International Journal of
Geo-Information

MDPI

Article

A Question and Answering Service of Typhoon Disasters Based on the T5 Large Language Model

Yongqi Xia [1], Yi Huang [1,2,3,*], Qianqian Qiu [4], Xueying Zhang [2,3], Lizhi Miao [1] and Yixiang Chen [1]

[1] School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; b21080626@njupt.edu.cn (Y.X.); miaolz@njupt.edu.cn (L.M.); chenyixiang@njupt.edu.cn (Y.C.)
[2] Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210023, China; zhangxueying@njnu.edu.cn
[3] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[4] Jiangsu Province Surveying & Mapping Engineering Institute, Nanjing 210019, China; b21080619@njupt.edu.cn
* Correspondence: huangyi@njupt.edu.cn

Abstract: A typhoon disaster is a common meteorological disaster that seriously impacts natural ecology, social economy, and even human sustainable development. It is crucial to access the typhoon disaster information, and the corresponding disaster prevention and reduction strategies. However, traditional question and answering (Q&A) methods exhibit shortcomings like low information retrieval efficiency and poor interactivity. This makes it difficult to satisfy users' demands for obtaining accurate information. Consequently, this work proposes a typhoon disaster knowledge Q&A approach based on LLM (T5). This method integrates two technical paradigms of domain fine-tuning and retrieval-augmented generation (RAG) to optimize user interaction experience and improve the precision of disaster information retrieval. The process specifically includes the following steps. First, this study selects information about typhoon disasters from open-source databases, such as Baidu Encyclopedia and Wikipedia. Utilizing techniques such as slicing and masked language modeling, we generate a training set and 2204 Q&A pairs specifically focused on typhoon disaster knowledge. Second, we continuously pretrain the T5 model using the training set. This process involves encoding typhoon knowledge as parameters in the neural network's weights and fine-tuning the pretrained model with Q&A pairs to adapt the T5 model for downstream Q&A tasks. Third, when responding to user queries, we retrieve passages from external knowledge bases semantically similar to the queries to enhance the prompts. This action further improves the response quality of the fine-tuned model. Finally, we evaluate the constructed typhoon agent (Typhoon-T5) using different similarity-matching approaches. Furthermore, the method proposed in this work lays the foundation for the cross-integration of large language models with disaster information. It is expected to promote the further development of GeoAI.

Keywords: typhoon disaster; question and answering; large language models; information retrieval

check for updates

Citation: Xia, Y.; Huang, Y.; Qiu, Q.; Zhang, X.; Miao, L.; Chen, Y. A Question and Answering Service of Typhoon Disasters Based on the T5 Large Language Model. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 165. https://doi.org/10.3390/ijgi13050165

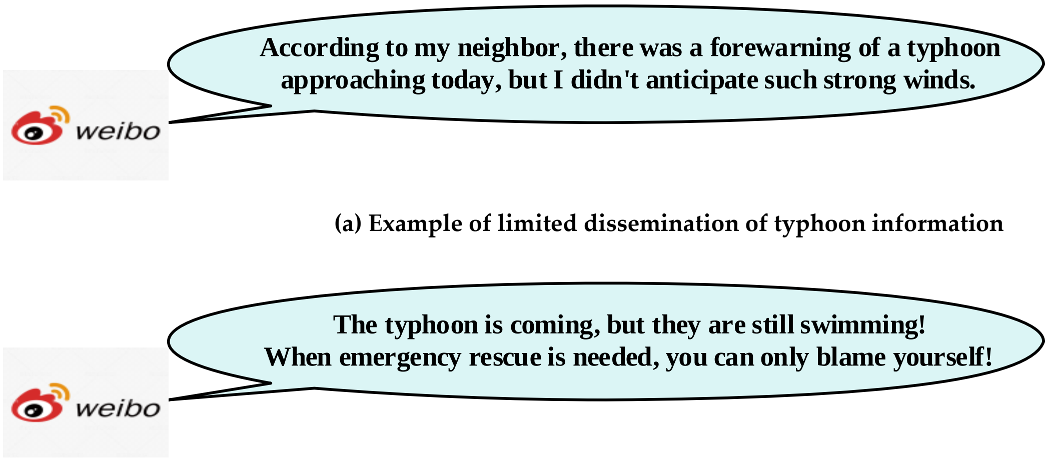
Academic Editors: Peng Peng, Shu Wang, Maryam Lotfian, Feng Lu, Yunqiang Zhu and Wolfgang Kainz

