Save It for the "Hot" Day: An LLM-Empowered Visual Analytics System for Heat Risk Management

Category: Research

Paper Type: Application

Abstract— The escalating frequency and intensity of heat-related climate events, particularly heatwaves, emphasize the pressing need for advanced heat risk management strategies. Current approaches, primarily relying on numerical models, face challenges in spatial-temporal resolution and in capturing the dynamic interplay of environmental, social, and behavioral factors affecting heat risks. This has led to difficulties in translating risk assessments into effective mitigation actions. Recognizing these problems, we introduce a novel approach leveraging the burgeoning capabilities of Large Language Models (LLMs) to extract rich and contextual insights from news reports. We hence propose an LLM-empowered visual analytics system, *Havior*, that integrates the precise, data-driven insights of numerical models with nuanced news report information. This hybrid approach enables a more comprehensive assessment of heat risks and better identification, assessment, and mitigation of heat-related threats. The system incorporates novel visualization designs, such as "thermoglyph" and news glyph, enhancing intuitive understanding and analysis of heat risks. The integration of LLM-based techniques also enables advanced information retrieval and semantic knowledge extraction that can be guided by experts' analytics needs. We collaborated with six domain experts to conduct a case study on the 2022 China Heatwave and an expert survey & interview, demonstrating the usefulness of our system in providing in-depth and actionable insights for heat risk management.

Index Terms—Heat risk management, climate change, numerical model, news data, large language model, visual analytics

1 Introduction

Extreme climate events [1], particularly those related to heat [33], have seen a marked increase in intensity and frequency in recent years, raising significant concerns globally. Heat risks are associated with excess mortality due to temperatures above long-term averages and specific extreme events like heatwaves [15]. The adverse impacts extend beyond personal health, as reduced productivity is observed due to heat-related health issues among employees [57]. They can further cause severe damage to infrastructure, including buildings, power grids, and roads, leading to disruptions in daily life and economic activities [49]. Developing effective strategies for heat risk management, therefore, becomes increasingly urgent [36].

This complex task requires a comprehensive understanding of the interplay between various factors, including meteorological, urbanization, demographic, and socioeconomic factors [15, 28, 69]. For example, short-sighted policy solutions like increased reliance on air conditioning during heatwaves, while essential for immediate relief, contribute to climate change and cause surges in energy consumption. This stresses power grids and amplifies the risk of failures [48].

Numerical models are dominantly used in assessing heat risk by environmental researchers [25, 70]. However, they have notable limitations. First, their sparse spatial and temporal resolution offers insufficient data-driven support for effective risk management [32,51]. For instance, the fifth-generation reanalysis (ERA5) data [26], one of the most used domain data, yields only one estimate for an area spanning approximately $27.75 \times 27.75 km^2$ per hour. This coarse-grained resolution can only predict normalized results and overlook extreme conditions [75]. Second, these approaches aim to predict meteorological variables (*e.g.*, temperature and humidity), failing to capture the complex risk dynamics involving human behaviors and social factors [63]. Third, the preparedness and response of society to the risk and the instructions for citizens in numerical models are absent.

News articles about environmental issues complement numerical models by providing detailed descriptions of extreme situations, documenting causes and consequences, and discussing city responses and specific advice for handling heat-related events [61]. This information offers retrospective evaluations of management strategies that experts can utilize to refine their approaches. However, efficiently retrieving relevant and valuable news articles remains a significant challenge. The absence of tools for integrating diverse data sources—such as textual news and numerical models—hinders decision-makers from fully leveraging the wealth of information available.

The emergence of large language models (LLMs) offers a new opportunity to address retrieval and integration challenges. LLMs have demonstrated abilities ranging from information extraction [10] to question answering and document retrieval [52]. Despite these strengths, integrating environmental news into heat risk management encounters technical challenges: the large volume of news articles challenges LLM token limits, impacting retrieval efficiency; the difficulty of fusing information from varied sources, such as numerical data and textual news, poses a barrier to obtaining a holistic view of heat risks; and the hallucination problem of LLMs may mislead experts.

To address those challenges, we develop *Havior* (Heat Savior), an LLM-empowered visual analytics system that integrates numerical data and textual news catering to the needs of environmental researchers and policy makers involved in heat risk management. *Havior* features a novel "thermoglyph" design that utilizes metaphorical representations to enhance experts' comprehension of meteorological conditions. It supports experts in efficiently retrieving, managing, and navigating a large volume of news articles within a human-in-the-loop retrieval process through hex bin visualizations of topics and news glyphs. To distill the news' semantic meaning and prevent hallucination, *Havior* further employs LLM techniques including prompt engineering and retrieval augmented generation (RAG) [41]. By combining the strengths of numerical models and the rich contexts from news, our system empowers stakeholders to make informed decisions and proactively mitigate heat risks. Our contributions are as follows:

- ◇ A novel LLM-empowered pipeline that supports human-in-the-loop retrieval and heterogeneous (numerical and textual) data integration in the context of heat risk management. The pipeline is with the potential to be applied to various types of textual documents, not limited to news articles.
- ◇ A visual analytics system, *Havior*, featuring "thermoglyph", news glyphs, and visualization of hex bins, allowing experts to explore and visualize heat risk insights interactively. We implement *Havior* through an open-source prototype system[1].
- ◇ Evaluation through a comprehensive case study, survey, and interview with six experts, showing that valuable insights can be obtained by integrating numerical and textual data using *Havior*.

[1] https://anonymous.4open.science/r/Havior-C58D

2 Related Work

2.1 Visual Analytics of Climate Risk Data

Climate risk refers to the potential for uncertain outcomes affecting valuable assets [56], quantified by the probability and impact of hazards like flooding [65, 66], rock cracking [17], and wildfires [9]. Visual analytics of climate risk primarily focuses on analyzing *weather simulation data* [4, 35, 60] to derive insights for mitigating these hazards. However, existing tools often lack the necessary granularity to address the spatial-temporal complexity of predictions with adequate explanations [11], making it challenging for stakeholders to capture detailed local variations and fully understand and respond to specific risks.

Several visual analytics systems [13, 43, 53] have emerged to explore the multifaceted spatial-temporal dynamics between meteorological variables. While these systems enhance numerical data analysis, they often neglect human-related factors essential for comprehensive heat risk management [66, 71, 76]. Bridging this gap is crucial for better decision-making in climate risk management.

Our approach integrates meteorological data with news to provide a novel perspective on risk analysis. This fusion of quantitative simulations with contextual news analysis offers a logical interpretation of quantitative variables and empirical support for subjective narratives [34]. Hence, our system can navigate users through both "Big" (quantitative) and "Small" (qualitative) data, which are mutually dependent and can enhance each other [5].

2.2 Risk Analysis of Text Data

Over the past decade, microblog data, notably Tweets, has been used for risk analysis [47], including studies on climate risks [66, 71]. Such data offer a direct glimpse into the evolution of public sentiment and suggestions, facilitating rapid reactions to emergencies [22, 66, 71]. However, the casual format and unstructured nature of microblogs present significant challenges. They often omit necessary context for analyzing periodic events and can amplify biases [37, 46], thus limiting their utility for analyzing broader meteorological phenomena.

News articles are more reliable in monitoring natural disasters than microblogs [54]. They provide in-depth descriptions of local conditions, consequences, and underlying reasons behind heat-related incidents. Authored by professionals, these articles are richer, more coherent, and less noisy in analytical discussions [61] (*e.g.*, expert opinions and suggested mitigation strategies), supporting retrospective evaluations of management strategies. However, the inherent complexity in these documents requires heterogeneous data sources and advanced techniques for effective data structuring and interpretation [72]. In addition, mapping news to specific climate events has traditionally relied on fuzzy logic [73] due to different detail levels between meteorological data and journalistic reporting. To connect news with climate events, textual descriptions can align with geospatial visualizations for spatial context. Direct references and visual cues within the text can enhance user understanding of geospatial context [30, 39, 40].

Inspired by these design guidelines, *Havior* highlights the key structured information extracted by LLM and connects them to numerical results through novel visual designs. These designs maintain a good balance between the precision of numerical models and the semantic richness of textual information.

2.3 Steering Documents Retrieval

Retrieving relevant content from large document corpora is a significant challenge in text mining due to the ambiguity of natural language [45]. Previous visual analytics systems have proposed solutions based on ranking [8], similarity [14], and removing overlaps [21] to steer document retrieval. However, these methods may overlook low-frequency topics, which are crucial in heat risk management. We introduce a hierarchical hexagon layout that prevents neglecting rare topics while maintaining a tight arrangement and preserving text similarity relations.

LLMs have shown promise in summarizing documents [64], but adapting them to specific domain tasks requires substantial computational resources. Efficient alternatives like prompt engineering are limited by token length constraints [18], making it challenging to analyze multiple documents simultaneously. Question-based retrieval systems like DocFlow [59] enhance search accessibility and efficiency using natural language queries. Inspired by these approaches, we explore using Retrieval-Augmented Generation (RAG), which integrates external knowledge bases. RAG effectively mitigates common LLM challenges, including hallucinations, reliance on outdated information, and lack of transparent reasoning processes [19]. By integrating LLMs for semantic analysis and news retrieval, *Havior* enhances machine reasoning and supports human-in-the-loop visual analysis.

