputs CSV files containing individual scores, weights, total risk, and tier (Low/Medium/High). Validation using the 2023 Gulf South pipeline landslide case showed high HS, PV, and PM scores, resulting in a High-risk classification that accurately reflected the accident mechanism and event facts. Overall, results demonstrate that the automated system is consistent with manual scoring, integrates environmental and expert data effectively, and provides a robust framework for large-scale geohazard risk assessment across pipeline networks.

The Rutgers university team focused on the development and implementation of a semi-quantitative hydrological hazard risk assessment module designed to evaluate water-related threats to pipeline integrity, including flooding, precipitation, soil moisture, snowmelt, and river interactions. A Python-based risk assessment engine was developed to ingest heterogeneous XML and GeoJSON data, normalize environmental parameters onto a standardized 0–5 risk scale, and aggregate point-level data to the segment level using worst-case and average strategies consistent with API RP-1187 and PHMSA guidance. A weighted scoring framework was implemented, prioritizing flood inundation depth, soil moisture, and precipitation, to compute total risk scores and classify pipeline segments into low, medium, or high-risk tiers. The system produces structured CSV outputs for batch processing and demonstrates reliable performance through threshold-based scoring logic that reflects known hydrological risk mechanisms. Database testing addressed data heterogeneity by integrating legacy XML and GeoJSON sources through a unified ingestion workflow and improving robustness against missing data. The model development and implementation were guided by PHMSA risk modeling documentation, API RP-1187 principles, and supporting internal references for hydrological data integration.

The Texas A&M team developed a corrosion risk–based assessment model that integrates publicly available pipeline database information with private data to evaluate corrosion damage at the pipeline segment level. Using soil condition parameters such as moisture, pH, silt and clay content, soil electrical potential, and resistivity, the team applied electrochemical and empirical equations to estimate corrosion rates, which were then used to calculate time to failure and assign discrete risk categories, with segments predicted to fail within two years classified as the highest risk. The current quarter focused on implementing and refining this methodology, resulting in improved risk characterization and support for prioritized inspection and mitigation planning. Testing of the PHMSA database showed improved usability and compatibility following updates from the UD team, though a key limitation remains the lack of time-dependent data, which restricts more accurate modeling of corrosion progression; this limitation is inherent to the publicly available dataset.

This report only provides a summary of the results. The detailed explanation and comparisons between predicted risk levels and observed impacts is provided in Appendix B. The implemented risk model program scripts are regularly uploaded to the GitHub repository with version control. Currently the GitHub repository is managed as a private repository as we are currently undergoing the software copy-right application process.

A meeting was also conducted with the Rosen Group and Integrity Solutions Field Services for a technical discussion on pipeline private data management practices, focusing on how to update the private data attribute schema alignment with industry standards. Following this discussion, the scope for testing the database with industry partners was planned to ensure the validation reflects real-world industry practices.

Task #5 Objective:

The main objective of Task 5 for this quarter is to update the developed database prototype with private data attributes and refine the database based on feedback from the university teams. As mentioned in the original proposal, Task 4 and Task 5 will be performed iteratively. To analyze findings from the review on LLMs to support integration into the database. A detailed explanation of the LLM integration, the implementation of soil survey private data, and the updates to the pipeline private attribute schema is provided in Appendix A.

Summary of work performed:

1) Exploration of LLM integration into the database

During the annual review meeting, the team proposed integrating an LLM model into the pipeline database to enable more effective querying capabilities and automation workflows. The idea was well received by PHMSA, and a non-cost extension was requested to pursue this initiative.

Pipeline integrity management relies heavily on analyzing complex datasets for pipeline risk-related data such as identified geohazard factors data and field inspection reports. Currently, these datasets are available in the database being created and often depend on Python script-based querying and pulling to the database. The proposed technical framework integrates data engineering, semantic retrieval, and intelligent reasoning for a simpler user-data interaction. The LLM can generate comprehensive summaries of pipeline risk-related data, bridging the gap between structured datasets and human intuitive understanding. The hybrid pipeline database combines relational and document-based structures (geoJSON and semi-structured XML) to manage spatial and temporal dimensions of pipeline data. Risk factors and textual information are embedded into high-dimensional vector representations for semantic search, enabling the LLM to retrieve relevant records efficiently. We would like to highlight that some lightweight, locally deployed models such as Mistral or LLaMA3 have potential to be used to ensure privacy-sensitive inference that are needed for critical infrastructure data management, with a user-facing chat interface allowing querying, visualization, and report generation. The system is also expected to include schema-driven validation, automatic unit conversion, and multi-format data visualization to ensure consistency and integrity of risk information.

Current efforts focus on reviewing and benchmarking existing LLMs for domain suitability, performance, and computational efficiency. Open-source models (LLaMA3, Mistral, Falcon, Mistral) are evaluated alongside API-based models (GPT-4, Claude 3, Gemini) for response time, accuracy, and adaptability to pipeline-specific datasets. Note that for API-based models, no PHMSA data is used by following the NDA. The evaluation also considers resource utilization and fine-tuning potential for secure, offline deployment in high-performance computing environments. The results will guide the selection of an optimal local LLM to support database management and decision support for pipeline integrity management.

2) Integration of Soil Survey Data into the Database's Private Attributes

Based on testing results of the database from the TAMU team, it was observed that several soil attributes obtained from field inspection data such as soil resistivity, carbonates, sulphates, etc. are missing. These attributes are important for evaluating corrosion-related risk affecting the pipeline integrity. Accordingly, the private schema has been updated to include these corrosion risk factors.

Pipeline-specific soil survey data is maintained as private due to its proprietary nature and ownership by pipeline operators. Since it is private, the survey data exists in different file formats according to the rules and practices of each pipeline operator. A standardization process was performed for the soil survey risk factors obtained in various data types, such as CSV and XLSX files. The data were standardized into XML format following the database design, referencing segment ID, pipeline ID, with spatial coordinates. The resulting XML data were then validated against the private attribute schema, which was designed to handle different soil survey data. The updated database schema and format conversion tool files have been uploaded to git, and the expert teams from TAMU, Rutgers, and U Cincinnati will proceed to test it using the implemented risk models.

3) Incorporation of field inspection survey data into the private schema

A comprehensive XML schema for private pipeline attributes was developed to standardize and manage diverse field inspection and soil survey data. The schema defines a root element containing subsections for Dates, Coordinate, Soil Survey, Integrity Survey, Pipeline General Information, and Pipeline Baseline Information. Each subsection includes detailed elements with specific data types and documentation annotations to capture key information, such as retrieval dates, pipeline and segment identifiers, spatial coordinates, chemical and physical soil properties, and integrity survey results from ILI, DCVG, and CIPS inspections. All elements include documentation annotations describing units or measurement standards. The schema supports standardization of heterogeneous data from CSV and XLSX formats into a unified XML format, organized by segment ID, pipeline ID, and spatial coordinates, and is validated against the private attribute schema to ensure consistent handling of pipeline-specific data. The resulting schema provides a structured framework for managing private pipeline inspection and soil survey information.

A technical meeting was held with ROSEN to discuss handling private pipeline inspection data according to industry practices, during which it was suggested to incorporate (Pipeline Open Data Standard) PODS attributes as the best practice. Implementing this recommendation, including support for multiple data formats, is planned as part of future work and will help maintain industry standards.

2. Project Financial Activities Incurred during the Reporting Period:

A cost breakdown list of the expenses during this quarter in each of the categories according to the budget proposal is provided below:

Prime Contract Number: 693JK32450004CAAP
Total Contract Value: $774,997.00
Total Funded Value: $774,997.00
Cost-share amount: $116,270

	Current Period Actual	Year To Date Actual	Contract To Date Actual
Salaries & Wages FT senior personnel	$7,258.98	$64,054.42	$67,779.09
Salaries & Wages Graduate Assistant	$6,000.00	$21,000.00	$21,000.00

Salaries & Wages Undergraduate student	$9,007.50	$16,236.24	$16,236.24
Benefit-Faculty/Staff	$1,712.39	$15,716.60	$16,660.06
Student Benefits -GA	$127.20	$1,967.60	$1,967.60
Student Benefits - undergraduate student	$190.96	$787.22	$787.22
Total Labor Cost	$24,297.03	$119,762.08	$124,430.21
Travel cost	$3,945.45	$6,268.94	$6,268.94
Conference Registration	$895.00	$1,795.00	$1,795.00
Subcontract Federal Cost	$26,822.85	$53,303.08	$53,303.08
GA Tuition Remission	$0.00	$17,900.00	$17,900.00
Lab Supplies	$200.00	$200.00	$200.00
Total Non-Labor Cost	$31,863.30	$79,467.02	$79,467.02
Total Indirect Cost	$28,192.48	$91,027.58	$93,370.98
UD Cost-share	$0.00	-$41,362.27	-$41,362.27
Total Federal Expense	$84,352.81	$248,894.41	$255,905.94

3. Project Activities with Cost Share Partners:

The meetings held with the cost share partners during the current quarter are listed below, the summaries provide overviews of key discussions and decisions/actions accomplished.

