npj | climate and atmospheric science
Article
Published in partnership with CECCR at King Abdulaziz University
https://doi.org/10.1038/s41612-025-01070-4

AirGPT: pioneering the convergence of conversational AI with atmospheric science

Check for updates

Jun Song[1,2] ✉, **Chendong Ma**[1,2] & **Maohao Ran**[1]

Large language models (LLMs) face significant limitations in specialized scientific domains due to their inability to perform data analysis and their tendency to generate inaccurate information. This challenge is particularly critical in air quality management, where precise analysis is essential for addressing climate change and pollution control initiatives. To bridge this gap, we present AirGPT, a computational framework that integrates conversational AI with atmospheric science expertise through a curated corpus of peer-reviewed literature and specialized data analysis capabilities. Through a novel architecture combining natural language processing and domain-specific analytical tools, AirGPT achieved higher accuracy in air quality assessments compared to standard LLMs, including GPT-4o. Experimental results demonstrate superior capabilities in providing accurate regulatory information, performing fundamental data analysis, and generating location-specific management recommendations, as validated through case studies in metropolitan areas such as Beijing.

The emergence of large language models (LLMs) such as ChatGPT and its derivatives has catalyzed a transformation across diverse disciplines, underpinned by their advanced natural language processing capabilities[1–5]. These sophisticated models, drawing from expansive datasets and intricate algorithms, have demonstrated a remarkable ability to emulate human-like text generation, rendering them invaluable in many applications. Nonetheless, deploying LLMs in highly specialized domains, including atmospheric science, is fraught with specific challenges.

Despite their linguistic virtuosity, LLMs conventionally encounter difficulties within this scientific domain. These challenges are primarily due to their lack of the latest specialized knowledge, their limited capacity to process complex data types, and their inclination to produce factually incorrect information[6–8]. Atmospheric science is critical for comprehending and mitigating climate change and air quality deterioration, demanding precise and dependable data analysis and presentation. In response, our research endeavors to amalgamate LLMs with authoritative air pollution analysis resources, aiming to exploit the LLMs' text generation prowess while ensuring content accuracy and veracity via scientific data integration.

Recent literature has witnessed a surge in research exploring the application of LLMs across various domains. LLMs have found widespread adoption in fields such as finance and social science analytics, where they are employed to extract information and inform decision-making processes[5,9,10]. Moreover, the integration of LLMs with other computational agents has been investigated in diverse areas[1,7,11,12]. For instance, in the transportation sector, LLMs have been utilized to enhance natural language understanding for autonomous vehicles, thereby improving their capability to interpret traffic conditions and navigate complex environments[11]. In the realm of personalized communication, LLMs have played a pivotal role in customized learning by facilitating adaptive assessments and content recommendations[13–16].

In the field of atmospheric analysis, LLMs have been applied to climate research, equipping policymakers with relevant knowledge by fine-tuning GPT-4o based on information from IPCC AR6 reports[17]. Lawson et al. (2024) have explored GPT-4o's capacity to interpret meteorological charts and communicate weather hazards through prompt engineering and the incorporation of reliable information, demonstrating the potential of LLMs in weather-chart analysis and the generation of weather hazard reports using natural language communication[18]. While these studies have showcased the ability of LLMs to communicate climate and weather information in natural language, they are not without limitations. In weather forecasting, the phenomenon of hallucination persists, leading GPT-4o to occasionally provide incorrect information in a coherent and confident manner. Furthermore, there remains a dearth of research exploring the capabilities of LLMs in air pollution analysis, management recommendations, and the provision of reliable knowledge to users. To address these gaps, we aim to develop an LLM-based GPT chatbot that can engage users in natural

[1]Department of Geography, Hong Kong Baptist University, 224 Waterloo Rd, Kowloon Tong, Hong Kong. [2]These authors contributed equally: Jun Song, Chendong Ma. ✉e-mail: junsong@hkbu.edu.hk

language communication about air pollution information. This chatbot will be designed to provide relevant knowledge, conduct pollution analysis based on imagery and other analytical tools, and offer reliable and precise recommendations for air pollution management, building upon and extending previous work in this domain.

In policy and decision-making, governmental entities frequently encounter challenges in accessing specialized knowledge and obtaining the most recent insights. This predicament is particularly pronounced in resource-constrained environments, where consultation with environmental experts may incur substantial costs and temporal delays. Consequently, there exists an urgent imperative to fortify the connection between governmental agencies and scientific institutions. This reinforcement can be achieved through the implementation of an innovative chatbot system, capable of providing immediate, authoritative, and literature-supported guidance on policy drafts and pronouncements. The adoption of such a system ensures that policymakers and frontline personnel have access to current and accurate information, thereby significantly enhancing the quality and efficacy of decision-making processes about environmental issues.

In this study, we present AirGPT, an innovative computational framework that pioneers the integration of Conversational Artificial Intelligence within the atmospheric science discipline. This advanced system functions as a specialized chatbot capable of engaging in natural language communication with individuals lacking professional expertise. AirGPT comprehensively responds to queries regarding the latest developments and specific AQ knowledge, such as recent regulatory updates and targeted air pollution information. Additionally, it facilitates rudimentary AQ data analysis, enabling government personnel to conduct rapid assessments of pollution scenarios. Furthermore, AirGPT offers tailored AQ management recommendations specifically addressing the situation in a specific location. This novel system transcends the capabilities of the LLaMA 2 model by assimilating a meticulously curated corpus of authoritative data, thereby establishing a new paradigm for the incorporation of specialized knowledge in AI-driven atmospheric analysis. AirGPT distinguishes itself through the systematic integration of peer-reviewed academic papers, ranked by the Chinese Academy of Sciences, ensuring that the model's knowledge base remains current with high-caliber research and providing users with the most up-to-date and reliable information. To evaluate the efficacy and performance of AirGPT, we solicited insights from atmospheric science experts and industry practitioners. Detailed findings from these validation tests are presented in subsequent sections of this paper.

The principal contributions of this research are enumerated as follows:

- Introduction of AirGPT, a state-of-the-art conversational AI model specifically designed for the domain of air pollution management, advancing the targeted application of conversational AI.
- Addressing the pervasive challenge of inaccurate information generation, commonly referred to as 'hallucinations,' and mitigating the use of outdated data in AI systems within the air quality sector.
- Development of a novel evaluation framework and associated benchmark for assessing the performance of LLMs in the context of air quality management, which, to the best of the authors' knowledge, represents the first such contribution in this domain.
- Provision of a refined LLM that consistently delivers precise and authoritative information, thereby equipping a diverse range of stakeholders with the requisite intelligence for informed decision-making in air pollution management.

The developed methodology is poised to provide decision-makers and the broader public with reliable insights, fostering more informed decision-making processes. It is imperative to emphasize, however, that this tool is not intended to supplant existing decision-making processes. Rather, it serves as a complementary resource in the multifaceted procedure of informed decision-making regarding air pollution management. Its primary function is to enhance information accessibility and expedite the review process in policy formulation and related scenarios. For instance, a more comprehensive understanding of the O_3 generation mechanism and associated chemical formations can lead to more targeted strategies for managing emissions and implementing mitigation measures. AirGPT aims to augment information accessibility while maintaining its role as an auxiliary tool in direct decision-making activities.

Results

Outline

Our experimental protocol comprised a systematic evaluation of multiple conversational models, including AirGPT and GPT-4o, through a structured set of ten queries collected from atmospheric specialists from our Metasequolia AI Lab. Four of these questions incorporated follow-up questions to enable a more granular analysis (Tables 1 and 2). The query set was developed in collaboration with environmental science experts from our research team, who identified both frequently encountered questions and critical queries requiring chatbot response capability. These questions were categorized into three domains: knowledge base provision (K), air pollution analysis (A), and air quality management tasks (M).