3 Informing the Design

To develop a system that effectively supports heat risk management, we conducted a year-long collaboration with two domain experts, **D(esigner)1** and **D2**. **D1** is a professor in environmental science with over thirty years of research experience, and **D2** is an environmental specialist with more than four years of experience. Additionally, **D1** is a representative in the 'My Climate Risk' scheme, launched by the World Climate Research Programme (WCRP) [58] to mitigate climate event risks. Through weekly meetings, we engaged in an iterative design process to identify the specific needs and challenges. Our discussions highlighted several limitations of existing numerical models. While these models provide quantitative analyses, they often fail to capture the complexities of urbanization, demographics, and socioeconomic factors crucial for developing effective heat risk management strategies. As **D1** noted, "*If you ask our numeric model how many people die, it has no answer since it only calculates some meteorological variables.*"

Recognizing the potential of news articles to provide contextual and human-centric insights, we aimed to integrate these textual sources with numerical data. However, integrating heterogeneous data sources posed significant challenges, particularly in creating a comprehensive and user-steerable decision-support pipeline. Based on the insights from our collaboration, we distilled six design requirements, grouped into three categories: `Numeric` understanding from climate data, `Semantic` understanding from news, and `Integration` of the two.

R1: Numeric **Analyze historical and future trends.** The system should support historical analysis and future forecasting functions. Experts should be able to load and analyze numerical data of interest at different time points, allowing them to examine past trends and forecast future scenarios effectively.

R2: Numeric **Examine spatial meteorological conditions.** The system should enable the analysis of spatial patterns using familiar visualization forms and analytical methods for domain experts. Features such as spatial zoom in/out and the ability to switch between different variables should be provided to facilitate insights.

R3: Semantic **Support human-in-the-loop news retrieval.** A multi-step retrieval approach should be developed to retrieve a suitable number of news articles that align with experts' interests. **D1** and **D2** stated, "*There are tons of news for me. I need a system to retrieve an appropriate amount of news in different stages of analysis.*"

R4: Semantic **Enhance management and navigation among large-scale news.** With the retrieved news list, efficient management should allow experts to filter and rank news based on numeric and semantic criteria. An easy navigation way among a large amount of news and similar news should be facilitated.

R5: Semantic **Extract insights from heat-related news.** The system should possess the abilities of structural information extracting, semantic understanding, and contextual question-answering to help experts gain semantic insights from heat-related news.

R6: Integration **Integrate numeric and semantic insights for decision-making.** The system should integrate the impact of news into the numeric model's results to generate insights.

4 LLM-Empowered Pipeline

We developed a novel pipeline (Fig. 1) that leverages the burgeoning capabilities of LLM to integrate numerical data analysis and semantic understanding to enhance heat risk management.

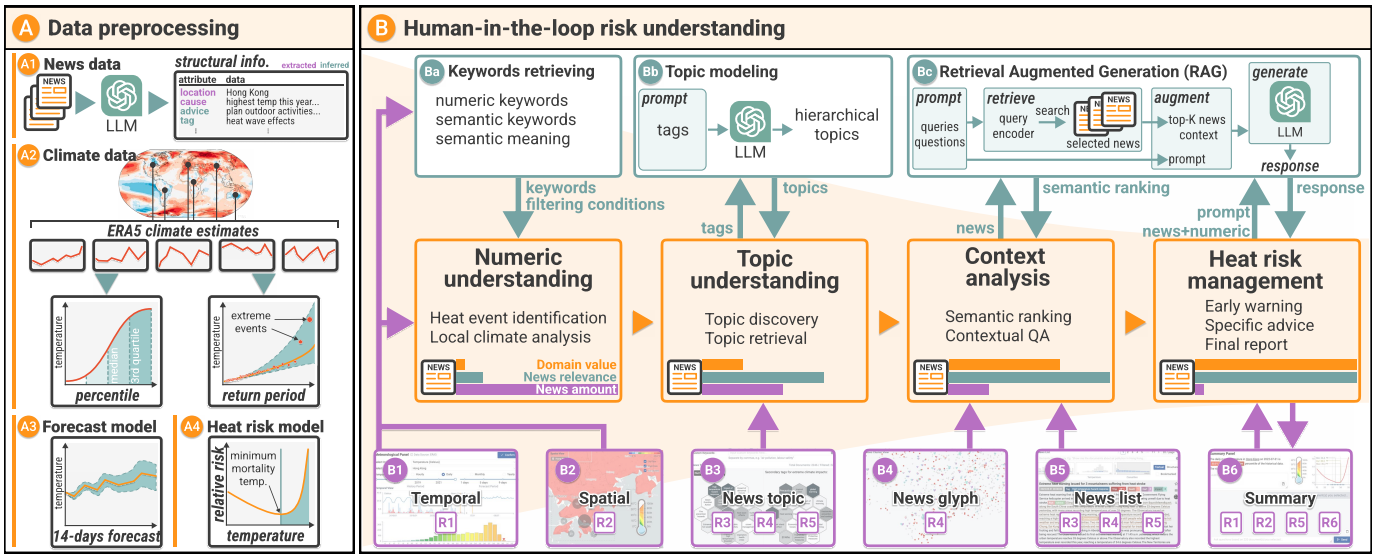


Fig. 1: The LLM-empowered pipeline contains two parts: data preprocessing (A) and human-in-the-loop risk understanding (B). The data preprocessing involves extracting structural information using LLM (A1) and calculating climate indices (A2-4). In human-in-the-loop risk understanding (B), heterogeneous understandings are integrated through keywords retrieving (Ba), topic modeling (Bb), and RAG (Bc). The interactive analysis process is supported by six views of *Havior* (B1-6) which fulfill the design requirements.

4.1 Data Preprocessing

Climate data were obtained from the ERA5 reanalysis dataset [26]. Its utilization and performance in heat risk research have been widely acknowledged for its quality, long-term availability, and accessibility [44]. The hourly data is from 2015 to 2023 and has a spatial resolution of approximately $27.75\ km \times 27.75\ km$. While it might seem intuitive to focus on daily temperature extremes (*e.g.*, max and min) to assess heat risks, studies have shown that *daily mean temperature* (hereinafter temperature) is more appropriate because it is more comprehensive in representing the day's temperature exposure [68]. The average of 24-hour (in hours UTC) temperatures within a day was utilized for the temperature for each location. Moreover, we adopted the Pangu-Weather [3] model with ERA5 temperature estimates to obtain temperature forecasts up to 14 days (Fig. 1-A3).

To contextualize magnitude-based metrics, we also incorporated probability-based indicators for a holistic analysis of heat risks.

- **Temperature percentiles** offer contexts into the local climatology of a specific region, enabling cross-regional and temporal comparisons [38]. This is particularly important when considering variations in local tolerance and preparedness levels [2].
- **Return period** is a statistical measure indicating the estimated average interval between the occurrence of heat events. This measure suggests the likelihood of a certain temperature threshold being exceeded at least once a year. It evaluates the frequency and intensity of heat events [36].

The heat risk model [20] examines the relationship between heat-related mortality and temperature fluctuations in 384 locations. After conducting a sensitivity analysis, the heat-related mortality is quantified using the daily count of deaths for non-external causes only (International Classification of Diseases [ICD]-9 0-799, ICD-10 A00-R99). The resulting patterns are often represented as "U" curves (Fig. 2-C1), representing the cold risk and heat risk, with the temperature range associated with the lowest mortality called the Minimum Mortality Temperature (MMT). Deviations from the MMT are generally associated with an exponential increase in the relative risk, highlighting the use of temperatures exceeding the MMT when evaluating heat risk. Since the local climate conditions (*e.g.*, tropical or temperate) have a strong influence on individual locations' MMT, our analysis adopts the corresponding heat risk models for each city (Fig. 1-A4).

Environmental news dataset was obtained from Wisers [67], which consisted of 7.7 million environmental (not limited to heat) news articles from Chinese news publishers with over ten years of experience, mainly covering East Asian regions. The dataset spans from July 2015 to June 2023. Each news article contains the title, content, character statistics, publishing date, publisher, and media type. The media types encompass both web and publication resources while excluding internet-based media sources primarily reliant on aggregating news reports from official news agencies. We extracted structural information from news articles by using LLM (GPT3.5 [7]), including extracted information (*e.g.*, location, time, risk description, and consequence) and inferred information (*e.g.*, advice and tag) (Fig. 1-A1). Due to the large dataset, it takes over two weeks to preprocess all news within one city, including keywords retrieval and structural information extraction. This structural information was developed together with our domain experts—the designers **D1** and **D2**. The prompts and examples can be found in the supplement material.

4.2 Human-in-the-loop Risk Understanding

4.2.1 Numerical Climate Data Understanding

Experts begin by analyzing the numerical climate data to gain a quantitative understanding of the risk (R1, R2). After the numerical data was processed by the method mentioned before and displayed in *numeric panel* (Fig. 2-A), experts can analyze both magnitude-based and probability-based indices for heat risk and identify heat hazards. Additionally, the temporal and spatial information of heat events can help interpret and understand meteorology. The extracted time and location from news sources are visualized in the *temporal view* (A1) and *spatial view* (A3), respectively. Integrating this information with the climate data facilitates experts' numerical understanding of heat risk, which was challenging in the experts' original workflow.

By examining the temporal trend (Fig. 1-B1), experts can identify noteworthy patterns and fluctuations in the meteorological variables and the occurrence of news events over time (R1). This analysis facilitates the detection of temporal correlations and the identification of significant events or trends within the meteorological context. Similarly, the spatial distribution of news events (Fig. 1-B2) provides valuable information regarding the geographical patterns and localized impacts of meteorological phenomena (R2). By visualizing the spatial distribution of news articles, experts can discern clusters and patterns of events resulting from heatwaves aiding in the identification of regions affected by specific risks. Furthermore, we extract numerical temperature from the news. Then we select the news that documents the highest number of casualties as the representative news for different temperature

(Fig. 2-C1), allowing us to provide semantic explanations for numerical risk levels. It transforms the risk level from a mere numerical value into a relatable example, enabling experts to develop a more concrete grasp of heat risks.

4.2.2 Topic Understanding & Context Analysis

To complement the numerical analysis, experts utilize a dataset of environment-related news to gain a semantic understanding of heat risk within the city's context.

