

VayuChat: An LLM-Powered Conversational Interface for Air Quality Data Analytics

Vedant Acharya*
vedant.acharya@iitgn.ac.in
Indian Institute of Technology
Gandhinagar
Gandhinagar, Gujarat, India

Abhay Pisharodi*
abhay.pisharodi@iitgn.ac.in
Indian Institute of Technology
Gandhinagar
Gandhinagar, Gujarat, India

Rishabh Mondal
rishabh.mondal@iitgn.ac.in
Indian Institute of Technology
Gandhinagar
Gandhinagar, Gujarat, India

Mohammad Rafiuddin
mohammad.rafiuddin@ceew.in
Council on Energy, Environment and
Water
New Delhi, Delhi, India

Nipun Batra
nipun.batra@iitgn.ac.in
Indian Institute of Technology
Gandhinagar
Gandhinagar, Gujarat, India

Abstract

Air pollution causes about 1.6 million premature deaths each year in India, yet decision makers struggle to turn dispersed data into decisions. Existing tools require expertise and provide static dashboards, leaving key policy questions unresolved. We present *VayuChat*, a conversational system that answers natural language questions on air quality, meteorology, and policy programs, and responds with both executable Python code and interactive visualizations. *VayuChat* integrates data from Central Pollution Control Board (CPCB) monitoring stations, state-level demographics, and National Clean Air Programme (NCAP) funding records into a unified interface powered by large language models. Our live demonstration will show how users can perform complex environmental analytics through simple conversations, making data science accessible to policymakers, researchers, and citizens. The platform is publicly deployed at <https://huggingface.co/spaces/SustainabilityLabIITGN/VayuChat>. For further information check out video uploaded on <https://www.youtube.com/watch?v=d6rklL05cs4>.

Keywords

Air Quality Analytics, Large Language Models, Conversational AI, Code Generation, Environmental Data

ACM Reference Format:

Vedant Acharya, Abhay Pisharodi, Rishabh Mondal, Mohammad Rafiuddin, and Nipun Batra. 2025. *VayuChat: An LLM-Powered Conversational Interface for Air Quality Data Analytics*. In *Proceedings of 13th International Conference on Data Science (CODS 2025)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CODS 2025, IISER, Pune, India

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Air pollution in India causes over 1.6 million premature deaths annually and reduces life expectancy by more than five years due to PM_{2.5} exposure [8, 12]. The Central Pollution Control Board (CPCB)¹ continuously collects nationwide measurements of various pollutants [4]. However, converting these raw datasets into timely, policy-relevant insights remains a challenge [5]. Even simple queries such as “How did PM_{2.5} vary last year in Delhi?” remain hard to answer. In contrast, policy questions like “Which cities reduced PM_{2.5} most relative to NCAP² funding?” require integrating pollution, funding, and population data. Existing tools often require expertise [2], offer only limited visualizations [6], or handle only narrow tasks, leaving stakeholders dependent on specialists. Large language models (LLMs) can map natural-language queries to executable analyses [1, 11], yet their capacity for domain-specific, multi-dataset air quality analytics remains untested. Recent advances in instruction following [11] and zero-shot tabular reasoning [14] show promise, but rigorous domain evaluation is needed for decision support. We present *VayuChat*, evaluated on our benchmark *VayuBench* (whose details are beyond the scope of this paper), a system designed to fill this gap by offering a conversational interface that allows users to ask questions in natural language and receive both visual results and executable Python code. The primary motivation for using the code generation capabilities of LLMs in *VayuChat* is to reduce hallucinations. Instead of giving the LLM the raw table, which can lead to hallucinations, we provide the dataset schema in the system prompt and have it generate Python code from the user’s query, ensuring reliable and transparent answers. A user can ask questions like “Which cities had the worst air quality in winter 2023?” or “Which meteorological factor has the strongest correlation with the reduction of PM_{2.5} levels?”, and *VayuChat* will return immediate, data-driven answers, accompanied by the underlying code. *VayuChat* integrates three complementary datasets: (i) continuous air quality and meteorological observations from the Central Pollution Control Board (CPCB), (ii) population and area

¹A statutory body under the Government of India responsible for monitoring and regulating water and air pollution, operating over 500 stations under the National Air Monitoring Programme (NAMP).

²Launched in 2019, NCAP allocated Rs. 9,650 crore to 131 non-attainment cities through FY 2019–24.