Received: 20 March 2024
Revised: 8 May 2024
Accepted: 11 May 2024
Published: 14 May 2024

1. Introduction

Natural disasters profoundly affect human productivity and lifestyles, which presents a significant challenge in disaster prevention and mitigation for humanity. With global disaster-related losses escalating, this issue has become increasingly urgent [1]. According to a report published by the United Nations Office for Disaster Risk Reduction (UNDRR) on 13 October 2016, over the last two decades, the world has witnessed more than 7000 disasters, which have claimed the lives of over 1.35 million people and resulted in annual economic losses of up to USD 300 billion [2,3]. As disaster threats grow, timely and effective access to disaster information and thorough research into the patterns of various disasters are crucial for enhancing disaster prevention measures and reducing disaster risks.

In the environment of big data, people are increasingly keen to use the internet to acquire knowledge about disasters [4–7]. At present, individuals commonly employ search engines to query relevant web pages and gather the required information [8,9]. Despite the efforts of various search engines to satisfy information retrieval needs, users are still limited to keyword searches and must sift through numerous search results. Furthermore, the complexity of disaster risks, coupled with the limitations of individual cognition, strengthens the public's challenge in accurately understanding natural disaster risks. The two examples shown in Figure 1 illustrate the above situation. First, the channels through which the public receives typhoon information are limited and often unreliable. As Figure 1a shows, even though there is a wealth of data and forecasts on typhoons, this information may not be widely disseminated or accessible to the public. Consequently, it is difficult for community members to obtain accurate and authoritative disaster information. Second, despite the existence of disaster prevention and mitigation measures, along with science popularization materials, insufficient disaster cognition still hinders accurate judgment during actual disasters. As Figure 1b shows, communities often fail to assess risks accurately and take appropriate actions in the face of disasters. Therefore, when a disaster approaches, it is first necessary to organize and sort disaster information orderly to ensure accuracy and authority. Furthermore, it is crucial to efficiently communicate disaster information to the public and remind them to take timely preventive preparations.


According to my neighbor, there was a forewarning of a typhoon approaching today, but I didn't anticipate such strong winds.
The typhoon is coming, but they are still swimming! When emergency rescue is needed, you can only blame yourself!

(a) Example of limited dissemination of typhoon information

(b) Example of inadequate public response measures

Figure 1. Two examples (**a**,**b**) illustrating the limits of public cognition from a Chinese social network platform.

With the advent of the big data era, users' needs have become more complex and diversified. Traditional methods of information search can no longer fully meet their needs. Q&A systems, as one of the vital solutions to this problem, primarily retrieve and analyze existing data, ultimately returning the answer and other relevant information [10,11]. Compared to traditional search engines, Q&A systems offer several significant advantages. First, they can understand user queries more accurately by employing semantic understanding techniques for information matching and inference. These techniques enable them to provide more precise and targeted responses. Second, Q&A systems can integrate multiple sources of information, including structured data, unstructured text, and knowledge graphs, to provide comprehensive and multidimensional solutions to user queries. Additionally, they support conversational interaction, which enables more natural and intelligent communication with users. This capability enhances user experience and satisfaction. In recent years, the emergence of generative large language models, the development of pretraining techniques, and the advancement of cloud computing technology have made the construction of intelligent Q&A systems based on LLMs a greater possibility for typhoon disaster emergency rescue [12–14]. In particular, generative large language models, represented by ChatGPT, have achieved remarkable success in this domain. These models are pretrained on large-scale text data to parse user questions and generate user-satisfying responses. Despite their strength, these models still struggle in vertical domains such as geoscience. They often exhibit hallucination issues that lead to decreased credibility. For example, when a user queries "Where did Typhoon 'In-Fa' land?", LLMs responses might present varied incorrect answers, such as "Typhoon 'In-Fa' landed in Fujian Province, China" or "Typhoon 'In-Fa' landed in Guangdong Province, China". To avoid generating incorrect answers that might mislead the public and lead to disastrous outcomes, this study uses textual knowledge from open-source encyclopedias like Wikipedia and Baidu Encyclopedia to create a corpus related to typhoon disaster information and format it into a training set. To fully leverage this knowledge to address typhoon-related questions, we first continuously pretrain the model on the training set. Typhoon disaster events are stored in the weights of the model in the form of parameterized information, which allows in-depth learning of typhoon disaster issues. Subsequently, the fine-tuning of the pretrained model is employed to enhance its performance on downstream typhoon disaster Q&A tasks. Considering that fine-tuning stores knowledge implicitly, which hinders response interpretation and knowledge traceability, we adopt the RAG technology to retrieve explicit knowledge stored in external knowledge sources [15–17]. This explicit knowledge has a certain semantic structure. It is expressed and stored in a specific organizational format, which constitutes structured knowledge. Such knowledge can be easily parsed and comprehended, thereby facilitating the completion of RAG tasks. Specifically, retrieval involves using the user's query to access relevant content from the typhoon knowledge base, enhancement includes incorporating the user's query and retrieved events into a prompt, and generation then feeds this enhanced prompt into the fine-tuned LLMs to produce the final typhoon response. As shown in Figure 2, through the combined use of fine-tuning and RAG, this approach can improve the model's response quality to typhoon disaster information. Thereby, it addresses the difficulty that users encounter in accurately judging the intensity of typhoon disasters.