Keywords retrieving (Fig. 1-Ba). The first step is to retrieve highly relevant news using keywords (R3), such as "Hong Kong," "prolonged," and "high temperature," and automatically filter the retrieved news using semantic meaning. These keywords can be automatically generated based on the numerical analysis results or suggested by experts who have analyzed the numerical data. For instance, a 97.5th percentile with ≥ 4 days duration [24] can lead to the inclusion of the keyword "heatwave." To ensure relevance, we automatically filter the retrieved news using structural information such as "is heat risk" and "location." Additionally, experts can apply numeric filters based on criteria like "time," "temperature," and "casualty," extracted from the news.

Topic modeling (Fig. 1-Bb). Leveraging the capabilities of LLMs, we extract and cluster tags (a type of structural information, Fig. 1-A1) to generate descriptive topics (R3, R4, R5). This approach surpasses traditional named entity recognition and clustering techniques [29], efficiently summarizing and categorizing the retrieved news (Fig. 1-B3, Fig. 1-B4). It serves two objectives: topic discovery and filtering.

First, using LLM, a wide range of comprehensive topics related to heat risk in a specific city can be identified. By exploring these topics, experts can gain an overview of the heat risk landscape and uncover unexpected issues. Second, since the number of news articles retrieved about a city often exceeds the experts' capacity to effectively read and analyze, leveraging topics as a criterion enables them to filter the news effectively. Consequently, they can focus their attention on specific areas of interest.

RAG (Fig. 1-Bc). To mitigate the hallucination problem in LLMs [62], we employ RAG, which allows experts to pose contextual questions and receive accurate answers by utilizing the retrieved news articles and numerical results as the knowledge source (Fig. 1-B6; R5). This integration broadens the scope of potential questions posed by experts. To continue enhancing the ability to delve into specific topics, we provide a ranking function based on semantic meaning (R4). Therefore, experts can use natural language (sentences or documents) to rank news (Fig. 1-B5). This is achieved by designating the retrieved news as the knowledge source and leveraging the semantic meaning to rank the news articles within this source through Embedchain[2]. This approach allows news that aligns closely with the experts' interests to be prioritized and placed higher in the ranked list.

4.2.3 Heat Risk Management

Havior integrates two essential perspectives of analysis: numerical and semantic (R6) for heat risk management. The numerical analysis offers quantitative insights into the magnitude, trends, and patterns of heat risk. Its results are utilized for subsequent tasks such as news retrieval, filtering, and comprehension. The semantic analysis retrieves, filters, clusters, ranks, and analyzes relevant news articles, enabling experts to delve into the context, impacts, and complexities surrounding the risk. To summarize, the integration of numerical analysis and semantic understanding allows for a more nuanced assessment, enabling experts to identify potential correlations, causal relationships, and interdependencies among various risk factors.

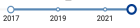
Havior also provides a summary functionality (Fig. 1-B6) that assists experts in consolidating their insights from both numerical and semantic analyses. Using LLMs, these insights are synthesized into a comprehensive final report, encompassing meteorological conditions, descriptions of heat risk scenarios, historical events or disasters, and advice for government entities and citizens. This final report serves as a valuable resource for facilitating informed and well-founded decision-making by experts, triggering more effective and rational risk management strategies. Decision-makers can develop proactive risk mitigation plans, allocate resources effectively, and implement targeted interventions to minimize the negative impacts on economies and the environment.

[2] https://github.com/embedchain/embedchain

5 Visual Design of Havior

The interface of *Havior* is shown in the Fig. 2. To exemplify the connection between views in the interface, let us consider an expert (Zoe) utilizing *Havior* for heat risk research. Initially, to check the numerical meteorological condition, Zoe selects the index, city, and temporal resolution from the top menus of the *meteorological panel*. The temporal (R1) and spatial (R2) climate information is then displayed in the *temporal view* and *spatial view*, respectively. After gaining the numerical understanding, the next step involves exploring the semantic aspect. Zoe selects the recommended keywords or types the customized keywords (R3) in the top menus of the *news panel*. The *news topic view* displays topics of the retrieved news using hex bins. Zoe can make positive or negative selections of hex bins (R3, R4), resulting in different scatter plots in *news glyph view* and news lists in the *news list*. Zoe can easily locate news of interest with the assistance of *news glyph view* and delve into the structural information or full-text of news within the *news list view* (R4). The human-in-the-loop retrieval process and the contextual in the *news panel* question-answering interface in the *summary panel* can help her understand heat-related news (R5). Moreover, insights or knowledge that experts wish to summarize for subsequent reviewing or report generation (R6) for decision-making can be pinned to the *summary panel*. With the integration of numerical results, *Havior* is able to generate an informative report for decision-making.

5.1 Temporal View

The *temporal view* (Fig. 2-A1) is for experts to understand the temporal trends and distribution of meteorological variables (R1). Considering the history period (2015-2023) is typically much longer than the future forecasting period (14 days), line charts for historical and future data are on the same y-axis, but they are separated on the x-axis. This separation helps experts distinguish between the known historical data and the projected future data. To zoom in on a particular timeframe, experts can drag the node of start or end (

).

Furthermore, they need to explore the temporal relationship between trends and the volume of news related to these parameters (R6). We use the bar chart (Fig. 2-A1) to show the number of news articles published. To save space and easily comparison between the relative magnitudes of changes in both meteorological data and news volume over time, we combine the bar chart and the line chart with the dual y-axis with both y-axes beginning from zero. This design helps in maintaining a visual consistency that can aid in understanding the relationship between the two datasets (Fig. 6-A, B) without overstating or understating the variations in either due to scaling issues [31]. In addition to the temporal trend, we provide a histogram (Fig. 2-A2) to display the frequency distribution of temperature (R1).

5.2 Spatial View

We provide the *spatial view* aiming to help experts visualize and comprehend the spatial distribution of meteorological variables (R2). To achieve this, we combine the citywide heat map of the variable (by averaging the data within each city) and the geography map to display the spatial distribution. In determining the color scheme, we draw inspiration from ERA5's color scheme [16], which is chosen based on its association with human perception of temperature. The spatial relationship between the distribution of meteorological variables and the geographic locations of news articles is vital (R6). Thus, we plot news on the map (Fig. 2-A3). They will be automatically aggregated in cases of close proximity when zooming in/out, allowing for a more concise visualization.

Thermoglyph. To effectively visualize both magnitude-based temperature and probability-based index, we have developed the "thermoglyph." They are presented in the city gallery (Fig. 3), alongside each city on the map (Fig. 2-A4), and in the summary panel, which caters to the varying needs of experts throughout different stages of analysis. The

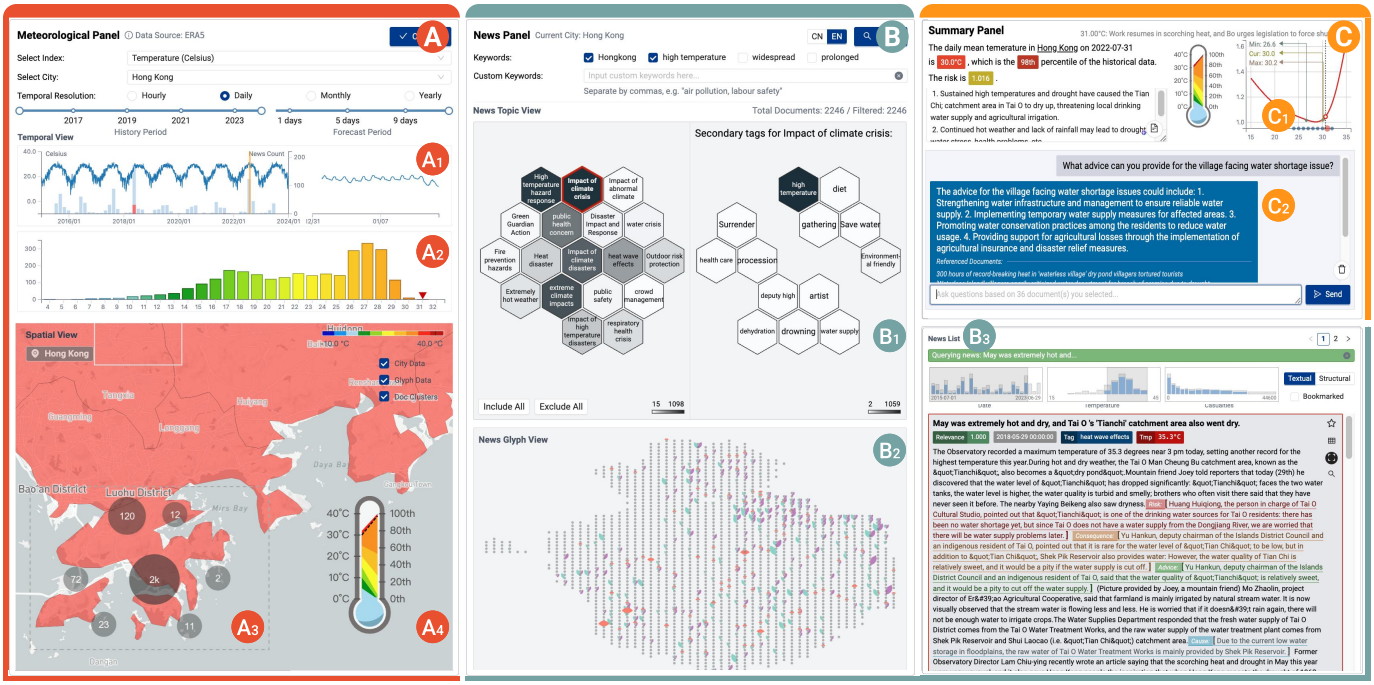


Fig. 2: The interface of *Havior* (Heat Savior). The Meteorological Panel (A) facilitates numerical understanding of meteorology, including temporal trends (A1), temporal distribution (A2), and spatial distribution (A3). The "thermoglyph" of Hong Kong (A4) intuitively shows the city-based pattern and correlation between temperature and percentile. The News Panel (B) supports human-in-the-loop news retrieval and enhancement in their semantic understanding, in terms of topic-based hierarchies (B1) and risk-based semantic proximity (B2) of retrieved news. The News List (B3) provides details of structural information in the retrieved news with supportive visual cues. The Summary Panel (C) enables experts to examine the integration of news and numeric risk model (C1), pose contextual questions (C2), and generate risk management reports.