1) Weekly progress meeting scheduled with Texas A&M team:

Data and time: 10:00 AM – 11:00 AM ET [10/01/2025], [10/08/2-25], [10/15/2025], [10/29/2025], [11/05/2025], [11/12/2025], [11/19/2025], [11/26/2025], [12/03/2025], [12/10/2025], [12/17/2025]

Attendees: Hui Wang, Homero Castaneda, Sreelakshmi Sreeharan, Kiranmayee Madhusudhan, Rebecca Crow, Sasha George.

Agenda discussed: The team reviewed challenges associated with soil survey data availability and standardization, recognizing that limited and inconsistent soil data can impact accurate pipeline risk assessments. Discussion on ongoing two review papers to be submitted soon, including one focused on integrating corrosion risk models with the database and another literature review of the overall database design and implementation. Discussed the integration of LLMs with the database to support querying, automation, and improved analysis of pipeline risk data.

Activities conducted (accomplishments)*:* The database handling private soil survey data has been updated and provided the testing team to validate the soil survey risk data. Progress was made on defining standardized field inspection data schemas, which will help structure and integrate inspection data. The findings from the LLM literature review were also presented, and the integration of LLMs with the database was discussed, including preliminary results on how LLM-based querying and retrieval could improve data accessibility, automation, and interpretation of pipeline risk information. The quarterly reports from all the teams are provided before the Christmas break for integration into the consolidated quarterly report.

2) PHMSA CAAP Progress Meetings with UC and RU:

Data and time*:* 10:00 AM – 11:00 AM ET [10/03/2025], [10/17/2025], [11/14/2025], [12/12/025]

Attendees: Hui Wang, Lei Wang, Hao Wang, Sreelakshmi Sreeharan, Kiranmayee Madhusudhan, Yating Yang, Tianjie Zhang.

Agenda discussed: The team discussed preparations for the last annual review meeting, including presentation schedules and time allocation. Ongoing progress on database development was reviewed, with emphasis on integrating private pipeline data, incorporating field inspection datasets. The testing results of the database prototype with geological and hydrological risk assessment models were discussed. The discussion also covered challenges related to soil survey data availability and standardization, approaches for quantitative pipeline reliability analysis, and strategies for linking the database with downstream risk models. In addition, the team reviewed plans for code integration into the project GitHub repository, coordination among team members for model testing and validation, and future dissemination activities, including conference and journal publications as well as patent filings. The team also discussed the integration of Large Language Models (LLMs), particularly for converting unstructured data into XML, enhancing database querying through RAG, generating standardized narratives, and supporting validation and reporting.

Activities conducted (accomplishments): The team advanced the development of the pipeline risk database and clarified the end-to-end workflow for integrating geological and hydrological risk models using standardized Python objects directly rather than exported spreadsheet-based files. Discussion that LLMs (e.g., Llama-based models) can successfully interpret XML schemas and assist with bidirectional conversion between unstructured data and structured XML formats. Progress was made on defining and updating XML schemas to accommodate new risk attributes with default values, and on integrating hydrological risk assessment outputs into the data container. The team also coordinated code stabilization and version control through GitHub, reviewed historical PHMSA incident data for ground truth validation, and outlined approaches for risk modeling.

<u>3) Ad hoc Discussion between UC and UD on DOTPHMSA CAPP Database project</u>

Data and time: 03:30 PM – 04:30 PM ET [10/17/2025]

Attendees: Hui Wang, Sreelakshmi Sreeharan, Kiranmayee Madhusudhan, Yating Yang

Agenda discussed: The team, including Yating, Sreelakshmi, Kiranmayee, and Hui reviewed the testing of the geological risk assessment model. Key points included ensuring that all necessary factors were available in the database, methods for geological data computation, averaging data over relevant time periods, generating geoJSON files for spatial visualization, and validating risk scores for each segment. The workflow for utilizing the updated database was reviewed, with plans to analyze the results after testing and addressing any issues that may arise.

Activities conducted (accomplishments Yating began to conduct the geological risk assessment model using the updated database, collecting the data from public data repository by running the developed tools to generate pipeline risk factor datasets. The team verified that all necessary factors required for the risk model were present in the database. Yating and Kiranmayee worked on averaging geospatial data over six-month periods and generating geoJSON files to represent spatial risk levels. Sreelakshmi provided guidance on validating model outputs and ensuring consistency with expected results. The team agreed to continue experimenting with the system as needed and to provide support to finalize the results for inclusion in the quarterly report.

4) PHMSA Annual meeting quick rehearsal

Data and time*:* 11:00 AM – 12:00 PM ET [10/28/2025]

Attendees*:* Hui Wang, Lei Wang, Hao Wang, Homero Castaneda, Sreelakshmi Sreeharan, Kiranmayee Madhusudhan, Yating Yang, Tianjie Zhang, Rebecca Crow, Sasha George

Agenda discussed*:* During the annual database project meeting rehearsal, teams presented updates on the database development, its downstream applications in pipeline risk assessment, hydrological and geohazard modeling, and corrosion risk evaluations, while simultaneously testing the database functionality. Each group shared progress on their models, data integration approaches, and visualization tools, including GIS-based pipeline risk mapping. The focus of the meeting was the live demonstration of the pipeline risk assessment and database tools, with careful alignment of timing to ensure each team had sufficient time for their presentations and demos. The discussion also covered project management topics such as financial tracking, progress reporting, slide preparation, and coordination across teams. Guidance was provided on structuring presentations with clear objectives, methodology, and key takeaways, emphasizing the need to compare database content against model requirements and highlight missing information for future work.

Activities conducted (accomplishments): The presentations highlighted the successful development of each team's work, showcasing progress on models, database integration, and visualization tools. Teams received constructive feedback to improve their models and presentations, including removing sensitive identifiers, enhancing slide clarity, and clearly stating assumptions. Recommendations were also made to add additional visualizations, broaden the coverage of hydrological and geohazard factors, and strengthen database usability and data completeness. Overall, the meeting confirmed the teams' progress, provided actionable guidance for refinement, and aligned all teams on the next steps for continued model development, testing, and database enhancement.

4. Project Activities with External Partners:

The meetings held with the external partners during the current quarter are listed below, the summaries provide overviews of key discussions and decisions accomplished.

1) Annual Report Review Meeting: Database development

Data and time*:* 10:30 AM – 12:00 PM ET [10/30/2025]

Attendees*:* Nusnin Akter [PHMSA], Meg O'Connor [PHMSA], David Bastidas [ROSEN], Kevin Cowan [Integrity Solutions], Stephen Jones [PHMSA], Hui (Jack) Wang [UD], Homero Castaneda [Texas A&M], Lei Wang [UC], Hao Wang [RU], Sreelakshmi Sreeharan (UD), Kiranmayee Madhusudhan (UD), Yating Yang [UC], Sasha George [Texas A&M], Rebecca Crow [Texas A&M], Tianjie Zhang [RU].

Agenda discussed*:* The meeting opened with Dr. Hui Wang providing updates on overall project status with Meg O'Connor and Dr. Nusnin Akter clarifying that PHMSA operations remained unaffected and outlining requirements for a potential no-cost extension. Dr. Wang also reviewed administrative items, annual reporting, patent activities, and upcoming conference presentations. Sreelakshmi and Kiranmayee presented detailed progress on the “Geopipe” database software, including prototype development, testing, documentation efforts, and plans for user interface development using QT, role-based access control, and incorporation of field inspection data. Kiranmayee demonstrated database testing using different pipeline regions and discussed ongoing work on LLM integration for natural-language querying and historical incident analysis. Technical updates on risk assessment modeling and downstream database applications were provided by multiple teams. Tianjie presented hydrological risk factors and model development, highlighting climate-driven hazards such as flooding, scouring, and soil erosion. Yating presented a semi-quantitative geohazard risk assessment framework aligned with industry standards, including validation using a Gulf South Pipeline case study and plans for spatial mapping and Google Maps integration. Contributions from Sasha, Rebecca, and Dr. Homero Castaneda’s group were discussed in the context of electrochemical corrosion and time-to-failure modeling. The meeting also included input from David and Kevin on industry practices for handling pipeline field inspection data and standards, with plans for follow-up discussions on data integration and interoperability. Updates from Texas A&M, Rutgers, and UC teams covered the integration of public and private data, assignment of risk categories, expansion of hazard coverage, and collection of user and industry feedback. The meeting concluded with a coordinated plan for next-quarter activities, continued collaboration, industry engagement, and preparation for project non-cost extensions and future publications.