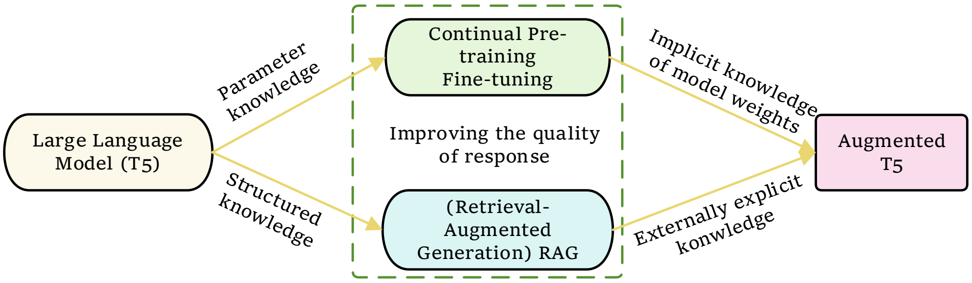

Large Language Model (T5)
Parameter knowledge
Structured knowledge
Continual Pre-training Fine-tuning
Improving the quality of response
(Retrieval-Augmented Generation) RAG
Implicit knowledge of model weights
Externally explicit konwledge
Augmented T5

Figure 2. A combination of fine-tuning and RAG.

The remainder of this work is structured as follows. Section 2 reviews relevant works on rescue efforts guided by typhoon disaster knowledge. Section 3 provides a detailed introduction to the methodology we adopted, while discussing the integration of typhoon information knowledge with the T5 model. Section 4 introduces the experimental design process aimed at evaluating the differential response outcomes between our method and others. Section 5 discusses the experimental results of this study and provides some analysis. Finally, Section 6 concludes this research.

2. Related Works

Wikipedia and Baidu Encyclopedia, known as the most widely used free online encyclopedias, offer extensive knowledge on typhoons and disaster prevention and mitigation fundamentals [18,19]. Furthermore, typhoon forecasts from major news websites and meteorological platforms have attracted significant attention. This knowledge is utilized for disaster response and situational research [20–22]. However, the vast volume of data from encyclopedic and news sources complicates effective processing and analysis. This increases challenges for emergency management personnel during rescue operations. Consequently, numerous studies aim to reduce data volume while preserving information quality to enhance the user-friendliness of encyclopedic data. Current research focuses on flood area identification [23], drought risk management [24], fire spread analysis [25], earthquake assessment [26], and typhoon disaster evaluation [27,28]. These studies concentrate on real-time monitoring, loss assessment, and spatiotemporal analysis of natural disasters, among other aspects. Typhoon research encompasses classifying disaster information, analyzing specific typhoons [29,30], tracking various stages of typhoon development, assessing the impact of geographical factors on typhoon formation, and conducting sentiment analysis on typhoon discussions in social media to understand public mood swings [31]. These fields are crucial for designing Q&A systems based on typhoon knowledge.

