LLM-AQRA: Leveraging LLMs for Efficient Urban Air Quality Research

Shanshan Jiang[1,*], Hanan Waheed[2], Dessislava Petrova-Antonova[3], Dumitru Roman[1,4]

[1] SINTEF AS, Norway – {shanshan.jiang,dumitru.roman}@sintef.no
[2] Oslo Metropolitan University, Norway – hawah7405@oslomet.no
[3] GATE Institute, Sofia University, Sofia, Bulgaria – dessislava.petrova@gate-ai.eu
[4] Bucharest University of Economic Studies, Romania
[*] Corresponding author

Keywords: Air Quality, LLM Research Assistant, Data Understanding and Exploration, Metadata Enrichment, Analytical Models

Abstract

Urban air quality research increasingly relies on integrating heterogeneous data from sources such as monitoring stations, satellite imagery, community sensors, built-environment datasets, and meteorological forecasts. However, researchers often face challenges in discovering, understanding, and integrating these datasets due to inconsistent metadata, mismatched spatial and temporal resolutions, diverse formats, and varying semantics. Moreover, effective use of air quality analytical models typically requires specialized expertise to interpret model requirements and identify suitable input datasets, leading to labor-intensive, inefficient workflows. This paper introduces LLM-AQRA, a prototype Large Language Model (LLM)-driven Air Quality Research Assistant designed to streamline these processes by automatically generating standard-compliant metadata, extracting model requirements, assessing dataset-model compatibility, and supporting map-based exploration of heterogeneous datasets. Initial user feedback indicates that LLM-AQRA is perceived as useful for dataset understanding and has the potential to reduce manual effort in analytical workflows. Future work will focus on retrieval augmented generation, ontology-guided constraints, expanded domain support, and broader user evaluations to improve reliability and applicability for real-world environmental analysis.

1. Introduction

Urban air quality research increasingly relies on a wide range of heterogeneous data sources, from governmental monitoring networks, satellite imagery, community sensors, built environment (e.g., roads, buildings, and green spaces), and meteorological forecasts. While many of these resources are openly available, researchers and practitioners routinely encounter barriers to discovering, understanding, assessing, and integrating them for analysis. In practice, the heterogeneity spans inconsistent or incomplete metadata, mismatched spatial and temporal resolutions, diverse data formats and semantics, and ambiguous or restrictive licensing (Fox et al., 2025, Singh and et al., 2025). These barriers complicate reproducible analysis and hinder the construction of coherent, cross-source urban knowledge.

Beyond the data acquisition and integration, there are expertise barriers that limit the translation of raw urban and environmental data into actionable insights. Researchers, particularly those outside data science, often lack the skills and tools needed to assess data quality, select appropriate analytical models, and identify relevant features (Zimmermann et al., 2025). This is particularly evident in air quality prediction models, whose correct usage typically requires deep knowledge of the model algorithm, inputs, and pre-processing pipelines, as well as an understanding of how to harmonize and integrate suitable input datasets (Fan et al., 2024, Karmoude et al., 2025). Consequently, scientific workflows are inefficient and effort-intensive, with substantial manual work spent on tasks that are repetitive yet require specialized knowledge, such as model understanding, dataset discovery, metadata curation, and dataset–model matching.

These challenges reveal several research gaps. First, there is a lack of robust, transparent methods for metadata enrichment that can harmonize heterogeneous sources and improve dataset findability and interoperability across space, time, and semantics (Zhang et al., 2025, Ahmed and Polini, 2025). Second, current tools insufficiently exploit geospatial context in both data and metadata to enable location-aware exploration and visualization that is meaningful for analysis. Third, while large language models (LLMs) show promise for information extraction, semantic alignment, and guided analytics (Ahmed and Polini, 2025, Tinn et al., 2025, Zimmermann et al., 2025), there is limited evidence on how to harness LLMs to (i) generate accurate and useful metadata for urban air quality datasets (Ahmed and Polini, 2025, Zhang et al., 2025), (ii) assist researchers in selecting and applying air quality prediction models (Fan et al., 2024), and (iii) automate tedious steps in the research workflow, such as dataset-model matching and data harmonization (Ahmed and Polini, 2025, Zhang et al., 2025). Addressing these gaps can substantially lower expertise barriers, improve research efficiency, and enhance interpretability in air quality analytics.

This paper aims to bridge these gaps by proposing an LLM-powered research assistant that can support researchers in exploring heterogeneous air pollution data together with other relevant urban datasets and respective analytical models used for air quality forecasting. The work focuses on the following research questions:

- *RQ1 (Metadata quality and utility)*: How can LLMs generate accurate and useful metadata for urban air quality datasets, thereby improving dataset discovery, comprehension, and interoperability?
- *RQ2 (Model literacy and automation)*: How can LLMs assist in analytical tasks and lower the expertise barrier,

e.g., by helping researchers understand air quality analytical models and automatically identify required datasets and their eligibility for running those models?

