npj | climate action

Article

https://doi.org/10.1038/s44168-025-00254-1

# MARSHA: multi-agent RAG system for hazard adaptation

Check for updates

**Yangxinyu Xie[1] ✉, Bowen Jiang[2], Tanwi Mallick[3] ✉, Joshua Bergerson[4], John K. Hutchison[4], Duane R. Verner[4], Jordan Branham[4], M. Ross Alexander[4], Robert B. Ross[3], Yan Feng[5], Leslie-Anne Levy[4], Weijie J. Su[1] & Camillo J. Taylor[2]**

Large language models (LLMs) are a transformational capability at the frontier of artificial intelligence and machine learning that can support decision-makers in addressing pressing societal challenges such as extreme natural hazard events. As generalized models, LLMs often struggle to provide context-specific information, particularly in areas requiring specialized knowledge. In this work, we propose a retrieval-augmented generation (RAG)-based multi-agent LLM system to support analysis and decision-making in the context of natural hazards and extreme weather events. As a proof of concept, we present WildfireGPT, a specialized system focused on wildfire hazards. The architecture employs a user-centered, multi-agent design to deliver tailored risk insights across diverse stakeholder groups. By integrating natural hazard and extreme weather projection data, observational datasets, and scientific literature through an RAG framework, the system ensures both the accuracy and contextual relevance of the information it provides. Evaluation across ten expert-led case studies demonstrates that WildfireGPT significantly outperforms existing LLM-based solutions for decision support.

Natural hazards and extreme weather events such as wildfires, floods, and hurricanes present significant operational and management challenges across sectors, particularly in the management of critical infrastructure systems[1–6]. From intensifying heatwaves[7] and floods[8] to more frequent and severe wildfires[9,10], natural hazard events are becoming increasingly disruptive to infrastructure networks, demanding more effective tools for risk assessment, planning, and response.

Although scientific research provides the building blocks for a wide range of solutions to address the increasing risk from natural hazards, the complexity and volume of published research often hinder the efficient translation of scientific knowledge into risk reduction policies and programs[11]. This disconnect is particularly problematic for professionals at the forefront of managing hazard-related risks, such as urban planners, emergency managers, and infrastructure operators, who may have limited access to the latest knowledge and methods to mitigate these threats. Recent advances in natural language processing, especially large language models (LLMs), present an innovative solution for democratizing science on natural hazard resilience and facilitating knowledge transfer[12–18]. LLMs possess the potential to process and synthesize vast amounts of textual information and explain them through conversations, making crucial information accessible to people from diverse backgrounds[19–26].

Recent research has explored the capabilities of LLMs in natural hazard science and management contexts. DisasterResponseGPT enables users to input disaster scenario descriptions and receive action plans as outputs[27]. To explore the potential for LLMs to revolutionize fire engineering, researchers evaluated LLM performance across several fire engineering scenarios, including structural design, prevention strategies, and regulatory compliance[28]. Building on these evaluations, researchers have explored multi-round prompting techniques that allow users to iteratively refine LLM responses with additional context, enhancing disaster management applications[27,29]. To address limitations in contextual understanding and domain-specific knowledge, ChatClimate[30] and ClimateGPT[31] incorporate assessment reports from the Intergovernmental Panel on Climate Change (IPCC). More sophisticated implementations connect LLMs with climate models such as the Model for the Assessment of Greenhouse Gas Induced Climate Change[32]. In the context of flood risk management, researchers have integrated relevant geospatial and demographic data[33]. Beyond academic solutions, commercial tools such as Perplexity AI[34] offer LLM-

[1]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. [3]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA. [4]Decision and Infrastructure Sciences Division, Argonne National Laboratory, Lemont, IL, USA. [5]Environmental Science Division, Argonne National Laboratory, Lemont, IL, USA.
✉e-mail: xinyux@wharton.upenn.edu; tmallick@anl.gov

integrated search capabilities that can retrieve and synthesize relevant hazard information from current scientific literature. These advancements demonstrate the potential of LLMs to bridge the gap between scientific research and practical natural hazard risk management applications.

Nonetheless, existing LLM research for natural hazard decision support faces fundamental limitations in personalization, data utilization, and evaluation. First, prior studies primarily focus on LLMs' performance on generic, one-off questions[26–33,35]. This approach assumes users possess sufficient expertise to articulate clear queries, disregarding the complex reality that professionals have varying backgrounds and levels of expertise and often require multiple rounds of interaction to capture their context-specific information needs. This issue is compounded by the tendency of current LLMs to provide hasty, generic responses without requesting clarifying details necessary for developing detailed, context-specific responses[36,37]. The combination of these factors—assuming user expertise and providing quick, undifferentiated answers—can lead to significant misunderstandings or oversimplifications. Second, the spatial heterogeneity of natural hazards and infrastructure vulnerability demands location-specific analyses at granular resolution. The pronounced lack of textual (i.e., scientific papers and reports) and data-based (i.e., projections and observational data) knowledge integration in existing research[22,23,26,30–32,38] hinders the ability to synthesize comprehensive evidence grounded in local contexts[39]. Our comparative analysis with existing alternatives such as ChatClimate[30] and Perplexity AI[34] shows that standard LLM applications fail to reliably interpret complex grid-structured data formats, critically limiting their utility for location-specific risk assessment and infrastructure planning. Third, systematic evaluation of LLMs at the intersection of the natural hazard and infrastructure domains presents significant methodological challenges. Conventional evaluation frameworks based on natural language processing predominantly rely on metrics that quantify lexical similarity or assess linguistic fluency approaches that are inadequate for domain-specific applications requiring actionable insights. While some studies have explored fact-checking of LLM-generated content[20,26,30], such verification represents only a preliminary step toward comprehensive evaluation. Critically absent from current assessment protocols is the systematic measurement of response utility within decision-making contexts specifically, the capacity of LLM outputs to provide contextually appropriate, implementable recommendations for infrastructure resilience that accurately reflect localized hazard conditions. This evaluation gap necessitates specialized expertise capable of assessing both factual accuracy and practical applicability within complex sociotechnical systems.

To address these limitations, we introduce an innovative multi-agent LLM prototype designed as a co-pilot for understanding natural hazards and developing adaptation strategies. Our approach encompasses several key contributions:

1. Human-Centered Personalization: The agents are designed to engage users in a series of questions to understand their professional background, location of interest, and specific concerns about natural hazards. Based on this input, the system develops a strategic plan for analysis, determining which datasets to examine, what literature to review, and the appropriate scope for recommendations.
2. Data Integration and Interactive Visualization: We combine location-specific projections, hazard history, census characteristics, and scientific literature to augment LLM responses using a retrieval-augmented generation (RAG) approach. Moreover, our system integrates interactive visualizations of geospatial data with textual outputs, enabling nuanced user exploration of multidimensional risk factors.
3. Three-Stage Evaluation: We assess the performance through 10 case studies covering a diverse range of topics and locations. Our evaluation comprises three complementary stages: (1) a modular comparison against ChatClimate and Perplexity AI for data and literature retrieval effectiveness, (2) a qualitative ablation study examining how profile specificity impacts response quality, and (3) a detailed assessment of case study responses through expert evaluation of utility and exploration of LLM-as-a-judge for automated assessment[40,41]. This evaluation process focuses on practical value and real-world performance. By incorporating LLM-as-a-judge, we explore the potential for scalable, automated evaluation that could maintain quality assurance during deployment without constant human oversight.

As a proof of concept, we develop WildfireGPT, an LLM tool to support decision-making surrounding wildfire risk and resilience within the United States. Projected environmental and land-use changes are expected to alter the frequency, severity, seasonality, and spread of wildfires across the United States[4,9,42,43]. Projected increases in lightning frequency are expected to increase the frequency of wildfire ignitions[44–46]. At the same time, the frequency, duration, severity, and seasonality of fire weather conditions are projected to increase, including increasing severity and frequency of drought conditions due to shifting precipitation patterns, increasing ambient and extreme temperatures, and changing wind patterns[47,48]. Beyond these environmental impacts, population growth, increased development in and adjacent to wildfire-prone areas, and modern fire suppression policies are expected to further increase wildfire risk[4,49]. This paper documents the development, methodology, and initial findings of the WildfireGPT prototype through comprehensive case studies and evaluation, offering a new paradigm for artificial intelligence (AI)-assisted decision-making in natural hazard risk management.

## Results

### The WildfireGPT experience

**User-centered multi-agent transition**. In our study, we employ GPT-4 Turbo[50] as the backbone of WildfireGPT, enabling it to dynamically invoke different agents via function calling through the OpenAI Assistant API[51]. The user interacts with WildfireGPT through a chat interface built on the Streamlit-based web app[52]. Behind the scenes, we designed WildfireGPT as a multi-agent system where specialized agents collaborate under the coordination of a task orchestrator that routes interactions appropriately throughout the workflow. The integrated functions and multi-agent system architecture, including the implementation details of each specialized agent, are detailed in the Methodology section.

From an end-user perspective, WildfireGPT linearly transitions through three main agents—the user profile agent, the planning agent, and the analyst agent—to guide users through a seamless, interactive, and personalized experience (Fig. 1). The user profile agent gathers information by asking a series of predefined questions. These questions cover the user's professional background, primary concerns, location of interest, timeframe for addressing concerns, and specific aspects of wildfire risks the user wants to explore. Then, the agent generates a detailed user profile and prompts the user to review and confirm the profile's accuracy. Based on the user profile, the planning agent formulates a step-by-step action plan outlining the datasets to be analyzed, the focus of the literature review, and how the recommendations will be formulated by the analyst agent. The user can provide feedback on the plan, allowing for refinements until it meets the user's expectations, ensuring transparency and oversight. The analyst agent guides the user through the analysis process outlined in the plan, presenting findings and recommendations, while actively addressing follow-up questions to ensure a comprehensive understanding and practical application of the results.

**Diverse data sources**. WildfireGPT combines natural hazard projections, observational data, socioeconomic indicators, and scientific literature to deliver a multifaceted analysis of wildfire risks. Location-specific datasets include Fire Weather Index (FWI)[53] projections from the ClimRR portal[54], recent wildfire incident records (2015 to 2023) from the Wildland Fire Interagency Geospatial Services Group[55,56], tree-ring and sediment-based fire history records from the International Multiproxy Paleofire Database[57], and census data on poverty rates and housing units from the 2022 American Community Survey 5-year estimates[58].

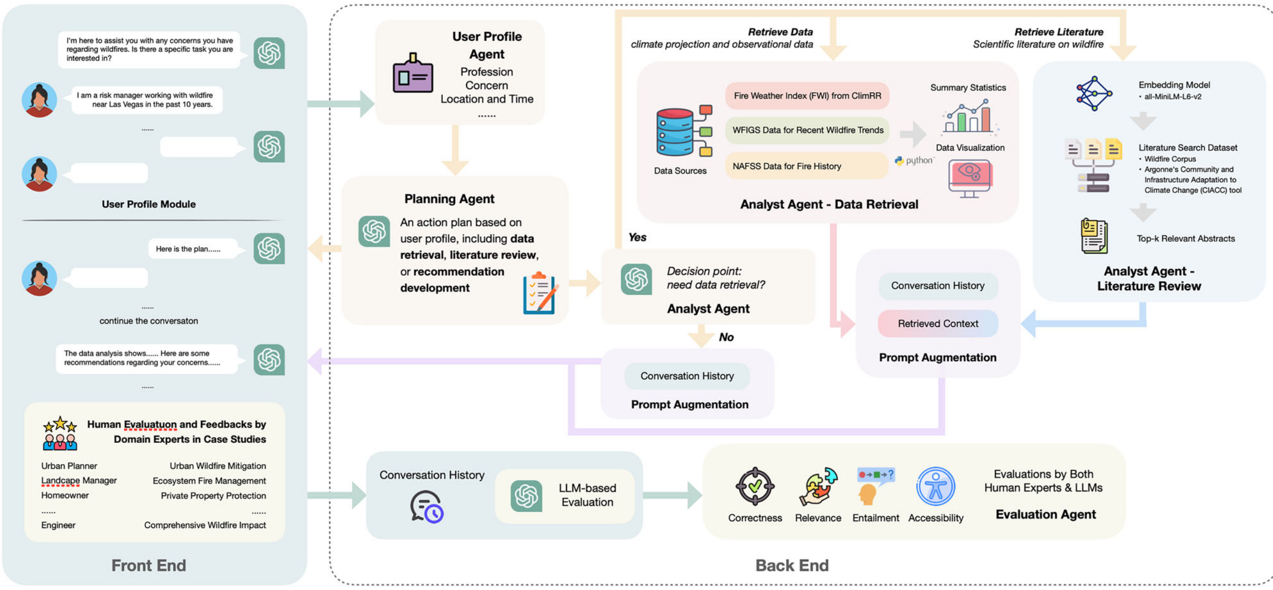


**Fig. 1 | Overview of WildfireGPT architecture integrated with the multi-agent RAG framework.** The WildfireGPT prototype focuses on enhancing consulting interactions using LLM agents stepping through a multistage approach. Its user profile agent engages the user with a tailored questionnaire to create a detailed profile; the planning agent formulates a customized action plan addressing the user's queries and concerns, ensuring alignment with their evolving needs; and the analyst agent aids in accessing and interpreting relevant data and literature and providing recommendations.

Furthermore, a vast collection of abstracts of scientific literature on wildfires, developed by Argonne National Laboratory for the Community and Infrastructure Adaptation to Climate Change tool, is also integrated[59,60].

**Interactive visualizations.** WildfireGPT provides interactive geospatial visualizations of the location-specific data within a 36 km radius of the user's area of interest, allowing the user to explore and analyze the data. For instance, the FWI displayed in Fig. 2 uses a color scale ranging from yellow (low risk) to red (very extreme). By hovering over a specific grid, the user can view the exact FWI value for that location; by selecting different seasons and time periods, the user can comprehend the changing landscape of the FWI. When visualizing recent wildfire history, the locations of all fire records are displayed as red dots on the map, and the temporal trend of wildfire incidents is shown in line graphs. Socioeconomic data, such as poverty rates and housing units, can be explored by hovering over census block groups. This comprehensive spatial visualization facilitates insights into the interplay among natural hazards, wildfire risk, and socioeconomic factors, enabling informed decision-making and risk assessment.