In these works, to better leverage research outcomes for addressing user needs, the adoption of Q&A systems as interactive mediums has seen significant advancement. Q&A systems have evolved through four development stages as follows: rule-based, based on statistical machine learning, based on knowledge graphs, and based on large language models. However, the diverse and complex nature of geographic information data and knowledge, with various geographical data intertwining, proposes challenges for rule-based systems in covering all cases. Formulating a large number of rules to cover all possible queries and scenarios is difficult [32,33]. Meanwhile, the implication of geographic information often varies due to regional, cultural, and historical influences, which means that the same geographic information may have different meanings in various contexts. Q&A systems based on statistical machine learning often perform poorly in handling semantic complexity and context dependency. They frequently fail to meet domain requirements [34,35]. Additionally, with the advancement of knowledge graphs, researchers have attempted to construct Q&A systems based on knowledge graphs. However, the construction and maintenance of geographic knowledge graphs are on a massive scale. This not only requires tremendous effort but also increases the difficulty in practical applications, because they cannot directly use natural language for querying [36]. With the rise of LLMs and their outstanding performance on multiple downstream tasks, the natural language processing (NLP) community has begun to consider that these models may contain the implicit knowledge in their parameters. Authors of [37] have demonstrated that LLMs can serve as knowledge bases. Because they store various types of knowledge in their parameters, including common sense, relational, and linguistic aspects [38,39]. References [40,41] have indicated that LLMs encode geospatial knowledge to some extent, and these models have preliminary geospatial awareness and geospatial reasoning potential. Geospatial knowledge encompasses factual understanding of geographic data such as location, time, and distance. Geospatial awareness involves the ability to perceive and comprehend geographical information in the temporal dimension. Geospatial reasoning is the process of making informed decisions using geospatial knowledge and awareness. However, the capability of LLMs in this regard is relatively limited. They can only be used for simple tasks, such as retrieving city coordinates and locations [40]. For more complex tasks, such as acquiring spatiotemporal information in disaster domains, different approaches are needed to explore and utilize the geospatial information stored in LLMs. Methods based on static data include “prompt engineering” and “RAG”. Methods based on data expansion include “continuous pretraining” and “fine-tuning”.

Open-source LLMs are used as a foundation to train domain-specific models [42]. In the medical sector, Google launched the Med-PaLM model in 2023, bringing AI applications to reality in medicine [43,44]. In the legal domain, Peking University’s team developed the open-source ChatLaw model by integrating external knowledge bases, which allows LLMs to offer comprehensive legal services to the public [45]. In the financial field, the FineGPT and BloombergGPT models, which have been successfully applied, democratize financial data access on a global scale [46,47]. In the geographic information field, the Baidu PaddlePaddle team introduced the ERNIE-GeoL model, which is a geo-linguistic, pretrained large model that integrates geographic and textual information. This integration achieves a dual emphasis on semantics and spatial data [48]. DAMO Academy and Gaode collaboratively launched the MGeo model. This model is a multitask, multimodal geographic text pretraining base model, which enhances performance across various downstream geographic text processing tasks [49]. In remote sensing, the Beijing Institute of Technology research team proposed the pioneering MLLM EarthGPT model, which unifies and integrates various sensor remote sensing interpretation tasks [50]. The Ant Group, in partnership with Wuhan University, introduced the SkySense model, which is a multimodal remote sensing base model with 2 billion parameters, applicable in key areas such as urban planning, forest protection, emergency rescue, green finance, and agricultural monitoring. The model advances the development of intelligent technology and applications in remote sensing [51]. Dilxat Muhtar and colleagues developed the large-scale remote sensing image-text data set LHRS-Align and the remote sensing–specific instruction data set LHRS-Instruct, and they subsequently introduced the LHRS-Bot model [52]. Despite the training of the aforementioned models in the geoscience field, their extensive parameters render them unfriendly to ordinary users, and their application capabilities in the disaster domain are not outstanding. Hu (2023) improved the precision in extracting location descriptions from disaster-related social media messages by fusing geographic knowledge and the Generative Pre-trained Transformer (GPT) model [48]. However, this work lacks contextual semantic information for disaster knowledge extraction and suffers from limited interactivity. Therefore, this work leverages the Q&A system as an interactive medium. It proposes integrating typhoon-related knowledge in the disaster domain into the T5 model and fine-tuning it on Q&A pairs. This aims to construct a Q&A model capable of responding to typhoon information.

3. Method

The core objective of our research is the integration of typhoon disaster knowledge into the T5 model, facilitated by a Q&A interaction to convey high-quality information. In this section, we will explore in detail the theoretical foundation of the T5 model, as well as how to cleverly integrate typhoon disaster knowledge within this theoretical framework. By embedding disaster information in the T5 model, we aim to augment the model's responsiveness to typhoon disaster information. This will also improve the interactive experience between users and the model. It will effectively carry out disaster prevention and mitigation tasks.