To answer the research questions, we designed and implemented an LLM-driven urban air quality research assistant (LLM-AQRA) that helps researchers discover air pollution datasets, understand air quality prediction models, and assess dataset–model compatibility.

The main contributions of the work are:

- A transparent, LLM-assisted data annotation and discovery pipeline that renders heterogeneous datasets comparable across space, time, and semantics through automated metadata enrichment and schema alignment.
- Integration of established analytical approaches, such as air quality prediction models, into an LLM-assisted workflow, which (i) guides model selection and usage; (ii) automates dataset–model matching based on model input requirements and dataset characteristics; and (iii) enables location-aware exploration and visualization.

The LLM-AQRA tool demonstrates the benefits of using an LLM to improve data understanding, metadata enrichment, and prediction model selection, thereby reducing manual workload and lowering barriers to domain and data science expertise. Initial user evaluation indicated the tool's usefulness and acceptance, as well as its potential to advance air quality research. The present study addresses the research questions primarily at the level of system design and feasibility; systematic assessment of metadata accuracy and matching quality is left for future work.

To our knowledge, LLM-AQRA is the first system that combines NGSI-LD–compliant metadata generation, geospatially grounded dataset exploration, LLM-based model literacy, and explicit dataset–model compatibility scoring in a single workflow for urban air quality research.

In the following, Section 2 presents related work. Section 3 describes the approach and main functionality of LLM-AQRA. Section 4 presents the validation of LLM-AQRA, including the prototype's user interface and results from an initial user evaluation. Finally, Section 5 concludes the paper.

2. Related Work

LLM-based metadata generation and dataset discovery: Recent work has demonstrated the potential of LLMs to automate metadata generation for open datasets. (Ahmed and Polini, 2025) fine-tuned T5 models to generate keywords and categories for datasets on the European Data Portal, demonstrating that such metadata can significantly improve dataset visibility. (Zhang et al., 2025) introduced AutoDDG, a framework for generating tabular dataset descriptions to enhance search-based discovery. (Tinn et al., 2025) proposed Pre-Meta, a domain-independent, retrieval-augmented pipeline that addresses LLM limitations in specialised domains such as biomedical genomics. Complementary research has explored DCAT-compatible metadata extraction using zero-shot and few-shot prompting (Busch et al., 2025) as well as RAG-based metadata enrichment for enterprise data catalogs (Singh and et al., 2025). (Gan et al., 2025) further demonstrated that LLMs can bridge the gap between natural-language queries and structured metadata fields in open data portals. (Fox et al., 2025) proposed a maturity model specifically for urban dataset metadata, implemented through a CKAN plugin. Despite this progress, no prior work evaluates LLM-generated metadata specifically for air quality datasets, where domain-specific semantics (e.g., pollutant codes, measurement methods, station coordinates, and temporal resolution) pose unique challenges for accuracy, consistency and interoperability. In contrast, LLM-AQRA focuses on these air-quality–specific semantics and integrates metadata generation with model analysis and dataset–model matching.

LLM-assisted analytical workflows and model literacy: A growing body of research investigates how LLMs can lower the expertise barrier for scientific data analysis. (Fan et al., 2024) proposed LLMAir, which adapts pre-trained LLMs for air quality prediction by incorporating spatio-temporal features, though the focus is on forecasting accuracy rather than user comprehension. (Stankov and Velinova, 2025) developed a RAG-based chatbot for querying air pollution documents, while (Duca et al., 2026) introduced MeteoChat, a semi-automatic framework that generates environmental reports from heterogeneous monitoring data using fine-tuned LLMs. (Gupta et al., 2025) demonstrated that LLMs can act as virtual annotators, directly converting raw sensor readings into AQI categories without manually coded rules. In related domains, S3LLM (Shaik et al., 2024) enables conversational exploration of large-scale scientific software through source code, metadata, and documentation, and (Zimmermann et al., 2025) catalogued 32 LLM applications spanning prediction, automation, and knowledge extraction in materials science. However, no existing system explains air quality prediction models to non-expert users and automatically links them to concrete datasets by assessing dataset eligibility against model requirements within a single workflow. In contrast, LLM-AQRA both extracts detailed model data requirements and computes transparent compatibility scores between those models and specific air-quality datasets, integrating these steps with metadata generation and geospatial exploration.

3. LLM-AQRA Approach

The LLM-AQRA leverages LLMs' capabilities to understand datasets and air-quality analytical models, and to automatically match datasets to their respective models. LLM-AQRA also harmonizes geospatial data and metadata (coordinates, regions, districts, etc.) by, e.g., converting non-standard field names according to standard schemas and incorporating them to support mapping, spatial comparison, and the integration of multiple datasets. Figure 1 illustrates the data pipeline for LLM-AQRA. The pipeline has four main components: *LLM-based NGSI-LD metadata generation, LLM-based model analysis, Dataset-model compatibility scoring*, and *Map-based visualization*. The components are described in detail in the following subsections. Referring to the research questions (RQs), RQ1 is addressed by the *LLM-based NGSI-LD metadata generation*, while RQ2 is addressed by the remaining components in combination.