## Evaluating WildfireGPT: a framework built on real-world case studies

Evaluating conversational systems in the natural hazard domain presents unique challenges, including validating information from diverse datasets, handling diverse user queries across varying levels of expertise and backgrounds, assessing hazard conditions and their impact on different critical infrastructure, ensuring the accuracy and reliability of responses based on the user's profile and background, and maintaining contextual relevance. While conversational systems are often evaluated by using lexical similarity metrics (e.g., BLEU[61], ROUGE[62]), they correlate poorly with human expert assessment of domain-specific content generation[63,64]. To address this limitation, we adopted an evaluation approach that prioritizes expert judgment over purely automated metrics. As part of this effort, we conducted 10 in-depth case studies with domain experts from Argonne National Laboratory specializing in natural hazard resilience, wildfire risk management, and infrastructure vulnerability assessment. These domain experts participated in one-hour structured interviews, during which they directly interacted with WildfireGPT through a facilitated interface. The experts either posed questions from their ongoing projects or represented stakeholder perspectives based on their regional collaborations, providing real-time feedback on system performance. These case studies encompassed a diverse range of wildfire-related topics and locations across the United States, with details summarized in Table 1.

Our evaluation of WildfireGPT employs three complementary stages. First, we conducted a modular comparison of the analyst agent's data and literature retrieval effectiveness against two alternatives ChatClimate[30] and Perplexity AI[34] using controlled simulations based on our case studies. Second, we performed a qualitative ablation study examining how different levels of user profile specificity impact the specificity and actionability of WildfireGPT's responses, also through controlled simulations derived from case study topics. Third, we evaluated the actual responses produced during expert case studies through: (1) expert assessment of response relevance, entailment, and accessibility and (2) exploration of scaling the human-in-the-loop evaluation using LLM-as-a-judge. (Our evaluation framework focuses on response quality and retrieval performance, as user intent identification is intrinsically validated through the user profile agent and planning agent's confirmation protocol with the user—a design choice that differs from traditional conversational systems[65]. Both profile and planning components incorporate explicit user verification of generated summaries before proceeding, embedding intent validation directly within the interaction pipeline.)

**Comparative evaluation against other models.** To evaluate the effectiveness of data retrieval and evidence-based query response capabilities, we compare WildfireGPT against two baseline alternatives practitioners might consider: ChatClimate and Perplexity AI. ChatClimate is a conversational AI platform developed to enhance understanding of natural hazard risks by providing accessible information grounded in the IPCC reports; Perplexity AI is an AI-powered answer engine that delivers real-time responses by summarizing information from multiple web sources. Specifically, we extract 20 prompts from our case studies to test whether each method is able to (i) retrieve and interpret location-specific wildfire data and (ii) provide evidence-based answers to domain-specific questions. The evaluation metrics for data analysis include data provision (successful retrieval of relevant data),

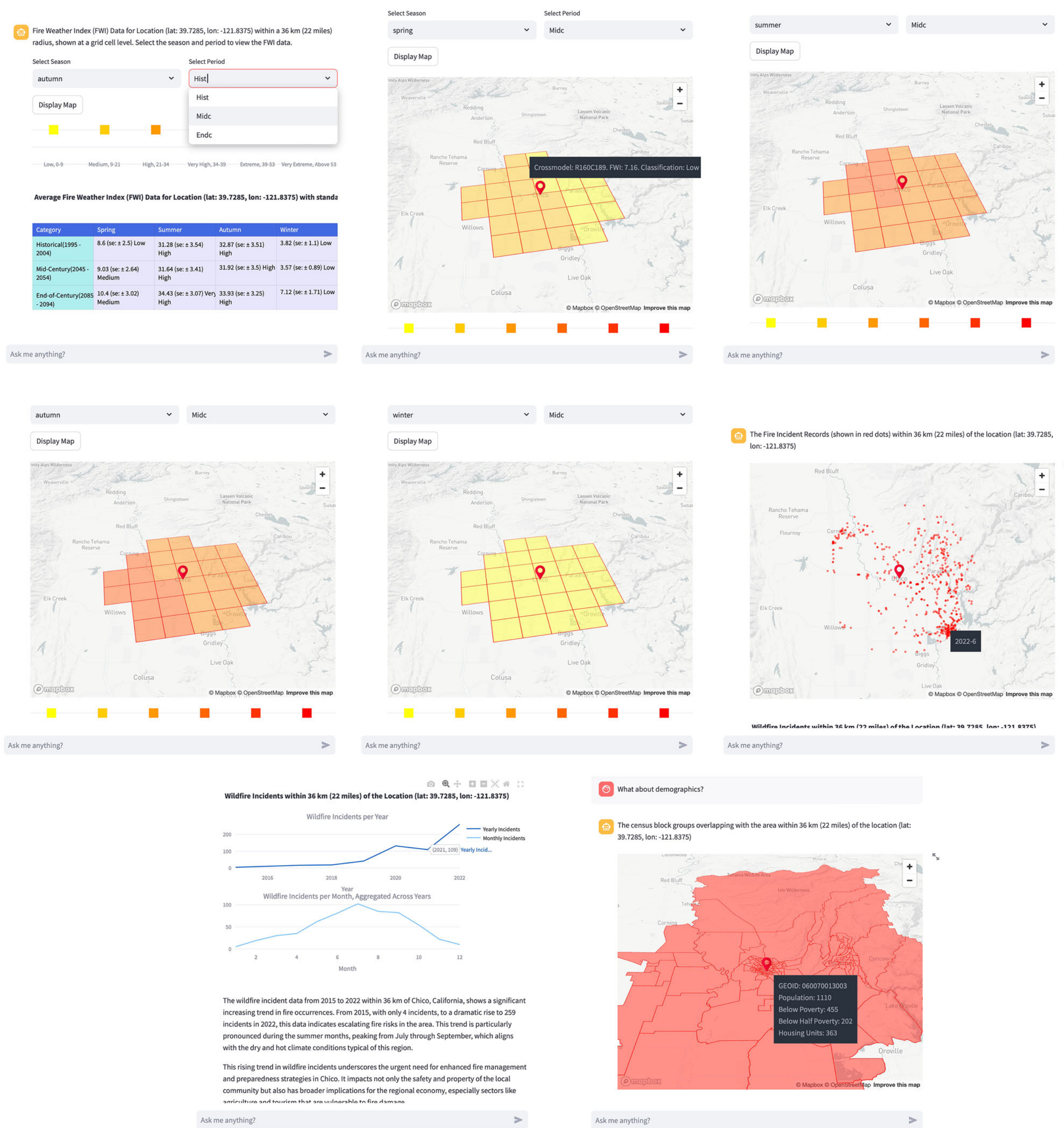


**Fig. 2 | Interactive visualizations in the WildfireGPT user experience.** This example is taken from the case study themed "Current Wildfire Risk Analysis." Users can select the season and time period to view the corresponding Fire Weather Index (FWI) map (left of first row), which displays risk levels using a color scale. Location-specific FWI values are accessible by hovering over the map (center of first row). By selecting different seasons and time period, the users can comprehend the changing landscape of the FWI (center of first row to center of second row). The wildfire incident map (right of the second row) shows the spatial distribution of recent fires, while the line graph (bottom left) presents the temporal trend of incidents. Socio-economic data is visualized through census block group overlays (bottom right), providing insights into poverty rates and housing units in each area.

location specificity (geographical precision of retrieved data), and data accuracy (alignment between retrieved data and prompt requirements). Similarly, we compare the integration of external knowledge, the citation practices, the authority of the source, and the contextual relevance of responses to domain-specific questions.

**Personalization.** To investigate the effects of the user profile agent, we conduct a qualitative, two-phase ablation study to systematically evaluate the impact of user profile granularity and diversity on WildfireGPT's recommendations. Inspired by the Private Property Protection case study, for both phases we use the prompt "Develop recommendations to enhance forest resilience against wildfires" to elicit a response from WildfireGPT. In the first phase, we systematically test three levels of user profile specificity provided to the agent: no profile information provided, location plus timeline, and the full profile of a power grid manager focused on maintaining transmission line clearance and grid resilience for power distribution reliability and access. In the second phase, we hold location and timeline constant and test five distinct professional profiles —homeowner, civil engineer, ecologist, emergency manager, and power grid manage—each with minimally distinct concerns and the respective

**Table 1 | Case studies conducted with domain experts to evaluate the performance and applicability of WildfireGPT across various professions, geographic locations, and wildfire management themes**

| Theme | Profession | Location | Description |
|---|---|---|---|
| Climate Impact on U.S. Wildfires | Atmospheric Scientist | Boston, MA | Exploring medium-term (10–30 years) impact of climate on wildfire occurrences and intensity in the U.S. Analyzing studies showing an increase in wildfire occurrences in Boston. Identifying and addressing uncertainties affecting future wildfire risks. |
| Comprehensive Wildfire Impact | Engineer | Mount Bigelow, AZ | Exploring long-term (50+ years) wildfire frequency, intensity, and impacts. Investigating mitigation strategies, ecological and economic effects, and infrastructure considerations. |
| Current Wildfire Risk Analysis | Data Analyst | Chico, CA | Analyzing current wildfire data, causes, and forest management practices. Exploring predictive models incorporating real-time variables. Gathering detailed demographic and socioeconomic profiles to assess risk and impact. |
| Ecosystem Fire Management | Landscape Manager | Naperville, IL | Analyzing wildfire frequency and impacts on vegetation with a focus on oak ecosystems. Developing strategies to balance fire use for oak ecosystem management with risk reduction for residential areas. |
| Hazard Mitigation Planning | Hazard Mitigation Planner | Mora County, NM | Assessing wildfire likelihood, frequency, intensity, and timing over the next 15 years for community hazard mitigation planning. |
| Infrastructure Wildfire Risk | Climate Change Risk Analyst | Denver, CO | Analyzing historical wildfire data and future projections to assess risks to energy sector infrastructure, considering land use changes and climate models (RCP 8.5 scenario). |
| Post-Wildfire Public Safety | Public Safety Manager | Sangre De Cristo Mountains, NM | Focusing on short-term (1–10 years) mitigation strategies after a high-intensity fire. Assessing emergency services readiness, community preparedness programs, and environmental restoration initiatives with the U.S. Army Corps of Engineers on erosion and flooding mitigations. |
| Private Property Protection | Homeowner | Near Covington, VA | Developing a 5–10-year plan for managing forest health, maximizing marketable species (focus on oak and cherry), and protecting properties from wildfires. |
| Urban Wildfire Mitigation | Urban Planner | Beaverton, OR | Developing a 30-year community plan focused on building codes and infrastructure resilience to mitigate fire-related hazard risks. |
| Wildland Urban Interface Impact | Risk Manager | Las Vegas, NM | Assessing medium-term (10–30 years) wildfire risks to housing and water resources in the wildland-urban interface. Exploring historical data, predictive models, current housing structures, future development plans, and wildfire impact on water quality and availability. Examining existing and proposed mitigation strategies. |

Each case study represents either a location-specific inquiry derived from experts' ongoing research projects or a stakeholder perspective based on established regional collaborations. The description column outlines the specific focus of each case. The studies span different time horizons and incorporate a range of data types and analytical approaches to assess wildfire risk, impact, and mitigation strategies. Case studies are listed in alphabetical order.

scopes. Table 2 summarizes the differences in primary concerns and scopes across these professional profiles, with more detailed profile descriptions available in the Methodology section. We conduct a controlled simulation for each profile by passing the profile directly into WildfireGPT's planning and analyst agents, allowing us to observe how the same prompt generated tailored recommendations across different user types. Outputs are analyzed across three dimensions: *plan generation*, *literature retrieval*, and *recommendation specificity*.

**Domain expert evaluation**. During the case studies, whenever WildfireGPT generates a response, domain experts from Argonne National Laboratory's Environmental Science Division and Decision and Infrastructure Sciences Division are asked to evaluate its relevance, entailment, and accessibility using a structured questionnaire. These experts are scientists and engineers specializing in natural hazards and wildfire control with different detailed concentrations or other related professions as listed in Table 1. Each question is rated on a three-point scale, with "Yes" receiving a score of 1, "Could be better" a score of 0.5, and "No" a score of 0. The valuation criteria and corresponding questions are presented in Table 3. Relevance assesses whether the model's responses appropriately address the user's last question and are relevant to the user's profession, concerns, location, timeline, and scope. Entailment evaluates the logical coherence between the model's analyses/recommendations and the provided data or literature. Accessibility examines the clarity and concision of the model's language, considering factors such as jargon, explanatory detail, and redundancy. We report the total scores for each criterion across all case studies. To complement the quantitative analysis, we also report qualitative insights from the expert evaluations to provide a more nuanced understanding of WildfireGPT's performance.

We additionally verify whether the reported data and citations in WildfireGPT's responses originate from the retrieved sources, and this result is provided in the Supplementary Materials.

**LLM-as-a-Judge evaluation**. To explore the potential for scalable, automated evaluation that could maintain quality assurance during deployment without constant human oversight, we implemented an LLM-as-a-judge approach. This method involves using a separate GPT-4[50] query to assess the quality of WildfireGPT's responses based on the same criteria used in the expert evaluation. We report the agreement of the LLM-as-a-judge evaluation with the expert evaluation to determine the feasibility of using this approach. Each case study involved a unique interaction between the user and WildfireGPT, resulting in varying uses of data sources and different numbers of rounds of follow-up interactions. Our evaluation framework was designed to accommodate this variability, ensuring that the assessment remained meaningful and relevant to each specific case.

### Comparative evaluation against baseline models

The comparative evaluation incorporated 20 prompts derived from the case studies: (i) location-specific data analysis tasks ($n = 10$) requiring retrieval and interpretation of wildfire risk data (Table 4); and (ii) evidence-based question-answering tasks ($n = 10$) necessitating domain-specific knowledge retrieved from the corpus of scientific literature (Table 5). The

**Table 2 | User profile variations and literature search queries in Phase 2 of the WildfireGPT personalization ablation study**

| Profession | Primary Concern | Scope | Search Query |
|---|---|---|---|
| Homeowner | Maximizing marketable species | Health and marketable species | "Strategies for managing forests to maintain health, maximize marketable species, and minimize wildfire risks in Virginia" |
| Civil Engineer | Ensuring structural and infrastructural resilience | Drainage efficiency and slope stability | "Wildfire risks and climate change impacts on forest management near Covington, VA; Strategies for enhancing drainage efficiency and slope stability; Structural resilience against wildfires in forested areas" |
| Ecologist | Maintaining biodiversity and ecosystem services | Ecological resilience and habitat connectivity | "Wildfire management and ecological resilience in forest ecosystems near Covington, VA" |
| Emergency Manager | Establishing defendable space and evacuation corridors | Emergency access and response capabilities | "Effective forest management practices, defendable space creation, evacuation protocols, and property protection measures against wildfires near Covington, VA" |
| Power Grid Manager | Maintaining transmission line clearance and grid resilience | Power distribution reliability and access | "Effective strategies for vegetation management, forest health maintenance, and wildfire risk mitigation around power grids near Covington, VA" |

We hold location and timeline constant and test five distinct professional profiles, with changes in the concerns and the respective scopes. The search query represents the resulting literature retrieval query automatically generated by WildfireGPT's analyst agent.