3.1. T5 Model Theory

T5, or "text-to-text transfer transformer", adopts the transformer's encoder–decoder structure. It is a universal framework proposed by Google in the pretrained model domain [53]. It employs a stack of self-attention layers instead of traditional RNNs or CNNs to handle variable-length input sequences. When provided with an input sequence, it is mapped to an embedding sequence in the encoder. These encoders share the same structure, each consisting of the following two subcomponents: a self-attention layer followed by a small feed-forward neural network. Layer normalization is applied to the input of each subcomponent, and residual skip connections add the input of each subcomponent to its output. Dropout is applied to the feed-forward networks, skip connections, attention weights, and the input and output of the entire stack. The decoder works similarly to the encoder as follows: after each self-attention layer, there is an additional attention mechanism to process the output of the encoder. The final output of the decoder block is passed through a dense layer to generate the output probability of the vocabulary. Unlike the general transformer model, the T5 model uses a simplified form of positional embedding, where each embedding is a scalar added to the corresponding logit used to compute the attention weights. As the authors pointed out, to improve efficiency, they share positional embedding parameters among all layers. T5 has the following two main advantages over other advanced models: (1) it is more efficient than RNNs because it allows parallel computation of the output layer and (2) it can detect hidden and distant dependencies between tokens without assuming that nearby tokens are more relevant than distant ones. This attribute is particularly important for tasks related to text generation, because words with the same semantics may be far apart.

The T5 model aims to create a high-quality, comprehensive pretrained language model through the use of more extensive data sets. Its core idea is to convert various natural language processing (NLP) tasks into a text-to-text format through a prefix task declaration, which means the input is text, and the output is also text. This approach enables the use of a singular model to address all NLP tasks. It provides a unified solution for diverse tasks without altering the loss function or training methodology. Figure 3 illustrates the text-to-text output format. The green part represents the translation task; the red and yellow parts represent the CoLA (The Corpus of Linguistic Acceptability) task and STS-B (Semantic Textual Similarity Benchmark) task, respectively; and the blue part represents the summary generation task. The left box shows the input example of T5, and the right box depicts the corresponding output. Under this framework, each task is considered an input. It only requires a task declaration prefix in front of the input data, which can guide the model to procedure the target text for a specific task.

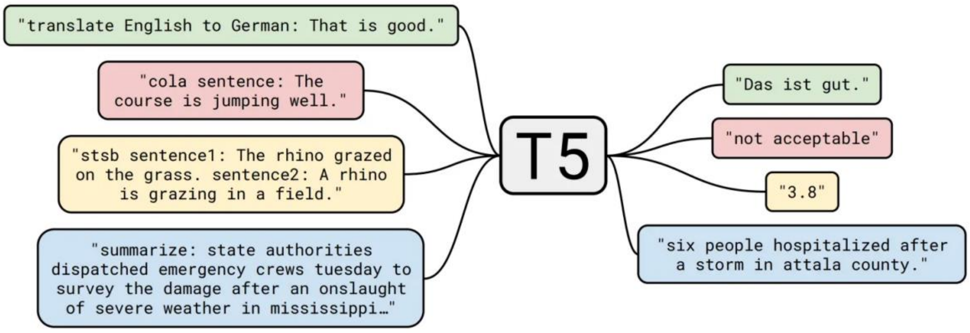

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 3. A diagram of the text-to-text framework.

In the T5 series, T5-large (https://huggingface.co/sentence-transformers/sentence-t5-large (accessed on 11 February 2024)) and T5-base (https://huggingface.co/sentence-transformers/sentence-t5-base (accessed on 11 February 2024)) are two common variants. T5-large denotes models with more parameters, whereas T5-base is smaller in scale. This tiered design strategy enables users to select the model best suited to the task's complexity and available computational resources. For example, in handling large data sets and complex tasks like text summarization and Q&A system development, T5-large, with its deeper network and broader parameters, typically yields more precise outcomes. Conversely, T5-base is preferable in scenarios demanding high real-time performance and limited computational resources, due to its reduced computational demands and faster response.