3.1 Schemas and LLM-assisted Metadata Generation

High-quality and rich metadata are required for efficient discovery of relevant datasets. Although portals such as OpenAQ provide valuable station-level metadata, they generally do not offer standardized, model-oriented metadata describing dataset suitability, analytical use cases, interoperability characteristics, or compatibility with forecasting models. Often, users still need

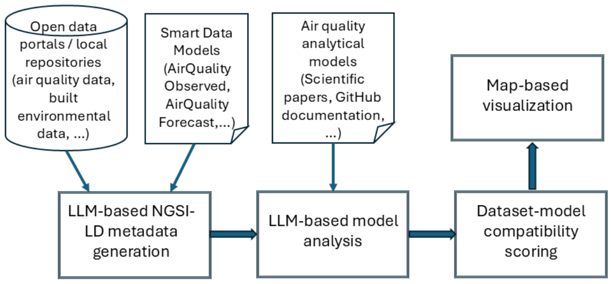

Figure 1. The LLM-AQRA pipeline.

substantial manual effort to assess whether a dataset satisfies specific modeling requirements.

Smart Data Models[1] are compatible, common data models used for replicable smart solutions across multiple sectors, including Smart Cities. They define common structures and semantics for representing domain-specific data, enabling interoperability and easier data sharing across systems and platforms. Each model is provided with a URI, a working URL with basic data about the attribute or entity, and examples of payloads for NGSIv2 and NGSI-LD[2].

In LLM-AQRA, LLMs are used to automatically generate metadata for both datasets and analytical models, based on schemas aligned with Smart Data Models & NGSI-LD. To generate metadata, LLM-AQRA combines air-quality station information (e.g., location, sensors, coordinates, and temporal coverage) with a sample of recent measurements and submits them to an LLM using a structured prompt grounded in these schemas. The prompt constrains the LLM to predefined vocabularies for key fields, such as dataset type, pollutants, measurement frequency, and recommended use cases, and instructs it to return a structured JSON description. This schema-guided approach improves consistency, interoperability, and comparability across heterogeneous datasets.

For air quality datasets, we use a data schema derived from the AirQualityObserved model[3] from Smart Data Models, extended with metadata such as the *list of air pollutants and recommended use cases*. The air quality analytical models are represented with a data schema for temporal forecasting models based on the AirQualityForecast model[4], extended with additional metadata including *model name and type, prediction horizon steps, past steps used, target pollutant, training defaults, evaluation metrics, and lead time hours*.

These schema extensions capture the details of the datasets and analytical models, and facilitate dataset discovery and dataset-model matching. Similarly, buildings, green areas and other urban environmental datasets can be described based on data schemas available for other Smart Data Models, e.g. for buildings[5] and green spaces[6].

3.2 Model Analysis

There are many analytical models for air quality, including traditional and statistical models such as Autoregressive Integrated Moving Average (ARIMA) and K-Nearest Neighbours (KNN), AI models based on Machine Learning, Deep Learning and Neural Networks, as well as hybrid approaches, merging the spatial feature extraction capabilities with the temporal dependencies. Expertise is often required to understand, select and effectively use an appropriate analytical model. For example, to properly interpret and apply an air quality forecast model, researchers need access to the following key information about the model's characteristics and requirements: the model's purpose and methodological approach, the target pollutants and prediction scope, the spatial and temporal resolution of data, and the required input datasets such as meteorological data, emission sources, monitoring observations, or Land Use and Land Cover (LULC) information. In addition, users must understand the necessary data preprocessing steps, expected output formats, and model performance metrics. Explaining these features is essential to enabling researchers, particularly non-data scientists, to correctly interpret the model's assumptions, identify suitable datasets, and apply the model within a reliable analytical workflow.

LLM-AQRA uses descriptions of an air-quality analytical model (e.g., a research paper or GitHub documentation) to automatically extract key information about the model, including its functionality, input data requirements, and expected outputs.

Table 1 provides key metadata for understanding the models, with example values. An NGSI-LD-compliant JSON description of the air quality forecast model (cf. Section 3.1) can be generated from these metadata.

To improve the accuracy of LLM output, established domain ontologies and taxonomies are used to guide the LLM generation, e.g., by constraining the output from the given allowable values from ontologies/taxonomies, as indicated in Table 1.

Metadata	**Example values**
Model type	Chemical transport model (CTM); land-use regression (LUR); dispersion model; statistical regression model; machine learning models (e.g., random forest, neural network). A taxonomy of the analytical model for air quality forecasting, based on (Karmoude et al., 2025), can be used.
Spatial resolution	Scale: global, regional, urban/city, street-level; grid resolution (e.g., 1 km, 10 km)
Temporal resolution	Annual, seasonal, daily, hourly, minute-level intervals
Prediction time step	1/3/12 hours
Target pollutants	$PM_{2.5}$, NO_2, O_3
Required input data	Air quality measurements, meteorological data, emission inventories, monitoring observations, land-use information.
Forecast horizon	Next 24-48 hours
Evaluation metrics	Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

Table 1. Examples of metadata for air quality prediction models.