**Table 3 | Domain expert evaluation criteria and corresponding questions used to assess WildfireGPT's performance in generating responses**

| Criteria | Questions |
|---|---|
| Relevance | (1) Does my response answer your last question? |
| | (2) Is my response relevant to your profession? |
| | (3) Is my response relevant to your concern? |
| | (4) Is my response relevant to your location? |
| | (5) Is my response relevant to your timeline? |
| | (6) Is my response relevant to your scope? |
| Entailment | (1) Do my analyses or recommendations logically follow from the information (data, literature) provided? |
| Accessibility | (1) Does my response contain too much jargon? |
| | (2) Does my response provide enough explanation? |
| | (3) Does my response contain redundant or useless information? |

Each question was rated on a three-point scale: "Yes" (1), "Could be better" (0.5), and "No" (0).

corresponding case study of each prompt is detailed in the Supplementary Materials.

The comparative data analysis in Table 4 reveals distinct performance patterns across the three models WildfireGPT, ChatClimate, and Perplexity AI in the categories of (1) data provision (successful retrieval of wildfire data); (2) location specificity (geographical precision of retrieved data); and (3) data accuracy (correctness of the retrieved data values when compared with reference datasets, evaluable only for WildfireGPT (with direct database access) and Perplexity AI (where external data files could be uploaded for analysis)). In the data provision category, WildfireGPT significantly outperforms the others, succeeding in 9 out of 10 prompts, indicating strong capability in retrieving relevant wildfire-related data. Perplexity AI follows with 7/10, while ChatClimate trails with just 5/10. In terms of location specificity, WildfireGPT again leads with 9/9, demonstrating precise geographic referencing in its responses. Perplexity AI achieves moderate performance with 4/7, whereas ChatClimate fails entirely in this category (0/6). For data accuracy, WildfireGPT shows a clear advantage, with 8/9 correct outputs compared with Perplexity AI's 2/8, indicating a substantial gap in the factual correctness of retrieved data. Note that while our initial evaluation framework included 10 test prompts, the effective sample size varies across categories because of dependencies between evaluation criteria. For instance, location specificity could be assessed only when data provision was successful, and accuracy could be evaluated only when the provided data contained specific locations. Overall, WildfireGPT consistently shows the highest performance across all three categories, making it the most robust tool for wildfire data analysis among those compared.

WildfireGPT demonstrated superior performance in location-specific data retrieval due to the system's user profile agent, which systematically verifies exact geographical coordinates with users prior to data retrieval attempts. This feature ensures precise location identification, enabling successful data retrieval whenever the requested information exists within the database. The failure occurred in the Infrastructure Wildfire Risk case study (Denver, CO), where WildfireGPT provided Fire Weather Index analysis using RCP 4.5 rather than the requested RCP 8.5 projections because RCP 4.5 was the only scenario available in the dataset, although this limitation was not acknowledged when WildfireGPT interacted with the user an issue that could be addressed in future iterations. ChatClimate exhibited significant analytical constraints, as its responses systematically defaulted to broad climatological generalizations without geographical precision, consistently failing to tailor information to the location in user queries. For example, in analyzing Boston's fire danger trends, ChatClimate inappropriately expanded its analytical scope to Arctic and Western U.S. fire patterns. This systematic geographical decontextualization significantly compromised the system's utility for location-specific wildfire risk assessment. This issue stems in part from ChatClimate's primary reliance on IPCC reports and a limited repository of studies within its database, which prioritize global or regional-scale climate insights over localized, granular data. Perplexity AI demonstrated reasonable performance in data provision but exhibited substantial limitations in location specificity when the corresponding data file was not provided by the user. The system primarily leverages publicly accessible information rather than conducting direct data analysis, resulting in significant geographical variance in performance based on public data availability. In regions with robust public documentation, such as Chico, CA, the system successfully integrated multiple data sources including local hazard mitigation plans. Conversely, analysis of areas with limited public records resulted in either null responses or inappropriate source application: for example, the system erroneously applied European wildfire danger reports to analyze local fire risk in Beaverton, OR. When the user uploads the data file, Perplexity AI incorrectly represents data extracted at geographical coordinates different from those specified in the query while claiming spatial correspondence or claims that the provided data does not include the requested information.

The Ecosystem Fire Management case study (Naperville, IL) provides a particularly instructive methodological comparison when confronting identical data limitations. When prompted to analyze 150-year fire history records for Naperville, all three systems encountered data unavailability. WildfireGPT's response demonstrated adaptability by explicitly acknowledging the absence of historical data and proposing alternatives—specifically recommending the user to explore recent fire incident data and FWI

**Table 4 | Comparison of wildfire analysis tools for data analysis**

| Prompt | Data Provision | | | Location Specificity | | | Data Accuracy | |
|---|---|---|---|---|---|---|---|---|
| | WGPT | CC | PAI | WGPT | CC | PAI | WGPT | PAI |
| (1) Analyze future climate data to understand potential trends in fire danger in Boston, Massachusetts | ✓ | ✓[1] | ✓ | ✓ | × | ✓ | ✓ | × |
| (2) Analyze the recent fire incident data for Mora County, New Mexico, reviewing wildfire frequency, intensity, and timing from 2015 to 2023 | ✓ | × | ✓ | ✓ | NA | ✓ | ✓ | × |
| (3) Review the long-term fire history records to assess the frequency and intensity of past wildfires over the last 50+ years in Mount Bigelow, Arizona | ✓ | × | × | ✓ | NA | NA | ✓ | ✓ |
| (4) Analyze demographic and socioeconomic profiles of the Chico area in California to understand which groups are most vulnerable in wildfire scenarios | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ |
| (5) Analyze long-term fire history records to assess wildfire events over the past 150 years in the Naperville, Illinois, region | × | × | × | NA | NA | NA | NA | NA |
| (6) Examine climate models and projections under the RCP 8.5 scenario to forecast potential future wildfire risks in the medium term (10–30 years) in Denver, Colorado | ✓ | ✓ | ✓ | ✓ | × | × | × | NA |
| (7) Analyze the recent fire incident data from the Wildland Fire Interagency Geospatial Services Group to assess the frequency, intensity, and locations of recent wildfires in Sangre de Cristo Mountains | ✓ | × | × | ✓ | NA | NA | ✓ | × |
| (8) Analyze the recent fire incident data from the last five years in Covington, Virginia | ✓ | × | ✓ | ✓ | × | ✓ | ✓ | × |
| (9) Analyze the FWI focusing on both current trends and projections for the mid-century period (2045–2054) to assess the potential increase in fire danger due to climate change in Beaverton, Oregon | ✓ | ✓ | ✓ | ✓ | × | × | ✓ | × |
| (10) Analyze projected wildfire risk data for the mid-century period (2045–2054) to understand the potential trends in fire danger in Las Vegas, New Mexico | ✓ | ✓[1] | ✓ | ✓ | × | × | ✓ | × |
| **Overall Success Rate** | **9/10** | **5/10** | **7/10** | **9/9** | **0/6** | **4/7** | **8/9** | **2/8** |

*WGPT* WildfireGPT, *CC* ChatClimate, *PAI* Perplexity AI

Data accuracy could not be evaluated for ChatClimate because there is no way to upload data files to its interface. Each row represents a specific prompt or task related to wildfire analysis. For each metric, a ✓ indicates that the tool met the criteria for that prompt, while a × indicates it did not. NA means the metric was not applicable or could not be evaluated, typically because of the previous criteria not being met; for example, when data is not provided, it is not applicable to discuss the specificity or accuracy of the data. The superscript [1] indicates that although no specific number was mentioned, ChatClimate nonetheless interprets the increase in fire risks in words.

projections to maintain analytical continuity despite the constraint. ChatClimate responded with generalization, discussing broad climatic shifts and anthropogenic interventions affecting wildfire patterns over the last 150 years without any Naperville-specific information, effectively abandoning the geographical aspects within the query. Perplexity AI acknowledged the data limitation but methodologically pivoted to institutional historiography, discussing the Naperville Fire Department's organizational history rather than wildfire patterns—a response that maintained geographical specificity while deviating from the analytical framework of the original query. These differential responses to identical data constraints reveal distinct epistemological approaches to knowledge gaps that systematically influence analytical utility.

For evidence-based question answering, we employed four metrics: (1) external knowledge integration incorporation of relevant information beyond vanilla language generation; (2) citation practices provision of complete, verifiable bibliographic information; (3) source authority utilization of peer-reviewed scientific literature versus non-academic sources; and (4) contextual relevance application of information to the specific geographical, ecological, and socioeconomic characteristics of query locations. Table 5 shows distinct patterns in scientific literature integration and source validation. By design, all methods demonstrated strong integration of external knowledge (7/10 to 10/10). When external knowledge was integrated, both WildfireGPT and Perplexity AI provided complete citations of the sources, while ChatClimate exhibited bibliographic deficiencies (3/7 in citation provision). Despite using authoritative sources, ChatClimate uses only in-text parenthetical citations author surname and year for scientific literature and standardized section references for IPCC documents (e.g., "IPCC_AR6_WGII_Chapter13, p.20")–without providing complete bibliographic lists. While readers familiar with climate science literature might recognize and locate IPCC citations, most scientific sources remain difficult to verify without full bibliographic information. WildfireGPT and ChatClimate relied on authoritative sources (7/7). Perplexity AI, on the other hand, demonstrated variable source authority (6/10), frequently incorporating non-peer-reviewed content including commercial publications, news media, non-profit organizations, and social media sources.

Contextual relevance varied substantially across systems. WildfireGPT maintained high contextual relevance (10/10), providing literature and analyses that shared characteristics with the specified locations, whether through direct geographical proximity (same state or region) or ecological similarity (comparable forest types, fire regimes, or climate patterns). For example, in responses to queries about Sangre de Cristo Mountains and Mount Bigelow, WildfireGPT identified studies conducted in comparable montane forest ecosystems with similar elevation profiles and fire history patterns. Perplexity AI similarly achieved high contextual relevance (10/10), although with variable levels of authority in source selection. The contextual relevance can be attributed to both WildfireGPT and Perplexity AI integrating a large set of sources to allow rich information to be retrieved, although we identify areas for improvement in the following sections. In contrast, ChatClimate demonstrated substantial contextual relevance limitations (4/10), frequently defaulting to continental or hemispheric climate trends without establishing clear applicability to the specified locations. This geographical generalization was particularly evident in responses to queries about Las Vegas, NM, and Covington, VA, where the system referenced broad North American climate patterns without demonstrating specific relevance to the distinctive fire regimes of the southwestern high desert or Appalachian forest ecosystems, respectively.

Overall, WildfireGPT demonstrated precise geographical analysis with methodological adaptation when confronting data limitations. ChatClimate exhibited geographical imprecision. Perplexity AI demonstrated comprehensive information retrieval with significant limitations in source validation and data interpretation accuracy. These performance patterns highlight fundamental distinctions in the analytical architecture of the three models. Furthermore, they demonstrate that having a wide but curated range of valid scientific resources with efficient retrieval mechanisms is critical for location-specific wildfire risk assessment and evidence-based domain-specific inquiry.

Table 5 | Comparison of Wildfire analysis tools for evidence-based question qnswering

| Prompt | External Knowledge Retrieved | | | Citation Provision | | | Authoritative Sources | | | Contextual Relevance | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WGPT | CC | PAI | WGPT | CC | PAI | WGPT | CC | PAI | WGPT | CC | PAI |
| Is FWI a robust estimate of the fire danger risk? | × | ✓ | ✓ | NA | ✓ | ✓ | NA | ✓ | ✓ | ✓ | ✓ | ✓ |
| What universities would make good partners to undertake studies on fire behavior and effective mitigation strategies in Mora County, NM? | × | × | ✓ | NA | NA | ✓ | NA | NA | ✓ | ✓ | ✓ | ✓ |
| What are the wildfire mitigation strategies, ecological impacts, and economic effects in forested areas similar to Mount Bigelow, AZ? | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ×[1] | ✓ | × | ✓ |
| Please provide sources on how to obtain data for Vegetation Type and Density, Previous Burn History, Human Activity Levels, Topography, Weather Conditions, Land Management Practices, Infrastructure and Accessibility, Socio-Economic Factors, and Climate Change Indicators | × | ✓ | ✓ | NA | ✓ | ✓ | NA | ✓ | ✓ | ✓ | ✓ | ✓ |
| It looks like March is the time folks implement a controlled burn in Naperville, IL. What do you think? | ✓ | × | ✓ | ✓ | NA | ✓ | ✓ | NA | ✓ | ✓ | × | ✓ |
| Let's dive deeper into Firebreaks and Buffer Zones. I'm interested in learning about approaches to establish firebreaks around the critical infrastructure and vegetation management practices. Can you provide more detail with evidences? | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ |
| Search for scientific literature on effective post-wildfire recovery strategies, focusing on public safety and environmental restoration in Sangre De Cristo Mountains. | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ×[2] | ✓ | × | ✓ |
| Is fire common in this region prior to 1900s in Covington, VA? What is the natural fire regime? | ✓ | × | ✓ | ✓ | NA | ✓ | ✓ | NA | ✓ | ✓ | × | ✓ |
| Are there supporting literature for Buffer Zones that communities can implement to manage wildfire risk? | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ×[3] | ✓ | ✓ | ✓ |
| Are there good examples of similar cities to Las Vegas, NM that might experience similar shifts in wildfire risk? What mitigation strategies are those cities implementing? | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ×[4] | ✓ | × | ✓ |
| **Overall Success Rate** | **7/10** | **7/10** | **10/10** | **7/7** | **3/7** | **10/10** | **7/7** | **7/7** | **6/10** | **10/10** | **4/10** | **10/10** |

*WGPT* WildfireGPT, *CC* ChatClimate, *PAI* Perplexity AI.

Each row represents a specific prompt or task related to wildfire analysis. For each metric, a ✓ indicates that the tool met the criteria for that prompt, while a × indicates it did not. NA means the metric was not applicable or could not be evaluated, typically because of the previous criteria not being met; for example, when external knowledge is not retrieved, no citation would be provided. For ChatClimate, when it refers to studies outside of IPCC reports, it fails to provide the full title/link to the cited study. Contextual relevance means that the response is related to the location to the best specificity possible and tries to address the prompt. The main reason behind ChatClimate's failures is that it often references general trends in large countries without a clear clue as to whether the information could be relevant/ transferred to the location-specific queries in the prompt. The superscript numbers indicate the following issues with PAI.

[1]Used articles from non-profit organizations.

[2]Included articles from news media.

[3]Used articles from commercial sources.

[4]Referenced a Reddit post.