Activities conducted (accomplishments): During this meeting, the team presented significant progress across database development, risk assessment, and dissemination activities. The project advanced through completion of Tasks 1–3, including a comprehensive literature review, identification of pipeline risk factors, and integration of public and private data sources. A functional Geopipe database prototype was developed and moved into the testing phase, and work was initiated on a provisional patent for the Geopipe system. We completed initial prototype validation using multiple pipeline datasets, established data validation rules, and demonstrated database functionality with several risk models. Progress was also made on incorporating field inspection data and defining the framework for a QT-based graphical user interface, along with role-based access control for secure data access. Tianjie from Rutgers University developed and implemented hydrological risk assessment components, identifying key climate- and water-related hazard factors and integrating model outputs with the database workflow. Yating completed a semi-quantitative geohazard risk assessment framework aligned with API RP 1180 and PHMSA guidance, validated through a real pipeline case study and prepared for spatial visualization and broader hazard coverage. Contributions from Sasha, Rebecca, and Homero strengthened the electrochemical corrosion and time-to-failure modeling components. The team also explored LLM integration to use retrieval-augmented generation for converting unstructured data to structured XML, enabling natural-language querying, and supporting historical incident report analysis. This work demonstrated the feasibility of using LLMs for data ingestion, validation, and report generation within the Geopipe framework. In terms of dissemination and outreach, the team successfully submitted papers and abstracts accepted for AMPP 2026, prepared presentations for upcoming conferences (including REX 2026 and AMPP 2026), and progressed toward a review journal manuscript. Preparations for annual and quarterly reports were completed, with documentation structured for integration into the comprehensive project report.

2) Technical discussion on pipeline private data management practice with ROSEN and Integrated Solutions

***Data and time**:* 03:30 PM – 04:00 PM ET [11/14/2025]

***Attendees**:* Hui Wang [UD], Sreelakshmi Sreeharan [UD], Kiranmayee Madhusudhan [UD], David Bastidas [ROSEN], Kevin Cowan [Integrity Solutions].

***Agenda discussed**:* The team reviewed progress on the database architecture for integrating private pipeline data, including XML-based handling of segment and inspection data. Discussions focused on standardizing database schemas for field inspection, soil survey, and inline inspection data. David suggested using PODS Model 7 as a reference with industry data practices. Discussed on incorporating language models for flexible data format conversion and ensuring compatibility with industry data practices. The team agreed to provide an existing XML schema and list all the private attributes following PODS standards for confirming on the private attributes which can be added to the schema. The team also discussed documentation needs, validation with subject matter experts, and coordination with industry partners for feedback and future integration.

***Activities conducted (accomplishments)**:* The team advanced the database architecture supporting both public and private data, finalized a preliminary schema structure for field inspection and soil survey data, and agreed to align required attributes with PODS Model 7. A clear approach was established for handling multiple data formats through XML and language-model-assisted conversion. Plans were defined for producing a tree-structured architecture document, sample XML and schema files for expert review, and structured feedback from industry collaborators, marking a key step toward industry-aligned, scalable pipeline risk database implementation.

5. Potential Project Risks:

The core of the pipeline database has been successfully developed, along with the framework for the overall system. Nonetheless, a few potential risks have been noted that must be managed before the database can be finalized. Testing of the database may experience minor delays due to the need for industry-level validation, ensuring compliance with operational standards, and coordinating schedules with industry partners to align on testing protocols and data verification. Additionally, the implementation of the proposed LLM integration, discussed during the annual meeting, may require extra time, as it involves model integration, automation, validation, and development of natural language querying capabilities while balancing pipeline standards and regulatory requirements. To accommodate the extended timeline, a non-cost extension has been requested, providing the flexibility to complete database testing, finalize LLM integration, and incorporate feedback from industry partners, while maintaining quality and consistency. We will follow up and plan accordingly given the length of the final approved non-cost extension.

6. Future Project Work:

The database prototype is completed, with the private schema to be updated to incorporate all relevant field inspection and soil survey attributes. Integration of a Large Language Model (LLM) will be implemented to enhance natural language querying, support flexible data format conversion, and assist with incident report analysis. Future work will focus on finalizing testing of the database with the integrated risk models, validating outputs, external testing and collecting feedback from industry partners, and updating the database for LLM deployment and documentation.

7. Potential Impacts to Pipeline Safety:

During this quarter, the team submitted a paper presentation to the AMPP 2026 Annual Conference, a presentation abstract accepted at PRCI REX2026, and a research-in-progress paper was accepted for presentation at the AMPP 2026 Conference. In addition, the team is drafting a review journal article that highlights key findings from the literature review, identifies existing challenges, and presents the proposed database framework.

Appendix A:

1) Review of LLM Integration into the database

Pipeline integrity management depends on the analysis of complex heterogeneous datasets, including geohazard factors and field inspection reports, to assess and mitigate pipeline risks. Currently, these datasets are available in the database being created and often depend on manual querying and connecting to the database. LLMs present a breakthrough opportunity to interpret technical language, correlate multi-source data, and provide explainable risk assessment data to the users. Recent advances in RAG (Retrieval-Augmented Generation) allow LLMs to reason over private databases securely. Despite progress in predictive maintenance, existing tools lack an intelligent interface that bridges human queries with structured pipeline data. Users will have a system that can answer questions through natural language processing without complex XML data searches and queries or manual aggregation. The LLM model works as an automated generation of comprehensive summary of risk assessment data through natural-language prompts.

(i)Literature reviewed

Recent research demonstrates the potential of LLM-based AI agents for industrial pipeline monitoring tasks. For instance, Wei et al. [1] explored the application of large language models to detect leakages in natural gas valve chambers, highlighting the feasibility of integrating LLMs for real-time anomaly detection and decision support in complex pipeline systems. An intelligent agent was implemented that uses audio and infrared data to diagnose leaks and generate corresponding response actions, with testing on real project data validating its effectiveness. This underscores the relevance of benchmarking domain-adapted LLMs for pipeline integrity management and risk assessment. Obi Chukwuemeka Nwokonkwo et al. [2] proposed a practical application of combining machine learning (Random Forest classifier) for real-time anomaly detection with time-series forecasting for maintenance prediction in pipeline systems. By leveraging these complementary models, the framework can both detect operational anomalies and forecast maintenance needs, enabling pipeline operators to transition from reactive to proactive integrity management, thus improving safety and efficiency in critical infrastructure monitoring. Amadhe, F. O. [3] highlights how machine learning and artificial intelligence are increasingly applied to pipeline integrity management as pipeline networks evolve toward intelligent and digitized infrastructure. The paper presented the study responses predicted by machine learning approaches for Probability of Failure (POF), Consequence of Damage/Failure (CDD), Failure Pressure, Defect Classification, and Corrosion Rate data classification in Figure 1.

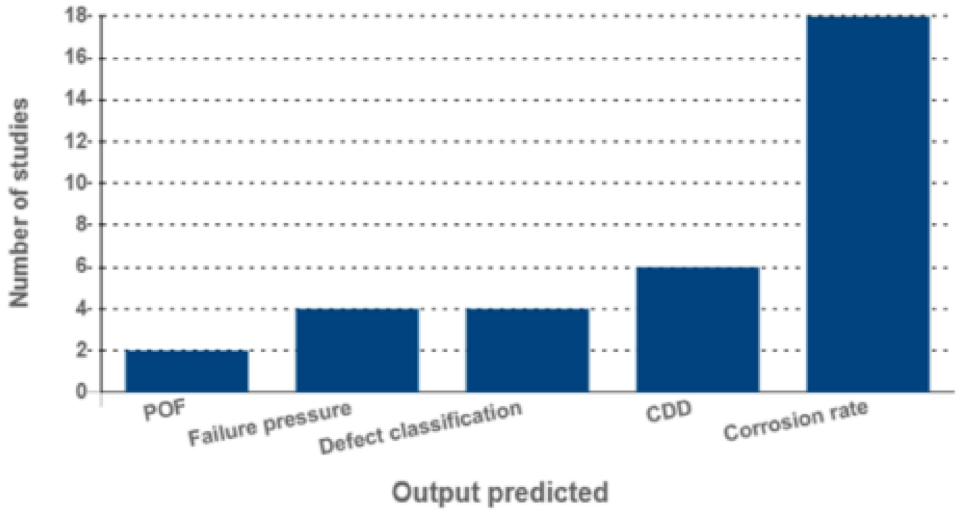


Figure 1: Studies anticipated using an AI model learning techniques for pipeline integrity.

To ensure optimal efficiency, scalability, and domain suitability, a comparative evaluation of various Large Language Models (LLMs) has been reviewed. The study focused on assessing performance, computational complexity, latency, and accuracy across a representative set of open-source and proprietary models. Open-source models such as LLaMA3&4, Mistral 7B, Falcon, and Mistral are benchmarked against API-based models like GPT-4, Claude 3, and Gemini to identify trade-offs between inference quality and system overhead. In addition, the LLM leaderboard emphasizes models optimized for fast inference, low latency, and cost efficiency, as illustrated in Figure 2. Evaluation metrics include response time, token generation throughput, context length, model parameter size, and domain-specific accuracy on pipeline risk and geohazard datasets.