3.3 Dataset-Model Matching with Explainable Compatibility Score

Traditionally, matching datasets to analytical models is challenging because researchers must manually verify whether datasets

[1] https://smartdatamodels.org
[2] https://ngsi-ld.org
[3] GitHub link for AirQualityObserved model: this link
[4] GitHub link for AirQualityForecast model: this link
[5] GitHub link for buildings model: this link
[6] GitHub link for green spaces model: this link

meet model requirements, such as spatial and temporal resolution, geographic coverage, variables, and formats, particularly when data come from heterogeneous sources with inconsistent metadata. While prior work has improved dataset discovery and interoperability through metadata standards, semantic models, and LLM-based information extraction, limited attention has been given to automatically linking dataset characteristics with model requirements and assessing their compatibility.

To support the practical use of air quality analytical models, LLM-AQRA compares metadata from available datasets with the input requirements extracted from model documentation. Based on this comparison, a *dataset-model compatibility score* is calculated that measures how well the dataset metadata matches the requirements extracted from the analytical model documentation. The highest-ranked datasets can then be explored through user interfaces that visualize temporal forecasts and spatial distributions, allowing users to assess the potential analytical outcomes before executing the full workflow.

The score is calculated using a set of compatibility criteria that capture key aspects of dataset usability, including required variables, spatial characteristics, temporal properties, and documentation quality. In general, the compatibility score C is computed as a weighted sum of individual criterion scores:

$$C = \sum_{i=1}^{n} w_i S_i \tag{1}$$

where S_i represents the normalized compatibility score for criterion i, and w_i denotes the weight assigned to that criterion. Each S_i takes a value between 0 and 1, where 0 indicates no compatibility and 1 indicates full compatibility with the model requirements. The weights reflect the relative importance of each criterion and satisfy the constraint

$$\sum_{i=1}^{n} w_i = 1 \tag{2}$$

ensuring that the final compatibility score lies between 0 and 1.

To demonstrate LLM-AQRA's usefulness, we instantiate this generic framework using seven criteria derived from common characteristics of air quality analytical models. These criteria include pollutant coverage, location match, measurement frequency, quality documentation, temporal range overlap, coverage duration, and geospatial availability. The weights assigned to these criteria are summarized in Table 2. These weights can be adapted to the needs of the users and the context.

This scoring approach allows the tool to rank candidate datasets transparently and guide users toward those most appropriate for training and execution of the selected air quality analytical model.

The overall compatibility score may not fully reflect a dataset's practical suitability. For example, even if the geographic location specified in the model documentation does not match the dataset location, the dataset may still be applicable, since the analytical model can be applied to regions different from those used in the original study. Therefore, the compatibility score should be interpreted as a guidance indicator rather than a strict requirement, allowing users to consider alternative datasets that satisfy most technical conditions.

For this purpose, LLM-AQRA provides individual matching scores (as demonstrated in the prototype in Section 4) and allows users to decide whether a mismatch is desired, unwanted, or even preferred.

3.4 Map-based Visualization

LLM-AQRA supports an interactive multi-layer map visualization of heterogeneous datasets (e.g., air quality observations, prediction outputs, built environmental data), facilitating spatial exploration and contextual analysis of urban environmental data. This capability is particularly important for urban air quality studies, where understanding pollution patterns often requires examining the relationships between environmental measurements and the built environment. By integrating heterogeneous datasets within a shared spatial context, LLM-AQRA enables more intuitive exploration of dataset suitability and potential analytical insights before running air-quality prediction models.

3.5 Key Features of LLM-AQRA

Based on the design considerations described in Section 3 and an initial exploratory user evaluation (see Section 4.2), a list of main features for an LLM-driven air quality research assistant is identified, as shown in Table 3.

3.6 Why LLMs?

Existing metadata standards, ontologies, and rule-based approaches provide important foundations for interoperability but depend on predefined schemas and manually engineered mappings. In contrast, information about air quality datasets and analytical models is often embedded in heterogeneous metadata records, scientific publications, and technical documentation that are difficult to process using fixed rules alone. LLMs are particularly valuable in this context because they can interpret unstructured text, identify relevant domain concepts, and transform them into structured, standards-compliant metadata. Therefore, LLM-AQRA uses LLMs as an intelligent semantic layer that complements existing standards by automating metadata generation, model interpretation, and dataset–model compatibility assessment across diverse sources.

4. Validation of LLM-AQRA

To validate LLM-AQRA and demonstrate its features, a basic workflow for air quality research is defined, which has the following three steps:

- *Dataset Exploration:* The researcher uses LLM-AQRA to search for available datasets with certain criteria (typically location and type of data) and explore the data through interactive map-based visualization.
- *Model Analysis:* The researcher identifies an air quality analytical model, analyzes the model and identifies data requirements from relevant research papers and model documentation.
- *Dataset-Model Matching:* The researcher determines how selected datasets satisfy the requirements of the identified prediction model.

https://doi.org/10.5194/isprs-annals-XII-4-W2-2026-73-2026 | © Author(s) 2026. CC BY 4.0 License.