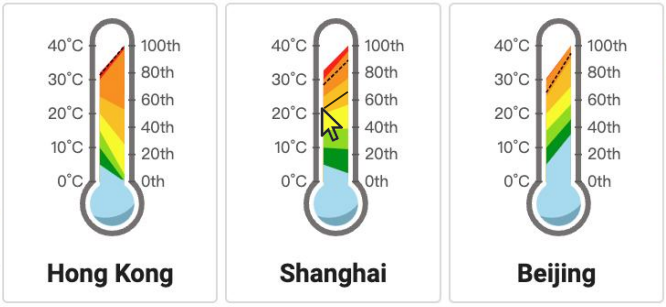


Fig. 3: The "thermoglyph" in the city gallery for selecting cities. They employ a metaphorical representation. The pattern of color blocks vividly depicts the relationship between temperature and percentile for each city. The black lines connect the current temperature (dashed) or the hovered temperature (bold) to its corresponding percentile.

"thermoglyph" resembles a thermometer, where the color gradually fills from the bottom to the top, representing the rising mercury in a traditional thermometer due to heat. The "thermoglyph" consists of two parallel axes: temperature and percentile. Different temperature ranges and the associated percentiles are linked and encoded using the same color scheme in the *spatial view*. Consequently, unique patterns emerge in different cities (Fig. 3). For instance, the "thermoglyph" for Hong Kong exhibits a concentrated pattern on the left side, indicating that the temperature in Hong Kong is more concentrated for the majority of the time. On the other hand, the "thermoglyph" for Beijing showcases parallel distribution, representing the broader temperature range and distinct seasonal characteristics observed in Beijing. A dashed back line indicates the current value. We also add a solid black line to accurately illustrate the link when the mouse hovers. This feature fulfills the experts' need for precise information.

Design alternatives. We consider a design alternative line chart (Fig. 1-A2) but find two issues: (1) It is challenging to differentiate between various patterns based on the changing slopes of the lines and (2) caused confusion among users because it implies a temporal change. Therefore, we opted for the design of the "thermoglyph," which intuitively conveys the relationships between temperatures and percentiles.

5.3 News Topic View

Topic generation of news can help efficiently filter and manage news (R3, R4) and analysis of news (R5). We use a case focusing on Shanghai in 2022 China heatwave to illustrate the design of *news topic view* (Fig. 4). We employ hex bins with text placed at their centers to represent topics for three reasons. Firstly, compared to a table list of topics, it has the advantage of conveying the information of overview. Secondly, the two-dimensional space of the hex bins preserves the relative spatial relationships between topics [50], enabling the keeping of semantic meaning relationships. Thirdly, in addition to being used as a container for displaying topics, hex bins are inherently well-suited to serve as buttons for experts to filter news. We encode the quantity of related news for each topic using a grayscale intensity scheme to avoid confusion with the color in the *spatial view*.

As mentioned in Sec. 4.1, we generated tags for each news. Then we employed the LLM to cluster those tags and generate the title for each cluster, resulting in a hierarchical structure. The first-level topics, which correspond to the cluster titles, provide an overview of the information in each cluster. For example, we know that the "power crisis" led to challenges during that particular heatwave in Shanghai. The second-level topics, which are tags of news, offer more specific information pertaining to each cluster. For instance, by clicking the hex bin for "power crisis", additional information such as "power shortage" and "electrical emergency repair" is revealed. The visual design for both first-level and second-level remains consistent.

By directly double-clicking on "Outdoor work protection", experts can choose to show or hide the relevant news in the *spatial view*, *news glyph view* and *news list*, enabling them to focus on specific aspects and in-depth analysis. Furthermore, when the mouse hovers over a topic, the associated bar (Fig. 2-A1) in *temporal view* and news glyph (Fig. 2-B2) in *news glyph view* will dynamically change color to red.

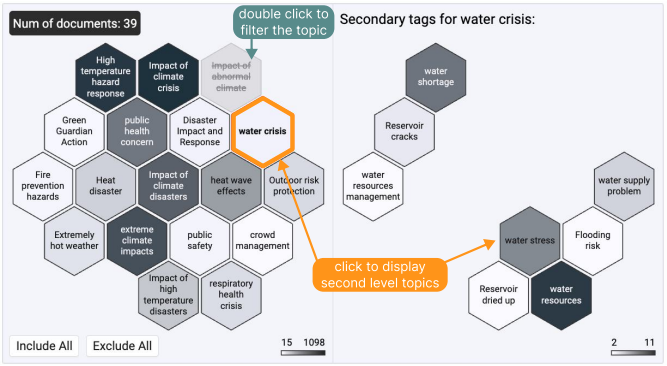


Fig. 4: The news topic view displays the hierarchical topics of the retrieved news articles. Left: the first-level topics. Right: the corresponding second-level topics upon clicking a first-level topic. Furthermore, double-clicking on a specific topic enables the filtering of news articles related to that topic in the subsequent analysis process.

5.4 News Glyph View

The *news glyph view* (Fig. 2-B2) aims to leverage experts' spatial memory for locating and navigating news effectively and enable experts to identify the news of interest quickly (R4) with each news represented as a glyph in a 2D space. In our design, we assume that news articles reporting higher casualties hold greater importance for experts (supported by **D1-2**). Thus, we encode the number of deaths, injuries and impacted individuals into the glyph. By incorporating this information visually, experts can quickly assess the severity and impact of each piece of news. For locating and navigating, they require a reasonable layout of news. To achieve this, each news is first encoded as a high-dimensional vector based on its semantic meaning and reduced to two dimensions using UMAP [50], maintaining semantic similarity. The comparable result of t-SNE can also be found in the supplementary materials. We also used a grid-based method [27] to avoid clutters and keep the relative distance of news glyphs. When the topic selection changes, the corresponding news will be shown or hidden without changing the position to keep the consistency of spatial memory.

We opt for dimension reduction and glyph visualization as opposed to ranking visualization methods like lineup [23], based on several considerations. Firstly, news articles encompass a diverse range of topics, and the absence of casualty information does not necessarily diminish their significance (**D2**). By organizing news articles based on their semantic meaning, we ensure that even those without casualty information are not overlooked entirely. Secondly, scalability was taken into account. Our glyph design, coupled with a 2D space, provides an effective solution for accommodating a large number of news articles. Lastly, the utilization of spatial memory aids in navigation when glyphs maintain interrelationships or distances while retaining intra-meaning through magnitude. The spatial relationships allow for better cognitive mapping and recall of specific articles. Thus, under the condition of a large number of news, we prioritized an intuitive and space-efficient design. As a result, we opted for a circular form of glyph to represent each news item for both inter and intra benefits.

Coxcomb glyph. For circular glyph form, we designed a modified version of the classical coxcomb visualization (Fig. 5-A) to represent the numbers of deaths, injuries, and impacts. To address cases where casualty information is absent, we added a grey node in the center. This modification ensures that each news article can be displayed, even if the casualty information is unavailable. The coxcomb glyph consists of three 120° sectors, each distinguished by a different color to represent deaths, injuries, and impacts. The number of casualties within each category is encoded by the length of the sector. By utilizing the coxcomb glyph, experts can easily identify important news that entails severe consequences. The news glyph is interconnected with both the *news list* and *spatial view*. Clicking on a glyph will synchronously center and highlight the corresponding news in the other two views.

Alternative designs. In addition to the coxcomb glyph, we have developed two alternative designs: the target glyph (Fig. 5-B) and the pie glyph (Fig. 5-C). However, there are certain limitations associated with them. The target glyph represents the number of deaths, injuries, and impacts using three individual arcs. It effectively conveys the information. However, a significant limitation is that it is spatially expensive. Regardless of whether a news article contains casualty information or not, the glyph occupies the same large size. The pie glyph employs size to encode the sum of deaths, injuries, and impacts. The angle ratio of the sectors is determined based on their respective numerical ratios. However, the pie glyph has been criticized for its lack of intuitiveness in expressing individual numbers within each category. Experts need to consider both size and angle to comprehend the magnitude of the numbers. Considering these factors, we have chosen the coxcomb glyph as the preferred design option.

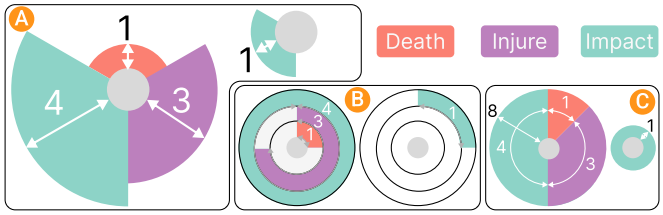


Fig. 5: We opt for the coxcomb glyph design (A) for the news glyph. Alternative design: target glyph design (B) and pie glyph design (C).

5.5 News List View

The *news list* (Fig. 2-B3) lists the headlines of news articles. It provides the functionality for filtering them based on criteria such as time, temperature (which is closely linked to results of domain models), and casualties (R3), ranking the news based on their semantic meaning (R4), and reading original text or structural information (R5). We plot three bar charts to display the number of news corresponding to each criterion. Then experts can directly brush the bar to apply the filters. The three bars are synchronized. When filtering is applied, a blue color emerges in all bar charts to indicate the number of remaining news.