Chatbots and inquiry assessment

Verification of response. The efficacy of the interactions between humans and the chatbot, AirGPT, is evaluated based on criteria established in prior research[19,20]. Following the methodology proposed by Wang et al. (2023)[21], our evaluation framework relies on three key indicators:

Table 1 | 10 atmospheric queries

Questions	Difficulty	Type[a]
Q1. "What are the air quality standards and regulations in Beijing today?"	1	K
Q2. "What are the main contributors to atmospheric pollution in Beijing?"	2	A
Q3. "How do different types of pollutants impact air quality and human health, and how can we deal with them differently?"	2	K&M
Q4. "What are the available technologies and strategies for reducing emissions and improving air quality, and how does they cost?"	3	K&M
Q5. "Are there any emerging pollutants or pollutants of concern that may require attention?"	3	A
Q6. "What was the longest continuous period of pollution in Beijing recently?"	2	A&M
Q7. "Which day had the most severe pollution in Beijing during the past two months?"	2	A&M
Q8. "What are the potential health risks associated with the identified pollutants?"	1	K
Q9. "Please provide some effective monitoring techniques and technologies for managing atmospheric pollution suitable to Beijing. "	2	K&M
Q10. "How can the analysis results be effectively communicated to stakeholders, policymakers, and the general public to promote awareness and support for pollution management efforts?"	2	M

[a]The study categorized questions into five difficulty levels: low (1), medium (2), and high (3). Additionally, each question was assigned a type to indicate AirGPT's capabilities in answering specific queries, such as K for atmospheric knowledge base, A for pollution data analysis, and M for air quality management.

Table 2 | Atmospheric queries with follow-up questions

Questions
Q1. "What are the air quality standards and regulations in Beijing today?"
Q1.1. "Please tell me more about the air pollution restrictions in Beijing."
Q1.2. "What about the traffic aspect?"
Q6. "What was the longest continuous period of pollution in Beijing recently?"
Q6.1. "Which station recorded the highest pollution levels during this period?"
Q6.2. "Using weather data including wind direction, analyze the pollution transport patterns for that most polluted station during this period."
Q6.3. "Based on the analysis, what control measures do you recommend?"
Q7. "Which day had the most severe pollution in Beijing during the past two months?"
Q7.1. "On that day, which station recorded the highest pollution levels?"
Q7.2. "Using weather data including wind direction, can you analyze the pollution transport patterns for that most polluted station on that day?"
Q7.3. "Based on your above analysis, what control measures would you recommend?"

- Quality (Q): The relevance and clarity of the chatbot's responses. High scores in this category indicate that the chatbot provides responses that are both pertinent to the user's query and coherent in a user-friendly manner.
- Specification (S): The specificity of the responses, which assesses the extent to which the answer is tailored to particular spatial and temporal scales, determining its practical applicability. High scores indicate that the response is adequately contextualized within specific geographical or chronological parameters, enhancing its relevance for real-world applications.
- Accuracy (A): User satisfaction and the chatbot's precision are quantified through accuracy criteria. This indicator assesses the chatbot's ability to provide factually correct information while minimizing the occurrence of "hallucinations" or fabricated content. It serves as a measure of the system's reliability in delivering accurate and trustworthy responses.

The indicators Q, S, and A collectively provide a comprehensive assessment of the chatbot's performance. Accuracy verification involves cross-checks by domain experts and evaluations by stakeholders. AirGPT's ability to cite sources from its training data enhances the transparency and credibility of its responses. For example, as shown in Fig. 1, AirGPT explicitly references its information sources when addressing a query about air pollution trends in Beijing, offering users a more transparent understanding of its outputs.

A comparative analysis of responses to Inquiries 1 and 2 from AirGPT and GPT-4o is presented in Table 3. Inquiry 1, "*What are the air quality standards and regulations in Beijing*?", assesses the breadth of knowledge accessible to users. GPT-4o accurately cites the Ambient Air Quality Standards (AAQS) and provides a concise explanation while advising users to verify the latest updates online. In contrast, AirGPT offers a detailed historical analysis of the evolution of these standards and an in-depth overview of current regulations, demonstrating its ability to provide context-rich, comprehensive responses.

To deepen this analysis, follow-up queries (Q1.1 and Q1.2) were introduced to explore specific elements of Beijing's air quality management framework (Table 2). These queries focused on regulatory standards and transportation-related restrictions. As detailed in Table 4, AirGPT produced precise, context-specific responses addressing regulatory frameworks and sector-specific interventions. Its answers included information on policy implementations, quantitative targets, and the temporal evolution of transportation-related regulations, showcasing its ability to navigate complex policy topics and deliver granular insights into urban air quality strategies.

In response to Inquiry 2, *"What are the main contributors to atmospheric pollution in Beijing?"*, GPT-4o provided a generalized summary of common urban pollution sources. AirGPT, however, furnished a quantitatively precise analysis, identifying specific source attribution percentages, temporal emission trends, and key sectoral contributors. Its response also highlighted the role of regional transport from neighboring provinces, a critical factor often overlooked in generic analyses. Furthermore, AirGPT incorporated socioeconomic factors, such as affluence levels and emission intensities, to deliver a comprehensive and nuanced understanding of Beijing's air quality dynamics. These findings underscore AirGPT's ability to synthesize specialized knowledge from peer-reviewed literature, offering insights that extend beyond generic urban pollution patterns.

Prompt engineering. Prompt engineering is a strategic approach used to guide large language models (LLMs) towards generating desired outputs[22]. The careful design of prompts is crucial for optimizing the performance of LLMs in natural language processing (NLP) tasks. In this study, we present two distinct prompts and their corresponding responses to Question 3: *"How do different types of pollutants impact air quality and human health, and how can we address them distinctly?"* (Table 5). The prompts consist of two key components: the *context*, which encapsulates information retrieved from the LLM's extensive memory base, and the *query*, representing the user's input.

In Prompt 1, AirGPT was instructed to generate responses solely based on the provided context, resulting in answers constrained by the information retrieved from the memory base. In contrast, Prompt 2 was enhanced with the directive: *"With your expert knowledge, please provide a detailed response to the question above. Feel free to include any additional insights that you believe are relevant to the discussion."* This augmentation encouraged the LLM to integrate its intrinsic language processing capabilities with the retrieved data from the memory base. Responses generated under Prompt 2 were more comprehensive, coherent, and addressed the query across multiple dimensions. These results validate our decision to adopt Prompt 2 for all subsequent comparisons within our experimental framework.

Hallucination. An analysis of model hallucination patterns revealed that AirGPT achieved enhanced accuracy through the implementation of Prompt 1, which constrained responses to information explicitly present in its knowledge base. This improvement was particularly evident in responses to queries about emerging pollutants of concern. For example, while GPT-4o generated potentially misleading information by suggesting inadequate regulation of VOCs and heavy metals in Beijing, AirGPT provided a more measured response focused on verified pollution control priorities specific to the region (Table 6).

A key feature of AirGPT's architecture is its transparent handling of information gaps. When queried about topics not covered in its peer-reviewed source materials, the system explicitly acknowledges these limitations by responding: *"The provided context does not include specific information."* This approach effectively mitigates hallucination risks by