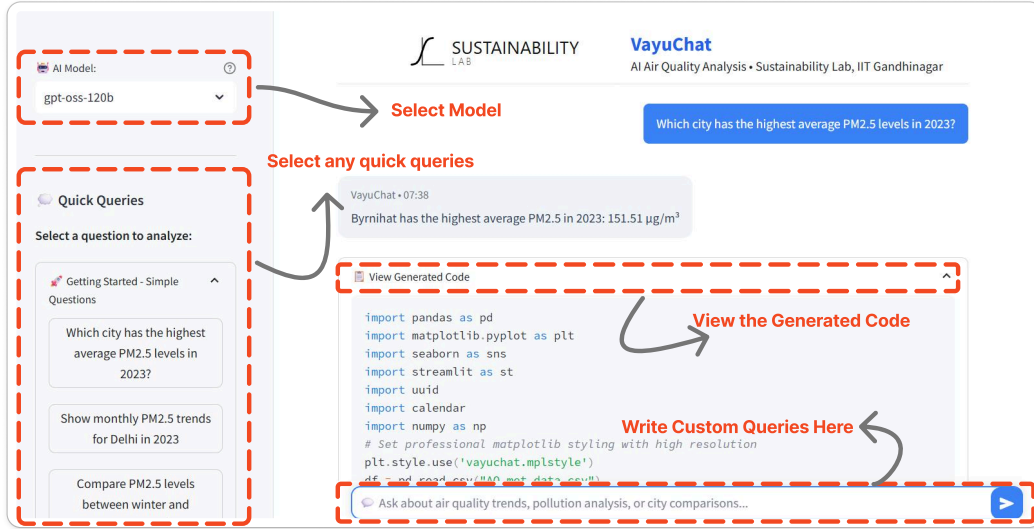


Figure 1: VayuChat Interface with key features: (1) **Select Model** - AI model dropdown (gpt-oss-120b), (2) **Quick Queries** - pre-defined air quality questions (e.g., PM2.5 comparisons, trends), (3) **Generated Code** - see generated Python code for analysis/visualization, and (4) **Custom Queries** - natural language input. Example query shows the highest PM_{2.5} level in 2023, with Byrnihat recording 151.51 $\mu\text{g}/\text{m}^3$.

at the state level, and (iii) funding allocations under the National Clean Air Programme (NCAP). The conversational layer is powered by multiple state-of-the-art large language models, enabling comparative evaluation of alternative AI approaches for analytical tasks. This design facilitates rigorous assessment of model capabilities while lowering the barrier for policymakers, researchers, and citizens to perform complex analyses on air-quality datasets.

2 System Description

In this section, we describe the user interface of *VayuChat* and its backend functionality.

2.1 User Interface

As shown in Figure 1, *VayuChat* provides a browser-based interface where users can either select from quick query categories or enter their own custom air-quality related queries. Users can also choose the models they wish to use. Once a query is submitted, *VayuChat* returns both the visual or textual answer and the corresponding code that generated the result.

2.2 Backend

As shown in figure 2, the backend of *VayuChat* is responsible for processing user queries using the system prompt, invoking the appropriate model for response generation, and executing the code produced by the LLMs within a sandboxed environment.

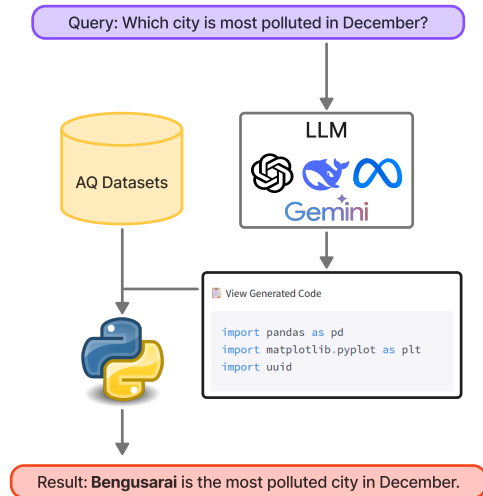


Figure 2: Flow diagram of *VayuChat* workflow: the user submits a query, the selected LLM generates code, which is executed in Python with the relevant dataset, and the output is presented.