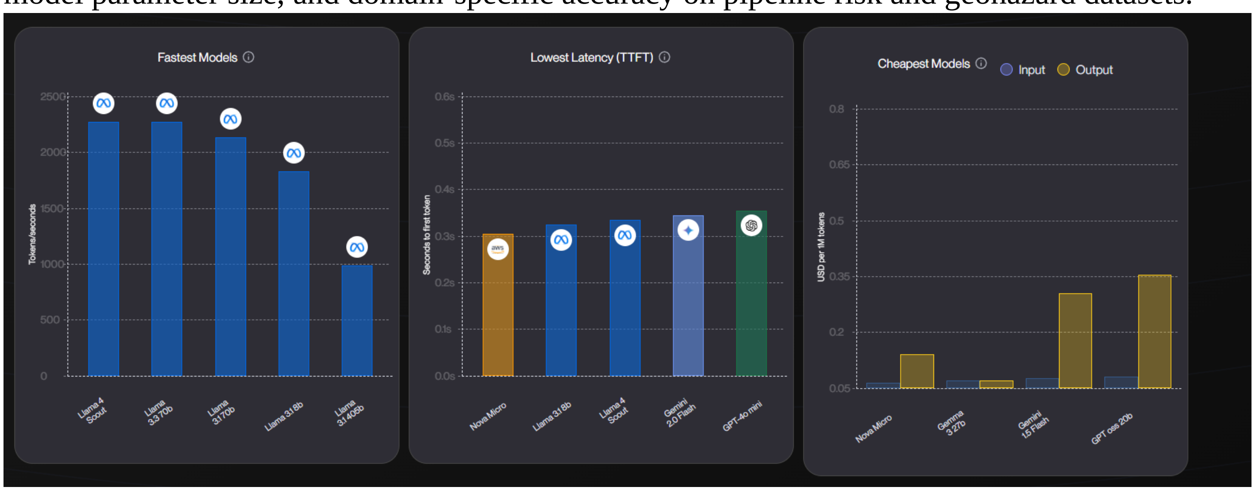


Figure 2: LLM leaderboard based on fast models, lowest latency, cheapest models [4].

A key research objective is to determine the most efficient *local model* capable of executing within the available high-performance computing environment. Performance profiling includes memory utilization, GPU acceleration efficiency, and scaling behavior under different load conditions. Additionally, model complexity has been analyzed through parameter count, architecture depth, and fine-tuning adaptability to proprietary pipeline data. The outcome of this comparative study has guided to the selection of an optimal local LLM that balances computational cost, interpretability, and accuracy, ensuring secure, offline deployment for sensitive pipeline infrastructure datasets.

(ii)Main Findings

Recent research highlights the use of advanced AI and machine learning techniques to enhance pipeline integrity management. The large language model (LLM)-based AI agents can detect leaks in natural gas valve chambers by combining infrared measurement data with a retrieval-augmented reasoning approach, creating a knowledge vector library for improved diagnosis and decision-making. A hybrid machine learning framework that integrates a Random Forest classifier for real-time anomaly detection with the prophet model for forecasting maintenance needs, using historical pipeline and inspection datasets to achieve high accuracy. The study shows that ML models, including supervised and unsupervised learning, hybrid approaches, and metaheuristic algorithms, significantly improve failure prediction, leak detection, and corrosion monitoring compared to traditional methods. Optimization techniques, such as genetic algorithms and particle swarm optimization, are increasingly used to fine-tune predictive models and support resource allocation and maintenance planning, balancing cost, safety, and reliability.

The challenges include that pipeline datasets are often fragmented across inspection reports, sensor readings, and legacy systems, with missing or inconsistent entries. Integrating complex algorithms like LLMs, random forest, and prophet models into existing pipeline databases requires robust data preprocessing, semantic alignment, and schema standardization to ensure compatibility. Interpreting model outputs for actionable decision-making requires domain expertise to validate predictions and avoid false positives or negatives, making human-in-the-loop oversight essential. The challenges remain in data quality, scalability, and real-world implementation, particularly across large or diverse pipeline networks. The review emphasizes future directions including integration of advanced hybrid ML systems, real-time data fusion from multiple sensors, and embedding AI into smart IoT platforms for dynamic monitoring and decision support, ultimately moving pipeline integrity management toward more proactive, intelligent, and cost-effective frameworks.

Selection of the LLM Model

Large Language Models (LLMs) are compared using a combination of benchmark datasets, evaluation metrics, and practical capability tests that aim to capture their reasoning, coding, comprehension, and generalization abilities. Common benchmarks include MMLU (for multi-tasking language understanding), GSM8K (for mathematical reasoning), Human Eval (for code generation), and Hella Swag (for common-sense reasoning). Models are evaluated both quantitatively, via accuracy, and qualitatively, through human preference or alignment tests. Comparative studies often include factors such as model size (parameters), architecture, context window length, and training data diversity, as these influence performance. Additionally, practical aspects like tool use, efficiency, inference latency, and robustness to prompt variations are considered. Leaderboards, such as Vellum [4], LLMDB [5], and Hugging Face's LLM leaderboards [6] aggregate these scores to rank models, enabling researchers and practitioners to identify state-of-the-art models and track progress in reasoning, coding, and general-purpose language capabilities. From the leaderboard (Table 1), Llama 3 was chosen due to its open-source nature and strong overall performance across benchmark tasks such as reasoning, language understanding, and generalization, while maintaining open-weight accessibility. Although other models demonstrate competitive performance, Llama 3 offers a balanced trade-off between accuracy and computational efficiency, making it well suited for scalable and reproducible database-driven applications. Alternative models may be considered in future work if performance or integration challenges arise.

Table 1: Comparison of LLM models based on the leaderboard sources

Model	Provider	Params	Benchmarks / Performance	Key Strengths
Llama 3 (405B)	Meta	~405B	Strong scores on general benchmarks like MMLU; close to top proprietary models; competitive in reasoning & math compared with GPT-4o & Claude 3.5 Sonnet.	Open-source, large context, powerful reasoning & coding
Llama 3 (70B)	Meta	~70B	Lower than 405B but still high on general tasks; similar performance to some mid-tier models; good tool use scores on Vellum leaderboard (e.g., ~81.1%).	Good balance of performance and cost for self-hosting
Mistral Large	Mistral AI	~32B (Large)	Moderately high performance; in some comparative tables below GPT & Claude on general benchmarks (e.g., GSM8K & human eval in one source).	Efficient architecture; open source
Claude 3.5 Sonnet	Anthropic	Proprietary	Often near top of benchmarks like GSM8K and general HELM/GPQA, e.g., 96.4% GSM8K	High reasoning & coding performance
Claude 3 Opus	Anthropic	Proprietary	High scores on benchmarks (e.g., Hella Swag); strong tool use & reasoning.	Broad capability across reasoning, coding, multilingual

Application design of the LLM Model for the database

To operate the integration of Large Language Models within the pipeline integrity management system, the following technical framework is proposed as shown in Figure 3. It combines data engineering, semantic retrieval, and intelligent reasoning into a cohesive architecture containing pipeline risk datasets.

1) Database: The Pipeline Risk Database Repository("geopipe"), currently under development, serves as the centralized data infrastructure for storing both public and private pipeline risk datasets. The data repository adopts a hybrid database architecture, integrating relational and document-based models to effectively manage the spatial and temporal dimensions of pipeline and environmental information. This hybrid design enables seamless querying, efficient data linkage, and scalable integration of multi-source datasets essential for comprehensive risk assessment and predictive analysis.
2) Embedding and Indexing (RAG Pipeline): Transforms critical textual and numerical features into high-dimensional vector embeddings using modern embedding models. These embeddings enable semantic search and contextual retrieval within the pipeline's existing database. With the existing developed database, which is hybrid-based model working on document-based database mode, i.e., semi-structured textual XML format and georeferenced relational database model, i.e., GeoJSON. It is easier for the semantic retrieval of database. Use LLM agents such as Lang Chain [7] /Llama Index [8] to retrieve relevant records based on user queries.
3) The model layer then acts as the cognitive core of the system. The retrieved passages, tables, or context are then fed into the LLM prompt so it can "see" this data before generating an answer. The LLM uses both its internal knowledge and the retrieved, current data to generate a grounded, accurate response. The lightweight, open-source local models such as Llama 3[9] are deployed locally for privacy-sensitive inference and domain fine-tuning, ensuring that proprietary integrity data never leaves the on-premises environment.
4) Interface Layer: Implement a user interface - chat dashboard for users to query, visualize.

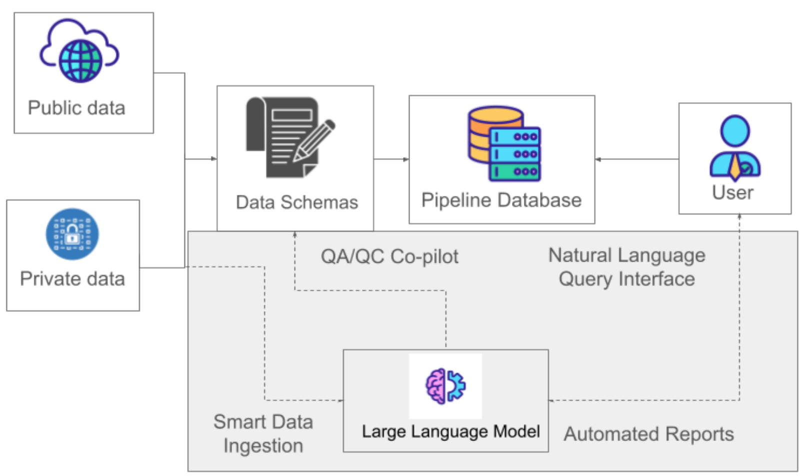


Figure 3: Incorporating LLM into the pipeline risk database

Once the LLM is established with the pipeline database, the system introduces intelligent automation features powered by Large Language Models (LLMs). It automates incident report classification and summarization, allowing rapid identification of high-risk events. The model further suggests corrective actions through contextual reasoning, leveraging historical data correlations and domain-specific knowledge to guide maintenance decisions. Because the dataset is governed by a well-defined schema, serving as a structural "rule book," the system ensures that any geohazard risk assessment data provided by users automatically confirms standardized data models. This schema-driven design allows the database to intelligently interpret, validate, and complete missing attributes, enabling seamless integration of new inputs. Furthermore, automatic unit conversion and multi-format data visualization, allowing users to view results consistently across different engineering and regulatory measurement systems. To enhance accessibility and data richness, a Document Digitization Module is integrated into the pipeline. This component converts historical incident reports into searchable digital formats and extracts structured data from scanned or PDF reports using optical character recognition (OCR) and LLM-based natural language understanding. By employing context-aware text interpretation, the system ensures that unstructured or legacy documentation becomes a functional part of the active knowledge base, enabling holistic and explainable risk intelligence across the entire pipeline network.