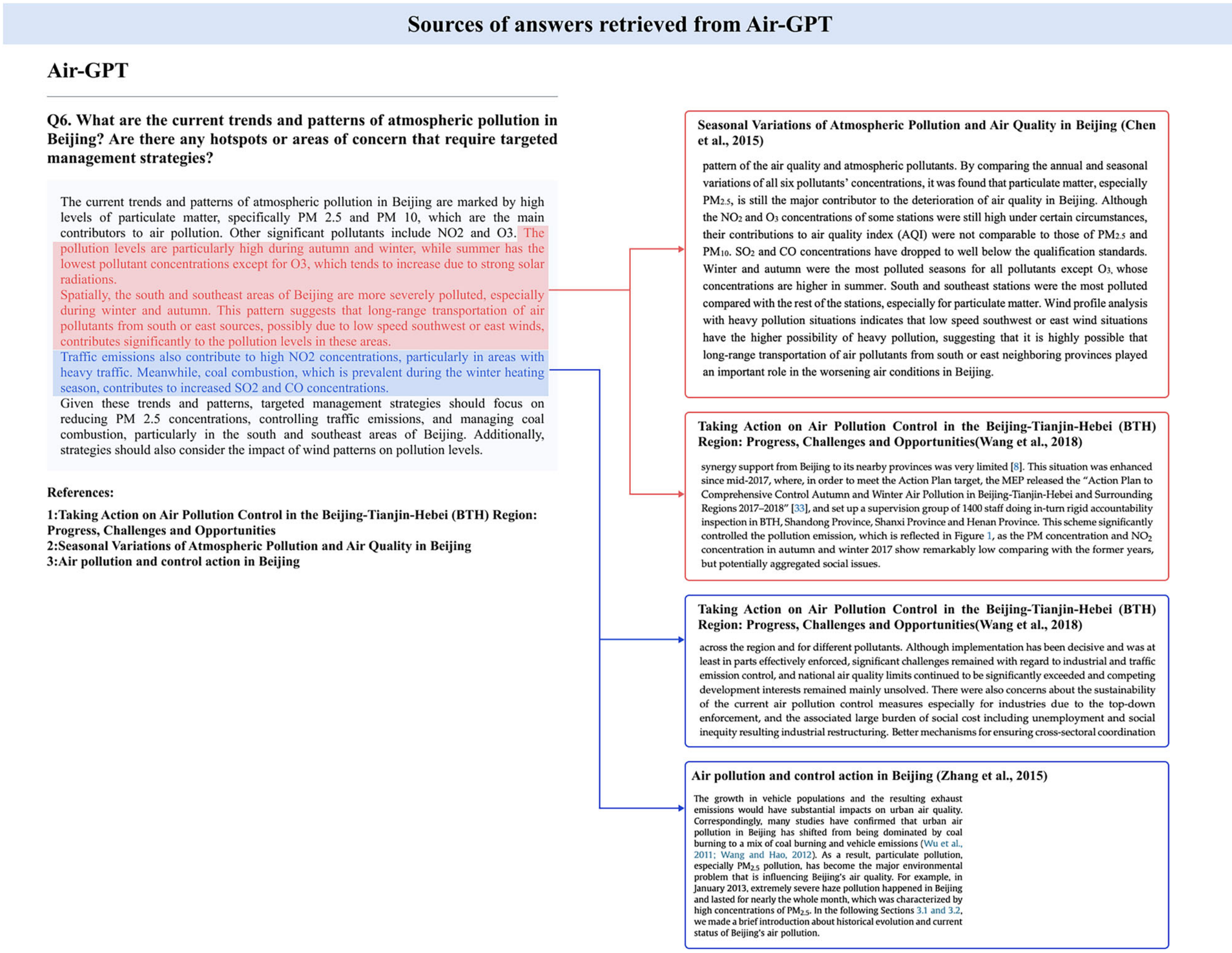

Fig. 1 | Demonstration of AirGPT's reference sources in response to air pollution trends in Beijing, showcasing the model's capability to cite sources from its training data.

Table 3 | Comparison of generated answers to questions from AirGPT and GPT-4o

QAs comparison	
AirGPT	**GPT-4o**
Q1. What are the current air quality standards and regulations in Beijing?	
In Beijing, air quality standards and regulations have evolved significantly over time. Initially, the Ambient Air Quality Standards (AAQS) in China, formulated in 1982 and amended in 1996, did not include $PM_{2.5}$. However, recognizing the importance of $PM_{2.5}$, the standards were updated in February 2012 to include this pollutant along with PM_{10}, SO_2, O_3, NO_2, and CO. These updated standards, known as NAAQS-2012, were adopted by Beijing in January 2013, allowing for real-time air quality monitoring. The Implementation Rules of the Action Plan on Prevention and Control of Air Pollution, issued in 2013, set-specific targets for reducing $PM_{2.5}$ levels by 2017, aiming for a 25% reduction in the Beijing–Tianjin–Hebei region compared to 2012 levels. Key tasks included integrated control of pollutants, traffic management, industrial restructuring, and coal consumption control. References: 1. Air pollution and control action in Beijing and 2. Seasonal variations of atmospheric pollution and air quality in Beijing.	Beijing's air quality standards and regulations have been updated over the years to address the increasing pollution levels. The Ambient Air Quality Standards (AAQS) in China, which also apply to Beijing, were first formulated in 1982 and have been amended several times since then. The first amendment in 1996 included three Grade standards (I–III) and recommended limit values for seven principal pollutants. The second amendment in 2000 removed the standard for NO_x pollutants and relaxed the limit values for NO_2 and O_3. The third amendment in 2010, made by the Ministry of Environmental Protection, added standards for $PM_{2.5}$ and Max. 8-h O_3 pollutants and tightened the guideline values for NO_2 and PM_{10}. The guideline values for Max. 8-h O_3, daily-$PM_{2.5}$ and annual-$PM_{2.5}$ are 100, 35 and 15 mgm/CO_3, respectively.
Quality: 3; Specification: 4; Accuracy: 4	**Quality: 3; Specification: 2; Accuracy: 3**
Q2. What are the main contributors to atmospheric pollution in Beijing?	
The main contributors to atmospheric pollution in Beijing are local emissions, with motor vehicles (31.1%), coal combustion (22.4%), industrial production (18.1%), and fugitive dust (14.3%) being the primary sources. Additionally, regional transport from neighboring areas contributes ~28–36% to Beijing's $PM_{2.5}$ levels. Affluence and emission intensity are also influential factors affecting air pollution in the city. References: 1:Air pollution and control action in Beijing 2:Identifying the main contributors of air pollution in Beijing	The main contributors to atmospheric pollution in Beijing are a combination of local and regional factors, influenced by industrial activity, urbanization, and geographical conditions. Below are the key contributors: 1. Coal Combustion, 2. Vehicle Emissions 3. Industrial Emissions 4. Construction and Dust...
Quality: 4; Specification: 5; Accuracy: 5	**Quality: 3; Specification: 2; Accuracy: 3**

Table 4 | Regulations and standards in Beijing

AirGPT
Q1. What are the air quality standards and regulations in Beijing?
In Beijing, air quality standards and regulations have evolved significantly over time. Initially, the Ambient Air Quality Standards (AAQS) in China, formulated in 1982 and amended in 1996, did not include $PM_{2.5}$. However, recognizing the importance of $PM_{2.5}$, the standards were updated in February 2012 to include this pollutant along with PM_{10}, SO_2, O_3, NO_2, and CO. These updated standards, known as NAAQS-2012, were adopted by Beijing in January 2013, allowing for real-time air quality monitoring. The Implementation Rules of the Action Plan on Prevention and Control of Air Pollution, issued in 2013, set specific targets for reducing $PM_{2.5}$ levels by 2017, aiming for a 25% reduction in the Beijing–Tianjin–Hebei region compared to 2012 levels. Key tasks included integrated control of pollutants, traffic management, industrial restructuring, and coal consumption control.
Q1.1. Please tell me more about the air pollution restrictions in Beijing.
Beijing has implemented several air pollution restrictions to address its severe air quality issues. Key measures include:
1. *Vehicle emissions control*: Beijing strengthened vehicle-emission standards and restricted the number of on-road vehicles using an even-odd license plate system during specific periods, such as the 2008 Olympics. Heavily polluting vehicles, known as yellow-label vehicles, were banned, and a significant portion of government vehicles was removed from the roads.
2. *Fuel and vehicle standards*: Enhancements in emission standards, improvements in fuel quality, and the promotion of public transportation and alternative-fuel vehicles were employed to reduce pollution.
3. *Regional cooperation*: Neighboring regions like Tianjin, Hebei, and others implemented similar measures, including traffic control and industrial restrictions, to support Beijing's efforts.
4. *Clean air action Plan*: Following severe haze episodes, Beijing formulated the "Clean Air Action Plan" and "Air Pollution Prevention and Control regulations" to specifically target $PM_{2.5}$ pollution and improve air quality.
These actions reflect a comprehensive approach to tackling air pollution, focusing on both vehicular emissions and regional cooperation.
References:
1: Air pollution and control action in Beijing
Q1.2. What about the traffic aspect?
The traffic aspect focuses on developing sustainable transport systems that positively influence environmental, social, and economic sustainability. These systems should meet basic access needs safely, be affordable, operate efficiently, offer diverse transport modes, and support a vibrant economy. They should also limit emissions and waste, minimize non-renewable resource consumption, and promote the reuse and recycling of components while minimizing land use and noise production.
References:
1: Implementation effects and integration evaluation of a selection of transport management measures in Beijing