Experts can use sentences to search and rank news so that they can easily access the news with similar semantic meanings. To get the details, they can expand the headlines to structural information or full-text. Additionally, visual cues are used to highlight relevant sentences pertaining to "risk," "cause," "consequence," and "advice" within full text, which improves the efficiency of reading original text (Fig. 2-B3).

5.6 Summary Panel

The summary panel was designed to integrate insights of both numeric and semantic to facilitate further decisions to minimize the losses of potential heat risk (R1, R2, R5, R6). To achieve this integration, we incorporate insights from one side to the other.

Numeric and semantic for numeric. To provide numerical understanding, the line chart (Fig. 2-C1) illustrates relationships between city risks and temperatures. Experts can read the numerical risk level for decision-making. We also select news with the highest number of deaths at each temperature as representative examples. This approach allows experts to match impacts with different risk levels, contextualizing a mere number. These representative news articles are plotted as scatter points on the x-axis, which show a tooltip when hovered.

Semantic and numeric for semantic. Any semantic insights found during the entire process can be pinned to this panel for the purpose of reviewing findings and generating the final report. In order to enhance the comprehension of risk, we implement a contextual question-answering interface (Fig. 2-C2) that helps experts pose contextual questions. Contextual answers will be generated based on selected news from the *news list* using RAG, which can help with the hallucination problem of LLM. To address the hallucination problem further from the visual analytic perspective, we link the answer generated by LLM to the source news that LLM references (Fig. 2-C2). By clicking the reference, experts can effortlessly access the original news article. The key results derived from the numerical model are rephrased in natural language for experts

to review. They are of utmost importance as they serve as references for generating the final report as well. By curating and maintaining the content within this panel, a comprehensive final report can be generated using LLM. The report becomes a valuable reference for experts working on future actionable plans.

6 Evaluation

For evaluation, we collaborated with experts possessing extensive research experience in environmental science. Besides **D1** (professor, 33 years experience) and **D2** (specialist, 4 years experience) who we closely collaborate with, we involved four more experts to conduct the case study. They are **E3** (professor, 10 years experience), **E4** (professor, 9 years experience), **E5** (professor, 10 years experience), and **E6** (the director of the city's observatory, 30 years experience). Regarding our roles [42], we positioned ourselves as fellow researchers exploring the same problem from different perspectives to facilitate open academic feedback from the professors. For the specialist and the observatory director, we presented ourselves as trustworthy, policy-oriented scholars to elicit practical comments and provide actionable solutions.

The experts studied the 2022 China extreme heatwave [74] in Hong Kong using *Havior*. The evaluation consisted of two sessions:

- **Training and Exploration** (60 minutes): We began with a training session where we demonstrated an interesting case to exemplify Havior's functionality and usability. During this period, the experts operated the system under our guidance. Following the demonstration, the experts engaged in a free exploration of the system to analyze the heatwave.
- **Survey and Interview** (30 minutes): We then conducted a survey that included a questionnaire and a semi-structured interview to gather ratings and detailed feedback from the experts.

6.1 Case Study - 2022 China Heatwave: Hong Kong

Hong Kong is an exemplary case to delve into the complexities of heat risks in light of the increasing occurrence of extreme heat events [28].

Analysis of Numerical Climate Data Firstly, the experts choose the index, city and temporal resolution as "temperature," "Hong Kong" and "daily" to check the meteorological condition. To research the 2022 China heatwave, the experts set the date (single click on Fig. 2-A1) to be "2022-07-24" since it has the highest temperature (Fig. 6-A).

The experts commenced their exploration by analyzing the *temporal view* (R1) and *spatial view* (R2), which effectively presented relevant meteorological information (**D1-2, E3-6**). **D2** commented, "*The geographical representations, color coding, and glyph design provide an intuitive and comprehensive means of understanding of the meteorological condition.*" They discovered that the temperature was anticipated to persist at an alarmingly high level of around 31 degrees Celsius, indicating a prolonged period of extreme heat (Fig. 6-A). Furthermore, the entire Greater Bay area was engulfed in high temperatures (Fig. 6-C). However, when they examined the temporal distribution of temperature, they found that temperatures between 26–29 degrees are normal in Hong Kong. This raised uncertainty about the severity of the situation.

Realizing the need for a deeper understanding, the experts investigated probability-based indices. Using the "thermoglyph" (Fig. 2-A4), they gained a holistic understanding of the relationship between temperature and percentile data (**D2, E3-4**). To their surprise, the current percentile for the temperature linked close to the 100th percentile, indicating an unusually severe situation. Seeking further validation, they delved into the detailed information on the return periods. The large return period (Fig. 6-B, D) signifies a low probability of the temperature occurrence, which raises concerns about the potentially severe consequences. The analysis of return periods provides additional support to the conclusion that Hong Kong is confronted with a substantial heat risk (**D1, E3, E5**).

Analysis of Textual News Data The experts turned to the news to enhance their understanding. They conducted a search using the recommended keywords, resulting in the retrieval of 2,246 news articles. By incorporating meteorological variables and news articles (R6), the researchers discovered that the return period yielded more intriguing insights (Fig. 6). The consistency observed between the outcomes derived from heterogeneous climate and news data enhances the rationale for utilizing news with climate data for heat risk research (**D1, E4-6**).

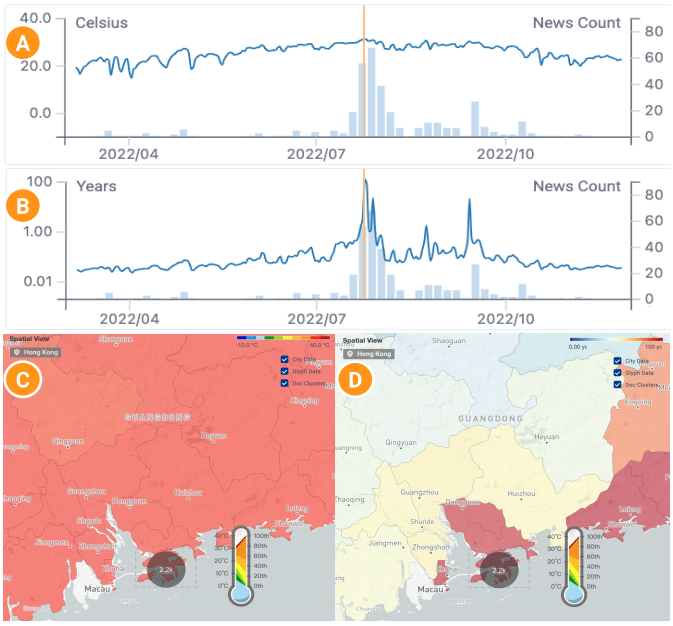


Fig. 6: The temporal trend (A) and the spatial distribution (C) of the temperature on 2022-07-24 in Hong Kong. The temporal trend (B) and the spatial distribution (D) of the return period on 2022-07-24 in Hong Kong. Bars in (A) and (B) are the number of news. Visualizations of the probability-based return period are better suited for studying extreme heat risk, as compared to magnitude-based temperature. The consistency of heterogeneous results (especially in (B)) enhances the interpretability of the system for heat risk research.

To obtain an overview of the potential heat risks (R5), the experts referred to *news topic view* (Fig. 2-B1). The topics of news are automatically generated by using LLM to cluster tags as mentioned in the topic modeling section. These heat risks encompassed various aspects, including well-known topics such as "high temperature hazard response," and "impact of climate crisis." It was not surprising that the majority of news articles were related to these aspects. However, the experts also encountered unexpected topics, such as the "water crisis," which received relatively less attention. They found this feature to be highly beneficial in expanding their awareness of previously unrecognized risk topics. They believed that it would enhance their considerations during decision-making processes, leading to more informed and rational decisions (**D1-2, E5**). **E5** commented, "*The topics found here are not preprogramming, which I think is crucial. When some parameters change or new sources are included, the result can be automatically updated. This is intrinsically natural to pick up new things.*"

To observe the temporal pattern of the number of news, the experts analyzed the bar plot showcasing the number of news, which is integrated into the meteorological panel (Fig. 2-A1). Notably, the number of news articles during the summers of 2018 and 2022 stood out significantly compared to other years. This observation aligned with the 2018 southern China heatwave [12] and the 2022 China heatwave [74]. Consequently, the experts (**E4-5**) could infer that a severe heatwave would likely occur if *Havior* were used in 2022.

Motivated by these findings, the experts aimed to explore how Hong Kong responded to heat risks and identify city-based features by deeply delving into specific topics (R5). This knowledge would enable the development of more appropriate strategies to assist in preparedness and provide guidance for governments and citizens alike (**D1-2, E3-6**). As **E6** remarked, "*The information from historical news is valuable for mitigating risk since we can reference them to make more suitable plans and take more actionable measures.*"

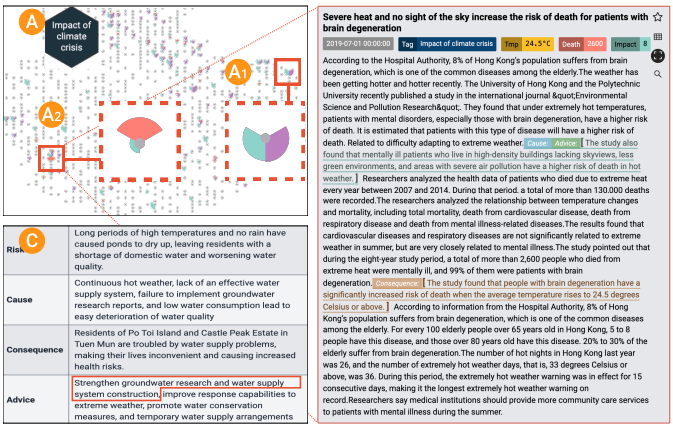


Fig. 7: News glyphs under the topic of "impact of climate crisis" (A). The largest news glyph (A1) and the news glyph with the largest number of deaths (A2) are likely to be selected. The structural information highlighted (B) in the original text helps the experts understand the news quickly. The structural information (C) of the only news with glyph under the topic of "water crisis" helps the experts understand the news easily. The advice generated based on summer's news is similar to what the government has taken [55] in 2018 autumn.