2.2.1 Datasets and System Prompt. In the system prompt, we provided descriptions of the datasets used. Below are the three datasets that we used in the system prompt:

CPCB Data: This dataset contains daily measurements of pollutants and weather variables for each CPCB monitoring station from

2017 to 2024 [3]. Additionally, it includes station metadata, such as city, state, and the station ID assigned by CPCB. The detailed descriptions of the measured environmental variables, along with their units, are provided below: *Pollutants*: PM_{2.5} ($\mu\text{g}/\text{m}^3$), PM₁₀ ($\mu\text{g}/\text{m}^3$), NO ($\mu\text{g}/\text{m}^3$), NO₂ ($\mu\text{g}/\text{m}^3$), NO_x (ppb), NH₃ ($\mu\text{g}/\text{m}^3$), SO₂ ($\mu\text{g}/\text{m}^3$), CO (mg/m³), Ozone ($\mu\text{g}/\text{m}^3$). *Weather variables*: Air Temperature (AT, °C), Relative Humidity (RH, %), Wind Speed (WS, m/s), Wind Direction (WD, deg), Rainfall (RF, mm), Total Rainfall (TOT-RF, mm), Solar Radiation (SR, W/m²), Barometric Pressure (BP, mmHg), and Vertical Wind Speed (VWS, m/s).

State population data: This dataset provides information on the population and area (km²) of 31 Indian regions according to the census data from 2011. It also includes a boolean column, `isUnionTerritory`, which indicates whether a region is a Union Territory.

NCAP Funding: This dataset provides financial year-wise records of funding released across Indian cities under the National Clean Air Program (NCAP) from 2019 to 2022. It includes columns for the total funds released and their utilization status as of June 2022.

2.2.2 Models. *VayuChat* integrates a range of state-of-the-art models, including GPT-OSS 20B [10] and 120B [10], Qwen3-32B [13], Llama variants, DeepSeek-R1 [7], and the latest Gemini models. The Gemini models are accessed through the Gemini API, while the other open-source models are accessed via the Groq Cloud API [9].

2.2.3 Response Generation. When a user submits a query, it is combined with the system prompt, which includes instructions that the dataframes are already loaded and passed to the selected model. The model then generates a Python code snippet based on the query. The generated code is stored and made accessible to the user, ensuring transparency and reproducibility of the results.

2.2.4 Code Execution. *VayuChat* executes the code generated by the LLM in a secure, sandboxed environment to prevent any potential system harm. The execution results are captured and integrated into the response. Depending on the nature of the query and the output produced, the results are presented to the user in different formats such as text, structured tables, or visualizations (e.g., plots and charts).

2.2.5 Plot Generation. The *VayuChat* system automatically generates visualizations in response to relevant analytical queries. Figure 3 and Figure 4 demonstrate examples of plot-based responses, where the system produces graphical outputs corresponding to user prompts requesting data visualization.

2.2.6 Capabilities of VayuChat. *VayuChat* analyzes comprehensive air-quality and meteorological data (2017–2024), including CPCB pollutant measurements, weather parameters, and state population and area information. Its capabilities can be grouped as follows:

- ✓ *Direct queries*: “Which city had the highest PM_{2.5} in 2023?”; “Show SO₂ levels for Delhi.”
- ✓ *Plot generation*: “Plot PM_{2.5} trends for Mumbai”, “Compare ozone levels between Punjab and Gujarat.”
- ✓ *Analytical queries*: “Analyze wind speed and PM_{2.5} correlation”; “Evaluate NCAP impact on air quality.”

The system generates both Python code and visualizations through code-generation LLMs, ensuring transparency and reproducibility.

3 VayuChat Case Study: Delhi Air Quality Analysis

We worked with air quality analysts to investigate the reasons behind Delhi’s severe pollution spike in December 2024 using our platform, *VayuChat*. In this section, we present the insights generated through *VayuChat* by asking targeted questions that break down the causes of the crisis. Our analysis uses CPCB data from 37 monitoring stations across Delhi, together with meteorological observations, to reveal significant trends. We evaluated *VayuChat* in two complementary ways: (i) by assessing LLM performance using *VayuBench* (details outside the scope of this paper), and (ii) by collaborating with experts and end users in real-world analytical scenarios. Our investigation proceeded through a systematic series of interconnected questions that progressively deepened the analysis:

Initial Identification: “Which days in Delhi were the most polluted in December 2024?” This pinpoints the crisis period.

Temporal Pattern Analysis: “Use a time series plot to compare pollution levels and wind speed during Delhi’s most polluted week in December 2024 to the preceding and following 15 days.” This examines short-term dynamics.

Historical Contextualization: “Plot and compare pollution levels and wind speed in Delhi during a polluted week in December 2024 with the data from the previous five years.” Assesses crisis severity relative to historical patterns.

Multi-Pollutant Correlation: “Analyze the correlation between CO, NO₂, and PM_{2.5} levels in Delhi for December from 2017 onwards.” Investigates pollutant interaction patterns.