Table 5 | Prompt engineering

QAs comparison	
Prompt 1	**Prompt 2**
As an environmental scientist, you are equipped with the knowledge to address a variety of environmental issues. Please review the provided context information and respond concisely to the question that follows: {*context_str*} Based on the above context, please address the following question with your expert insight: {*query_str*} Your response should be direct and succinct, avoiding unnecessary elaboration.	As an environmental scientist, you are equipped with the knowledge to address a variety of environmental issues. Please review the provided context information and respond concisely to the question that follows: {*context_str*} Based on the above context, please address the following question with your expert insight: {*query_str*} With your expert knowledge, please provide a detailed response to the question above. Feel free to include any additional insights that you believe are relevant to the discussion. Your response should be direct and succinct, avoiding unnecessary elaboration.
Q3. How do different types of pollutants impact air quality and human health, and how can we deal with them differently?	
Different types of pollutants, such as particulate matter, SO_2, NO_2, CO, and O_3, can have varying impacts on human health. These pollutants are associated with respiratory, cardiovascular, and cerebrovascular diseases. Short-term exposures to these pollutants can exacerbate symptoms or cause acute forms of diseases, while long-term exposures can lead to chronic diseases. For instance, an increase in these pollutants can worsen respiratory and cardiovascular health, reduce lung function, weaken immune function, and increase the prevalence of malignant tumors. To deal with these pollutants, it is important to monitor air quality and understand the health effects of each pollutant. This can be done using tools like the Baidu indices, which assess daily query frequencies of keywords related to air pollution-related diseases, and air quality monitoring data. A health Air Quality Index (AQI), or HAQI, can also be used to better indicate the effects of air pollutants on diseases compared to the traditional AQI. Furthermore, research and studies can help in understanding the relative risks of different pollutants and their impact on various diseases. This knowledge can then be used to develop and implement effective policies and measures to reduce air pollution and its health impacts.	Different types of pollutants have varying impacts on air quality and human health. Particulate matter (PM_{10} and $PM_{2.5}$) can penetrate the respiratory system, leading to respiratory diseases and even lung cancer. Sulfur dioxide (SO_2) can cause respiratory problems and contribute to the formation of acid rain. Nitrogen dioxide (NO_2) can cause inflammation of the airways and contribute to respiratory diseases. Carbon monoxide (CO) can reduce oxygen delivery to the body's organs and tissues, while ozone (O_3) can cause respiratory problems and other health issues. To deal with these pollutants, it's crucial to implement strategies that reduce emissions from their primary sources. For particulate matter and SO_2, this often means regulating industrial emissions and transitioning to cleaner energy sources. For NO_2 and CO, reducing vehicle emissions is key, which can be achieved through promoting electric vehicles and improving fuel efficiency. For O_3, reducing emissions of volatile organic compounds and nitrogen oxides, which contribute to ozone formation, is important. Public awareness and behavioral changes can also play a significant role in mitigating the impact of these pollutants. This includes actions like reducing energy consumption, using public transportation, and staying indoors on days with poor air quality. Lastly, robust air quality monitoring systems and health impact assessments are crucial for understanding the scale of the problem, informing policy decisions, and tracking progress over time.

strictly adhering to verified scientific literature rather than generating speculative responses. By grounding its outputs in empirical evidence and peer-reviewed research, AirGPT ensures higher reliability for environmental decision-making support.

Similarity hyperparameter. The similarity hyperparameter plays a critical role in aligning user queries with relevant information from the memory base. To evaluate its impact, we analyzed the effects of varying this parameter (see Table 7) by setting the number of nearest vectors (k)

Table 6 | Hallucination

QAs comparison	
AirGPT	**GPT-4o**
Q4.What are the available technologies and strategies for reducing emissions and improving air quality, and how does they cost?	
The available technologies for reducing emissions and improving air quality include desulfurization technologies, denitrification technologies, PM control technology, and energy-saving technologies. These technologies are primarily used in the power, steel and iron, and cement sectors. Desulfurization technologies are used for SO_2 removal, denitrification technologies for NO_x removal, and PM control technology for soot or dust removal. Energy-saving technologies, on the other hand, are often preferred due to their ability to reduce production costs and generate economic profits. The cost of these technologies varies. For instance, the SO_2 unit abatement cost ranges from 1000 to 400,000 CNY/ton of SO_2, and the NOx unit abatement cost varies from 1000 to 1,400,000 thousand CNY/ton of NO_x. The soot unit abatement cost ranges from 200 to 1500 CNY/ton of soot. However, the unit pollutant abatement cost of energy-saving technologies is generally higher than the cost of technologies used specifically for pollutant removal. This is because energy-saving technologies generally abate less pollutant compared to the specific pollutant control technologies. In terms of strategies, the implementation of these technologies is often preferred in plants due to their effect in reducing production costs or in generating economic profits by selling recycled resources. Therefore, a plant is more likely to invest in energy-saving technologies than in pollutant-removal technologies.	There are several technologies and strategies available for reducing emissions and improving air quality in urban areas like Beijing. Some common approaches include: 1. Vehicle Emissions Control: … 2. Industrial Emission Controls: … 3. Clean Energy Transition: … 4. Residential Heating Upgrades: … 5. Green Infrastructure: … 6. Public Awareness and Behavior Change: … While implementing technologies and strategies to reduce emissions and improve air quality may involve upfront costs, the long-term benefits of cleaner air, improved public health, and a more sustainable environment often outweigh the initial investments. Governments, businesses, and communities can work together to implement cost-effective solutions that address air quality challenges and create a healthier environment for all.
Quality: 4; Specification: 4; Accuracy: 5	**Quality: 4; Specification: 2; Accuracy: 2**
Q5. Are there any emerging pollutants or pollutants of concern that may require attention?	
The context highlights concerns about $PM_{2.5}$, NO_2, and O_3 as pollutants requiring attention. While $PM_{2.5}$ control has been a focus, NO_2 has shown a slight decline, and O_3 levels could potentially increase, indicating a need for precautionary strategies to manage these pollutants effectively. Additionally, the influence of meteorological conditions on air pollution formation and diffusion suggests that emerging pollutants may also be linked to changing climate patterns, such as those related to ice cap melting. Therefore, continuous monitoring and adaptive strategies are essential to address these pollutants of concern. References: 1. Construction and countermeasure discussion on government performance evaluation model of air pollution control: A case study from Beijing–Tianjin–Hebei region. 2. Taking Action on Air Pollution Control in the Beijing–Tianjin–Hebei (BTH) Region: Progress, Challenges and Opportunities	In Beijing, as in many urban areas, there are emerging pollutants and pollutants of concern that may not be currently regulated but are increasingly recognized as potential threats to public health and the environment. Some of these pollutants include: 1. $PM_{2.5}$ and PM_{10}: Particulate matter (PM) with a diameter of 2.5 micrometers or less ($PM_{2.5}$) and 10 micrometers or less (PM_{10}) are already regulated in Beijing due to their significant impact on air quality and public health. However, emerging research continues to highlight the health risks associated with exposure to ultrafine particles and their potential sources, such as vehicle emissions, industrial activities, and construction dust. Further efforts to reduce PM pollution and address sources of ultrafine particles may be needed to improve air quality in Beijing. 2. Volatile Organic Compounds (VOCs): … 3. Heavy Metals: Heavy metals such as … 4. Microplastics: Microplastics are small plastic particles… 5. Emerging Contaminants: Emerging contaminants such as… Addressing emerging pollutants and pollutants of concern in Beijing requires a multi-faceted approach that involves collaboration among government agencies, research institutions, industry stakeholders, and the public. By identifying specific sources of pollution, implementing targeted control measures, and raising awareness about environmental health risks, Beijing can work towards a cleaner and more sustainable urban environment for its residents.

to 3, 5, 15, and 25. Higher values of k allow the language model (LLM) to generate more precise and comprehensive responses. For instance, when responding to Question 9—*"Please provide some effective monitoring techniques and technologies for managing atmospheric pollution suitable for Beijing"*—a value of $k = 3$ results in no relevant answer. Increasing k to 5 produces a response that includes references to permanent air quality monitoring stations. At $k = 15$, the response becomes significantly more detailed, incorporating information about Beijing's specific monitoring stations, advanced monitoring techniques, and their integration with ground sensors. This demonstrates how adjusting k can substantially influence the quality of the generated responses.