**Table 6 | Comparative analysis across three profile specificity tiers**

| Aspect | No Profile Information | + Location and Timeline | + Power-Grid Manager Profile |
|---|---|---|---|
| Vegetation Management | "Thinning and Pruning: Reduce forest density through thinning and pruning [...] Controlled Burns: Implement controlled burns to reduce available fuels." | "Develop and promote fire-adapted ecosystems by incorporating native species [...] Practice selective cutting to reduce fire risk." | "Establish routine schedules for pruning and removing trees that *pose a risk to power lines*. Utilize drones or satellite imagery to monitor vegetation growth for timely intervention." |
| Technology | "Employ technologies like satellite imaging and drones [...] Use artificial intelligence to predict fire patterns." | "Invest in enhanced surveillance systems [...] satellite imagery, drones, and remote sensors to detect early signs of fire." | "Utilize Geographic Information Systems (GIS) and remote sensing [...] Install sensors and advanced monitoring systems along the grid to detect temperature changes, smoke, or increased humidity levels indicating the risk of fire. Use drones for real-time surveillance [...] of *power lines, especially in hard-to-reach areas*." |
| Role-Specific Actions | "International cooperation [...] Reduce carbon footprints." | "Encourage Responsible Logging Practices [...] Climate Adaptation Strategies." | "*Expand the Right-of-Way (ROW) clearance around transmission lines* beyond standard regulations [...] Implement an IVM program that combines mechanical, biological, and chemical methods to manage vegetation" |

WildfireGPT responses demonstrate progressive domain adaptation from generic principles to stakeholder-specific protocols when provided with increasing user profile granularity. Representative outputs shown for vegetation management, technology implementation, and role-specific interventions demonstrate systematic enhancement of recommendation precision with profile enrichment. Italicized text highlights power-grid-specific recommendations emergent only with comprehensive user profiling. The full table can be found in the Supplementary Materials.

## Personalization

Our two-phase ablation study explores the impact of user profile granularity and diversity on WildfireGPT's recommendations. In the first phase, we progressively test three levels of user profile specificity provided to the agent: no profile information provided, location plus timeline, and the full profile of a power grid manager focused on maintaining transmission line clearance and grid resilience for power distribution reliability and access. In the second phase, we hold location and timeline constant and test five distinct professional profiles—homeowner, civil engineer, ecologist, emergency manager, and power grid manage—each with minimally distinct concerns and the respective scopes. We report the results from both phases below.

**Hierarchical specificity from profile granularity**. Table 6 demonstrates the tiered progression in recommendation specificity provided by WildfireGPT in the first phase of our ablation study. The table focuses on three key aspects of wildfire resilience—vegetation management, technology implementation, and role-specific action—and demonstrates how recommendations become increasingly targeted and context-aware as input specificity increases. For example, a generic prompt without profile information yields broad guidance such as thinning and pruning or the use of satellite imaging and drones. When location and timeline details are included, the system generates more ecologically informed and time-sensitive strategies, such as "develop and promote fire-adapted ecosystems by incorporating native species." With a complete power grid manager profile, the recommendations become more specific and operational, with recommendations prioritizing "sensors and advanced monitoring systems along the grid" and "expand the Right-of-Way (ROW) clearance around transmission lines," synthesizing ecological approaches with sector-specific operational requirements. Most notably, technological recommendations progressed from general "satellite imaging and drones" to grid-specific applications for "real-time surveillance of power lines, especially in hard-to-reach areas." More aspects are reported in the Supplementary Materials. This stratification demonstrates how increasing profile granularity introduced by the user profile agent enables WildfireGPT to align general resilience principles with the operational needs of distinct professional roles, particularly in critical infrastructure contexts where vegetation management intersects with operational reliability requirements.

**Professional identity as a recommendation filter**. For the second phase, we summarize WildfireGPT's adaptation to different professional roles and priorities across three dimensions: *plan generation*, *literature retrieval* and *recommendation specificity*. The full experimental results are included in the Supplementary Materias.

First, our analysis shows that the planning agent adapts to the user's professional role and concerns. For example, when a power grid manager expresses primary concern about maintaining transmission lines and grid resilience, WildfireGPT proposes to search for literature about vegetation management and wildfire protection strategies around power grids. Similarly, when a homeowner indicates a focus on marketable species, WildfireGPT targets research about economically valuable tree species and property protection methods. This pattern continues across all profiles—emergency managers receive evacuation-focused plans, civil engineers get infrastructure-centered recommendations, and ecologists receive biodiversity-oriented strategies. In each case WildfireGPT maintains the same basic framework (analyzing fire data, reviewing literature, developing recommendations) but tailors the specific content to match the user's professional needs and objectives. This alignment between user profiles and generated plans is by design, as the planning agent's primary function is to establish clear expectations and a structured pathway through which the analyst agent can deliver personalized insights relevant to each stakeholder's specific context.

Second, the analyst agent uses search queries calibrated to professional priorities when conducting literature retrieval, with details provided in Table 2. For each profession, the search queries effectively capture their unique concerns and scope homeowners' focus on marketable species, civil engineers' emphasis on structural resilience, ecologists' focus on ecological resilience, emergency managers' priority on defendable space, and power grid managers' concentration on grid resilience. These search queries distinctly different literature being retrieved, and we report the literature retrieval results based on these queries. When a power grid manager's profile emphasized transmission line clearance and grid resilience, the system yielded studies such as Hvenegaard's (2014)[66] analysis of fuel treatment monitoring and Dale's (2006)[67] critique of fire suppression policies. Conversely, when responding to a civil engineer's profile, the system retrieved Stevens et al.'s (2020)[68] findings on vegetation structure and soil moisture dynamics. This pattern of query-response alignment extends across all professional identities: emergency management queries yielded Ager et al.'s (2019)[69] work on community wildfire exposure, while the ecologist's profile resulted in Schoennagel et al.'s (2017)[70] research on adaptive resilience and Waltz et al.'s (2014)[71] findings on ecological impacts of fuel reduction treatments.

This adaptation pattern extends to recommendation development. As shown in Table 7, WildfireGPT generates distinct recommendation sets across the five professional profiles while maintaining core wildfire management principles. The table uses checkmarks to indicate which recommendations appear in each stakeholder's profile for example, all profiles receive recommendations for regular thinning

**Table 7 | Distribution of Recommendation Categories Across Five Stakeholder Profiles**

| Recommendation Category | Emergency Manager | Ecologist | Civil Engineer | Homeowner | Power Grid Manager |
|---|---|---|---|---|---|
| *Fuel Management* | | | | | |
| Regular thinning/pruning/fuel reduction | ✓ | ✓ | ✓ | ✓ | ✓ |
| Controlled/prescribed burns | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Infrastructure and Planning* | | | | | |
| Defendable space creation | ✓ | | | ✓ | ✓ |
| Fire-resistant materials/infrastructure | ✓ | ✓ | | ✓ | ✓ |
| Emergency access/water sources | ✓ | | ✓ | ✓ | ✓ |
| *Community Engagement* | | | | | |
| Evacuation planning | ✓ | ✓ | | ✓ | |
| Public education/awareness programs | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Collaboration and Management* | | | | | |
| Cross-boundary coordination | ✓ | ✓ | | | ✓ |
| Monitoring systems/technology | ✓ | ✓ | ✓ | ✓ | ✓ |
| Research | ✓ | ✓ | | | ✓ |
| *Ecological Considerations* | | | | | |
| Fire-resistant species promotion | | | ✓ | ✓ | |
| Natural fire regime maintenance | | ✓ | | | |
| Ecological monitoring | | ✓ | ✓ | | ✓ |

We use ✓ to indicate which recommendations appear when given each stakeholder's profile.

and prescribed burns, while only the ecologist receives natural fire regime maintenance recommendations. Although certain fundamental strategies such as regular thinning/pruning and controlled burns are universally recommended across all profiles, significant variations emerge in other categories based on professional priorities. For civil engineers, recommendations pivot toward drainage efficiency and slope stability considerations, incorporating Stevens et al.'s (2020)[68] findings on soil moisture dynamics. Notably, civil engineers are the only profession not recommended to implement fire-resistant materials/infrastructure, instead focusing on ecological monitoring and emergency access solutions. Emergency managers receive the most comprehensive recommendation set, with checks in nearly all categories and exclusive emphasis on defendable space creation and evacuation planning. Ecologists uniquely receive recommendations for natural fire regime maintenance, reflecting their biodiversity conservation priorities, while being the only group besides civil engineers not advised on defendable space creation. Emergency management profiles elicit recommendations heavily weighted toward evacuation planning and defendable space creation, drawing on Ager et al.'s (2019)[69] community exposure analyses. Homeowners and power grid managers show similar recommendation patterns in infrastructure planning but diverge significantly in community engagement homeowners receive evacuation planning advice while power grid managers do not. Interestingly, power grid managers share the research recommendation with emergency managers and ecologists, highlighting the technical expertise common to these roles.

This pattern of differential adaptation suggests that user profiles serve as effective filters for distilling broadly applicable wildfire management principles into professionally actionable, context-specific recommendations. The profile-specific recommendations reflect the granularity in both tactical precision and implementation feasibility, aligned with each stakeholder's operational priorities while maintaining scientific rigor in the underlying management principles.

Though the user profile agent help successfully retrieve thematically-relevant literature at a profession-specific level, the specificity of matches invites further refinement that goes beyond the user profile agent. First, when matching papers to user concerns, it might not align exactly with the specific scope. For example, when the power grid manager requested information about protecting power lines, the system identified Hvenegaard's (2014)[66] paper on fuel treatment. This is valuable for fire management, but does not specifically address power grid infrastructure. Second, regarding location-specificity, searches for Covington, Virginia yielded Schoennagel et al.'s work in western North America or Hansen's research in Grand Teton National Park[72] which could still offer applicable insights despite the geographic mismatch. These patterns reflect two areas for enhancement for the analyst agents: (i) the finite scope of the available literature database where closely aligned research may not exist for highly specialized user needs (such as power grid protection in specific geographic contexts); and (ii) inherent limitations in literature search mechanism as a part of the analyst agent. We explore these limitations and potential solutions in greater detail in the Discussion section.

### Domain expert evaluations: relevance, entailment, and accessibility of WildfireGPT's responses

**Relevance:** Domain experts generally found WildfireGPT's responses contextually appropriate (a success rate of 96.67% to 100% across all relevance categories, Table 8), but there are areas for improvement. For the "Last Question" and "Profession" categories (Table 3, Q1 & Q2), domain experts noted occasional misalignments, such as offering broad advice when specific zoning changes are requested or making impractical recommendations such as suggesting that homeowners host community workshops. The "Location" category revealed potential limitations in the literature corpus (Table 3, Q4), such as offering recommendations for homeowners in Virginia based on studies from the West Coast. WildfireGPT sometimes failed to specify which strategies could be reasonably implemented within given timeframes (Table 3, Q5).

**Entailment:** With an overall success rate of 92.86% (Table 9), WildfireGPT generally remained logically consistent; but experts noted some issues when WildfireGPT provided recommendations, such as a lack of transparency in tracing the sources of its recommendations and a failure to present recommendations in a logical hierarchy from most important to least important. These situations could lead to long-winded responses that lack focus.

**Table 8 | Relevance scores for each case study, representing the percentage of responses that appropriately addressed the user's context and needs**

| Case Study | Last Question | Profession | Concern | Location | Time | Scope |
|---|---|---|---|---|---|---|
| Climate Impact on U.S. Wildfires | 17/18 | 18/18 | 18/18 | 18/18 | 18/18 | 18/18 |
| Comprehensive Wildfire Impact | 6/6 | 7/7 | 7/7 | 7/7 | 7/7 | 7/7 |
| Current Wildfire Risk Analysis | 5/5 | 6/6 | 6/6 | 6/6 | 5.5/6 | 6/6 |
| Ecosystem Fire Management | 3/3 | 2.5/3 | 3/3 | 3/3 | 3/3 | 3/3 |
| Hazard Mitigation Planning | 5/5 | 7/7 | 7/7 | 7/7 | 7/7 | 7/7 |
| Infrastructure Wildfire Risk | 7/7 | 7/7 | 7/7 | 7/7 | 7/7 | 7/7 |
| Post-Wildfire Public Safety | 3/3 | 4/4 | 4/4 | 3.5/4 | 4/4 | 4/4 |
| Private Property Protection | 6.5/7 | 8/9 | 9/9 | 8/9 | 9/9 | 9/9 |
| Urban Wildfire Mitigation | 3.5/4 | 4/4 | 4/4 | 4/4 | 4/4 | 4/4 |
| Wildland Urban Interface Impact | 2/2 | 4/4 | 4/4 | 4/4 | 3.5/4 | 4/4 |
| **Average across All Percentages** | 97.48% | 97.22% | 100.00% | 97.64% | 97.92% | 100.00% |
| **Overall Success Rate** | 96.67% (58/60) | 97.83% (67.5/69) | 100.00% (69/69) | 97.83% (67.5/69) | 98.55% (68/69) | 100.00% (69/69) |

Raw scores are shown (score/total times the domain expert opinion was collected to evaluate a response).

**Table 9 | Entailment and accessibility scores for each case study**

| Case Study | Entailment | Accessibility | | |
|---|---|---|---|---|
| | | No Jargon | Enough Explanation | No Redundancy |
| Climate Impact on U.S. Wildfires | 11/13 | 17/17 | 16/17 | 14.5/17 |
| Comprehensive Wildfire Impact | 7/7 | 7/7 | 7/7 | 7/7 |
| Current Wildfire Risk Analysis | 5.5/6 | 6/6 | 5.5/6 | 6/6 |
| Ecosystem Fire Management | 3/3 | 3/3 | 3/3 | 3/3 |
| Hazard Mitigation Planning | 6/6 | 7/7 | 7/7 | 7/7 |
| Infrastructure Wildfire Risk | 6.5/7 | 7/7 | 7/7 | 7/7 |
| Post-Wildfire Public Safety | 3/3 | 4/4 | 4/4 | 4/4 |
| Private Property Protection | 6/6 | 8/8 | 7/8 | 7/8 |
| Urban Wildfire Mitigation | 1.5/2 | 4.5/5 | 2.5/5 | 5/5 |
| Wildland-Urban Interface Impact | 2.5/3 | 4/4 | 4/4 | 4/4 |
| **Average Across All Percentages** | 92.75% | 99.00% | 92.33% | 97.28% |
| **Overall Success Rate** | 92.86% (52/56) | 99.26% (67.5/68) | 92.65% (63/68) | 94.85% (64.5/68) |

Entailment indicates the percentage of responses where the model's analyses and recommendations logically followed from the provided information (data, literature). Accessibility assesses the clarity and concision of the model's language. Scores represent the percentage of responses that avoided jargon, provided sufficient explanation, and minimized redundancy. Raw scores are shown (score/total times the domain expert opinion was collected to evaluate a response).