<u>Preliminary results</u>

The developed database was integrated with a Llama Index retrieval-augmented generation (RAG) agent and the Llama 3 large language model, leveraging an XML-based risk factor folder structure. The LLM was deployed on the system configuration illustrated in Figure 4, while the LLaMA 3 model parameters configured using Ollama are presented in Figures 5 and 6.

System parameters:

Figure 4: The system configuration for the local llama3 deployed.

Model parameters:

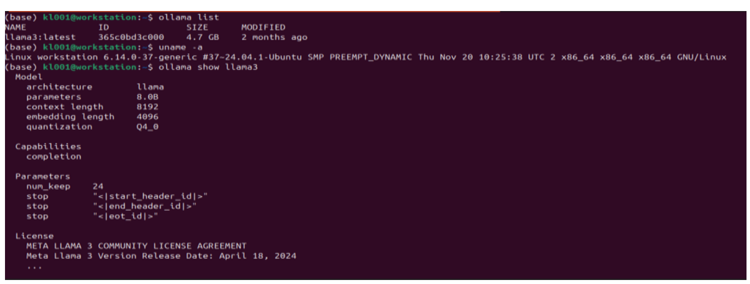


Figure 5: The deployed local LLM- llama3 model parameters.

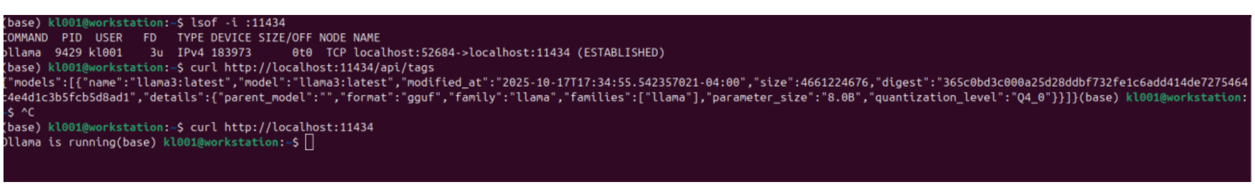


Figure 6: The deployed local LLM- llama3 model running at local host port 11434.

Results obtained using one structured pipeline segment data file:

The database was constructed using a pipeline dataset where spatial location details were intentionally altered and anonymized to meet NDA and data privacy requirements. The database was created for the location Utah Pipeline dataset (OPID-12876) with 17,994 pipeline segments at a spatial resolution of 50 m. The results obtained from this integrated framework are presented in Figure 7 - 8.

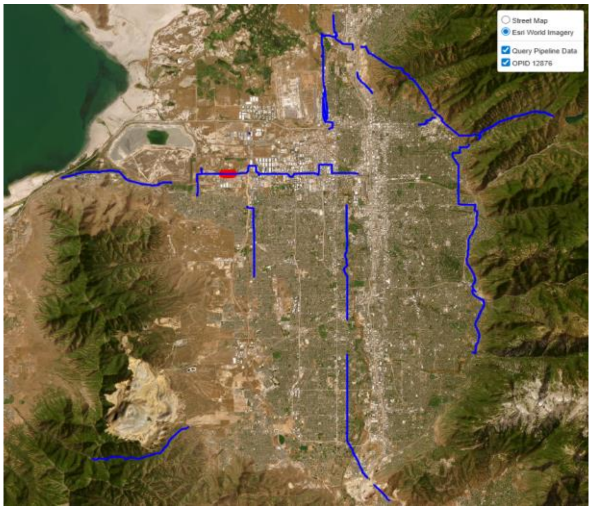
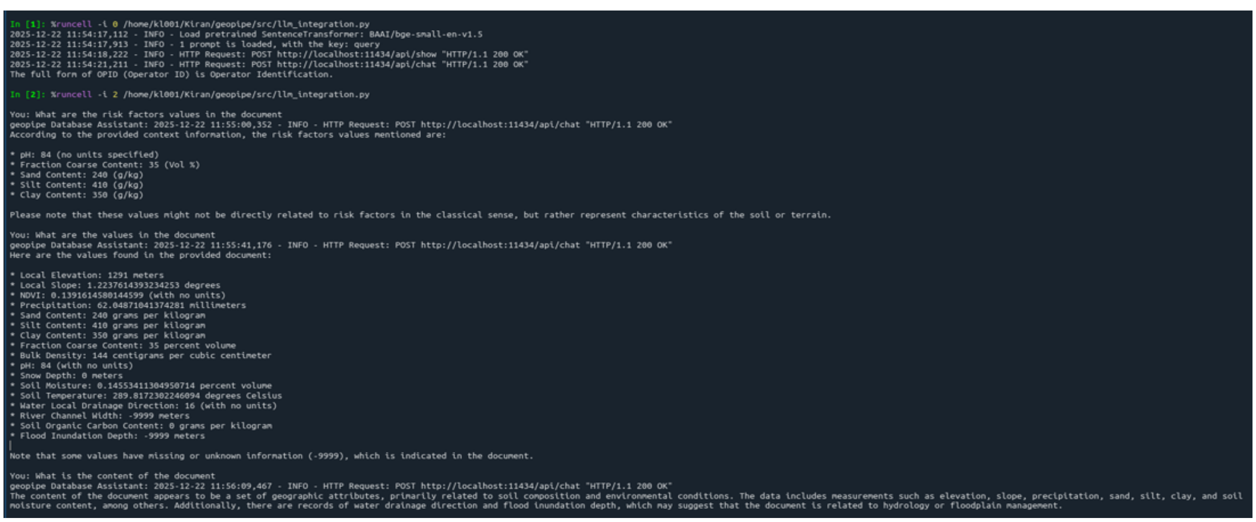


Figure 7: The developed database for the Utah pipeline (left) and the structured XML file containing pipeline segment risk data (right).