However, larger values of k may also introduce irrelevant information. For example, when k is increased to 25, the response includes unrelated elements such as general summaries and policy recommendations, which dilute the relevance of the answer. These observations underscore the importance of balancing precision and relevance when selecting the similarity hyperparameter. Based on our experiments, we standardized the similarity hyperparameter to $k = 15$, as it provided the optimal trade-off between response detail and relevance.

Pollution data analysis and follow-up prompts. This study also investigates pollution data analysis by utilizing AirGPT to facilitate interactions with other intelligent agents. Within this framework, AlphaAir performs analytical tasks based on directives provided by human users. For a comprehensive case study, air quality data collected from the literature was employed. Using the LLM, human instructions were systematically decomposed into discrete tasks, which were subsequently executed by invoking appropriate data processing modules. The processed data and results were then synthesized and presented by the LLM.

An illustrative example of this stepwise, agent-based LLM workflow is shown in Table 8, which highlights the systematic decomposition and execution of tasks. This approach demonstrates the potential of integrating LLMs with modular agents to enhance the efficiency and accuracy of complex environmental data analyses. Furthermore, it aligns with our vision for advancing this domain by integrating multi-agent systems with AI-driven models to support scalable and dynamic pollution monitoring solutions. By leveraging the synergy between LLMs and specialized data processing modules, such as those in AlphaAir, this methodology represents a significant advancement in automating and streamlining environmental data analysis. Future work will focus on refining this system, with an emphasis on real-time data integration, adaptive task allocation, and improved interpretability of results to support decision-making processes in urban air quality management.

Retraining LLM. A comparative analysis demonstrates that AirGPT produces responses with greater accuracy than the unmodified GPT-4o model. Detailed results for questions Q1–Q10 are provided in the Supplementary Information section. While retraining the LLM with a comprehensive set of relevant documents represents another approach to

Table 7 | Comparison of parameters

QAs comparison	
***k*—3**	***k*—5**
Q9. Please provide some effective monitoring techniques and technologies for managing atmospheric pollution suitable to Beijing.	
The document does not provide specific information on the effective monitoring techniques and technologies for measuring atmospheric pollution that are suitable for Beijing.	The document discusses the use of *permanent air quality monitoring stations as an effective technique for measuring atmospheric pollution in Beijing*. These stations monitor six key pollutants: $PM_{2.5}$, PM_{10}, SO_2, O_3, NO_2, and CO. The data is collected at one-hour intervals, providing real-time air quality information. This method is suitable for Beijing due to the city's significant air pollution challenges. The monitoring stations are divided into urban, suburban, background, and traffic categories, allowing for a comprehensive understanding of pollution levels across different areas of the city. However, the document does not provide specific technologies used in these stations.
Accuracy: 0	**Accuracy: 2**
***k*—15**	***k*—25**
The effective monitoring techniques for measuring atmospheric pollution in Beijing include the deployment of air quality monitoring stations and the use of real-time web platforms for data collection. The city *has 35 permanent air quality monitoring stations* that are divided into urban, suburban, background, and traffic monitoring stations. These stations collect hourly air quality data for six pollutants: $PM_{2.5}$, PM_{10}, SO_2, O_3, NO_2, and CO. These techniques are suitable for Beijing because they allow for comprehensive and continuous monitoring of air quality across different parts of the city. They also enable the collection of large amounts of data, which can be used for detailed analysis of air quality trends and the effectiveness of pollution control measures. *In addition to these techniques, advanced technologies such as satellite remote sensing and air quality modeling can also be used to complement ground-based monitoring, providing a more comprehensive picture of atmospheric pollution in Beijing*.	Beijing utilizes a network of *permanent air quality monitoring stations and real-time online platforms* to effectively monitor atmospheric pollution. The city has set up 35 permanent monitoring stations, which are divided into urban, suburban, background, and traffic monitoring stations. These stations collect hourly data on six key pollutants: $PM_{2.5}$, PM_{10}, SO_2, O_3, NO_2, and CO. The data is then made available to the public in real-time via an online platform. This method is effective because it provides a *comprehensive and timely overview of the city's air quality, which aids in better evaluation and policy-making to combat air pollution*. Furthermore, the real-time data helps in quickly identifying equipment malfunctions or internet transfer errors, ensuring the reliability of the data.
Accuracy: 5	**Accuracy: 4**

obtaining domain-specific responses, this process incurs significant computational costs and a substantial carbon footprint. Moreover, our findings reveal that the marginal improvements in response quality do not justify these additional costs when compared to the proposed memory-based approach. Consequently, we adopted the memory-based methodology in this study.

Evaluation and metrics

A comprehensive framework was developed to evaluate AirGPT's capabilities across multiple dimensions, following the methodologies outlined by Bi et al. (2023)[23]. AirGPT addressed a wide range of topics, including providing an air quality (AQ) knowledge corpus, analyzing AQ conditions through emission source tracking, and managing AQ in urban contexts. Responses that were inconsistent or of low quality were promptly refined or excluded to ensure the overall reliability and accuracy of the outcomes.

Manual evaluation. The manual evaluation phase involved human judges, including some of our co-authors specializing in atmospheric and environmental sciences, who assessed the retrieved answers. Responses from both AirGPT and GPT-4o were evaluated based on three key indicators—Quality, Specification, and Accuracy—along with overall satisfaction, using a 1–5 Likert scale.

Automatic evaluation. The automatic evaluation phase leveraged GPT-4o to assess response quality using metrics derived from prior research[21,24]. GPT-4o evaluated the satisfactoriness of responses, providing a benchmark for quality comparisons. To minimize bias, the order of response sequences was alternated in subsequent prompts, and the synthesized results were compiled into a comprehensive evaluation report.

Results. We conducted a detailed manual evaluation of the Quality, Specification, and Accuracy indicators, categorized by question type: knowledge base, analysis, and management (Fig. 2). The results indicate that AirGPT outperformed in Specification and Accuracy across most question categories. Although Quality scores were comparable, GPT-4o occasionally delivered more detailed responses due to its broader knowledge base. Task-specific performance highlighted AirGPT's strengths in knowledge base and analysis queries, which can be attributed to its reliance on focused academic data sources. However, AirGPT underperformed in management-related tasks, such as Question 10 on communication strategies, where general-purpose LLMs like GPT-4o excelled due to their extensive and diverse datasets.

Our research is centered on the development of specialized chatbots for analyzing air pollution data. The first iteration, AirGPT, demonstrated improved accuracy when answering a set of 10 test questions. To minimize errors from external datasets, we constructed a knowledge base using a randomly selected set of recent research papers on air pollution. However, due to token limitations in the system, the knowledge base was restricted to a maximum of 20 papers[25–27], potentially limiting the comprehensiveness of its responses. While the random selection of papers yielded satisfactory results, it did not always guarantee optimal answers. The accuracy and completeness of AirGPT's responses depend on several factors, including the *k* parameter in top-*k* retrieval settings, the structure of the prompts, and the selection of data sources.

AirGPT generates responses based on the 15 most relevant text passages retrieved, which may result in the omission of critical information from other parts of its knowledge base. To address this, we plan to develop an automated system for continuously incorporating new scientific findings into the knowledge base. Additionally, we aim to improve AirGPT's accessibility for general users, requiring minimal training for effective use. Given ongoing discussions about pausing LLM development until appropriate regulatory frameworks are implemented, we emphasize the importance of regulating external data incorporation and prompt construction in these systems. Furthermore, we recognize the environmental implications of training large language models, particularly their carbon footprint[28]. However, we note that inference tasks, such as question answering, consume significantly less energy compared to the initial training process.