Deep Dive into Specific Topics They explored the most well-known topics first by filtering news articles (R3) under the categories of "impact of climate crisis," excluding others (Fig. 2-B1). Then they utilized the news glyph to identify the largest one (Fig. 7-A1) and clicked on it to access the details (R4). This particular news discussed the Marathon held in high temperatures that resulted in numerous athlete injuries. While severe, this outcome was somewhat expected. The experts bookmarked the news article for future review and pinned important sentences (R4) to the summary panel, contributing to the final report. Subsequently, they came across another news glyph that indicated the highest number of deaths (Fig. 7-A2). With the help of visual cues for reading the full text (Fig. 7-B), they can easily gain insights (R5). E3 pointed out, "*The visualization that maps structural information to the original news text is remarkably neat and useful. You must have dedicated considerable thought to this design.*". They found that this news article highlighted the correlation between high temperatures and increased mortality among mental health patients, particularly in Hong Kong for its lack of sky view and green space. The experts found this insight unexpected, as it would have been challenging to identify this correlation solely based on numerical models.

Exploration of Unexpected Topics The experts proceeded to investigate the unexpected topic, "water crisis" (R5). They ranked the list of news based on the semantic meaning of the selected news (Fig. 2-B3). By examining the structural information of the first few news in the ranked *news list*, the experts quickly discovered that heatwaves can give rise to water supply challenges, even in a prominent city like Hong Kong. By referring to the red proportion depicted in the bar (Fig. 2-A1), they observed that all news articles concerning tap water supply problems were reported solely during the previous heatwave in summer of 2018 (R6). **E6** remarked, *"Very good. I remember the drought in 2018 as the observatory director. It should be June, right? (It is June in the system.) The rainfall that year was extremely limited."* Noting the government is criticized due to the absence of the tap water supply system for remote villages, the experts posed the question for advice and a considerable answer derived based on the 2018 summer news selected (Fig. 2-C2).

Based on the observation that news regarding the unavailability of tap water was only reported during the summer of 2018, the experts inferred that the government had taken measures to tackle the issue. Our subsequent investigation proved this assumption, as evidenced by a government document [55]. Additionally, the measures taken by the government in 2018 autumn are very similar to the advice given by *Havior* (Fig. 2-C2) and structural information (Fig. 7-C). This realization also dispels **E6**'s misconception that the problem been solved due to the rainfall returning to normal after 2018. He is now aware that the government has taken measures as well. It verifies the ability of *Havior* to discover risk topics and facilitate informed decision-making (**D1-2, E3-6**).

Identifying Unresolved Risks On the contrary, some types of disasters were not inadequately handled, such as crop risks due to the heatwave. The experts uncovered another unexpected topic centered around "crop loss" and "crop damage," revealing incidents of crop death and economic loss resulting from unforeseen insect infestations (tropical butterflies have migrated to Hong Kong) triggered by high temperatures and drought (R5). These risks were witnessed in 2018, 2021, and recurred in 2022. The lack of attention and proactive preventive measures towards these issues may explain their persistence. However, specific measures are expected to have a positive impact on the risk. Seeking advice from *Havior*, they received valuable suggestions such as "strengthening pest control," "improving the irrigation system," and "using shade nets to reduce plant heat stress." With *Havior*, they not only identify unresolved heat-related risks but also devise informative strategies to effectively address them (**D1-2, E3-5**).

Summary and Integration of heterogeneous insights To integrate the heterogeneous insights for decision-making (R6), the experts compiled the semantic insights they discovered in the *summary panel*. Alongside these insights, the numerical conclusions were presented. Additionally, the experts examined the representative news related to the current (31°) temperature and identified the need for heightened attention to the risks faced by outdoor workers. By combining the numerical conclusions with the semantic insights from the news, the experts generated a final report on the heat risk in Hong Kong. This report provided them with a comprehensive understanding of the heat risk and helped them to make informed decisions and take appropriate actions (**D1-2, E4-6**). It can be found in the supplementary material. E2 pointed out, "*The convergence of all heterogeneous data and model generates many insights to me. It is efficient to use LLM to integrate and summarize them. The report is a valuable reference for informing subsequent strategy considerations.*"

6.2 Expert Survey and Interview

To evaluate the effectiveness and usability of *Havior*, we adapted a questionnaire for the experts based on the goals and requirements of the system and SUS questionnaire [6]. It includes ten 7-point Likert-scale questions (1 = strongly disagree and 7 = strongly agree) questions and is divided into 2 sections: functionality (assessing usefulness) and usability (evaluating ease of use). After the case study, the six experts were asked to rate *Havior* from various perspectives using each question. The result is shown in Tab. 1.

6.2.1 Survey Result

Overall assessment All the experts highly rated their experience with *Havior* (all mean scores ≥ 6). It is worth noting that the experts are meticulous and cautious in their rating. Particularly, **E6** thought that for some questions he could not rate immediately and preferred to do so after using the system for more days. The high score indicates the effectiveness of *Havior*, showcasing its ability to manage risks.

Functionality The experts are satisfied with the functionalities offered by *Havior*, especially the news panel, which could help them gain different types of insights compared to numerical models, including textual impact and advice (highest mean score and lowest SD). One interesting finding is that while the meteorological panel is appreciated by the experts (mean = 6.33), the expert with the longest tenure in environmental research (33 years) assigns the lowest score (score = 4) to the meteorological panel. Despite our efforts to simplify meteorological visualizations, he perceives them as somewhat complex and divergent from his accustomed charts. It indicates that the experts in the domain exhibit a certain "inertia" regarding visualization and highlight areas where we can further enhance.

Usability It was evident that the experts highly valued the visual designs (mean = 6.33) and interactions (mean = 6.50) of *Havior*, perceiving them as intuitive. Furthermore, they found *Havior* to be easy to learn and operate, with a mean score of 6.00 in this regard. As a result,

Table 1: Ratings on *Havior* on a 7-point Likert scale (with 1 = strongly disagree and 7 = strongly agree). Questions 1-6 relate to the functionality, and 7-10 relate to the usability of *Havior*.

	Ratings	Mean	SD
1	The meteorological panel helps me understand heat from the quantitative perspective.	6.33	1.21
2	The news panel helps me retrieve the news of interest.	6.17	0.75
3	The news panel helps me manage the news of interest.	6.00	0.89
4	The news panel helps me understand the impact, reasons, and advice of the heat risk.	6.50	0.84
5	The summary panel helps me better understand the risk with the integration of numeric and news.	6.00	0.89
6	The system has the potential to help me make better decisions for heat risk if I am a decision-maker for the city's policy.	6.00	1.26
7	The visual designs in the system are intuitive.	6.33	0.52
8	The interactions in the system are intuitive.	6.50	0.55
9	It was easy to learn the system with the demonstration and training session.	6.00	0.63
10	I will use this system again.	7.00	0

all the experts gave a full score (mean = 7.00, SD = 0) for their intention to use *Havior* in the future. The result not only underscores the effectiveness of *Havior*, but also indicates a strong inclination towards its adoption for real-world impacts.

6.2.2 Implication

Following the survey, we conducted semi-structured interviews with the domain experts to gather deeper insights into Havior's impact and potential for future developments. Their feedback provided valuable guidance for future research directions in visualization and pipeline design for heat risk management and related fields.

Yesterday's news informs tomorrow's risks. The utility of incorporating historical news for dealing with unprecedented extreme events has sparked discussions among both our team and domain experts. A consensus emerged that while historical insights may have limitations when applied to new or intensely amplified risk scenarios, they can still aid in identifying risks with greater magnitude and tracking unresolved risks. **E5** believed that "there's nothing new under the sun," suggesting that new heat extremes still echo those risks in the news. He expressed that identifying existing risks helps *"extend my scope of consideration by imagining how they scale.*" The LLM's risk summaries enable experts to anticipate and develop preemptive strategies for more severe extremes. The recurring risks in the news indicate unresolved problems that need more attention, as in the first case study. On the other hand, **D1** praised the system's capability to draw connections between seemingly irrelevant events and reveal new compounded risks. This analysis, spanning various spatial and temporal dimensions, alerts experts to emerging risks and suggests proactive measures. **D1** noted that the significant changes brought by several climate events highlight the potential for risks to converge and intensify, like a chain reaction, eventually creating more significant issues. He further elaborated, *"These risks have been hidden and cannot be foreseen, but with your system, we now know that they exist and how they evolve."*

Non-numerical factors in decision-making. In the past, the inclusion of non-numerical factors in climate analysis and heat risk management was hindered by the challenge of modeling these unstructured and non-numeric elements, which stand in stark contrast to the domain's usual numerical models. However, our LLM-empowered pipeline represents significant progress, enabling the seamless incorporation of these factors and their contextualization with the climate model results. **E5** expressed, "*By analyzing highly relevant news, the risk is no longer an abstract number to me. These multifaceted insights help me make informed decisions.*" Our findings reveal that the extensive documentation (*e.g.*, in news and official documents) is now accessible for various critical applications, enhanced by the efficiency of LLMs and the effectiveness of visual analytics.

Applicability and Generalizability The experts specifically appreciated the vivid "thermoglyph" and the valuable insights they obtained through interacting with *Havior*. In particular, **E4** expressed the satisfaction of effectiveness with *Havior* and complained about the low efficiency and uninspired results of the manual investigation in their original workflow. **D1** and **E6** expected the collaboration of the implementation of *Havior* at the city's observatory to introduce impactful value. For generalizability, *Havior* provides a novel and feasible pipeline for integrating numeric results and textual insights. The text is not limited to news, other documents, such as government reports, can seamlessly integrate into the pipeline of *Havior*. *Havior* can serve as a source of inspiration for tasks in other domains hurdled by heterogeneous data, especially numeric and textual.