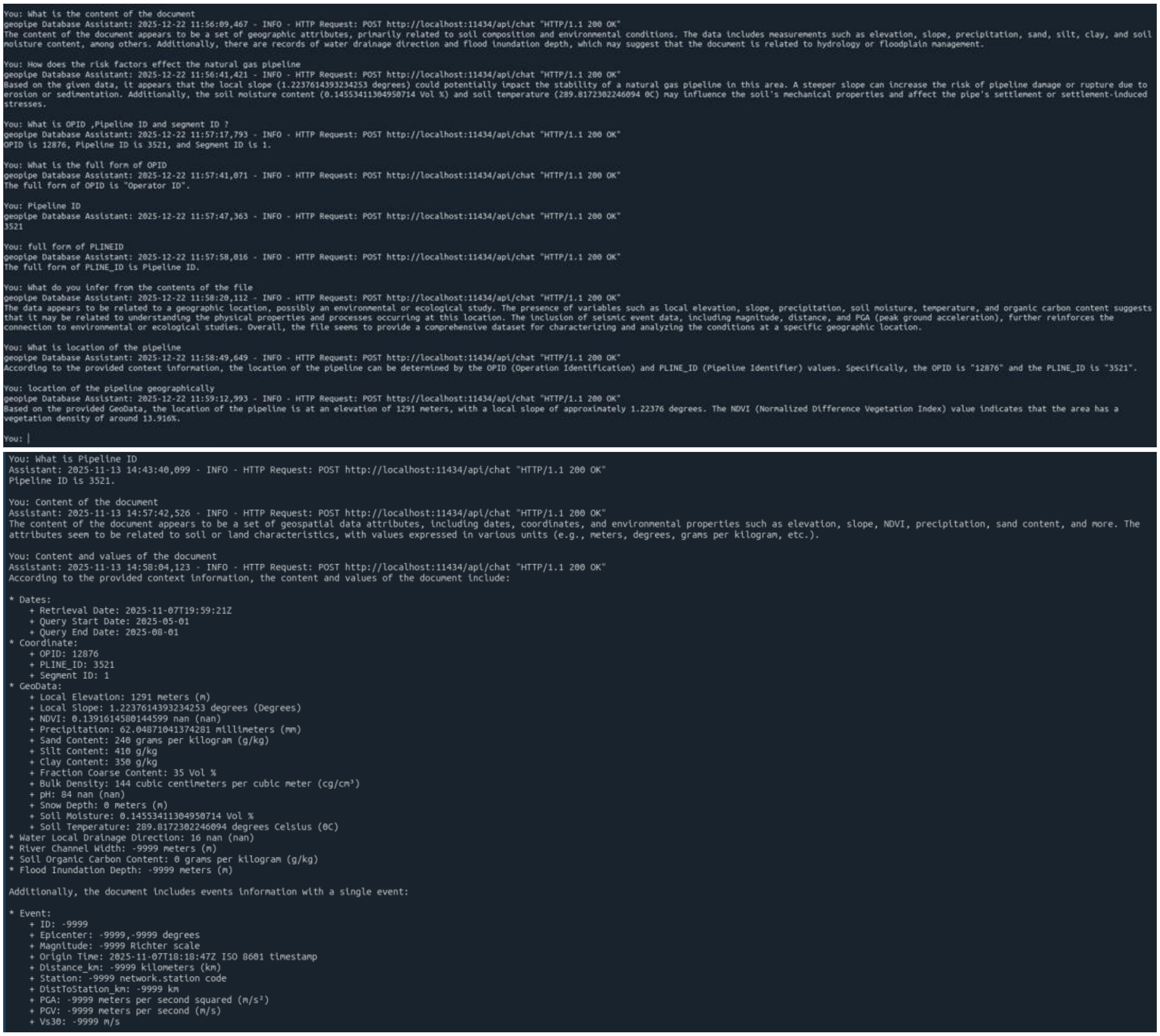


Figure 8: The results obtained from the LLM after integration with the database xml file.

Based on the results obtained, the use of an LLM demonstrates strong potential as an effective approach. However, the current model captures only specific aspects of the source documents, indicating the need for further fine-tuning and development. The next steps include validating the XML against a defined schema and generating standardized XML files in accordance with that schema.

2) Integration of Soil Survey Data into the Database's Private Attributes

During testing of the schema and database using datasets provided by the TAMU team, several soil attributes critical to corrosion assessment such as soil resistivity, carbonates, and sulphates were identified as missing in some records. The schema was updated accordingly to ensure these corrosion-relevant soil attributes are consistently captured. The schema should support ingestion of heterogeneous soil survey datasets originally stored in CSV and XLSX formats, which are transformed into validated XML files organized by spatial coordinates and segment identifiers. This standardized, unit-aware soil survey schema enables consistent integration of corrosion-relevant soil attributes to downstream risk modeling and pipeline integrity analyses. On this note, a comprehensive XML schema for private pipeline attributes was developed to standardize and manage diverse field inspection and soil survey data. Each subsection includes detailed elements with specific data types and documentation annotations to capture key information, such as retrieval dates, pipeline and segment identifiers, spatial coordinates, chemical and physical soil properties, and integrity survey results from ILI, DCVG, and CIPS inspections. The soil survey section incorporates attributes like pH, redox potential, soil resistivity at multiple depths, carbonates, chlorides, bicarbonates, and sulphates. The integrity survey section captures inspection dates, defect names, defect types, and measurement parameters. Pipeline general and baseline information such as contractor, installation date, material, diameter, and wall thickness are also included. All elements include documentation annotations describing units or measurement standards. The schema developed xml supports standardization of heterogeneous data from CSV and XLSX formats into a unified XML format, organized by segment ID, pipeline ID, and spatial coordinates as shown in Figure 9-10, and is validated against the private attribute schema to ensure consistent handling of pipeline-specific data. The resulting schema provides a structured framework for managing private pipeline inspection and soil survey information.

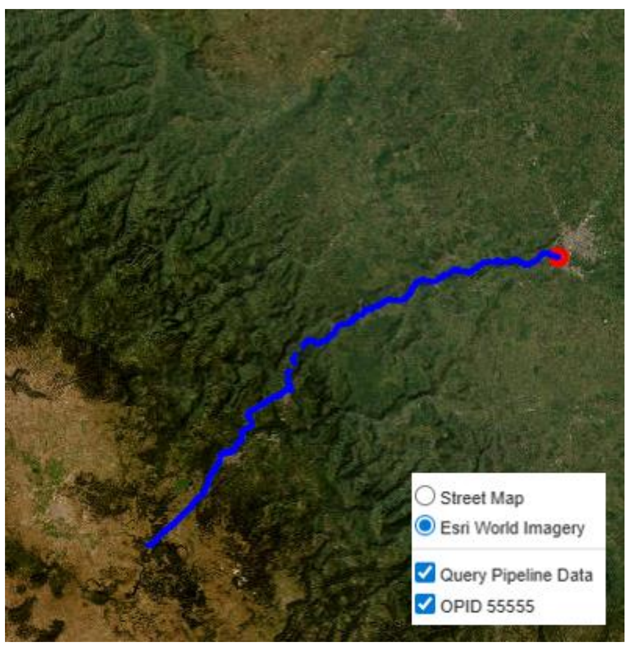


Figure 9: An example pipeline dataset for testing the field soil survey data.

The database, developed to manage private pipeline attributes, including soil survey data provided in various operator-specific formats, has been shared with other teams for testing. Furthermore, a technical meeting was conducted with the ROSEN group and Integrated Solutions Field Services to ensure the database aligns with established industry practices.

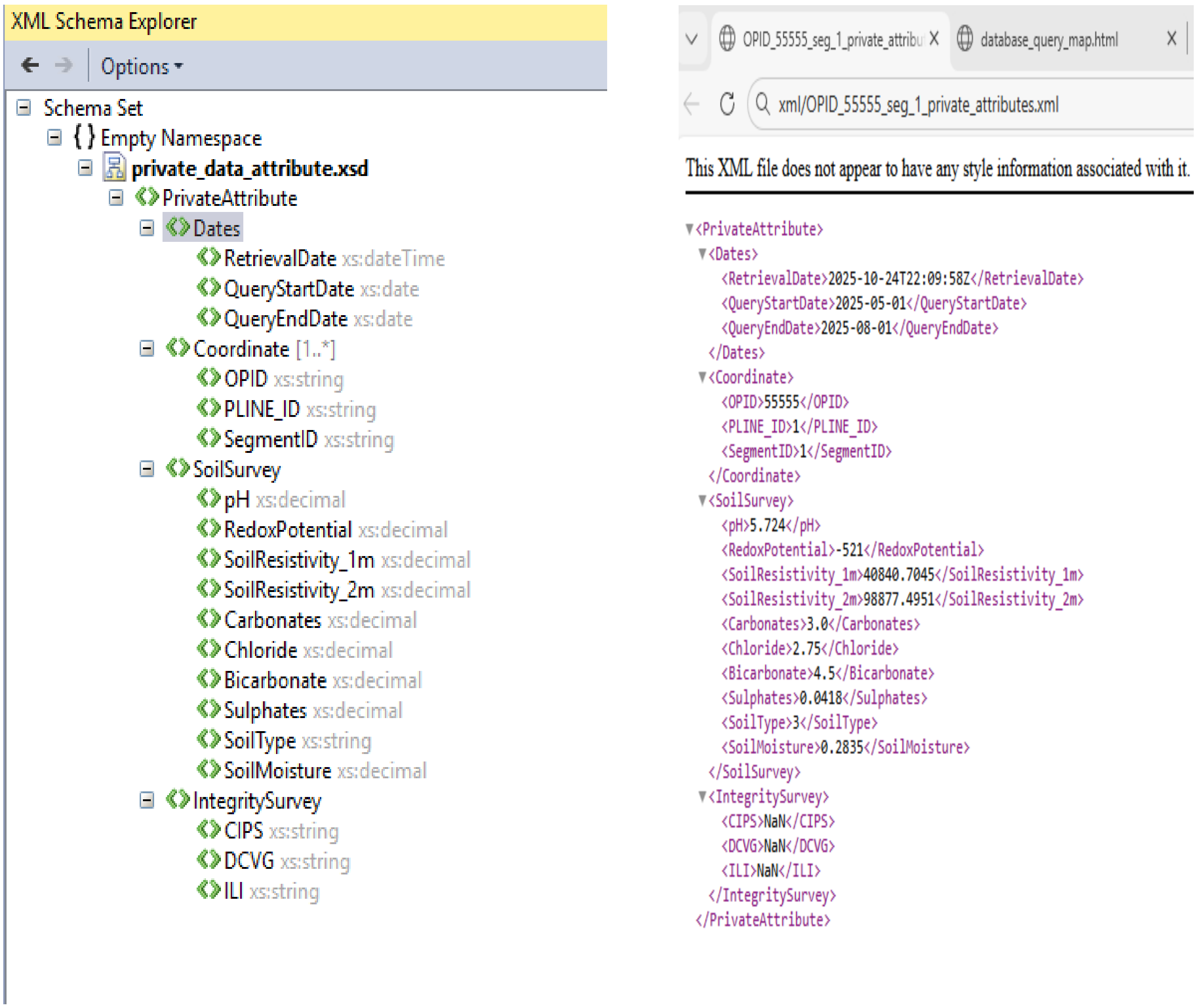


Figure 10: The schema design and xml generated based on schema for the soil survey private attributes.