Criterion	Weight	Description
Pollutant coverage	0.30	Availability of the pollutants required by the model (e.g., $PM_{2.5}$, NO_2, O_3).
Location match	0.20	Match between the dataset's geographic coverage and the study area described in the model documentation.
Measurement frequency	0.15	Alignment between the dataset's temporal resolution (e.g., hourly, daily) and the input frequency required by the model.
Quality documentation	0.10	Availability and completeness of metadata describing measurement reliability, completeness, or uncertainty indicators. Note that this criterion measures whether quality information is documented, not the quality level itself.
Temporal range overlap	0.10	Degree of overlap between the dataset time span and the time range required for model training or evaluation.
Coverage duration	0.10	Length of the dataset record and whether sufficient historical data is available for the modeling task.
Geospatial availability	0.05	Presence of spatial metadata such as coordinates or geometry enabling spatial analysis and visualization.

Table 2. Example criteria used to calculate the dataset-model compatibility score in the prototype.

ID	Feature description
F1	Automatic NGSI-LD metadata generation from air quality station data (e.g., OpenAQ or local sources) and built environmental data.
F2	Automatic extraction of model data requirements (e.g., pollutants, temporal frequency, spatial coverage).
F3	AI-assisted model analysis, including model overview, algorithm description, and input/output interpretation.
F4	Dataset-to-model compatibility scoring with detailed dimension breakdown.
F5	Support for multiple LLM providers (e.g., Groq, OpenAI, Gemini).
F6	Natural language query interface for searching and exploring datasets.
F7	LLM-generated dataset summaries explaining variables, pollutants, and potential research uses.
F8	Explainable AI outputs showing how dataset compatibility scores are calculated.
F9	Interactive dataset exploration across heterogeneous data sources with map-based visualization.

Table 3. Key features of the proposed LLM-based air quality research assistant.

The LLM-AQRA implements the above workflow leveraging LLMs to automatically analyze analytical models, extract data requirements from the models, and enrich metadata descriptions of datasets. In addition, the LLM-AQRA prototype generates standard-compliant metadata in NGSI-LD JSON format and provides basic analytical visualizations. Moreover, the tool includes a configuration module that allows users to select LLM models and workflow settings. This prototype has implemented and demonstrated the identified features as shown in Table 3.

In the following, Section 4.1 describes the prototype tool that supports this workflow and Section 4.2 presents the results from an initial user evaluation of the prototype.

4.1 User Interfaces

4.1.1 Dataset Exploration and Visualization Users can explore datasets from open portals or private sources and automatically generate rich, ontology-aware metadata that follows NGSI-LD standards, leveraging the selected LLM (**Features F1 and F7**). The *Search OpenAQ* feature searches for air quality data on the OpenAQ portal using a city name, country code, bounding box, or point+radius. Such search functionality is provided via the portal's API. The *Search Environmental Data* feature is used to search for built-environment data, such as buildings and green spaces. The *Generate Metadata* feature sends a sample of recent measurements from the user-selected datasets to the LLM, which then generates a JSON description that complies with the NGSI-LD *AirQualityObserved* schema. The JSON description includes metadata on dataset type (e.g., government station, community sensor), location (city), a list of pollutants and their units, measurement frequency, temporal coverage, data quality score, description, and recommended use cases. This metadata is displayed in a formatted table and is stored for later matching. Figure 2 shows the results from the search and the NGSI-LD metadata generated, as well as a natural language query interface for exploring the datasets (**Feature F6**).

LLM-AQRA supports interactive map-based visualization with multiple layers (**Feature F9**), enabling users to simultaneously display heterogeneous datasets (such as air quality observations and buildings) and explore their spatial relationships. Figure 3 shows an interactive map with two datasets: a selected air quality monitoring station and a dataset of nearby streets. Selecting a station reveals detailed metadata, such as coordinates, pollutants, and recommended use cases (not shown in the figure).

4.1.2 Model Analysis Given a research paper and optionally GitHub documentation (such as the README file), LLM-AQRA automatically extracts information about the analytical model's purpose and functionality, its type, target pollutants, spatial and temporal resolution, required input data, generated output, use cases and other details (**Features F2 and F3**). Figure 4 shows an example result of the analytical model analysis, using the SATADL deep learning model as a test case.

4.1.3 Dataset-Model Matching In the *Matching & Analysis* page, available datasets and model requirements are summarized at the top. To find the best-matching datasets, LLM-AQRA computes a compatibility score for each dataset using a weighted combination of pollutants, frequency, location, temporal overlap, geospatial availability, coverage duration, and data quality, as explained in Section 3.3 (**Feature F4**). The results are shown in a table, sorted by score.

https://doi.org/10.5194/isprs-annals-XII-4-W2-2026-73-2026 | © Author(s) 2026. CC BY 4.0 License.