Discussion

This study presents AirGPT, an innovative framework that represents a significant advancement in the application of large language models (LLMs)

Table 8 | Detailed analysis and follow-up prompts

AirGPT
Q6. Which day had the most severe pollution in Beijing during the past 2 months?
I'll help you analyze the air pollution data for Beijing over the past 2 months to find the day with the most severe pollution. I'll use the daily city air quality data for this analysis. Based on the analysis of the past two months' air quality data in Beijing, here are the key findings:
1. The day with the highest pollution level was September 19, 2024, with an AQI of 94
2. Key characteristics of the most polluted day:
- AQI: 94 (Good level)
- Main pollutant: Ozone (O_3) at 152.0 μg/m^3; $PM_{2.5}$: 30.0 μg/m^3; PM_{10}: 55.0 μg/m^3. It is worth noting that even during the most polluted day, the air quality remained in the "Good" category, indicating relatively well-controlled air pollution levels during this period. Ozone (O_3) was the primary pollutant on most of the higher AQI days, which is typical during warmer months due to increased photochemical reactions.
Q6.1. On that day, which station recorded the highest pollution levels?
I'll analyze the station-level air quality data for September 19, 2024, to identify the station with the highest pollution levels. Based on the analysis, on September 19, 2024, here are the findings: 1. The Dingling station recorded the highest pollution levels with:
- AQI: 105 (lightly polluted)
- Main pollutant: Ozone (O_3) at 209.0 μg/m^3
- $PM_{2.5}$: 52.0 μg/m^3–PM_{10}: 68.0 μg/m^3
2. This was followed closely by two stations in the Huairou area, both with:
- AQI: 103 (lightly polluted)
- O_3 levels around 206.0 μg/m^3
The data shows a clear pattern of elevated ozone levels in the northern parts of Beijing, with the highest concentrations recorded at stations in relatively less urbanized areas. This is typical for ozone pollution, which often peaks in suburban areas due to complex photochemical processes and air mass transport.
Q6.2. On that day, which station recorded the highest pollution levels?
I'll analyze the pollution transport patterns by examining the weather conditions and pollution levels throughout the day at the Dingling station, along with wind directions. Analysis of Pollution Transport at Dingling Station on September 19, 2024:
Peak pollution period (15:00–17:00):
- Wind direction: consistent south wind
- Wind scale: 2 (moderate breeze)
- Highest pollution levels at 17:00:
* O_3 peaked at 209 μg/m^3
* $PM_{2.5}$ reached 52 μg/m^3
* PM_{10} increased to 68 μg/m^3
Transport pattern analysis
1. The pollution episode was characterized by a typical ozone formation pattern, with southerly winds bringing pollutant precursors from urban areas.
2. The Dingling station's location north of Beijing's urban center made it susceptible to accumulation of pollutants transported from the city.
3. The combination of high temperatures, strong solar radiation (indicated by the time of peak), and consistent southerly winds created ideal conditions for ozone formation.
4. The evening shift to easterly winds brought some relief in ozone levels but maintained elevated particle pollution. This pattern suggests that the pollution episode was largely influenced by urban emissions being transported northward to the Dingling area, where meteorological conditions favored photochemical ozone formation.
Q6.3. Based on your above analysis, what control measures would you recommend?
Based on the pollution transport patterns and characteristics observed, I'll recommend targeted control measures focusing on both short-term and long-term strategies: 1. Time-specific control measures: Early morning (6:00–9:00): *Restrict heavy-duty diesel vehicle movement in southern urban areas. *Control VOC emissions from industrial facilities during facility startup. *Strengthen supervision of paint and solvent use in auto repair and furniture manufacturing—peak formation period (10:00–17:00): *Implement flexible work hours or remote work on high-risk days to reduce traffic. *Temporarily reduce operation loads of high-VOC emission industries. *Control construction dust more strictly during high-temperature periods ...

to air pollution analysis. By integrating the LLaMA 2 architecture with a curated repository of peer-reviewed literature, AirGPT addresses key challenges related to domain expertise gaps and information accuracy in environmental analysis. Rigorous evaluations conducted by government and industry experts demonstrate that AirGPT outperforms conventional LLMs, including GPT-4o, in generating accurate and contextually relevant responses.

Our research makes several notable contributions to the field. AirGPT introduces a novel framework for combining domain-specific expertise with analytical precision in air pollution management, while also advancing methodologies for evaluating LLM-based conversational systems. This study highlights the effectiveness of sophisticated prompt engineering and hyperparameter optimization in adapting LLM outputs to specific domain requirements without compromising computational efficiency. Furthermore, AirGPT broadens access to specialized knowledge on air pollution, serving as a valuable resource for policymakers and fostering public understanding of atmospheric science. However, it is important to emphasize that AirGPT is designed to complement, rather than replace, localized expertise and tailored solutions in air quality management.

The implications of this research extend beyond the immediate domain of air pollution. AirGPT establishes a transferable framework for integrating domain-specific knowledge with conversational AI, which can be applied across various scientific disciplines with similar information challenges. Nevertheless, certain limitations remain. The finite capacity of prompt tokens restricts the memory base to a maximum of 20 articles, which may limit the comprehensiveness of its responses. Additionally, reliance on randomly sampled data introduces

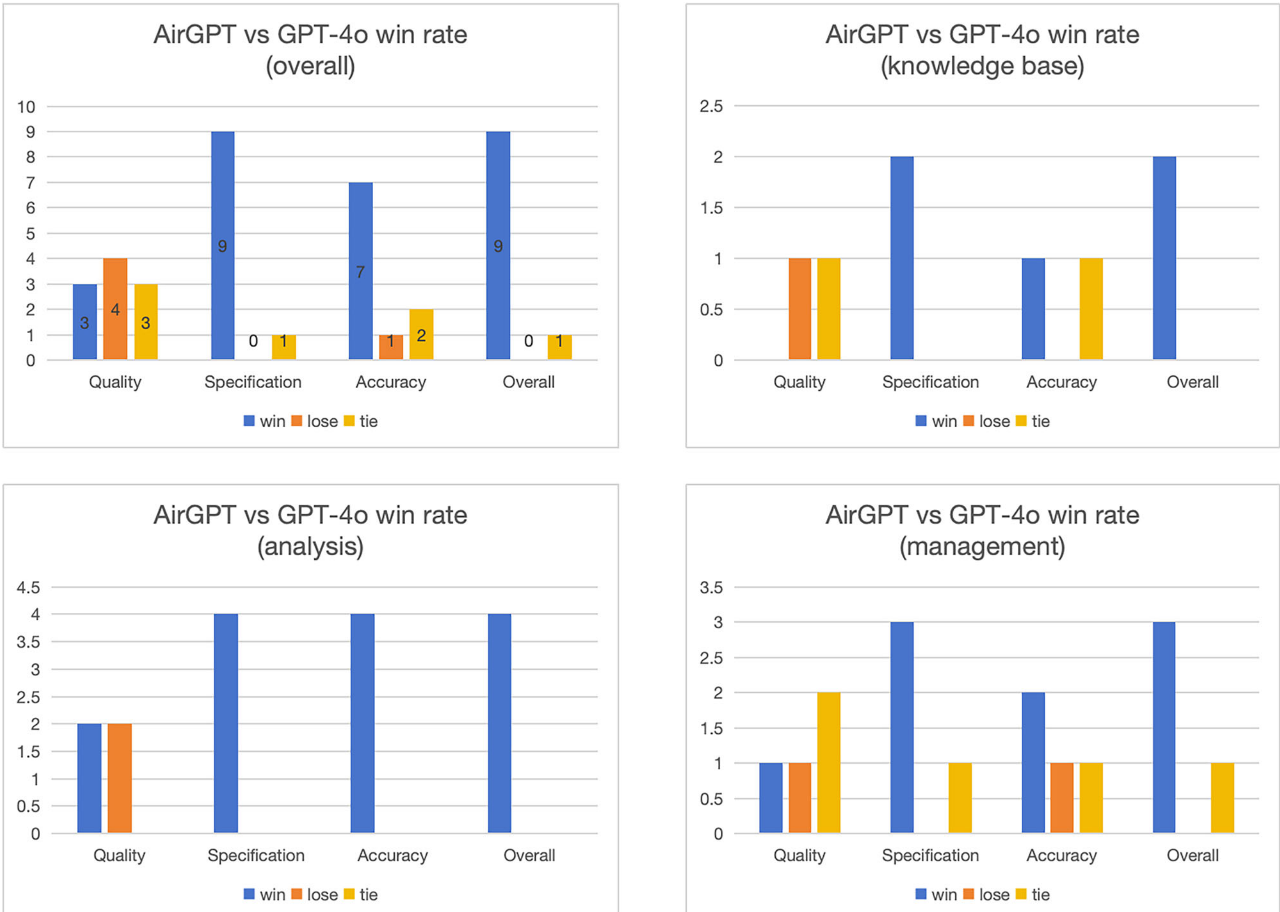

Fig. 2 | The comparative winning rates of AirGPT and GPT-4o across different question types: knowledge base, analysis, management, and overall satisfaction.

variability in response quality and the risk of omitting critical information. Future work will focus on systematically expanding and updating the memory base with the latest scientific literature, refining data selection processes, and improving accessibility for non-expert users. Ethical and environmental considerations, particularly those related to the energy consumption of LLM training and deployment, will also be prioritized.