7 Limitations and Future Work

While *Havior* demonstrates significant potential in enhancing heat risk management, there are several limitations that need to be addressed to further improve its effectiveness and usability.

The limitation of novel visualization design is that while new visualizations are beneficial, domain experts may exhibit "inertia." We stroke for simple, yet efficient visualizations recommended by experts during the collaboration. However, experts' feedback still indicates that the visual design is helpful but requires time to be adapted. **E3** commented, "*I have never seen visualizations like the 'thermoglyph' before. However, after a brief introduction, I understand its efficiency.*" The lesson we learned is the importance of considering the prevalent visualization used in the domain and the underlying reasons. This awareness enables us to design the most suitable visualizations that minimize the transition cost.

The limitations in LLM, like accuracy, hallucination, robustness, domain expertise, and timeliness, also affect *Havior*'s efficiency in providing precise analysis and up-to-date advice. During the design phase, we took into account the hallucination problem associated with LLM. To enhance the reliability of LLM, we implemented techniques such as RAG, human-in-the-loop analytics, and reference-based Q&A (Fig. 2-C2) to improve the reliability of LLM. As insights are generated during the analytical process and experts have direct access to the source news, the reliability has significantly improved, as noted by experts. **E6** commented, "*I think the human-in-the-loop exploration process is better than automatically generating a report that I may not trust entirely. It allows me to verify insights in detail by examining the source directly if necessary.*" However, we still find some inaccurate instances like "real estate market overheating" from heat risks. Our next step is to refine a domain-specific LLM, aiming for a deeper comprehension of heat-related risks.

Visual scalability issue arises when the news glyph view displays the topic with a large volume of news articles. To address this issue, we have designed a modified coxcomb glyph and implemented a grid-based algorithm to alleviate the problem of visual clutter. In future research, the exploration of more advanced visualization techniques and filtering methods holds the potential to further mitigate this limitation.

Unexplored modalities, such as crowd-sourced damage photos [3], satellite imagery and video footage about disasters suggested by **E6**, also contain information on risk events. We seek to incorporate multimodality capabilities in *Havior* for risk management.

8 Conclusion

In this study, we have undertaken the characterization of the risk management problem, with a specific focus on the integration challenges posed by the heterogeneity of numerical results from domain models and risk insights derived from news sources. We then developed *Havior*, an LLM-empowered VA system, guided by the domain-characterized requirements. *Havior* aims to enhance the analysis of heat risk, improve heat risk management strategies, and mitigate heat-related threats. The evaluation involved conducting a case study, followed by an expert survey and an interview with six experts. Their positive feedback and insights serve as evidence of the usefulness and efficiency of *Havior*. Significantly, the evaluation results demonstrate the potential for integrating quantitative model results with heterogeneous insights of risk derived from news reports to enhance the management of heat risks.

[3] https://www.hko.gov.hk/en/cwsrc/index_mangkhut.html

REFERENCES

[1] G. Accarino, D. Elia, D. Donno, F. Immorlano, and G. Aloisio. A machine learning-powered digital twin for extreme weather events analysis. Technical report, Copernicus Meetings, 2023. doi: 10.5194/egusphere-egu23-6060 1

[2] B. G. Anderson and M. L. Bell. Weather-related mortality: how heat, cold, and heat waves affect mortality in the united states. *Epidemiology*, 20(2):205, 2009. doi: 10.1097/EDE.0b013e318190ee08 3

[3] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023. doi: 10.1038/s41586-023-06185-3 3

[4] A. Biswas, G. Lin, X. Liu, and H.-W. Shen. Visualization of time-varying weather ensembles across multiple resolutions. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):841–850, 2017. doi: 10.1109/TVCG.2016.2598869 2

[5] A. Blok and M. A. Pedersen. Complementary social science? quali-quantitative experiments in a big data world. *Big Data & Society*, 1(2):2053951714543908, 2014. doi: 10.1177/2053951714543908 2

[6] J. Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2), 2013. 8

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. doi: doi.org/10.48550/arXiv.2005.14165 3

[8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 143–152, 2012. doi: 10.1109/VAST.2012.6400557 2

[9] C. A. T. Cortes et al. Analysis of wildfire visualization systems for research and training: Are they up for the challenge of the current state of wildfires? *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–20, 2023. doi: 10.1109/TVCG.2023.3258440 2

[10] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024. 1

[11] C. V. F. de Souza et al. Prowis: A visual approach for building, managing, and analyzing weather simulation ensembles at runtime. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):738–747, 2024. doi: 10.1109/TVCG.2023.3326514 2

[12] K. Deng, S. Yang, D. Gu, A. Lin, and C. Li. Record-breaking heat wave in southern china and delayed onset of south china sea summer monsoon driven by the pacific subtropical high. *Climate dynamics*, 54:3751–3764, 2020. doi: 10.1007/s00382-020-05203-8 7

[13] Z. Deng et al. Airvis: Visual analytics of air pollution propagation. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):800–810, 2020. doi: 10.1109/TVCG.2019.2934670 2

[14] A. G. Dias, E. E. Milios, and M. C. F. de Oliveira. Trivir: A visualization system to support document retrieval with high recall. In *Proc. DocEng*. ACM, New York, 2019. doi: 10.1145/3342558.3345401 2

[15] K. L. Ebi et al. Hot weather and heat extremes: health risks. *The Lancet*, 398(10301):698–708, 2021. doi: 10.1016/S0140-6736(21)01208-3 1

[16] ECMWF. 2 m temperature and 30 m wind. `https://charts.ecmwf.int/products/medium-2mt-wind30/`, 2024. [Online; Accessed 15-January-2024]. 4

[17] T. Feng, J. Yang, M.-C. Eppes, Z. Yang, and F. Moser. Evis: Visually analyzing environmentally driven events. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):912–921, 2022. doi: 10.1109/TVCG.2021.3114867 2

[18] Y. Feng, X. Wang, W. Kam-Kwai, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen. Promptmagician: Interactive prompt engineering for text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):295–305, 2024. doi: 10.1109/TVCG.2023.3327168 2

[19] Y. Feng, X. Wang, B. Pan, W. Kam-Kwai, Y. Ren, S. Liu, Z. Yan, Y. Ma, H. Qu, and W. Chen. Xnli: Explaining and diagnosing nli-based visual data analysis. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–14, 2023. doi: 10.1109/TVCG.2023.3240003 2

[20] A. Gasparrini et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The lancet*, 386(9991):369–375, 2015. doi: 10.1016/S0140-6736(14)62114-0 3

[21] E. Gomez-Nieto et al. Similarity preserving snippet-based visualization of web search results. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):457–470, 2014. doi: 10.1109/TVCG.2013.242 2

[22] K. Gorro, J. R. Ancheta, K. Capao, N. Oco, R. E. Roxas, M. J. Sabellano, B. Nonnecke, S. Mohanty, C. Crittenden, and K. Goldberg. Qualitative data analysis of disaster risk reduction suggestions assisted by topic modeling and word2vec. In *2017 International Conference on Asian Language Processing (IALP)*, pp. 293–297, 2017. doi: 10.1109/IALP.2017.8300601 2

[23] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013. doi: 10.1109/TVCG.2013.173 6

[24] Y. Guo et al. Heat wave and mortality: A multicountry, multicommunity study. *Environmental Health Perspectives*, 125(8):087006, 2017. doi: 10.1289/EHP1026 4

[25] S. Hajat, S. C. Sheridan, M. J. Allen, M. Pascal, K. Laaidi, A. Yagouti, U. Bickis, A. Tobias, D. Bourque, B. G. Armstrong, and T. Kosatsky. Heat–health warning systems: A comparison of the predictive capacity of different approaches to identifying dangerously hot days. *American Journal of Public Health*, 100(6):1137–1144, 2010. PMID: 20395585. doi: 10.2105/AJPH.2009.169748 1

[26] H. Hersbach et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803 1, 3

[27] G. M. Hilasaca, W. E. Marcílio-Jr, D. M. Eler, R. M. Martins, and F. V. Paulovich. A grid-based method for removing overlaps of dimensionality reduction scatterplot layouts. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–14, 2023. doi: 10.1109/TVCG.2023.3309941 6

[28] J. Hua, X. Zhang, C. Ren, Y. Shi, and T.-C. Lee. Spatiotemporal assessment of extreme heat risk for high-density cities: A case study of Hong Kong from 2006 to 2016. *Sustainable Cities and Society*, 64:102507, 2021. doi: 10.1016/j.scs.2020.102507 1, 7

[29] R. Y. Huang and C. R. Small. Cafellm: Context-aware fine-grained semantic clustering using large language models. In *Generalizing from Limited Resources in the Open World*, pp. 66–81. Springer Nature Singapore, Singapore, 2024. 4

[30] N. Ingulfsen, S. Schaub-Meyer, M. Gross, and T. Günther. News globe: Visualization of geolocalized news articles. *IEEE Computer Graphics and Applications*, 42(4):40–51, 2022. doi: 10.1109/MCG.2021.3127434 2

[31] P. Isenberg, A. Bezerianos, P. Dragicevic, and J.-D. Fekete. A study on dual-scale data charts. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2469–2478, 2011. doi: 10.1109/TVCG.2011.160 4

[32] Y. Jiang, K. Yang, C. Shao, X. Zhou, L. Zhao, Y. Chen, and H. Wu. A downscaling approach for constructing high-resolution precipitation dataset over the tibetan plateau from era5 reanalysis. *Atmospheric Research*, 256:105574, 2021. doi: 10.1016/j.atmosres.2021.105574 1