**Table 10 | Agreement between human evaluation and LLM-as-a-Judge evaluation of WildfireGPT responses**

| Category | Agree | Disagree | Yes vs Could be better |
|---|---|---|---|
| Relevance | 97/154 (62.99%) | 57/154 (37.01%) | 51/57 (89.47%) |
| Entailment | 15/20 (75.00%) | 5/20 (25.00%) | 2/5 (40.00%) |
| Accessibility | 56/84 (66.67%) | 28/84 (33.33%) | 22/28 (78.57%) |

Most misalignments occur when domain experts respond with "Yes" while GPT-4 tends to give a more critical judgment, saying "Could be better" instead of a larger disagreement like "Yes" vs "No."

**Accessibility:** WildfireGPT generally performed well in avoiding jargon (99.26% overall success rate) and redundancy (92.65% overall success rate), with most responses providing sufficient explanation (94.85% overall success rate); see Table 9. However, the interpretation of these scores varied by audience. Urban planners and homeowners frequently requested more detailed explanations and examples as well as clearer definitions of technical terms and contextual information for data interpretation. In contrast, climate scientists preferred more concise responses, with less detail and more direct answers to their questions. These findings point to the challenge of balancing detail and concision to meet diverse user needs.

Overall, the domain expert evaluation demonstrates that WildfireGPT is capable of generating relevant, logically sound, and accessible responses to user queries across a range of wildfire-related scenarios. However, the results also highlight potential areas for improvement, such as improving the relevance of responses to the user's last question in certain cases and ensuring differing levels of explanation to maximize clarity and utility for different domain experts.

**Domain expert evaluations: qualitative feedback on WildfireGPT's recommendations.** To better understand the nuance and practical value of WildfireGPT's recommendations, we sought qualitative feedback from domain experts as they evaluated the system's responses via the structured questionnaire reported above.

**Fig. 3 | Ecosystem fire management.** WildfireGPT demonstrates its ability to integrate data analysis and domain knowledge to provide actionable recommendations for ecosystem fire management. By analyzing wildfire incident data, retrieving relevant literature, and examining seasonal FWI trends, WildfireGPT suggests optimizing the timing of controlled burns to minimize risks while maintaining oak ecosystem health. The domain expert's positive feedback highlights WildfireGPT's nuanced approach and its potential to support informed decision-making in wildfire management.

Case Study: Ecosystem Fire Management

**Context:** WildfireGPT analyzed recent wildfire incident data and found that from 2018 to 2021, the months with recorded incidents were March (2 incidents) and July (1 incident), suggesting that March may be when controlled burns are typically implemented. To confirm this point, WildfireGPT retrieved a study by Stan, Rigg, and Jones (2006) [73], which highlighted the importance of reintroducing fire in March for white oak regeneration in northeastern Illinois. Additionally, WildfireGPT analyzed seasonal FWI data, identifying lower-risk periods during spring.

**User Query:** Please suggest specific adaptations to fire management strategies (to balance the utilization of fire for managing oak ecosystems and reducing the risk to residential and human developments.)

**WildfireGPT Response (Snapshot):**

*Optimized Timing for Controlled Burns*:

- Continue to utilize controlled burns in March when the FWI (Fire Weather Index) historically shows lower values, minimizing the risk of uncontrollable wildland fires.
- Schedule additional controlled burns during early spring or late autumn to take advantage of similarly low FWI scores, ensuring burns are conducted under optimal conditions for controlling spread and smoke dispersal.

**Expert Comment:** The answer effectively addresses a key concern: whether to continue conducting controlled burns in March. It confirms that March remains a suitable time for burns due to historically lower Fire Weather Index (FWI) values, which reduce the risk of uncontrolled fires. The response also provides a nuanced suggestion by recommending additional burn periods in early spring and late autumn. This approach takes advantage of similarly low FWI scores during these times, allowing for burns at both the beginning and end of the dormant season. This strategy demonstrates a nuanced understanding of fire management, as it balances the need for regular burns with the importance of conducting them under optimal weather conditions for safety and control. The suggestion to utilize multiple burn windows throughout the year shows a thoughtful approach to ecosystem management while minimizing risks to nearby developments.

One notable example of WildfireGPT's effectiveness in integrating data analysis with domain-specific knowledge is demonstrated in the Ecosystem Fire Management case study (Fig. 3). When asked for suggestions on adapting fire management strategies to maintain oak ecosystems while minimizing risks to residential areas, WildfireGPT analyzed recent wildfire data, revealing that controlled burns typically occur in March. It validated this response by citing literature such as that of Stan, Rigg, and Jones (2006)[73], which emphasizes the importance of March fires for white oak regeneration. WildfireGPT also analyzed seasonal FWI data, identifying lower-risk periods in early spring and late autumn. Synthesizing these insights from more than one data source, it recommended timing controlled burns to coincide with historically low FWI values, balancing ecological management goals with safety. The domain expert praised WildfireGPT's nuanced approach, emphasizing its ability to balance ecological management goals with safety considerations.

The Wildland Urban Interface Impact case study (Fig. 4) further demonstrates WildfireGPT's practical utility. When the domain expert asked questions from the perspective of a risk manager from Las Vegas, NM, WildfireGPT provided recommendations on water resource protection that aligns well with actual challenges faced in the field. Furthermore, it identifies California as a relevant comparison due to similar fire characteristics. The retrieved study by Moritz and Stephens (2008)[74] recommends strategies such as risk-based frameworks, controlled burns, and reevaluating urban planning—approaches the domain expert recognizes as potentially valuable for Las Vegas, NM.

Overall, the qualitative feedback from domain experts underscores WildfireGPT's ability to provide nuanced recommendations that align with real-world challenges and existing research. We also provide more examples of qualitative feedback in the Supplementary Materials.

**Scaling up evaluation by LLM-as-a-Judge.** To evaluate GPT-4's[40,50] feasibility as an automated judge for WildfireGPT, we compared GPT-4 assessments with human judgments on relevance, entailment, and accessibility. We applied LLM-as-a-judge to evaluate WildfireGPT's responses only immediately after data or literature retrieval to manage context length and maintain performance (Table 10). Agreement rates varied: relevance (62.99%), entailment (75.00%), and accessibility (66.67%). Disagreements primarily occurred when experts rated responses positively ("Yes"), whereas GPT-4 suggested improvements ("Could be Better"), reflecting GPT-4's stricter evaluative criteria.

The moderate 62.99% agreement in WildfireGPT relevance evaluations is largely due to nuanced differences rather than fundamental disagreements, often due to the data availability issues since retrieved data may not exactly match user-specified timelines or locations, whereas experts were more lenient, acknowledging these constraints.

In most entailment disagreements, domain experts found responses satisfactory ("Yes"), while GPT-4 flagged them as needing improvement ("Could be better" or "No"). These discrepancies often stemmed from WildfireGPT extrapolating beyond available data. For example, based on FWI data showing minimal change over time, WildfireGPT described wildfire risk as "relatively stable"; GPT-4 criticized this, arguing that it can be misleading and that there are possibilities for wildfire risk to increase. In another case, when asked about fire controls for cherry trees, the retrieved literature did not directly address this specific species but rather the same family of trees in general; experts found this reasonable given the limitations of the available literature, but GPT-4 criticized it as unsupported. Experts appeared more forgiving of such extrapolations, likely because of the experts filling in gaps with their own background knowledge.

Accessibility disagreements highlighted differing standards: GPT-4 recommended simpler wording or more detailed explanations. However, experts' satisfaction varied with their specific expertise and familiarity with the topic. For example, In areas of deep knowledge they may prefer concise, technical responses, while in less familiar areas they require more detailed explanations. GPT-4's critiques do not fully account for this spectrum of expertise.

In summary, the results reveal moderate agreement between GPT-4 and human evaluations. While GPT-4 can provide scalable and cost-effective evaluations, it may not fully capture the nuanced expectations and information needs of domain experts, leading to discrepancies in relevance, entailment, and accessibility assessments.

## Discussion

This section highlights the core contributions and insights derived from the design and evaluation of WildfireGPT. It focuses on the system's strengths, particularly its human-centered personalization capabilities, and evaluates how integrating user context enhances the relevance and utility of LLM-generated recommendations. Additionally, this section reflects on the broader implications for domain-specific LLM adaptation and outlines promising directions for future research in AI-assisted climate hazard decision-making.

**Fig. 4 | Wildland urban interface impact.** WildfireGPT demonstrates its effectiveness in addressing the complex challenges posed by the wildland-urban interface. Wildfire's recommendations regarding water quality protection align well with the actual challenges faced by domain experts in the field. WildfireGPT's ability to identify relevant case studies from other areas with similar shifts in wildfire risk and to highlight pertinent mitigation strategies showcases its potential to support informed decision-making in wildfire risk management and urban planning.

Case Study: Wildland-Urban Interface Impact

**Context:** In Las Vegas, NM, a risk manager developing urban plans for 2050 is concerned about the increasing wildfire risks and their impact on water resources. Previous analysis of Fire Weather Index (FWI) data indicated an increasing trend in fire danger levels, particularly during the spring season.
**User Query:** Proceed with the recommendations (on urban planning strategies to minimize wildfire risks and methods to protect water resources from the adverse effects of wildfires)
**WildfireGPT Response (Snapshot):**
*Water Resource Management*

- Water Supply Resilience: Increase the resilience of water supply systems to ensure availability during and after wildfire events. This could involve constructing new reservoirs, enhancing existing water treatment facilities, and establishing emergency water supply locations.
- Protect Water Quality: Develop strategies to protect water quality from the effects of wildfires, such as increased sedimentation and contamination. These might include erosion control measures and enhanced filtration systems at water treatment facilities.

**Expert Comment:** The recommendation to protect water quality from increased sedimentation and contamination is highly relevant. This issue was a central concern when we initially developed this case study, and it remains a significant challenge for the area. The local water treatment plant is struggling to handle the fine-grained sediment that is now flowing downstream in much larger quantities than before. The plant's processes are overwhelmed by the volume of material, which far exceeds what they typically dealt with from normal erosion. It's very good that the answer highlights the need to address water quality issues, specifically mentioning sedimentation and contamination. The suggestion to implement erosion control measures and enhance filtration systems directly addresses the problems we encountered. In our modeling work, we aimed to predict the sediment load to help design appropriate mitigation measures. The recommendation aligns well with the actual challenges faced in the field and the solutions we were exploring.
**User Query:** Are there good examples of similar cities that might experience similar shifts in wildfire risk? What mitigation strategies are those cities implementing?
**WildfireGPT Response (Snapshot):**
Here's a look at some of the approaches highlighted in recent research:[...] California is a prime example of a region adapting to increased wildfire risks. The study by Moritz and Stephens (2008) emphasizes the need for a sustainable coexistence with wildfire through policy, planning, and management changes. Key strategies include the following:

- Risk-Based Framework: Implementing a risk-based framework for fire-related decisions to address natural hazards systematically.
- Reintroduction of Fire: Using controlled burns to reintroduce fire in a managed way to fire-prone ecosystems.
- Urban Planning Reevaluation: Reevaluating urban planning and building in fire-prone locations to make future cities less susceptible to wildfire damage.

*References*:
Moritz, M., & Stephens, S. (2008). Fire and sustainability: considerations for California's altered future climate. https://doi.org/10.1007/s10584-007-9361-1
**Expert Comment:** The WildfireGPT response effectively identifies California as a region experiencing similar shifts in wildfire risk to Las Vegas, NM. Despite some differences in fire characteristics, California remains a relevant comparison. The strategies highlighted, such as implementing a risk-based framework, reintroducing controlled burns, and reevaluating urban planning, could provide valuable guidance for Las Vegas.

## Human-centered personalization: a key strength of WildfireGPT

Prior efforts to leverage LLMs for climate action have primarily focused on generating factual responses to generic, one-off questions[26,30–32,35]. However, this approach can fail to capture the nuanced expertise required in complex domains, since LLMs tend to provide hasty, generic responses without considering the user's specific context and needs[36,37,75,76].

In contrast, WildfireGPT's key strength lies in its ability to deliver personalized, context-aware recommendations. By integrating user-specific information through its user profile agent and generating transparent action plans through its planning agent, WildfireGPT tailors its responses to the user's background, location, and specific concerns about natural hazards. This personalization capability was systematically validated through our ablation study demonstrating how the same query ("Develop recommendations to enhance forest resilience against wildfires") yields distinctly different literature retrieval strategies and recommendations when filtered through various professional lenses. The effectiveness of this approach is quantitatively confirmed by the high relevance and accessibility scores reported, where domain experts consistently rated the system's recommendations as highly pertinent, not only to their questions, but also to the contexts of their concerns in the natural hazard domain. The alignment between user needs and system outputs is a critical advantage of WildfireGPT, ensuring that the information provided is not only scientifically accurate but also practically useful for professionals working in wildfire management.

Future research should build on the approach demonstrated by WildfireGPT, exploring methods for context-aware domain adaptation of LLMs. By developing techniques to efficiently integrate domain-specific knowledge and problem context into LLM training and inference, researchers can create AI systems that can bridge the critical gap between complex natural hazard information and actionable, site-specific strategies for a wide range of related challenges[11,77].

## Transforming localized data into actionable insights

Natural-hazard-informed decision-making requires the integration of multiple location-specific data sources and scientific findings to capture its multifaceted nature[78–81]. However, navigating and synthesizing this information can be cumbersome for professionals; using Geographic Information Systems software is time-consuming[82], writing custom data analysis code requires specialized skills not always available in these roles, and science literature is complex to comprehend[11]. Prior attempts to integrate LLMs for this task have either limited themselves to scientific reports[22,23,26,30,31], which can lack local context, or focused solely on data analysis[32,38], missing the transfer of scientific insights for recommendation development.

Unlike baseline alternatives such as ChatClimate and Perplexity AI, which demonstrated significant limitations in geographical precision and data accuracy, respectively, WildfireGPT addresses these challenges by integrating location-specific projections and observational data, socioeconomic indicators, and scientific literature from trusted sources to deliver an extensive analysis of wildfire risks. WildfireGPT consistently outperforms alternatives in data provision, location specificity, and data accuracy, while maintaining high contextual relevance through literature that shares ecological or geographical characteristics with specified locations. When confronted with data limitations, WildfireGPT demonstrates methodological adaptation by explicitly acknowledging constraints and proposing alternative analytical resources rather than defaulting to geographical generalizations. A prime example of WildfireGPT's capabilities is shown in Fig. 3, where WildfireGPT synthesizes insights from analyzing fire weather indices, historical wildfire data, and scientific literature to provide suggestions on the timing for controlled burns to mitigate wildfire risks.