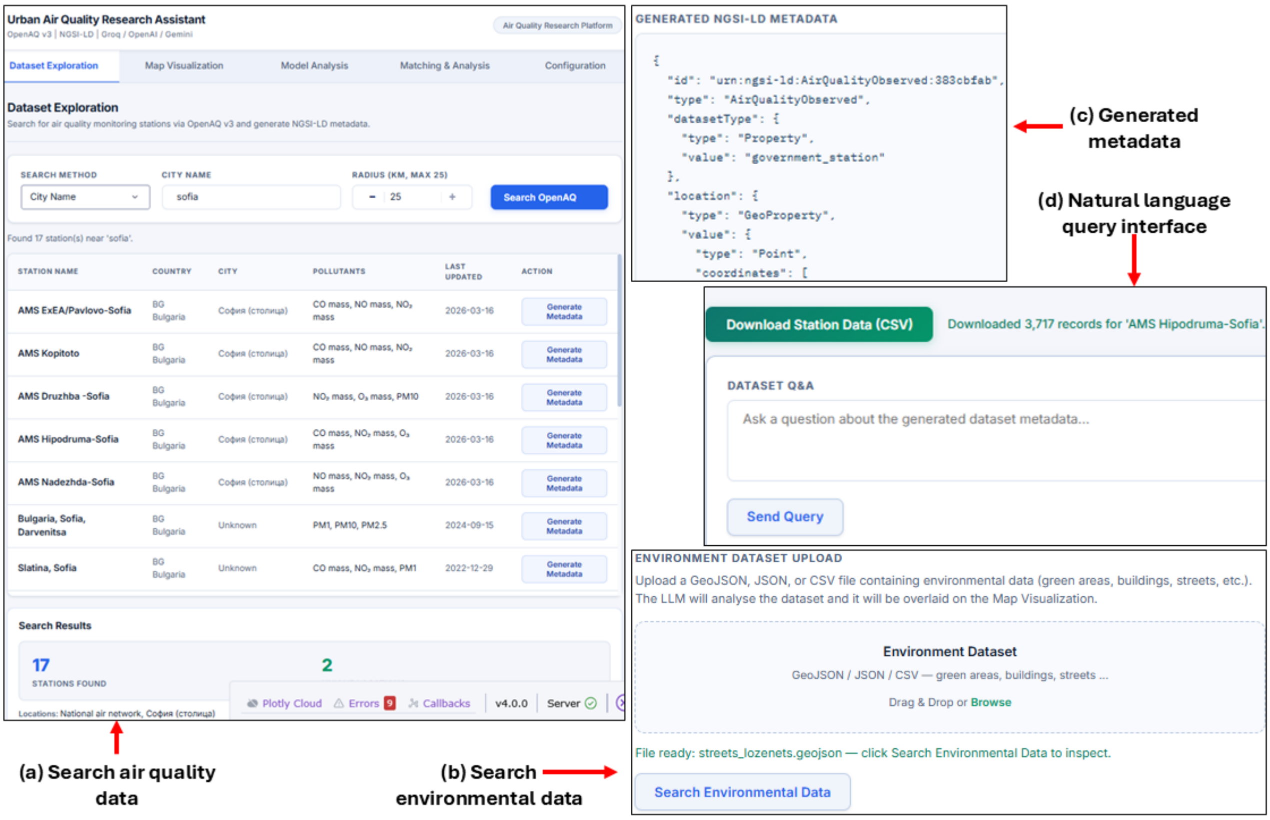

Figure 2. Dataset exploration: (a) results from *Search OpenAQ*, (b) upload and search environmental datasets, (c) generated metadata for a selected dataset, (d) natural language query interface for exploring datasets.

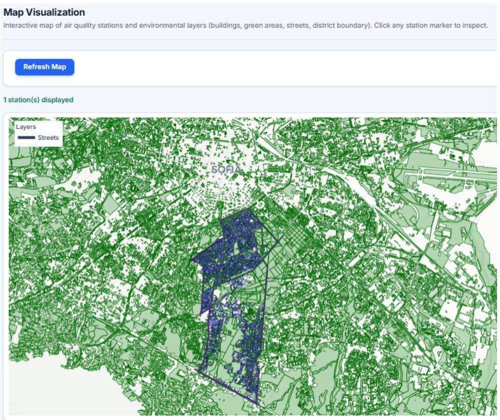

Figure 3. Interactive map visualization: datasets of air quality stations and streets shown on the same map.

To support **Feature F8** for matching analysis and explainability, LLM-AQRA allows users to select a dataset from the result table and then choose one of the analysis tabs, as exemplified in Figure 5:

- *Compatibility Breakdown* shows how the selected dataset fulfils each model requirement across seven dimensions.
- *Temporal Overview (temporal analysis)* provides an illustrative visualization consisting of a simulated 48-hour historical trace plus a mock 24-hour projection generated by simple pattern repetition) for the selected dataset. It does not constitute a model-based forecast.
- *Pollutant Coverage (spatial analysis)* provides an illustrative bar chart of pollutant levels across e.g. a district, intended to preview the type of spatial output an analytical model could produce.