<u>3) Incorporation of field inspection survey data into the private schema</u>

A technical meeting was held with the ROSEN group and Integrated Solutions Field Services to discuss handling private pipeline inspection data according to industry practices. During this discussion, it was recommended that the schema incorporates the Pipeline Open Data Standard (PODS) attributes to align with industry and improve interoperability. Implementing this recommendation, including support for multiple operator-specific data formats and additional PODS attributes is planned as part of future work. These updates will extend the schema to ensure compliance with industry standards while maintaining the current capability to manage private soil survey and inspection data effectively. The current XML schema has been developed to standardize and manage private pipeline attributes, including diverse field inspection and soil survey data. The schema defines a root element containing subsections for Dates, Coordinate, Soil

Survey, Integrity Survey, Pipeline General Information, and Pipeline Baseline Information. Each subsection includes detailed elements with specific data types and documentation annotations to capture key information, such as retrieval dates, pipeline and segment identifiers, spatial coordinates, chemical and physical soil properties, and integrity survey results from ILI, DCVG, and CIPS inspections. The Pipeline General Information section captures metadata about the pipeline, including attributes such as contractor or operator name, installation date, fabrication date, and material type, providing context on ownership, construction, and operational characteristics. The Pipeline Baseline Information section contains technical specifications and design parameters, such as pipeline diameter, wall thickness, and design pressure, which serve as reference points for integrity management, risk assessment, and inspection analysis. All the subsections in the schema include a Survey Date element, which is critical for evaluating temporal variations in soil properties such as resistivity, pH, and moisture content, as these can directly affect corrosion risk assessments. The Integrity Survey subsection captures inspection-specific attributes for each technique. For ILI, attributes include Pigging Date, Defect Name, Defect Type, and MetalLoss_percentage, which quantifies internal corrosion. For DCVG, attributes include Inspection Date, Defect Name, Voltage_OL_RE, Signal_OL_RE, Total_voltage, and DefectType, which characterize coating defects and potential cathodic protection issues. For CIPS, attributes include Inspection Date, Defect Name, On and Off pipe potentials, and DefectType, which measure corrosion current density and coating integrity. All attributes are standardized with units and validated within the schema to ensure consistency and accurate integration with downstream risk and corrosion models. Together, these sections provide a comprehensive overview of both the administrative and engineering aspects of each pipeline segment. All elements include documentation annotations describing units or measurement standards.

Figure 11: The updated schema design for the private attributes handling different field inspection data.

Appendix B:

1)Database Testing Results from UC team

The focus of this quarter's work from the UC team is on the construction of a semi-quantitative geohazard risk assessment system and the development and validation of an automated database development. Aim to develop a repeatable, scalable, and batch-runnable risk-scoring system that complies with API RP-1187 and PHMSA guidelines, capable of consistently determining risk across multiple pipeline sections. This stage mainly completed the model logic design, Python-based automated script development, structured processing of input data, and preliminary effectiveness verification using typical accident cases.

Overall framework of the risk assessment system

This system employs a semi-quantitative scoring method by dividing risk into two core parts:

(1) Likelihood, consisting of three components:

Hazard Susceptibility (HS)
Pipeline Vulnerability (PV)
Monitoring & Detection (MD)

(2) Consequence, consisting of two components:

Preventive Measures (PM)
Exposure & Consequence (EC)

Each score is rated on a 0-5 scale and assigned different weights (HS 30%, PV 25%, MD 20%, PM 15%, EC 10%), ultimately forming a total risk score and classification (Low/Medium/High). This structure considers in-site data, environmental data, and expert judgment, conforming to the "multi-factor, quantifiable, and traceable" risk assessment concept required by API RP-1187 and PHMSA expectation for semi-quantitative assessment.

Automated input data structuring and XML data workflow

To achieve large-scale assessment, the generation and processing module of automated geospatial environment data was completed this quarter. Using upstream scripts, environmental parameters for each pipe section (such as slope, NDVI, rainfall, flood inundation depth, soil moisture, particle composition, etc.) were written to standardized XML files. The Python main program would batch-read these XML files and extract the value to form a unified environmental feature dictionary for subsequent scoring calculations.

This kind of structured data enables the model to automatically and stably conduct risk assessments on multiple pipe sections, significantly reducing the subjective errors caused by manual input, and laying the foundation for future GIS mapping and database integration.

Development of Automated Hazard Susceptibility (HS) Scoring Logic

The automated algorithm for geological hazard sensitivity (HS) was completed this quarter. The scoring logic comprehensively considers environmental factors such as slope, flood inundation

depth, rainfall, soil moisture, NDVI, clay content, and coarse particle ratio, and produces a risk trend consistent with expert judgment. The scoring example logic is as follows:
Slope is a fundamental indicator of HS. While slopes greater than 30° are classified as high susceptibility (5 points), and those between 15° and 30° are considered moderate (3.5 points). The depth of the flood, as a parallel criterion, can enhance the HS score when it exceeds 0.1-3m. High soil moisture content (over 0.35) or heavy rainfall (over 200 mm) increases landslide susceptibility. A lower NDVI (less than 0.2) indicates insufficient vegetation cover, further enhancing HS. If the proportion of coarse particles in the soil is relatively large, the risk is slightly reduced (- 0.5 points). Parts of the code are shown in Figure 12.

```
# base scores
slope_score = 5 if (slope is not None and slope >= 30) else \
              3.5 if (slope is not None and slope >= 15) else \
              2.5 if (slope is not None and slope >= 8) else \
              1.0 if (slope is not None) else 0.0

flood_score = 5 if (flood is not None and flood >= 3) else \
              3.5 if (flood is not None and flood >= 1) else \
              2.5 if (flood is not None and flood >= 0.1) else 0.0

hs = max(slope_score, flood_score)

# adjustments
if (moist is not None and moist >= 0.35) or (rain is not None and rain >= 200):
    hs += 0.5
if ndvi is not None:
    hs += 0.5 if ndvi < 0.2 else (0.25 if ndvi < 0.5 else 0.0)
if (clay is not None and sand is not None) and (clay >= 400 and sand <= 200):
    hs += 0.5
if coarse is not None and coarse >= 20:
    hs -= 0.5
```

Figure 12: Automated rule for HS scoring based on slope, flood, and environmental conditions

The advantages of this method lie in its quantifiability, repeatability, scalability, and its ability to automate sensitivity by scoring across the entire pipeline network when connected to upstream data.

<u>An interactive collection mechanism for inputs of PV, MD, PM, and EC</u>

Since some attributes cannot be directly obtained through remote sensing or public data (such as the age of pipelines, welding defects, IMU monitoring conditions, treatment effects, and population exposure levels, etc.), the script has designed an interactive question-and-answer module to collect expert judgments. The program uses a unified input and applies the corresponding scores to all pipeline sections in this batch processing, which is suitable for the overall assessment of sections with the same geological conditions, consistent pipeline attributes, or continuous sections.
For example, also shown in Figure 13:

(1) PV (Pipeline vulnerability): If the pipeline was constructed before 1980 or has weld defects, the PV rises to 4.5 points.
(2) MD (Monitoring & detection): If IMU monitoring is available, the MD score is reduced

to 1 point (lower risk).

(3) PM (Preventive measures): If recent effective governance measures are collected, PM can be rated at the lowest risk level of 1 point.

(4) EC (Exposure & consequence): High population density sections will significantly increase the EC score (up to 4.5 points).

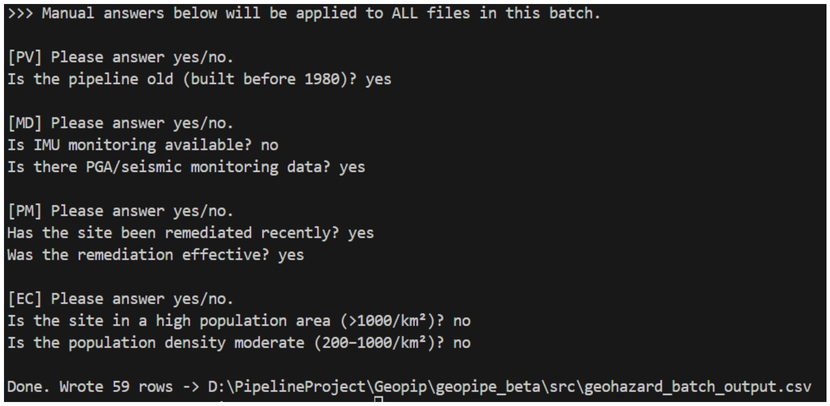


Figure 13: Collects expert yes/no answers for missing parameters

This mechanism maintains the consistency of the assessment while retaining the flexibility of expert judgment.

Batch processing and risk output results

The main Python program supports batch analysis of any number of XML files and automatically generates CSV files containing each risk factor, weight, total score, and risk level. The output file includes the following key fields, as shown in Table 2.

Table 2. Key fields

Field	Description
HS, PV, MD, PM, EC	Five core semi-quantitative scores
weight XX	Weight of the score
total	Comprehensive risk score
tier	Low/Medium/High

The verification results show that the automated scoring is highly consistent with the manual benchmark scoring, indicating that the model accurately reflects the risk logic of API RP-1187 and is suitable for promotion in larger-scale geological disaster risk assessment work.