As natural hazards intensify, the need for tools that can rapidly synthesize complex data into actionable strategies becomes increasingly critical[11,78,83]. The framework and design principles behind WildfireGPT are general and readily adaptable to more data sources. For example, future development could explore the integration of real-time sensor data and community-reported information, creating a more dynamic, responsive system for wildfire risk assessment[81,84]. Moreover, more extensive data integration could enable WildfireGPT to provide insights into other climate hazards, such as heatwaves[7] and floods[8]. Furthermore, as the integration of data sources grows more complex, going beyond text-only models to incorporate multimodality and code generation for more nuanced analysis and flexibility could be a promising direction for future research[85–87].

### Systematic evaluation of domain-specific conversational tools

Evaluating conversational systems in the natural hazard domain presents unique challenges beyond conventional lexical similarity metrics such as BLEU and ROUGE, which correlate poorly with expert assessment of domain-specific content generation. Moreover, while existing benchmarking frameworks with well-defined, one-off questions can provide a starting point for assessing AI-generated responses[20,26,35], they fail to capture the nuanced, multifaceted nature of real-world decision-making. Our case studies highlight that effective climate decision-making is inherently interactive, requiring systems to engage users in ongoing dialogue while synthesizing information from diverse data sources and literature in a personalized manner tailored to their specific professional context and needs.

To address these challenges, we developed a three-stage evaluation framework that prioritizes expert judgment over purely automated metrics. First, we conducted a comparative evaluation against ChatClimate and Perplexity AI, assessing data retrieval and evidence-based response capabilities. Second, we performed an ablation study examining how different levels of user profile specificity impact response relevance and actionability across five distinct professional profiles. Third, we evaluated actual responses through domain expert assessment using standardized criteria for relevance, entailment, and accessibility. This comprehensive approach captures the practical applicability and ease of understanding of the information provided. Future research can build on this framework and explore additional evaluation metrics that capture the real-world utility of LLMs in natural hazard decision-making scenarios. In our case studies, domain experts noted instances where WildfireGPT's recommendations aligned with their existing knowledge. However, quantifying this alignment and its impact on decision-making remains an open research question. Going beyond this, future metrics can assess the system's ability to accelerate decision-making processes, improve decision quality, or introduce users to new resources and insights.

Another critical direction is to develop more sophisticated automated evaluation techniques to assess the utility of AI tools in natural hazard decision-making at scale. Our preliminary efforts to incorporate LLM-as-a-judge highlight the potential of using AI systems to evaluate AI-generated responses. However, the moderate level of agreement between LLM-as-a-judge and human evaluators suggests that LLM-as-a-judge currently struggles to capture the nuanced understanding of data limitations and context-specific, utility-driven considerations that human evaluators bring to the table. A promising approach is prediction-powered inference[88], which enables valid statistical inference by supplementing small-scale expert-labeled datasets with machine learning predictions without making assumptions about the underlying algorithms. This framework could quantify uncertainty in model-based evaluations and provide confidence intervals that narrow with more accurate predictions, potentially enabling more efficient and rigorous assessment of LLM-based judges against limited human evaluation data. This technique has particular relevance for LLM evaluation, as demonstrated in recent work that constructs statistically valid rank sets from a combination of human and LLM-provided pairwise comparisons[89]. This approach could be particularly valuable for evaluating WildfireGPT, allowing us to make reliable conclusions about system performance by combining limited expert feedback with more abundant GPT-based assessments. Future research may explore approaches such as fine-tuning on domain-specific data[90], leveraging reinforcement learning techniques[91], and employing more sophisticated prompt engineering[92,93].

### Opportunities for Improvement

**Enhancing semantic and spatial information retrieval through integrated knowledge frameworks**. To deliver precise, location-aware results, we observe three areas for improvement in semantic processing and geographic contextualization. First, as explained in the Methodology section, we apply semantic search using BERT models to match the semantic similarity between the search query for literature and the abstract of each paper in the database[94] However, in semantic processing, the reliance on general-purpose language models such as BERT leads to mismatches between user queries and retrieved content for example, suggesting fuel treatment studies for power grid protection inquiries due to insufficient domain-specific nuance. Implementing a hybrid hierarchical retrieval framework would help address these limitations by interleaving results from sparse retrievers (capable of processing lengthy documents without structural metadata) with dense retrievers (optimized for semantic similarity)[95]. Ontology-enhanced semantic search frameworks[96] can also help formalize relationships between concepts and spatial entities, enabling precise differentiation between management strategies with varying degrees of regional appropriateness. Second, geographic contextualization failures arise when the system retrieves studies from ecologically dissimilar regions (e.g., applying West Coast strategies to Virginia) or overlooks location-specific data gaps (e.g., missing RCP 8.5 projections for Denver). Constructing and integrating spatially informed knowledge graphs, such as KnowWhereGraph[97], with its 29 billion RDF triples spanning environmental, hazard, and demographic domains, would help enable explicit spatial reasoning and improve the final response. Third, the limited scope of the underlying scientific literature database restricts the system's ability to retrieve relevant studies even when semantic and geographic logic are sound. Moreover, it is worth noting that abstracts might not capture the full details of papers, potentially causing relevant content to be overlooked. While improving search methods helps maximize existing resources, systematically expanding the knowledge base remains essential for comprehensive coverage. This expansion should particularly target underrepresented regions and niche professional needs where current gaps are most acute.

**Developing adaptive communication frameworks for diverse stakeholders**. The significant variation in information requirements across stakeholder groups from the technical precision demanded by climate scientists to the contextual clarity needed by urban planners underscores the necessity for adaptive communication frameworks in wildfire decision-support systems. WildfireGPT already collects critical user context at the outset of interactions, including profession, primary concerns, and scope of inquiry. Leveraging this metadata to dynamically tailor responses such as adjustment of technical depth, prioritization of recommendations, or presentation format remains an area for further development. The Human-Robot Teaming Framework with Multi-Modal Language Feedback (HRT-ML) proposed by Liu et al.[98] offers a promising architectural template for this enhancement. Specifically, its dual-module structure a coordinator for high-level user intent inference (e.g., inferring that a civil engineer needs slope stability insights rather than general recommendations) and a manager for task-specific adaptation (e.g., prioritizing recommendations in improving slope stability) could be adapted to WildfireGPT's workflow. By explicitly mapping user-provided context (profession, location, timeframe) to HRT-ML's tiered support levels, WildfireGPT could proactively modulate responses: for example, suppressing tangential literature for time-constrained emergency responders or augmenting explanations with local regulatory precedents for planners. This promising approach could address current gaps in personalization, such as overly generic advice for homeowners or insufficient technical rigor for scientists, while maintaining the system's core strengths in data fidelity and location specificity.

**Limitations of GPT-4 in WildfireGPT**. Our case studies revealed two significant limitations of GPT-4 that can negatively impact user experience: hallucinations in the form of typos and distractions. Typos can lead to misspelled words or functions, resulting in invalid function calls. These can cause unsuccessful module transitions or data retrieval failures. Distractions can also cause failed module transitions. For example, in the Community Hazard Mitigation Planning case study, WildfireGPT failed to transition from the user profile module to the planning module after all questions were answered because it hallucinated that the conversation was complete. These limitations often require manual intervention from the facilitator to correct the issues and move the conversation forward. We note that these issues are not unique to WildfireGPT[99,100], and future improvements should focus on enhancing the language model to mitigate hallucinations. Another challenge is the tendency of GPT-4 to generate lengthy and generic outputs, potentially overwhelming users and obscuring key information. This may be attributed to current Reinforcement Learning from Human Feedback (RLHF) techniques that encourage longer outputs[101] and the lack of domain expert involvement in the human labeling component of RLHF. The absence of specialized professionals in the feedback loop can lead to responses that fail to capture the nuanced expertise required in complex domains like wildfire management.

## Methodology

An overview of the WildfireGPT system is shown in Fig. 1. In this section we provide further details of the multi-agent setup within WildfireGPT and the evaluation framework. We developed a multi-agent RAG[102] system as a collaborative framework designed to streamline complex decision-making tasks by leveraging specialized agents for different stages of the workflow. Each agent in the system has a distinct role. For instance, the user profile agent collects user inputs and refines the task scope, the planning agent formulates actionable plans based on user-defined objectives, and the analyst agent conducts in-depth analyses to provide insights. These agents work collaboratively under the coordination of the task orchestrator agent, which dynamically routes interactions to the appropriate agent in sequence. In this framework the orchestrator does not act as a decision-maker; rather, it facilitates the flow of information by passing outputs from one agent to the next. Additionally, it supports resuming conversations by replicating the thread history into a new thread, facilitating continuity in long-running or multisession interactions. This modular and scalable design allows the system to efficiently handle diverse and evolving tasks while ensuring seamless transitions between agents.

### Task orchestrator

We designed the task orchestrator agent to serve as the central coordinator in the multi-agent system, ensuring seamless interaction between the user and the specialized agents for efficient task execution. The orchestrator includes key functions to enable its role through the OpenAI function calling API. The `get_response` function processes user inputs, invokes the appropriate agent based on the context, and manages outputs or follow-up actions to keep the interaction smooth and efficient. The `update_assistant` function allows the orchestrator to dynamically transition between agents, ensuring that the most relevant agent is assigned to handle the current task based on user input and workflow requirements. The `resume_conversation` function is designed to restore context in cases of interrupted or paused conversations by replicating the thread's messages into a new thread, allowing for seamless continuation. Together, these functions are designed to empower the task orchestrator agent to manage workflows, prioritize user-centric engagement, and deliver a cohesive experience across various stages of decision-making and analysis.

### User profile agent

The user profile agent is designed to collect and refine user-specific inputs for wildfire risk assessment through a structured, interactive, and iterative approach. This agent helps WildfireGPT produce outputs tailored to the unique needs of each user in a later stage of the interaction. To achieve this, we employed prompting techniques to guide users through individualized questions presented one at a time, enabling a systematic completion of a checklist covering personalized information about the user. Once all questions are answered, the system compiles the responses into a structured checklist and presents a summary to the user for verification, ensuring both accuracy and alignment with the user's expectations. The design also incorporates flexibility, allowing users to respond with "I don't know" if they are uncertain about any question, accommodating varying levels of expertise and familiarity with wildfire risk topics. The agent operates in two distinct stages, predefined inquiry and summary verification, working together to provide understanding of the user's background and concerns.

The first stage of the user profile agent in WildfireGPT involves a structured dialogue designed to gather essential information to prepare for generating actionable wildfire risk insights later. This stage begins with five predefined questions that capture critical user details: profession, concern, location, time, and scope.

- **Profession:** The profession question collects the user's professional background and expertise. This information will act as a proxy for WildfireGPT to tailor its recommendations and technical depth to the user's knowledge level, whether engineer, urban planner, emergency manager, or infrastructure operator.
- **Concern:** The concern question identifies the user's key motivations or queries surrounding wildfire risks, such as community safety, infrastructure resilience, or ecological preservation. This information will be used to help the system target the relevant aspects of wildfire management to the user.
- **Location:** The location of interest question identifies the geographic area for wildfire risk assessment, allowing users to input locations as place names, geographic coordinates, or descriptive details. When a user specifies a location in natural language, the system is designed to leverage the LLM to convert the input into geographic coordinates (latitude and longitude). To ensure accuracy, we integrated a map-based verification system that enables users to visually confirm or refine their input. The specified coordinate is then saved to help retrieve structured data and enable robust and geographically grounded wildfire risk assessments.
- **Timeframe:** The timeframe question defines the temporal scope of interest, offering options for short-term (1–10 years) mitigation strategies, medium-term (10–30 years) resilience planning, or long-term (30–80+ years) resilience adaptation. Historical analysis is also available, including recent (1–10 years), past (10–50 years), or long-term (50+ years) fire patterns.
- **Scope:** The scope of interest question refines the focus of the risk assessment, covering areas such as infrastructure vulnerability, emergency preparedness, ecological impacts, and insurance planning, with guided examples provided for users who are unsure. After gathering responses, WildfireGPT compiles the information into a structured checklist, which is shared with the user for review to ensure accuracy and alignment with their expectations.

Upon completing and confirming the finalized checklist, the agent saves the information and transitions to the planning agent, where WildfireGPT begins generating actionable insights and recommendations. This transition marks the shift from information gathering to solution generation, ensuring that the outputs are tailored to the user's specific needs and the context of their wildfire risk assessment.

### Planning agent

We designed the planning agent prompt in WildfireGPT to guide the system in creating a structured and user-centered wildfire risk assessment plan. By embedding step-by-step instructions into the prompt, we enable the planning agent to systematically address user concerns, integrate user feedback, and leverage relevant datasets effectively. The prompt for the planning agent comprises the following components:

- Step-by-Step Engagement Plan: The prompt instructs the system to create a short, clear plan to engage with the user effectively. This plan includes leveraging the most relevant datasets to address the user's concerns through data analysis and a literature search. The system is guided to select datasets based on the user's specific needs such as analyzing trends in wildfire risk, understanding historical patterns, or focusing on immediate mitigation strategies.
- Feedback Integration: Once the plan is drafted, the system is prompted to pause and share the plan with the user. This ensures that the user is fully informed about the proposed approach and has the opportunity to provide feedback or request modifications. The prompt explicitly instructs the system to ask whether the user would like to include additional information or datasets, while also clarifying that the wildfire analysis is limited to the three available datasets: FWI, long-term fire history records, and recent fire incident data. This step ensures transparency and active collaboration between the system and the user.
- Finalization and Execution: After incorporating user feedback and securing the user's agreement, the system transitions to finalizing the plan. The prompt ends with a specific instruction to call the function `plan_complete()` with the completed plan, signaling the readiness to move to the execution phase.

The planning prompt also leverages one-shot learning by including an example plan that act as a demonstration, guiding the model to understand the desired structure and format of the output. In addition to the example, the prompt includes detailed dataset descriptions, outlining the characteristics and scope of the three available wildfire data sources to constrain the model's data selection. The prompt also integrates the user profile for the agent to make informed decisions.

During the user interaction, WildfireGPT shares brief descriptions of the available datasets, before presenting the proposed plan. From the user's perspective, this plan comprises three key steps: (1) data retrieval: WildfireGPT identifies the most relevant wildfire datasets from the three available sources and explains the reasoning behind the selection; (2) literature review: WildfireGPT outlines the thematic focus of the literature to be examined; and (3) recommendation development: based on the data analysis and literature review, WildfireGPT develops personalized recommendations.