This sequential questioning approach allowed analysts to build understanding from crisis identification to mechanistic insights. In future, we envision the system to automatically reason out such sub-questions to provide complete investigation workflows. A striking finding is the strong inverse relationship between wind speed and PM_{2.5}: even a 0.6 m/s drop trapped pollutants, shifting air quality from “Very Poor” to “Severe” within hours. These results show how subtle weather changes can turn Delhi’s winter air into a hazardous blanket for 30 million residents, offering insights for both emergency planning and long-term policy.

3.1 Most Polluted Dates in December 2024 Delhi

User Prompt: Which days in Delhi are the most polluted in December 2024?

VayuChat Response: Based on our analysis of December 2024 data, the most severely polluted dates in Delhi are presented in **Table 1**.

Date	PM _{2.5} ($\mu\text{g}/\text{m}^3$)
18-12-2024	344.59
19-12-2024	341.46
17-12-2024	330.25
20-12-2024	291.46
22-12-2024	285.98

Table 1: (Generated by VayuChat) Most Polluted Dates in December 2024 Delhi

3.2 Wind Speed Analysis: Most Polluted Week vs. Surrounding Periods

User Prompt: Use a time series plot to compare the pollution levels and wind speed during Delhi’s most polluted week in December 2024 to the previous and after 15 days?

VayuChat Response: The most polluted week was identified in December 2024 (see Fig. 3).

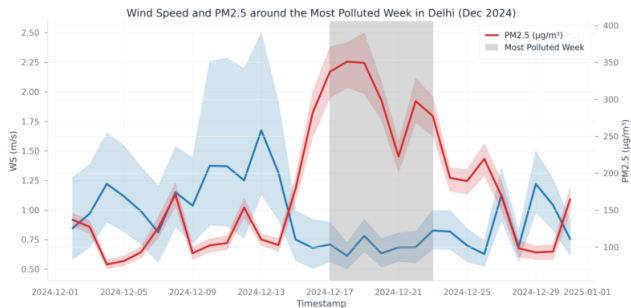


Figure 3: (Generated by VayuChat) Wind Speed vs PM_{2.5} Concentration - December 2024 Critical Week

Key Observation: A clear inverse relationship exists between wind speed and PM_{2.5}. When wind speeds dropped below 1.0 m/s, PM_{2.5} levels exceeded 300 µg/m³, as stagnant conditions trapped pollutants near the surface. Even small decreases in wind speed rapidly pushed air quality from “Very Poor” to “Severe,” showing the sensitivity of Delhi’s winter pollution to meteorological variations.

3.3 Five-Year Comparison: December Pollution and Wind Speed Trends

User Prompt: Plot and compare the pollution levels and wind speed in Delhi during a polluted week in December 2024 with the data from the previous five years.

VayuChat Response: Comparison of December 2024 with previous years (2019–2023), as shown in Fig. 4.

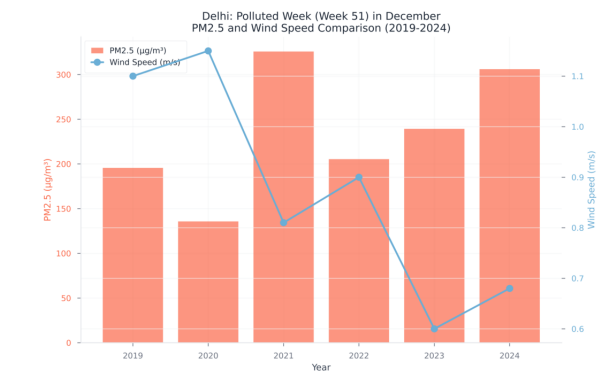


Figure 4: (Generated by VayuChat) Five-Year December PM_{2.5} and Wind Speed Trends (2019-2024)

Key Observation: In 2024, wind speeds showed marginal improvement compared to previous years. While we observe a strong

negative correlation between PM_{2.5} and ind speed during 2019 and 2020, a more comprehensive analysis requires considering both temperature and wind speed. The lowest wind speeds (0.6 m/s) were recorded in 2023, while 2021 had the highest PM_{2.5} levels at 325 µg/m³.

3.4 Correlation Analysis: CO, NO₂, and PM_{2.5} (December 2017-2024)

User Prompt: Analyze the correlation between CO, NO₂, and PM_{2.5} levels in Delhi during December from 2017 onwards, as presented in Table 2.