LLM-AQRA does not perform air quality forecasting. The temporal and spatial views described above are illustrative previews only. The tool's role is to support metadata generation, model interpretation, and dataset-model matching prior to running an external analytical model. Consequently, forecasting performance and predictive accuracy are outside the scope of this study.

Figure 6 shows the configuration page for selection of LLMs and relevant settings (**Feature F5**).

4.2 User Evaluation

The perceived usefulness of LLM-AQRA was evaluated in a demonstration session with 21 participants, including undergraduate and postgraduate students, academic researchers, and domain experts. Feedback was collected through a guided discussion with all attendees and an individual survey completed by six participants. The evaluation concerned an earlier version of the prototype; the refinements described in this paper were implemented in response to this feedback.

The evaluation results indicate a generally positive perception of the proposed tool among participants. Most respondents reported that the interface was clear and easy to understand, and all participants indicated that the workflow, comprising dataset exploration, model analysis, and dataset–model matching, was logical and easy to follow. The most valued functionalities were dataset exploration with automatic metadata generation, model

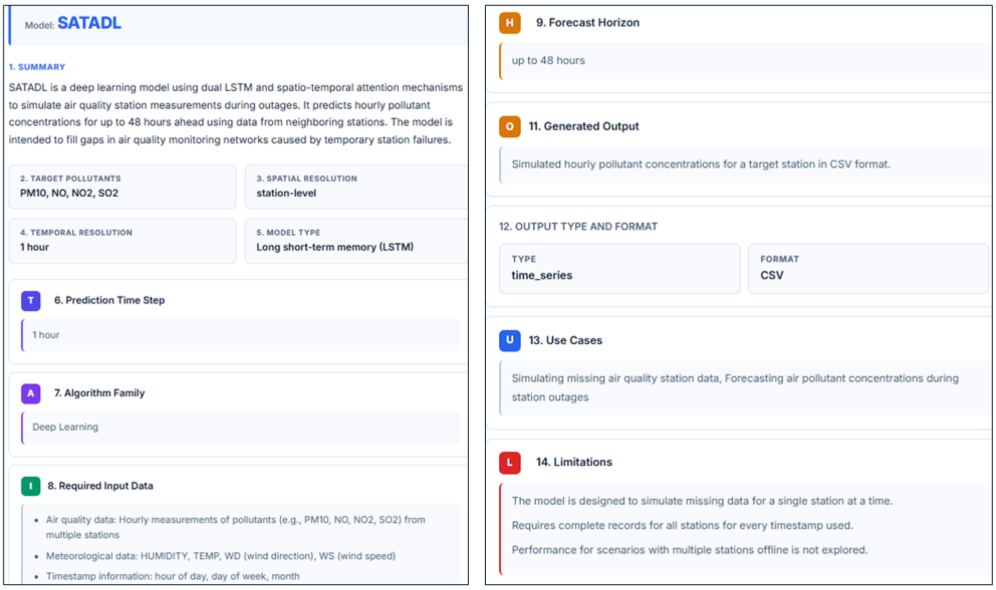

Figure 4. Model details: results from model analysis.

Figure 5. Matching results with compatibility scores across seven dimensions.

analysis, and dataset–model compatibility matching. Participants expressed strong interest in capabilities related to explainable compatibility scoring, interactive map-based dataset exploration, and support for multiple LLM providers, and also suggested that transparency, spatial analysis, and system flexibility are important aspects for future development. They indicated that the tool has the potential to save time in dataset discovery and model understanding, and most would consider using a tool like LLM-AQRA in their own research. Feedback emphasized the value of fast, structured model summaries and automatic extraction of model requirements. Identified areas for improvement included enhancing analytical depth, clarifying the interpretation of the compatibility score, and adding capabilities such as anomaly detection, temporal alignment, and richer model summaries.

Limitations include a small number of survey responses (6 of 21 attendees) and a focus on user perceptions rather than objective performance measures. Metadata accuracy, model requirement extraction, compatibility scoring, and comparisons with alternative approaches were not assessed. Therefore, the results should be interpreted as preliminary evidence of usefulness rather than technical validation.