### Analyst agent

The analyst agent leverages the RAG framework[103,104] to conduct analyses by integrating diverse data sources and domain knowledge. This agent maintains the summarized user profile and action plan within its prompt context to ensure that WildfireGPT adheres to the plan and addresses user follow-ups, offering to explore specific topics or proceed to the next step. The analyst agent functions as an information processor, ensuring that outputs are grounded in scientific research articles and wildfire data. Its structured approach enables it to process volumes of information efficiently while maintaining accuracy and relevance. The analysis consists of three main stages:

- *Wildfire data retrieval and visualization*: The data retrieval and visualization functionality in our system is designed to assess wildfire risk by leveraging geospatial data and fire trends. The process begins with location-based data retrieval, based on the location information collected by the user profile agent. For all datasets, data within a 36-kilometer radius of the specified coordinate is retrieved and presented. The results are visualized through interactive maps, tables, and statistical plots. The map-based visualization employs GeoJsonLayer rendering to display the data points. A pin-layer visualization is also integrated to mark the exact user-specified location. To facilitate decision-making, the system dynamically generates a structured report summarizing the statistics. *Recent Wildfire Incident Record Data:* Historical wildfire locations (2015–2023) are visualized as individual red markers on the GeoJsonLayer, with each incident precisely geolocated within a 36-kilometer analysis radius. In addition, two complementary line plots are presented for temporal analysis: an annual frequency plot revealing incident patterns over the years and a monthly distribution highlighting seasonal variations. The report contextualizes total incident frequency per year and aggregated monthly distributions into risk assessments, analyzing temporal clustering patterns to identify periods of elevated fire activity. *Tree-Ring and Sediment-Based Fire History Record Data:* The system retrieves and ranks the three nearest paleofire study locations using a geodesic distance algorithm, integrating site-specific metadata from the International Multiproxy Paleofire Database. Each identified site is presented as an individual red marker on the GeoJsonLayer. Additionally, for each site, the system provides the user with the research metadata including site names, precise coordinates, and associated publication records. *Fire Weather Index Data:* Unlike the other data sources, the FWI data is structured on a grid system and referenced by using Crossmodel indices. Thus, we developed a process to map the user-specified geographic coordinates to the corresponding Crossmodel reference. This process begins by transforming the latitude and longitude into a spatial reference system compatible with the grid model. The geographic coordinates, initially expressed in degrees, are converted into the coordinate reference system used by the FWI data, typically employing projected coordinates for spatial accuracy. Using spatial operations such as buffering and intersecting, the transformed geographic point is compared against the grid cells in the model. The system retrieves the associated Crossmodel indices if they fall within a radius of 36 kilometers of the specified coordinate. This mapping establishes a direct connection between the user-defined location and the corresponding grid cell in the FWI data. By designing this mapping mechanism, we ensured that WildfireGPT can seamlessly integrate user-specified locations with grid-based FWI data, providing precise and scientifically accurate insights tailored to the specified area. The retrieval process integrates historical (1995–2004), mid-century (2045–2054), and end-of-century (2085–2094) projections to offer a temporal perspective on fire weather trends. The retrieved FWI values are aggregated and analyzed to compute mean wildfire indices and standard deviations across seasons, ensuring a statistically robust representation of fire risk. The system then categorizes FWI values into six risk classes (low, medium, high, very high, extreme, and very extreme) based on the Canadian Forest Fire Weather Index classification system[105]. The GeoJsonLayer displays fire weather indices across spatial grids, with each grid cell color-coded according to its FWI classification. Additionally, the module provides a table presenting the seasonal comparisons of historical, mid-century, and end-of-century projections, enabling users to track long-term fire risk evolution. Lastly, the module is prompted to provide a structured summary report detailing past and projected wildfire risks in the region. *Census Data for Socioeconomic Analysis:* If location-specific wildfire data is retrieved, users are informed of an additional analysis capability: demographic and socioeconomic impact assessments from census datasets. This can help provide an understanding of wildfire risks beyond environmental factors, by incorporating population vulnerabilities, economic conditions, and housing density. If wildfire data is not available, the system shares a prewritten cautionary message, emphasizing the preliminary nature of its recommendations and encouraging users to seek further investigation or expert advice before implementing significant changes. The census data retrieval functionality extracts demographic and socioeconomic statistics for regions surrounding a user-specified location, with each census block group data retrieved from the American Community Survey (ACS5) dataset. Key indicators such as total population, poverty distribution, and housing unit counts are aggregated to generate a demographic summary, highlighting the number of individuals below the poverty line and those with income less than half the poverty threshold. Each census block group is visualized on the GeoJsonLayer for interactive spatial understanding. Additionally, a structured table presents aggregated population and housing statistics, allowing users to explore socioeconomic factors in detail. By integrating real-time geospatial data analysis, interactive visualization, and RAG techniques, WildfireGPT ensures that wildfire risk assessments are data-driven, context-aware, and easily interpretable for end users.

- *Literature retrieval*: For scientific literature, WildfireGPT queries a corpus developed by Argonne National Laboratory for the CIACC tool[59,60]. All abstracts of the papers in this corpus are pre-embedded using the `all-MiniLM-L6-v2 embedder` from the SentenceTransformers library[94] and stored in a Faiss vector store[106]. The literature search query is processed by the same embedding model and undergoes a k-nearest neighbor search, identifying the top k-most relevant abstract based on cosine similarity between the query vector and stored paper embeddings in the FAISS vector database. To enhance the reliability of the retrieved results, the system attempts to validate DOIs (Digital Object Identifiers) by cross-referencing them with CrossRef metadata. This validation process includes retrieving DOIs based on the paper title, checking for title and author consistency, and computing a similarity score to confirm accuracy. If discrepancies such as title mismatches or incorrect author attributions are detected, the DOI is discarded to maintain data integrity and prevent misinformation. The final output is structured as a ranked list of the three most relevant papers, displaying each paper's title, authors, publication year, and abstract, with a direct DOI link if verified. This integration of FAISS-based vector search, semantic similarity ranking, and DOI validation ensures the provision of high-quality, research-backed insights for scientists, policymakers, and decision-makers in wildfire risk assessment and mitigation planning.
- *Generation of tailored recommendations*: To generate tailored recommendations, the system constructs an augmented prompt that incorporates context from the ongoing conversation to provide actionable recommendations. If additional clarification or exploration is required, the system iterates the retrieval and response process, continuously refining the output to address the user's evolving needs. In this way, the system contextualizes data retrieval and recommendation generation based on the user's profession, ensuring that risk assessments, mitigation strategies, and policy suggestions align with their specific safety priorities and operational responsibilities.

### Details of the evaluation framework

**Personalization**. We carry out a two-phase ablation study to evaluate the impact of user profile granularity on WildfireGPT's responses. In the first phase, we use the prompt 'Develop recommendations to enhance forest resilience against wildfires' to elicit recommendations from WildfireGPT, with three levels of user profile details:

1. **Generic**: No contextual details.
2. **Location + Timeline**: Added geospatial (Covington, VA; 37.7935°N, 79.9939°W) and temporal (5–10 year implementation window) parameters.
3. **Full Profile**: Integrated profession-specific attributes (e.g., *power grid manager*), operational concerns (e.g., grid resilience, transmission line clearance), and scope (e.g., infrastructure protection).

In the second phase, we hold location and timeline constant and test five distinct professional profiles (*homeowner, civil engineer, ecologist, emergency manager, power grid manager*). For each profile, we modify three variables—profession, concern, and scop—in the following template user profile:

```
1. Profession:  {profession} in Virginia.
2. Concern:  Managing the forest, keeping it healthy, while {concern}, and
   protecting properties from potential wildfires.
3. Location:  Near Covington, VA with Latitude 37.7935 and Longitude
   -79.9939.
4. Time:  Recommendations to be implemented within the next 5 to 10 years.
5. Scope:  Management of the forest and properties to maximize {scope}, and
   protect against potential wildfires.
```

**LLM-as-a-Judge**. We prompt, in a zero-shot manner, GPT-4 to assess the relevance, entailment, and accessibility of the WildfireGPT responses using the same questions posed to human experts in Table 3 with minor modifications such as adjusting the pronouns—replacing "*my*" with "*the*" and "*your*" with "*the user's*." Additionally, GPT-4 as judge is instructed to provide one of four possible judgments—Yes, No, Could be better, or Not applicable—and to offer explanations for each assessment. For all evaluations, it receives the user profile and concerns, the user's previous queries, retrieved scientific literature and data, and the model's responses as inputs.

## Data availability

The FWI projections from ClimRR are available for download at https://anl.app.box.com/s/hmkkgkrkzxxocfe9kpgrzk2gfc4gizp8. Wildland Fire Incident Locations data can be accessed at https://data-nifc.opendata.arcgis.com/datasets/nifc::wildland-fire-incident-locations/about. The North

American Tree-Ring Fire Scar Synthesis dataset can be downloaded from https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=noaa-fire-34853. Census ACS5 data is accessible via the Python API package at https://github.com/Census-ACS/census. The scientific literature data developed by CIACC can be downloaded from https://anl.box.com/s/b4m2mnt5wa4z9l71qioz05cb5qpduj2j. All data are available for download at https://anl.box.com/s/wm888zovyapyou1txae7g75ghpc7sxre.

## Code availability

The code for WildfireGPT is publicly available on GitHub at https://github.com/Xieyangxinyu/WildfireGPT. The repository contains all the necessary components for reproducing the functionality of WildfireGPT, including the prompts used to query GPT-4, functions for identifying the location of interest (through a pinpoint mapping tool), visualizing data, and enabling transitions between modules based on the user's input. These functions are integrated with function calling support provided by the OpenAI Assistant API[51]. The software is implemented by using Python 3.11, and the front-end user interface is built as a Streamlit-based web application[52].

Received: 30 September 2024; Accepted: 28 April 2025;
Published online: 19 July 2025

## References


1. Tabari, H. Climate change impact on flood and extreme precipitation increases with water availability. *Sci. Rep.* **10**, 13768 (2020).
2. Mora, C. et al. Over half of known human pathogenic diseases can be aggravated by climate change. *Nat. Clim. Change* **12**, 869–875 (2022).
3. National Academies of Sciences. et al. *Attribution of Extreme Weather Events in the Context of Climate Change* (National Academies Press, 2016).
4. Task Committee on Future Weather, Climate Extremes, Tye, M. R. & Giovannettone, J. P. *Impacts of Future Weather and Climate Extremes on United States Infrastructure: Assessing and Prioritizing Adaptation Actions* (American Society of Civil Engineers, 2021).
5. Lu, B. et al. OxyGenerator: reconstructing global ocean deoxygenation over a century with deep learning. In: *Forty-first International Conference on Machine Learning* https://openreview.net/forum?id=0HUInAsdoo (2024).
6. Intergovernmental Panel on Climate Change Climate change 2021: The physical science basis. contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change https://www.ipcc.ch/report/ar6/wg1/ (2021).
7. Lopez, H. et al. Early emergence of anthropogenically forced heat waves in the western United States and Great Lakes. *Nat. Clim. Change* **8**, 414–420 (2018).
8. Hirabayashi, Y. et al. Global flood risk under climate change. *Nat. Clim. Change* **3**, 816–821 (2013).
9. Senande-Rivera, M., Insua-Costa, D. & Miguez-Macho, G. Spatial and temporal expansion of global wildland fire activity in response to climate change. *Nat. Commun.* **13**, 1208 (2022).
10. Brown, P. T. et al. Climate warming increases extreme daily wildfire growth risk in California. *Nature* **621**, 760–766 (2023).
11. Adshead, D. et al. A mission-driven approach for converting research into climate action. *npj Clim. Action* **2**, 13 (2023).
12. Callaghan, M. et al. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nat. Clim. Change* **11**, 966–972 (2021).
13. Webersinke, N., Kraus, M., Bingler, J. & Leippold, M. ClimateBERT: a pretrained language model for climate-related text. Available at SSRN 4229146 (2022).
14. Planas, J. et al. Beyond modeling: NLP pipeline for efficient environmental policy analysis. Preprint at https://arxiv.org/abs/2201.07105 (2022).
15. Rolnick, D. et al. Tackling climate change with machine learning. *ACM Comput. Surv. (CSUR)* **55**, 1–96 (2022).
16. Spokoyny, D., Laud, T., Corringham, T. & Berg-Kirkpatrick, T. Towards answering climate questionnaires from unstructured climate reports. Preprint at https://arxiv.org/abs/2301.04253 (2023).
17. Schimanski, T., Bingler, J., Kraus, M., Hyslop, C. & Leippold, M. ClimateBERT-NetZero: detecting and assessing net zero and reduction targets. In: Houda, B., Juan, P. & Kalika, B. (eds) *Proc. 2023 Conference on Empirical Methods in Natural Language Processing,* pp. 15745–15756 (Association for Computational Linguistics, 2023).
18. Schimanski, T. et al. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Financ. Res. Lett.* **61**, 104979 (2024).
19. Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A. & Sauerland, U. Risks and benefits of large language models for the environment. *Environ. Sci. Technol.* **57**, 3464–3466 (2023).
20. Bulian, J. et al. Assessing Large Language Models on climate information. Proceedings of Machine Learning Research (PMLR) (2024).
21. Chen, S., Long, G., Jiang, J., Liu, D. & Zhang, C. Foundation models for weather and climate data understanding: a comprehensive survey. Preprint at https://arxiv.org/abs/2312.03014 (2023).
22. Toetzke, M., Probst, B. & Feuerriegel, S. Leveraging large language models to monitor climate technology innovation. *Environ. Res. Lett.* **18**, 091004 (2023).
23. Stagnol, L. et al. Answering clean tech questions with large language models. Available at SSRN 4663447 (2023).
24. Atkins, C., Girgente, G., Shirzaei, M. & Kim, J. Generative AI tools can enhance climate literacy but must be checked for biases and inaccuracies. *Commun. Earth Environ.* **5**, 226 (2024).
25. Muccione, V. et al. Integrating artificial intelligence with expert knowledge in global environmental assessments: opportunities, challenges and the way ahead. *Regional Environ. Change* **24**, 1–8 (2024).
26. Hsu, A., Laney, M., Zhang, J., Manya, D. & Farczadi, L. Evaluating ChatNetZero, an LLM-Chatbot to demystify climate pledges. In: Dominik S. et al. (eds) *Proc. 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pp. 82–92 (Association for Computational Linguistics, 2024).
27. Goecks, V. G. & Waytowich, N. R. DisasterResponseGPT: large language models for accelerated plan of action development in disaster response scenarios. Preprint at https://arxiv.org/abs/2306.17271 (2023).
28. Hostetter, H. et al. The role of large language models (AI chatbots) in fire engineering: An examination of technical questions against domain knowledge. *Nat. Hazards Res.* **4**, 669–688 (2024).
29. Chen, W. & Fang, J. Optimizing AI-driven disaster management through LLMs. Preprint at https://doi.org/10.20944/preprints202407.1446.v1 (2024).
30. Vaghefi, S. A. et al. ChatClimate: grounding conversational AI in climate science. *Commun. Earth Environ.* **4**, 480 (2023).
31. Thulke, D. et al. ClimateGPT: Towards AI synthesizing interdisciplinary research on climate change. Preprint at https://arxiv.org/abs/2401.09646 (2024).
32. Cao, C., Zhuang, J. & He, Q. LLM-assisted modeling and simulations for public sector decision-making: Bridging climate data and policy insights. In: Ryan, S., Hong, S., Sera, L., Lei L. & Fei, F. (eds) *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design* (Association for the Advancement of Artificial Intelligence, 2024).
33. Martelo, R., Ahmadiyehyazdi, K. & Wang, R.-Q. Towards democratized flood risk management: An advanced ai assistant enabled by gpt-4 for enhanced interpretability and public engagement. *arXiv preprint arXiv:2403.03188* (2024).