Validation of effectiveness based on typical accident cases

This quarter, model validation was also conducted based on the Gulf South pipeline landslide rupture accident in Jackson, Mississippi in 2023. The accident was triggered by the lateral sliding of the artificial embankment, which led to the weld failure and complete fracture. By inputting the measured environmental parameters of this case into the model, HS received a high sensitivity score, and both PV and PM were at high risk. The final risk level was determined as High, which effectively reflected the accident risk mechanism and verified the rationality and reliability of the model.

This quarter, significant progress was made in developing a semi-quantitative pipeline geohazard risk assessment system in line with API RP-1187. The upstream XML data generation process and the Python module for automated processing were completed, enabling scalable and repeatable risk scoring. An automated HS (Hazard Susceptibility) scoring algorithm was constructed, integrating multiple environmental factors such as slope, flood depth, rainfall, NDVI, and soil properties. An interactive expert input mechanism was implemented to capture PV (Pipeline Vulnerability), MD (Monitoring & Detection), PM (Preventive Measures), and EC (Exposure & Consequence) scores. Additionally, a batch-executable Python script was developed to process multiple pipeline segments and output standardized CSV reports. The validity of the model was further confirmed through testing on typical accident cases, demonstrating its ability to reproduce real-world risk mechanisms accurately. The next stage of the project will focus on promoting and enhancing the system. This includes collecting additional feedback from industry users to refine the rule base and weighting system, visualizing risk results on platforms such as Google Maps or GIS for spatial representation, and expanding the model to cover additional disaster types, including riverbank erosion, earthquakes, and ground subsidence. The ultimate goal is to develop a sustainable, updatable, and comprehensive pipeline geological disaster risk platform capable of supporting broad applications across multiple pipeline networks.

2)Database Testing Results from TAMU team

The TAMU team focused on implementing a corrosion risk-based assessment model using data provided through the Department of Transportation (DOT) pipeline database. Publicly available DOT records were integrated with private, user- input parameters to evaluate corrosion damage across different pipeline segments. Risk scores were calculated by combining multiple corrosion parameters and were categorized into defined risk levels. Finally, a time-to-failure was estimated to support prioritized inspection and mitigation planning.

Design of the risk model being implemented in the current quarter

The risk-based methodology developed by the TAMU team focused on evaluating key corrosion parameters to assess the risk of individual pipeline segments. Publicly available database inputs including soil moisture, pH, silt and clay content, soil electrical potential, and soil resistivity were incorporated to estimate corrosion rate. Each parameter was integrated into governing electrochemical equations to derive local electric potential and an overall corrosion rate for the pipeline. These calculated corrosion rates formed the basis for segment risk characterization.

Implementation and Validation of the database

Technical details and results incorporated in this work include the use of publicly available corrosion-related parameters obtained from the developed database. These parameters were used as model inputs and incorporated into a series of electrochemical and empirical equations to calculate the overall corrosion rate for each pipeline segment. The resulting corrosion rate was then used to estimate the time to failure, defined as the time required for corrosion pit growth to result in critical material loss. Based on the estimated time to failure, pipeline segments were assigned discrete risk categories; segments predicted to fail within two years were classified under the highest risk category (Risk Level 5).

Feedback from using the developed database

The extraction of publicly available data was tested using the GitHub repository. The database was first integrated into TAMU workflow in the past quarter, and since then, updates implemented by the UD team have resolved several usability improving compatibility with TAMU risk-assessment model. The primary limitation identified is the lack of time-dependent variables within the available dataset. Because corrosion rate is time-dependent, the absence of data restricts the ability to model corrosion progression more accurately. However, it is to be noted that this limitation is understood to be an inherent constraint of the publicly available data rather than a deficiency of the pipeline database itself.

3)Database Testing Results from RU team

This quarter centered on the development and implementation of a dedicated Hydrological Hazard Risk Assessment Model. This module is specifically engineered to evaluate water-related threats to pipeline integrity, including seasonal flooding, heavy precipitation, snowmelt runoff and so on. The primary objective was to obtain a granular, multi-factor hydrological analysis for each pipeline segment. This involved the creation of a Python-based calculation engine capable of ingesting heterogeneous data from XML and GeoJSON sources, defining specific risk indicators, and establishing a weighted scoring algorithm to classify segments into risk tiers.

Design of the risk model being implemented in the current quarter

The risk model was designed as a semi-quantitative scoring system that aligns with the principles of relative risk assessment outlined in API RP-1187 and PHMSA guidelines. The methodology operates on a "Likelihood of Failure" basis driven by environmental loads. The process begins with data ingestion, where the system parses batch files containing geospatial attributes (GeoData) for pipeline coordinates. These raw environmental values such as flood depth in meters or precipitation in millimeters are then normalized onto a standardized 0-5 Consequence Index Scale, where 0 represents negligible risk and 5 represents critical risk.

To determine the segment-level risk, point-level data is aggregated using specific strategies. The model utilizes a Worst-Case (Maximum) aggregation for critical drivers like Flood Depth and River Width to ensure the highest risk point governs the segment score, while an Average (Mean) aggregation is applied to distributed factors like Precipitation. These aggregated scores are then processed through a weighted sum formula. The Total Risk Percent (R_{total}) is calculated as follows:

$$R_{total} = \sum_{i=1}^{n} (W_i \times \frac{S_i}{5})$$

Where W_i represents the weight of the factor and S_i represents the assigned score (0-5). The weighting system prioritizes factors based on their direct impact on pipeline stability, assigning Flood Inundation Depth (30%) the highest weight due to its correlation with buoyancy and scour threats. This is followed by Soil Moisture (20%) and Precipitation (20%), River Channel Width (15%), Snow Depth (10%), and Drainage Direction (5%).

Implementation and Validation of the database

The technical implementation relies on specific threshold-based logic to assign risk scores derived from the input data. For Flood Inundation Depth, the model assigns a critical score of 5.0 when depths exceed 2.0 meters, and a score of 4.0 for depths between 1.0 and 2.0 meters, reflecting the exponential increase in buoyancy forces and scour potential. Similarly, heavy rainfall is treated as a triggering event; the Precipitation logic dictates that rainfall exceeding 200mm results in a maximum score of 5.0, while levels above 100mm yield a score of 3.5. Soil Moisture is also evaluated, where saturation levels above 40% indicate a significant reduction in shear strength and are scored at the maximum level. The execution of the model generates a structured CSV output containing segment identification and the raw factor scores for all six hydrological indicators. The final Total Risk Percent allows for the classification of segments into distinct tiers: Low Risk (0 - 30%), Medium Risk (30% - 60%), and High Risk (> 60%). For example, a pipeline segment experiencing a 2.5m flood depth and 150mm rainfall would trigger high component scores, accurately reflecting the compounded threat of a major storm event at a river crossing.

A major challenge regarding data heterogeneity was resolved by implementing a unified merging function that successfully integrates legacy XML data with modern GeoJSON sources, ensuring no data loss during ingestion.

Additionally, the utility functions were refined to robustly handle missing or null tags by assigning safe default values (0.0), which prevented calculation failures during batch processing.

References

[1] Wei, Q., Sun, H., Xu, Y., Pang, Z., & Gao, F. (2024). Exploring the Application of Large Language Models Based AI Agents in Leakage Detection of Natural Gas Valve Chambers. Energies, 17(22), 5633. https://doi.org/10.3390/en17225633.

[2] Obi Chukwuemeka Nwokonkwo, Nwankwo Uchechukwu Samuel, Udoka Felista Eze, Adetokunbo MacGregor John-Otumu (2024). Machine Learning Framework for Real-Time Pipeline Anomaly Detection and Maintenance Needs Forecast Using Random Forest and Prophet Model. Automation, Control and Intelligent Systems, 12(2), 22–34.

[3] Amadhe, F. O., Anjorin, R. O., & Uwoghiren, F. O. (2024). Advancements in Machine Learning for Pipeline Integrity Management: A Comprehensive Review of Predictive and Optimization Techniques. Cognizance, 04(11). https://doi.org/10.47760/cognizance.2024.v04i11.011

[4] Vellum AI. (2025). LLM Leaderboard 2025. Vellum. Retrieved December 18, 2025, from https://www.vellum.ai/llm-leaderboard

[5] LLM Benchmarks. (2025). Performance comparison of Claude 3.5 Sonnet and Claude 3 Opus across benchmarks. Retrieved from LLMDB: https://llmdb.com/benchmarks

[6] Hugging Face. (2023). Open LLM Leaderboard (E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall & T. Wolf). Hugging Face. Retrieved from https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[7] LangChain. (2023). LangChain: Building applications with large language models. Retrieved from https://www.langchain.com

[8] LlamaIndex. (2023). LlamaIndex: Data framework for large language models. Retrieved from https://www.llamaindex.ai

[9] Meta AI. (2024). LLaMA 3: Open foundation and instruction-tuned large language models. Meta Platforms, Inc. Retrieved from https://ai.meta.com/llama/