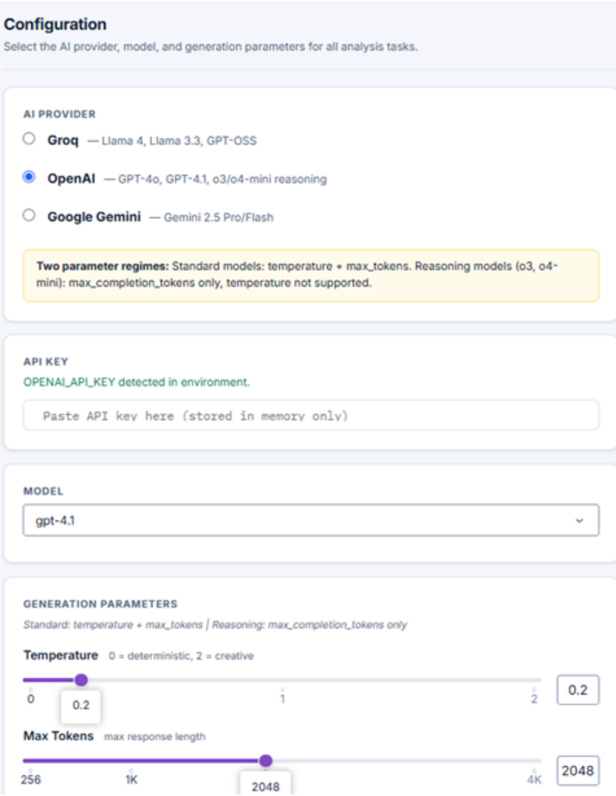

Figure 6. Configuration page

The evaluation results suggest strong user interest and indicate the potential of LLM-AQRA to support and streamline workflows in air quality research, while also revealing areas that require further improvement before wider adoption. Based on survey responses and feedback from the demonstration session, the list of desired features was refined, and the prototype was further improved. These updates include enhanced model summaries and other refinements described in this paper.

5. Conclusion and Future Work

This paper introduced LLM-AQRA, an LLM–based research assistant designed to support urban air quality research by improving metadata generation, dataset exploration, model interpretation, and dataset–model compatibility assessment. The main contributions are: (1) an LLM-assisted pipeline for generating structured, NGSI LD–compliant metadata for heterogeneous environmental datasets; (2) automated extraction of analytical model requirements and transparent model literacy support; and (3) an interpretable dataset–model matching approach that ranks datasets based on multiple compatibility criteria. Initial user feedback suggests that these capabilities are useful for dataset understanding and have the potential to support more efficient analytical workflows.

Future work will focus on improving the robustness and reliability of LLM outputs through RAG, constrained decoding, ontology-guided prompting, and additional validation layers. Expanding support for broader urban domains and enhancing geospatial analytics and harmonization capabilities are also key priorities. In addition, incorporating uncertainty estimation, interactive refinement workflows, and more extensive user studies and performance evaluations will further strengthen LLM-AQRA's applicability to real-world environmental analysis.

Acknowledgments. This work was funded through the projects DataPACT (HE 10189771), CauseFinder (PNRR 760049). DataTwin (BG16RFPR002-1.009-0005), 857155 (H2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme), and BG16RFPR002-1.014-0010-C01.

References

Ahmed, U., Polini, A., 2025. Enhancing Open Data Findability: Fine-Tuning LLMs (T5) for Metadata Generation. *Conference on Digital Government Research*, 26.

Busch, L., Tebernum, D., Velarde, G., 2025. Exploring llm capabilities in extracting dcat-compatible metadata for data cataloging. *Proc. of DATA 2025*, SCITEPRESS, 299–309.

Duca, A. L., Duca, R. L., Marinelli, A., et al., 2026. Semi-Automated Reporting from Environmental Monitoring Data Using a Large Language Model-Based Chatbot. *ISPRS International Journal of Geo-Information*, 15(2), 80.

Fan, J., Chu, H., Liu, L., Ma, H., 2024. LLMAir: Adaptive Reprogramming Large Language Model for Air Quality Prediction. *ICPADS*, 423-430.

Fox, M. S., Gajderowicz, B., Lyu, D., 2025. A Maturity Model for Urban Dataset Meta-data. *ArXiv*, abs/2402.05211.

Gan, L.-Y., Das, A., Walker, J., Simperl, E., 2025. Keywords are not always the key: A metadata field analysis for natural language search on open data portals. *ArXiv*, abs/2509.14457.

Gupta, R., Mirza, A. A., Danilov, C., et al., 2025. Llm-powered data annotation for bridging the semantic gap in air quality monitoring. *Proc. of BuildSys'25*, 394–399.

Karmoude, M., Munhungewarwa, B., Chiraira, I., et al., 2025. Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks. *Science of The Total Environment*, 1002, 180593.

Shaik, K., Wang, D., Zheng, W., et al., 2024. S3LLM: Large-Scale Scientific Software Understanding with LLMs using Source, Metadata, and Document. *ArXiv*, abs/2403.10588.

Singh, M., et al., 2025. Leveraging Retrieval Augmented Generative LLMs For Automated Metadata Description Generation to Enhance Data Catalogs. *ArXiv*, abs/2503.09003.

Stankov, S., Velinova, D., 2025. Application of a large language model for air pollution analysis by using a rag system. *Proc. of ICST'25*, 2, 345–351.

Tinn, P., Sørbø, S., Jiang, S., et al., 2025. Pre-Meta: priors-augmented retrieval for LLM-based metadata generation. *Bioinformatics*, 41(10), btaf519.

Zhang, H., Liu, Y., Santos, A., et al., 2025. AutoDDG: Automated Dataset Description Generation using Large Language Models. *ArXiv*, abs/2502.01050.

Zimmermann, Y., Bazgir, A., Al-Feghali, A., et al., 2025. 32 examples of LLM applications in materials science and chemistry: towards automation, assistants, agents, and accelerated scientific discovery. *Mach. Learn.: Sci. Technol.*, 6.