Looking ahead, our ongoing development efforts center on enhancing AirGPT's capabilities and ensuring seamless integration with established atmospheric modeling systems, such as GEOS-Chem and CMAQ. The ultimate objective is to support evidence-based environmental policy formulation and implementation, thereby contributing to global initiatives in air quality management and environmental protection.

Methods

AirGPT framework

We implemented a retrieval augmented generation (RAG) model[29] to enhance LLM performance. Our process converts PDF documents into vector representations as follows. Let $D = d_1, d_2, \ldots, d_n$ represent our set of PDF documents. Each document d_i is transformed into a vector $\mathbf{v}i$ using the embedding function $f$embed: $\mathbf{v}i = f\mathrm{embed}(d_i)$. The resulting vectors are organized into clusters $C = \mathbf{v}1$: 100, $\mathbf{v}101$: 200, … and stored in vector database $\mathcal{V}$.

$$\mathbf{v}_i = f_{\mathrm{embed}}(d_i), \quad \forall d_i \in D \tag{1}$$

We utilize OpenAI's text embedding model to optimize these data clusters. When a user poses a query q, it is converted into a vector $\mathbf{q}$, using the same embedding function:

$$\mathbf{q} = f_{\mathrm{embed}}(q) \tag{2}$$

Subsequently, we employ a k-nearest neighbors (k-NN) approach to find the most similar vectors from the database $\mathcal{V}$. The similarity score between the query vector $\mathbf{q}$ and document vector $\mathbf{v}_i$ is computed using a cosine similarity measure $s(\mathbf{q}, \mathbf{v}_i)$.

$$s(\mathbf{q}, \mathbf{v}_i) = \frac{\mathbf{q} \cdot \mathbf{v}_i}{\| \mathbf{q} \| \| \mathbf{v}_i \|}, \quad \mathbf{v}_i \in \mathcal{V} \tag{3}$$

For each query, we select the top k vectors with the highest similarity scores exceeding a pre-defined threshold τ.

$$\mathcal{R} = \{\mathbf{v}_i \in \mathcal{V} | s(\mathbf{q}, \mathbf{v}_i) > \tau\} \tag{4}$$

The selected vectors $\mathcal{R}$ are then used to form a prompt that includes the background information needed for the LLM, specifically LLaMA 2, to generate a response. Let G_{LLaMA} be the generative function of LLaMA 2, the final response $\mathbf{r}$ is derived as follows:

$$\mathbf{r} = G_{\mathrm{LLaMA}}(\mathrm{prompt}(\mathcal{R})) \tag{5}$$

An iterative feedback loop is employed where the generated responses $\mathbf{r}$ are evaluated and refined through GPT until they align with the user requirements. The final refined answer $\mathbf{r}^*$ is provided to the user. The entire workflow is graphically represented in Fig. 3.

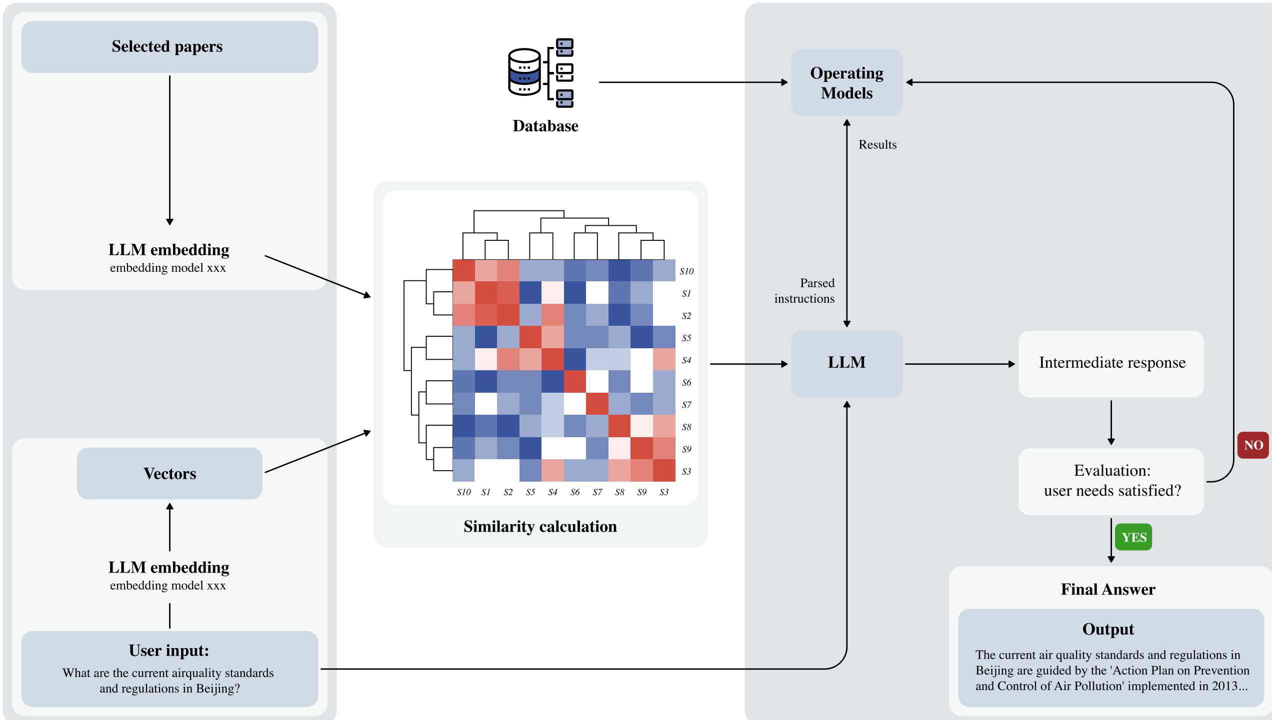

Fig. 3 | A schematic overview of the AirGPT framework. Illustrating the workflow of the AirGPT framework, which incorporates RAG for enhanced information retrieval and response accuracy in the context of air pollution analysis. The process commences with the embedding of selected academic PDF documents into vector representations, subsequently grouped and stored in a vector database, informed by the literature cited in the references. Utilizing OpenAI's text embedding and search models, we optimized these data blocks for semantic search to meet specific user queries.

Baseline comparison

In this study, we evaluated three distinct methodologies for answer generation in conversational AI systems. The first methodology employed an integrated approach that combined large language models (LLMs) with a curated knowledge database. The second methodology relied solely on database retrieval mechanisms without LLM augmentation, while the third utilized a standalone LLM configuration with no external database integration.

The core architecture of our conversational system focused on extracting and utilizing relevant information from peer-reviewed research papers stored within the knowledge database. In the integrated approach, the system first performed an information retrieval process to extract relevant data from the database's persistent memory. This retrieved contextual information was then combined with the user's query to enable a detailed semantic analysis using the LLM. In contrast, the database-centric approach generated responses exclusively from the stored research corpus, relying on query-document relevance metrics to retrieve the most pertinent information. To establish a comparative baseline, we deployed GPT-4o in its standard configuration, which relied solely on its internal semantic processing capabilities without external knowledge augmentation.

Ethical considerations

We emphasize that AirGPT is primarily intended as an analytical research tool, and its outputs should be considered within this context. We have implemented comprehensive safeguards to ensure transparency in the model's limitations and capabilities, explicitly acknowledging when responses are based on peer-reviewed literature versus analytical inference. The system is designed to avoid hallucination through strict adherence to verified scientific sources, particularly crucial when addressing air quality management policies that impact public welfare. While AirGPT demonstrates robust analytical capabilities, we emphasize that it serves as a decision-support tool rather than a replacement for expert judgment. Users are advised to exercise professional discretion and verify critical recommendations, especially in contexts affecting public health and environmental policy. Additionally, we have incorporated safeguards against potential biases in the training data and continuously monitor the system's outputs to ensure they align with established scientific consensus in environmental science and air quality management.