[33] P. Jyoteeshkumar reddy, S. E. Perkins-Kirkpatrick, and J. J. Sharples. Intensifying Australian Heatwave Trends and Their Sensitivity to Observational Data. *Earth's Future*, 9(4):e2020EF001924, 2021. doi: 10.1029/2020EF001924 1

[34] W. Kam-Kwai, Y. Luo, X. Yue, W. Chen, and H. Qu. Prismatic: Interactive multi-view cluster analysis of concept stocks. *arXiv preprint arXiv:2304.05011*, 2024. doi: 10.48550/arXiv.2402.08978 2

[35] C. P. Kappe, M. Böttinger, and H. Leitte. Exploring variability within ensembles of decadal climate predictions. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1499–1512, 2019. doi: 10.1109/TVCG.2018.2810919 2

[36] H. Kunreuther, G. Heal, M. Allen, O. Edenhofer, C. B. Field, and G. Yohe. Risk management and climate change. *Nature climate change*, 3(5):447–450, 2013. doi: 10.1038/nclimate1740 1, 3

[37] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 591–600. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1772690.1772751 2

[38] W. Lass, A. Haas, J. Hinkel, and C. Jaeger. Avoiding the avoidable: Towards a European heat waves risk governance. *International journal of disaster risk science*, 2:1–14, 2011. doi: 10.1007/s13753-011-0001-z 3

[39] S. Latif, S. Chen, and F. Beck. A deeper understanding of visualization-text interplay in geographic data-driven stories. *Computer Graphics Forum*, 40(3):311–322, 2021. doi: 10.1111/cgf.14309 2

[40] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2022. doi: 10.

1109/TVCG.2021.3114802 2

[41] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proc. NeurIPS*. Curran Associates Inc., New York, 2020. doi: 10.48550/arXiv.2005.11401 1

[42] L. Li. How to tackle variations in elite interviews: Access, strategies, and power dynamics. *Qualitative Research*, 22(6):846–861, 2022. doi: 10.1177/1468794121994475 7

[43] D. Liu, K. Veeramachaneni, A. Geiger, V. O. K. Li, and H. Qu. Aqeyes: Visual analytics for anomaly detection and examination of air quality data. *CoRR*, abs/2103.12910, 2021. 2

[44] J. Liu et al. Mortality burden attributable to high and low ambient temperatures in China and its provinces: results from the global burden of disease study 2019. *The Lancet Regional Health–Western Pacific*, 24, 2022. doi: 10.1016/j.lanwpc.2022.100493 3

[45] S. Liu et al. Bridging text visualization and mining: A task-driven survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2482–2504, 2019. doi: 10.1109/TVCG.2018.2834341 2

[46] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 359–367, 2011. 2

[47] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 181–190, 2011. doi: 10.1109/VAST.2011.6102456 2

[48] C. J. Maller and Y. Strengers. Housing, heat stress and health in a changing climate: promoting the adaptive capacity of vulnerable households, a suggested way forward. *Health Promotion International*, 26(4):492–498, 02 2011. doi: 10.1093/heapro/dar003 1

[49] S. Mathew, B. Zeng, K. K. Zander, and R. K. Singh. Exploring agricultural development and climate adaptation in northern australia under climatic risks. *The Rangeland Journal*, 40(4):353–364, 2018. doi: 10.1071/RJ18011 1

[50] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426 5, 6

[51] B. McNicholl, Y. H. Lee, A. G. Campbell, and S. Dev. Evaluating the reliability of air temperature from era5 reanalysis data. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi: 10.1109/LGRS.2021.3137643 1

[52] L. Nan, E. Zhang, W. Zou, Y. Zhao, W. Zhou, and A. Cohan. On evaluating the integration of reasoning and action in LLM agents with database question answering. *arXiv preprint arXiv:2311.09721*, 2023. doi: 10.48550/arXiv.2311.09721 1

[53] A.-A.-R. Nayeem, H. Lee, D. Han, M. Elshambakey, W. J. Tolone, T. Dobbs, D. Crichton, and I. Cho. Dcpviz: A visual analytics approach for downscaled climate projections. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 291–300, 2022. doi: 10.1109/BigData55660.2022.10020827 2

[54] A. Noviello et al. Guiding environmental messaging by quantifying the effect of extreme weather events on public discourse surrounding anthropogenic climate change. *Weather, Climate, and Society*, 15(1):17 – 30, 2023. doi: 10.1175/WCAS-D-22-0053.1 2

[55] S. D. C. S. (Oct). Remote rural water supply plan - shatian meilin village. https://www.districtcouncils.gov.hk/st/doc/2016_2019/sc/committee_meetings_doc/DHC/13889/st_dhc_2018_047_tc.pdf, 2018. Publication ID: DH 47 /201 8. 8

[56] M. Oppenheimer, M. Campos, R. Warren, J. Birkmann, G. Luber, B. O'Neill, K. Takahashi, M. Brklacich, S. Semenov, R. Licker, et al. Emergent risks and key vulnerabilities. In *Climate change 2014 impacts, adaptation and vulnerability: part a: global and sectoral aspects*, pp. 1039–1100. Cambridge University Press, 2015. 2

[57] L. A. Parsons, D. Shindell, M. Tigchelaar, Y. Zhang, and J. T. Spector. Increased labor losses and decreased adaptation potential in a warmer world. *Nature communications*, 12(1):7286, 2021. doi: 10.1038/s41467-021-27328-y 1

[58] W. C. R. Programme. https://www.wcrp-climate.org/my-climate-risk. https://www.wcrp-climate.org/my-climate-risk, 2021. [Online; Accessed 09-January-2024]. 2

[59] R. Qiu, Y. Tu, Y.-S. Wang, P.-Y. Yen, and H.-W. Shen. Docflow: A visual analytics system for question-based document retrieval and categorization. *IEEE Transactions on Visualization and Computer Graphics*, 30(2):1533–1548, 2024. doi: 10.1109/TVCG.2022.3219762 2

[60] M. Rautenhaus et al. Visualization in meteorology—a survey of techniques and tools for data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics*, 24(12):3268–3296, 2018. doi: 10.1109/TVCG.2017.2779501 2

[61] M. Roberts. What are the heat exhaustion and heatstroke symptoms? https://www.bbc.com/news/health-62120167, 2023. [Online; Accessed 08-January-2024]. 1, 2

[62] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021. 4

[63] N. P. Simpson et al. A framework for complex climate change risk assessment. *One Earth*, 4(4):489–501, 2021. doi: 10.1016/j.oneear.2021.03.005 1

[64] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021. Curran Associates, Inc., 2020. 2

[65] L. Styve, C. Navarra, J. M. Petersen, T.-S. Neset, and K. Vrotsou. A visual analytics pipeline for the identification and exploration of extreme weather events from social media data. *Climate*, 10(11), 2022. doi: 10.3390/cli10110174 2

[66] D. Thom, R. Krüger, and T. Ertl. Can twitter save lives? a broad-scale study on visual social media analytics for public safety. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1816–1829, 2016. doi: 10.1109/TVCG.2015.2511733 2

[67] Wisers. Wisers: best media monitoring. https://login.wisers.net/, 2024. [Online; Accessed 01-January-2024]. 3

[68] Z. Xu, J. Cheng, W. Hu, and S. Tong. Heatwave and health events: A systematic evaluation of different temperature indicators, heatwave intensities and durations. *The Science of the total environment*, 630:679–689, 02 2018. doi: 10.1016/j.scitotenv.2018.02.268 3

[69] J. Yang et al. Projecting heat-related excess mortality under climate change scenarios in China. *Nature communications*, 12(1):1039, 2021. doi: 10.1038/s41467-021-21305-1 1

[70] W. Yi and A. P. Chan. Effects of temperature on mortality in Hong Kong: a time series analysis. *International journal of biometeorology*, 59:927–936, 2015. doi: 10.1007/s00484-014-0895-4 1

[71] F. Yuan, M. Li, and R. Liu. Understanding the evolutions of public responses using social media: Hurricane matthew case study. *International Journal of Disaster Risk Reduction*, 51:101798, 2020. doi: 10.1016/j.ijdrr.2020.101798 2

[72] W. Zhang, W. Kam-Kwai, Y. Chen, A. Jia, L. Wang, J.-W. Zhang, L. Cheng, and W. Chen. ScrollTimes: Tracing the provenance of paintings as a window into history. *IEEE Transactions on Visualization and Computer Graphics*, 2024. To appear. doi: 10.48550/arXiv.2306.08834 2

[73] M. Zhizhin, E. Kihn, V. Lyutsarev, S. Berezin, A. Poyda, D. Mishin, D. Medvedev, and D. Voitsekhovsky. Environmental scenario search and visualization. In *Proc. GIS*, pp. 1–10. ACM, New York, Nov. 2007. doi: 10.1145/1341012.1341047 2

[74] B. Zhou, S. Hu, J. Peng, D. Li, L. Ma, Z. Zheng, and G. Feng. The extreme heat wave in China in August 2022 related to extreme northward movement of the eastern center of SAH. *Atmospheric Research*, 293:106918, 2023. doi: 10.1016/j.atmosres.2023.106918 7

[75] J. Zou, N. Lu, H. Jiang, J. Qin, L. Yao, Y. Xin, and F. Su. Performance of air temperature from era5-land reanalysis in coastal urban agglomeration of southeast china. *Science of The Total Environment*, 828:154459, 2022. doi: 10.1016/j.scitotenv.2022.154459 1

[76] J. Zuo, S. Pullen, J. Palmer, H. Bennetts, N. Chileshe, and T. Ma. Impacts of heat waves and corresponding measures: a review. *Journal of Cleaner Production*, 92:1–12, 2015. doi: 10.1016/j.jclepro.2014.12.078 2