34. Perplexity AI. Perplexity AI: An answer engine for knowledge discovery. Available at https://perplexity.ai (Association for the Advancement of Artificial Intelligence, 2024).
35. Pirozelli, P. et al. Benchmarks for Pirá 2.0, a reading comprehension dataset about the ocean, the Brazilian coast, and climate change. *Data Intell.* **6**, 29–63 (2024).
36. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
37. Li, S. S. et al. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *Adv. Neural Inf. Process. Syst.* **37**, 28858–28888 (2024).
38. Kraus, M. et al. Enhancing large language models with climate resources. Preprint at https://arxiv.org/abs/2304.00116 (2023).
39. Larosa, F. & Wickberg, A. Artificial intelligence can help loss and damage only if it is inclusive and accessible. *npj Clim. Action* **3**, 59 (2024).
40. Liu, Y. et al. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Houda, B., Juan, P. & Kalika, B. (eds) *Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing,* pp. 2511–2522 (Association for Computational Linguistics, 2023).
41. Zheng, L. et al. Judging LLM-as-a-Judge with MT-bench and Chatbot arena. *Adv. Neural Inf. Process. Syst.* **36**, 46595–46623 (2023).
42. Prestemon, J. P. et al. Projecting wildfire area burned in the South-Eastern United States, 2011–60. *Int. J. Wildland Fire* **25**, 715–729 (2016).
43. McKenzie, D. & Littell, J. S. Climate change and the eco-hydrology of fire: Will area burned increase in a warming western USA? *Ecol. Appl.* **27**, 26–36 (2017).
44. Romps, D. M., Seeley, J. T., Vollaro, D. & Molinari, J. Projected increase in lightning strikes in the United States due to global warming. *Science* **346**, 851–854 (2014).
45. Krause, A., Kloster, S., Wilkenskjeld, S. & Paeth, H. The sensitivity of global wildfires to simulated past, present, and future lightning frequency. *J. Geophys. Res.: Biogeosci.* **119**, 312–322 (2014).
46. Hessilt, T. D. et al. Future increases in lightning ignition efficiency and wildfire occurrence expected from drier fuels in boreal forest ecosystems of western North America. *Environ. Res. Lett.* **17**, 054008 (2022).
47. Richardson, D. et al. Global increase in wildfire potential from compound fire weather and drought. *NPJ Clim. Atmos. Sci.* **5**, 23 (2022).
48. Gamelin, B. L. et al. Projected US drought extremes through the twenty-first century with vapor pressure deficit. *Sci. Rep.* **12**, 8615 (2022).
49. Pausas, J. G. & Keeley, J. E. Wildfires and global change. *Front. Ecol. Environ.* **19**, 387–395 (2021).
50. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
51. OpenAI. OpenAI Assistant API (2023). https://platform.openai.com/docs/guides/assistant. Accessed 21 June 2024.
52. Streamlit. https://www.streamlit.io/. Accessed 21 January 2024.
53. National Wildfire Coordinating Group. Fire weather index (fwi) system (2021). https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-fwi-system. Accessed 30 September 2024.
54. Argonne National Laboratory. Climate risk & resilience portal (ClimRR) (2022). https://disgeoportal.egs.anl.gov/ClimRR/. Accessed 21 January 2024.
55. National Interagency Fire Center (NIFC). Wildland fire incident locations. NIFC Open Data. https://data-nifc.opendata.arcgis.com/datasets/nifc::wildland-fire-incident-locations/about. Accessed 21 January 2024.
56. Wildfire.gov. Integrated reporting of wildfire information (irwin). Wildfire.gov. https://www.wildfire.gov/application/irwin-integrated-reporting-wildfire-information. Accessed 21 January 2024.
57. Margolis, E. & Guiterman, C. NOAA/WDS Paleoclimatology—North American Tree-ring Fire Scar Synthesis (NAFSS). NOAA National Centers for Environmental Information (2021). https://doi.org/10.25921/pef0-zz47. Accessed 21 January 2021.
58. Bureau, U. C. American community survey 5-year data (2018-2022) (2023). https://www.census.gov/data/developers/data-sets/acs-5year.html. Accessed 13 January 2024.
59. Mallick, T. et al. Understanding the impact of climate change on critical infrastructure through NLP analysis of scientific literature. *Sustain. Resilient Infrastruct.* 1–18 (2024).
60. Mallick, T. et al. Analyzing regional impacts of climate change using natural language processing techniques. Preprint at https://arxiv.org/abs/2401.06817 (2024).
61. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In: Pierre B., Eugene, C. & Dekang, L. (eds) *Proc. 40th Annual Meeting of the Association for Computational Linguistics,* pp. 311–318 (Association for Computational Linguistics, 2002).
62. LIN, C. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summariation Branches Out, Post-Conference Workshop of ACL 2004* (2004).
63. Liu, Z., Zhou, K. & Wilson, M. L. Meta-evaluation of conversational search evaluation metrics. *ACM Trans. Inf. Syst. (TOIS)* **39**, 1–42 (2021).
64. Deriu, J. M., Tuggener, D., von Däniken, P. & Cieliebak, M. Probing the robustness of trained metrics for conversational dialogue systems. In: Smaranda, M., Preslav, N. & Aline, V. (eds) *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Dublin, Ireland, 22-27 May 2022*, Vol. 2, pp. 750–761 (Association for Computational Linguistics, 2022).
65. Farshidi, S. et al. Understanding user intent modeling for conversational recommender systems: a systematic literature review. *User Modeling User-Adapted Interaction* **34**, 1643–1706 (2024).
66. Hvenegaard, S. Long-term monitoring programs and data-collection protocols for fuel treatment sites, a literature review, final report. https://library.fpinnovations.ca/link/fpipub9450. (2014).
67. Dale, L. Wildfire policy and fire use on public lands in the United States. *Soc. Nat. Resour.* **19**, 275–284 (2006).
68. Stevens, J. T. et al. Forest vegetation change and its impacts on soil water following 47 years of managed wildfire. *Ecosystems* **23**, 1547–1565 (2020).
69. Ager, A. A. et al. Wildfire exposure to the wildland urban interface in the western US. *Appl. Geogr.* **111**, 102059 (2019).
70. Schoennagel, T. et al. Adapt to more wildfire in western North American forests as climate changes. *Proc. Natl Acad. Sci.* **114**, 4582–4590 (2017).
71. Waltz, A. E. et al. Effectiveness of fuel reduction treatments: assessing metrics of forest resiliency and wildfire severity after the wallow fire, AZ. *For. Ecol. Manag.* **334**, 43–52 (2014).
72. Hansen, W. D., Abendroth, D., Rammer, W., Seidl, R. & Turner, M. G. Can wildland fire management alter 21st-century subalpine fire and forests in Grand Teton National Park, Wyoming, USA? *Ecol. Appl.* **30**, e02030 (2020).
73. Stan, A. B., Rigg, L. S. & Jones, L. S. Dynamics of a managed oak woodland in northeastern Illinois. *Nat. Areas J.* **26**, 187–197 (2006).
74. Moritz, M. A. & Stephens, S. L. Fire and sustainability: considerations for California's altered future climate. *Climatic Change* **87**, 265–271 (2008).
75. Chen, X. et al. Challenges and contributing factors in the utilization of large language models (LLMs). Preprint at https://arxiv.org/abs/2310.13343 (2023).

76. Wang, W. et al. Learning to ask: When LLMs meet unclear instruction. Preprint at https://arxiv.org/abs/2409.00557 (2024).
77. Chaiir, S., Charrad, M. & Saoud, N. B. B. Using recommendation systems in disaster management: A systematic literature review. *Procedia Comput. Sci.* **225**, 1738–1747 (2023).
78. Kirchhoff, C. J., Carmen Lemos, M. & Dessai, S. Actionable knowledge for environmental decision making: broadening the usability of climate science. *Annu. Rev. Environ. Resour.* **38**, 393–414 (2013).
79. Tomaszewski, B. *Geographic Information Systems (GIS) for Disaster Management* (Routledge, 2020).
80. Vásquez, F., Cravero, A., Castro, M. & Acevedo, P. Decision support system development of wildland fire: a systematic mapping. *Forests* **12**, 943 (2021).
81. Abid, S. K., Chan, S. W., Sulaiman, N., Bhatti, U. & Nazir, U. Present and future of artificial intelligence in disaster management. In: Varlamov, O. & Dolgopyatova, A. (eds) *2023 International Conference on Engineering Management of Communication and Technology (EMCTECH)*, 1–7 (IEEE, 2023).
82. Kurniawan, D., Indah, D. R., Sari, P. & Alif, R. Understanding the landscape of usability evaluation in geographic information systems: A systematic literature review. *J. Appl. Sci. Eng. Technol. Educ.* **5**, 35–45 (2023).
83. Ali, A. H. & Thakkar, R. Climate changes through data science: understanding and mitigating environmental crisis. *Mesop. J. Big Data* **2023**, 125–137 (2023).
84. Ahmed, M. R. Forest fire danger/risk forecasting: a remote sensing approach. *Forest* 03 University of Calgary (2020).
85. Ma, C. et al. WeatherQA: Can multimodal language models reason about severe weather? Preprint at https://arxiv.org/abs/2406.11217 (2024).
86. Kadiyala, L. A., Mermer, O., Samuel, D. J., Sermet, Y. & Demir, I. The Implementation of Multimodal Large Language Models for Hydrological Applications: A Comparative Study of GPT-4 Vision, Gemini, LLaVa, and Multimodal-GPT. *Hydrol.* **11**, 148 (2024).
87. Jiang, B. et al. Towards Rationality in Language and Multimodal Agents: A Survey. In: Luis, C. Alan, R. & Lu, W. (eds) *Proc. of the 2025 Conf. of the Nations of the Americas Chapter of the Assoc. for Comput*. Linguistics: Hum. Lang. Technol. (Vol. 1: Long Papers), pp. 3656–3675 (Association for Computational Linguistics, 2025).
88. Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. & Zrnic, T. Prediction-powered inference. *Science* **382**, 669–674 (2023).
89. Chatzi, I., Straitouri, E., Thejaswi, S. & Rodriguez, M. Prediction-powered ranking of large language models. *Adv. Neural Inf. Process. Syst.* **37**, 113096–113133 (2024).
90. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: efficient finetuning of quantized LLMs. *Adv. Neural Inform. Process. Syst.* **36** (2024).
91. Carta, T. et al. Grounding large language models in interactive environments with online reinforcement learning. In: Andreas, K. et al. (eds) *International Conference on Machine Learning,* pp. 3676–3713 (PMLR, 2023).
92. Lester, B., Al-Rfou, R. & Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In: Xuanjing, H., Lucia, S. & Wen-tau, Y. S. (eds) *Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing,* pp. 3045–3059 (Association for Computational Linguistics, 2021).
93. Guo, Q. et al. Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers. In The Twelfth International Conference on Learning Representations (2024).
94. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using siamese bert-networks. In: Kentaro, I., Jing, J., Vincent, N.G. & Xiaojun, W. (eds) *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992 (2019).
95. Arivazhagan, M. G. et al. Hybrid hierarchical retrieval for open-domain question answering. In: *Findings of the Association for Computational Linguistics: ACL 2023*, 10680–10689 (2023).
96. Naqvi, S. M. R., Ghufran, M., Varnier, C., Nicod, J.-M. & Zerhouni, N. Enhancing semantic search using ontologies: a hybrid information retrieval approach for industrial text. *J. Ind. Inform. Integration* **45**, 100835 (2025).
97. Zhu, R. et al. The knowWhereGraph: a large-scale geo-knowledge graph for interdisciplinary knowledge discovery and geo-enrichment. Preprint at https://arxiv.org/abs/2502.13874 (2025).
98. Liu, S., Shrutika, F., Zhang, B., Huang, Z. & Qian, F. Effect of adaptive communication support on human-ai collaboration. Preprint at https://arxiv.org/abs/2412.06808 (2024).
99. Shi, F. et al. Large language models can be easily distracted by irrelevant context. In: Andreas, K. (eds) *Int. Conf. on Machine Learning*, 31210–31227 (PMLR, 2023).
100. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
101. Singhal, P., Goyal, T., Xu, J. & Durrett, G. A long way to go: Investigating length correlations in R:HF. In: Choi, Y., Zhou, D., Das, D., Chen, D., Artzi, Y. & Fan, A. (eds) *Proceedings of COLM 2024 First Conference on Language Modeling* (2024).
102. Singh, A., Ehtesham, A., Kumar, S. & Khoei, T. T. Agentic retrieval-augmented generation: a survey on agentic rag. Preprint at https://arxiv.org/abs/2501.09136 (2025).
103. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
104. Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. Preprint at https://arxiv.org/abs/2312.10997 (2023).
105. Van Wagner, C. E. Development and structure of the Canadian Forest Fire Weather Index System. Forestry Technical Report 35, Canadian Forestry Service, Headquarters, Ottawa (1987).
106. Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **7**, 535–547 (2019).



## Acknowledgements
This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DEAC02-06CH11357. Y. Xie and W. J. Su acknowledge support from the NSF HDR TRIPODS award (CCF-1934876). B. Jiang and C. J. Taylor acknowledge support from the NSF grant CCF-2112665 (TILOS). M. R. Alexander, J. Branham, and Y. Feng contributed significantly to the case studies. The authors also thank Jane S. Frantz, Parfait Gasana, Vijay P. Ramalingam, and Charles Doktycz for their valuable contributions to the case studies. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.


## Author contributions
Y.X., B.J., T.M., and J.K.H. designed the study and developed the code base. Y.X., T.M., M.R.A., and J.B. prepared the data used in the study. Y.X., T.M., and J.K.H. facilitated the case studies. M.R.A., J.B., and Y.F. contributed domain knowledge to the case studies. Y.X., B.J., T.M., and

J.D.B. wrote the main manuscript text. All authors reviewed, edited and improved the manuscript.

## 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/s44168-025-00254-1.

**Correspondence** and requests for materials should be addressed to Yangxinyu Xie or Tanwi Mallick.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.