Data availability

The data that support the findings of this study are drawn from multiple sources. The foundational knowledge base for AirGPT was constructed using peer-reviewed literature, primarily from Zhang et al. (2016), Li et al. (2017), Li et al. (2020), Wang et al. (2013), Huang et al. (2018), Zhou et al. (2019), Xu et al. (2017), He et al. (2016), Wu et al. (2015), Chen et al. (2015), Wang et al. (2018), Zhong et al. (2019), Sun et al. (2021), Wang et al. (2021)[30–43] and associated supplementary materials. Historical air quality measurements and analytical datasets used in Questions 6 and 7 were accessed through the National Earth System Science Data Center's Data Sharing Platform[44]. Any restrictions on data availability will be subject to the terms and conditions of the original data providers.

Code availability

The code corresponding to this paper is publicly accessible at https://github.com/acodercat/AirGPTunder the commit ID bc871d1. Detailed instructions for deployment and implementation are provided in the accompanying README file available in the repository.

Received: 23 May 2024; Accepted: 16 April 2025;
Published online: 13 May 2025

References

1. Epstein, Z. et al. Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).

2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019).
3. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
4. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
5. Grossmann, I. et al. Ai and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
6. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
7. Qiu, J. et al. Llm-based agentic systems in medicine and healthcare. *Nat. Mach. Intell.* **6**, 1418–1420 (2024).
8. Sandmann, S., Riepenhausen, S., Plagwitz, L. & Varghese, J. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nat. Commun.* **15**, 2050 (2024).
9. Zhang, X. & Yang, Q. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management,* (eds Frommholz, I. & Hopfgartner, F.), 4435–4439 (2023).
10. Fan, L. et al. A bibliometric review of large language models research from 2017 to 2023. *ACM Trans. Intell. Syst. Technol.* **15**, 1–25 (2024).
11. Zhang, S., Fu, D., Zhang, Z., Yu, B. & Cai, P. Trafficgpt: viewing, processing and interacting with traffic foundation models. *Transp. Policy* **150**, 95–105 (2024).
12. Yin, W., Xu, M., Li, Y. & Liu, X. Llm as a system service on mobile devices. arXiv preprint arXiv:2403.11805 https://arxiv.org/abs/2403.11805 (2024).
13. Sallam, M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Edited by Daniele Giansanti. In *Healthcare* Vol. 11, 887 (MDPI, 2023).
14. Baidoo-Anu, D. & Ansah, L. O. Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI* **7**, 52–62 (2023).
15. Oh, Y. et al. Llm-driven multimodal target volume contouring in radiation oncology. *Nat. Commun.* **15**, 9186 (2024).
16. Singhal, K. et al. Toward Expert-level Medical Question Answering with Large Language Models. *Nat. Med.* **31**, 943–950 (2025).
17. Vaghefi, S. A. et al. Chatclimate: grounding conversational AI in climate science. *Commun. Earth Environ.* **4**, 480 (2023).
18. Lawson, J. R. et al. Pixels and predictions: potential of GPT-4V in meteorological imagery analysis and forecast communication. *Artif. I. Earth Syst.* **4**, 240029 (2025).
19. Abdar, M., Tait, J. & Aleven, V. The impact of chatbot characteristics on user satisfaction and conversational performance. *J. Educ. Psychol.* **112**, 667–683 (2020).
20. Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P. & Mazurek, G. In bot we trust: a new methodology of chatbot performance measures. *Bus. Horiz.* **62**, 785–797 (2019).
21. Wang, X. et al. Mint: evaluating LLMs in multi-turn interaction with tools and language feedback. arXiv preprint arXiv:2309.10691 (2023).
22. Schick, T. & Schütze, H. Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676 (2021a).
23. Bi, Z. et al. Oceangpt: a large language model for ocean science tasks. arXiv preprint arXiv:2310.02031 (2023).
24. Taloni, A. et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci. Rep.* **13**, 18562 (2023).
25. Pichai, S. *Our Next-generation Model: Gemini 1.5* https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note (2024).
26. Anthropic. *Long Context Prompting for Claude 2.1* https://www.anthropic.com/news/claude-2-1-prompting (2023).
27. OpenAI. *Rate limits—openAI API* https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-free (2024).
28. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
29. Hopsworks. *What is Retrieval Augmented Generation (RAG) for LLMs?* https://www.hopsworks.ai/dictionary/retrieval-augmented-generation-llm (2024).
30. Zhang, H. et al. Air pollution and control action in Beijing. *J. Clean. Prod.* **112**, 1519–1527 (2016).
31. Li, S., Feng, K. & Li, M. Identifying the main contributors of air pollution in Beijing. *J. Clean. Prod.* **163**, S359–S365 (2017).
32. Li, Y., Huang, S., Yin, C., Sun, G. & Ge, C. Construction and countermeasure discussion on government performance evaluation model of air pollution control: a case study from Beijing–Tianjin–Hebei region. *J. Clean. Prod.* **254**, 120072 (2020).
33. Wang, J.-F., Hu, M.-G., Xu, C.-D., Christakos, G. & Zhao, Y. Estimation of citywide air pollution in Beijing. *PLoS ONE* **8**, e53400 (2013).
34. Huang, J., Pan, X., Guo, X. & Li, G. Health impact of China's air pollution prevention and control action plan: an analysis of national air quality monitoring and mortality data. *Lancet Planet. Health* **2**, e313–e323 (2018).
35. Zhou, L. & Chen, W. Impact of district-level decomposition policies to achieve a post-fossil carbon city: a case study of Beijing, China. *J. Clean. Prod.* **215**, 1371–1381 (2019).
36. Xu, M., Grant-Muller, S. & Gao, Z. Implementation effects and integration evaluation of a selection of transport management measures in Beijing. *Case Stud. Transp. Policy* **5**, 604–614 (2017).
37. He, G., Fan, M. & Zhou, M. The effect of air pollution on mortality in China: evidence from the 2008 Beijing Olympic games. *J. Environ. Econ. Manag.* **79**, 18–39 (2016).
38. Wu, D., Xu, Y. & Zhang, S. Will joint regional air pollution control be more cost-effective? An empirical study of China's Beijing–Tianjin–Hebei region. *J. Environ. Manag.* **149**, 27–36 (2015).
39. Chen, W., Yan, L. & Zhao, H. Seasonal variations of atmospheric pollution and air quality in Beijing. *Atmosphere* **6**, 1753–1770 (2015).
40. Wang, L. et al. Taking action on air pollution control in the Beijing–Tianjin–hebei (BTH) region: progress, challenges and opportunities. *Int. J. Environ. Res. Public Health* **15**, 306 (2018).
41. Zhong, S., Yu, Z. & Zhu, W. Study of the effects of air pollutants on human health based on Baidu indices of disease symptoms and air quality monitoring data in Beijing, China. *Int. J. Environ. Res. Public Health* **16**, 1014 (2019).
42. Sun, L. et al. Reducing energy consumption and pollution in the urban transportation sector: a review of policies and regulations in Beijing. *J. Clean. Prod.* **285**, 125339 (2021).
43. Wang, X. et al. Sensitivities of ozone air pollution in the Beijing–Tianjin–Hebei area to local and upwind precursor emissions using adjoint modeling. *Environ. Sci. Technol.* **55**, 5752–5762 (2021).
44. National Earth System Science Data Center. *National Earth System Science Data Center—Earth System Science Data Sharing Platform* https://www.geodata.cn/data/index.html?publisherGuid=126744287495931&categoryId=2 (2023).

Acknowledgements

This work is supported by the Smart Society Lab at Hong Kong Baptist University.

Author contributions

Jun Song conceptualized and led this research project, established the fundamental methodology, provided strategic guidance throughout the investigation, and reviewed the final manuscript. Chendong Ma performed the experimental design and implementation, conducted data analysis and interpretation, created visualizations, and took primary responsibility for manuscript preparation and revision. Maohao Ran developed and implemented the computational framework and performed the programming tasks essential to the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-025-01070-4.

Correspondence and requests for materials should be addressed to Jun Song.