Pollutants	CO (mg/m ³)	NO ₂ (µg/m ³)	PM _{2.5} (µg/m ³)
CO (mg/m ³)	1	0.3	0.47
NO ₂ (µg/m ³)	0.3	1	0.34
PM _{2.5} (µg/m ³)	0.47	0.34	1

Table 2: (Generated by VayuChat) Correlation Matrix: CO, NO₂, and PM_{2.5} (December 2017–2024).

Key Observation: The analysis shows strong inter-correlation among all three pollutants in December, with PM_{2.5} and CO most related ($r = 0.47$). This suggests common sources such as vehicles, stubble burning, and industrial emissions drive synchronized pollution episodes.

4 Potential Applications

VayuChat serves diverse communities in three ways: (1) *Government and policy*, enabling officials to evaluate programs, compare outcomes, and prioritize interventions (e.g., “Which states improved most after NCAP funding?”); (2) *Journalism and public engagement*, helping media create data-driven stories and citizens interpret local air trends without technical expertise; and (3) *Education and community monitoring*, serving as both a teaching resource in data science and environmental studies and a tool for organizations to track conditions, identify sources, and advocate interventions.

5 Future Work

In the future, we aim to enhance VayuChat in four ways: (1) *Live dataset integration*, incorporating real-time air quality streams via APIs; (2) *Expanding datasets*, adding ERA5 reanalysis, satellite products, and land-use or emission inventories for richer spatial and temporal coverage; (3) *Model finetuning*, tailoring LLMs for accurate code generation and domain-specific query interpretation; and (4) *Automated reasoning workflows*, enabling decomposition of complex queries into sub-questions and multi-step analyses, as shown in our Delhi pollution case study.

6 Conclusion

VayuChat advances accessible environmental analytics while maintaining scientific rigor, coupling conversational AI with integrated datasets to enable complex analyses without technical skills. Transparent code generation, multiple models support, and domain-specific design make it suitable for policy, research, public engagement, and education, demonstrating that natural language queries can yield reproducible analyses and actionable insights on complex air-quality datasets.

References

- [1] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D. M., WU, J., WINTER, C., HESSE, C., CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHES, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., AND AMODEI, D. Language models are few-shot learners, 2020.
- [2] CARSLAW, D. C., AND ROPKINS, K. openair — an r package for air quality data analysis. *Environmental Modelling Software* 27–28, 0 (2012), 52–61.
- [3] CENTRAL POLLUTION CONTROL BOARD. Central control room for air quality management - all india. <https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing/caaqm-data-repository>, 2025.
- [4] CENTRAL POLLUTION CONTROL BOARD. Cpcb - overview and functions. <https://cpceb.nic.in/>, 2025.
- [5] CLARITY MOVEMENT. How open access air pollution data is paving the way for greater air quality awareness, Feb. 2022. Accessed: 2025-05-15.
- [6] CPCB. National air quality index real-time data, 2025. Accessed: 2025-05-15.
- [7] DEEPSEEK-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- [8] ENERGY POLICY INSTITUTE AT THE UNIVERSITY OF CHICAGO (EPIC). Aqli india fact sheet. EPIC Report, August 2023, https://aqli.epic.uchicago.edu/wp-content/uploads/2023/08/India-FactSheet-2023_Final.pdf, 2023. Accessed: 2025-08-13.
- [9] INC., G. Groqcloud™ documentation. <https://console.groq.com/docs>, 2025.
- [10] OPENAI. gpt-oss-120b gpt-oss-20b model card, 2025.
- [11] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C. L., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., SCHULMAN, J., HILTON, J., KELTON, F., MILLER, L., SIMENS, M., ASKELL, A., WELINDER, P., CHRISTIANO, P., LEIKE, J., AND LOWE, R. Training language models to follow instructions with human feedback, 2022.
- [12] PANDEY, A., BRAUER, M., CROPPER, M. L., BALAKRISHNAN, K., MATHUR, P., DEY, S., TURKGULU, B., KUMAR, G. A., KHARE, M., BEIG, G., ET AL. Health and economic impact of air pollution in the states of india: the global burden of disease study 2019. *The Lancet Planetary Health* 5, 1 (2021), e25–e38.
- [13] TEAM, Q. Qwen3 technical report, 2025.
- [14] WEI, X., CUI, X., CHENG, N., WANG, X., ZHANG, X., HUANG, S., XIE, P., XU, J., CHEN, Y., ZHANG, M., JIANG, Y., AND HAN, W. Zero-shot information extraction via chatting with chatgpt, 2023.