3.2. Model Enhancement

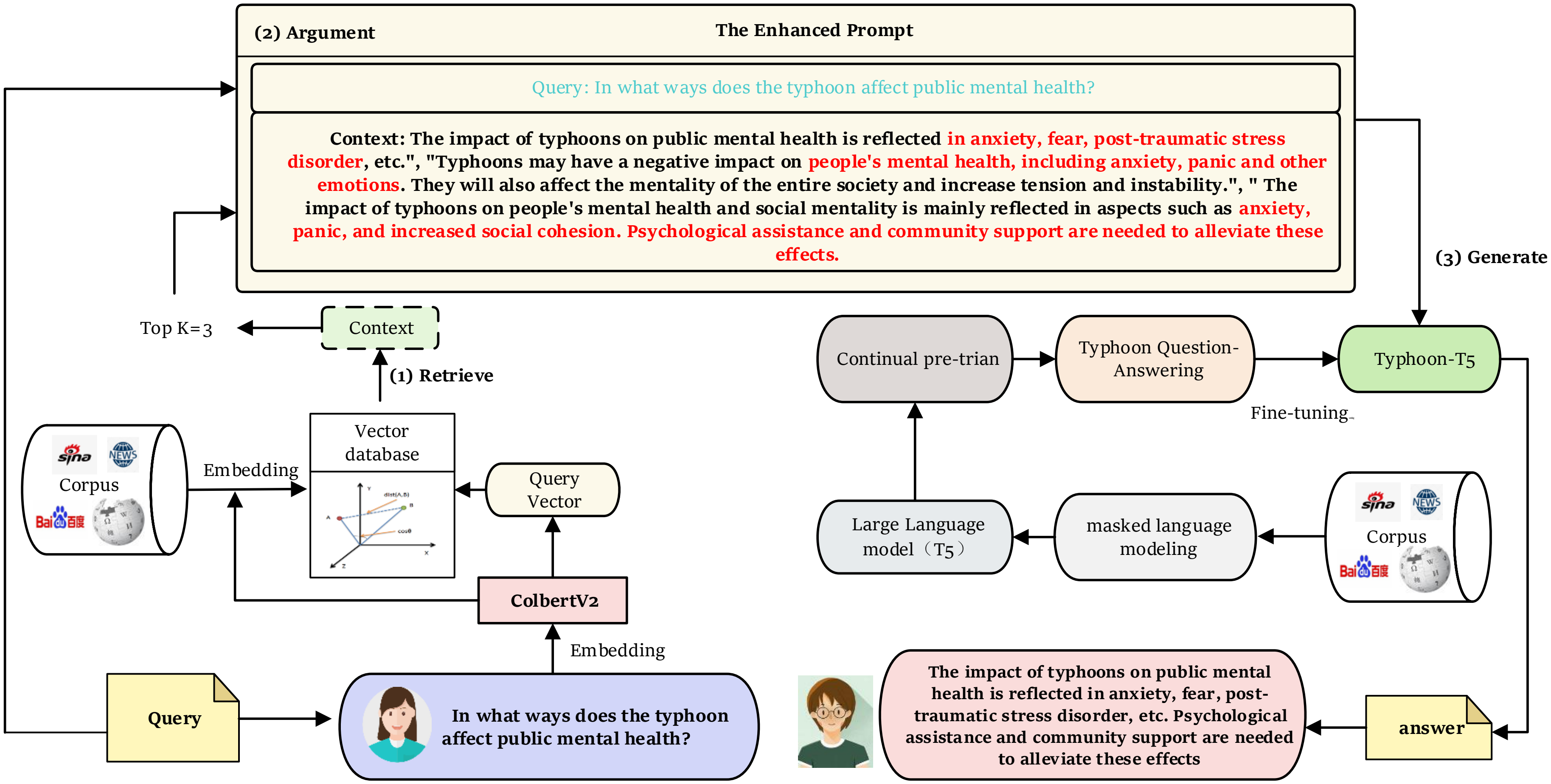
In this research, we committed to incorporating extensive typhoon disaster knowledge into the T5 model. Our goal is to optimize the model's capability for processing text information related to natural disasters, which augments its efficacy in disaster response applications. This is related to but different from the geographic information retrieval field's geospatial parsing problem. Geospatial parsing aims to identify and parse time and place names in text [54]. However, focusing solely on spatial information is insufficient for typhoon emergency disaster response. For example, in a news report stating "This year's Typhoon No. 6 'In-Fa' is expected to land in Zhoushan Putuo District, Zhejiang Province around 12:30 on 25 July, with a central pressure of 965 hPa, and a maximum wind speed of 38 m/s!", we need to consider not only the location information of "In-Fa" but also the provided time information. Therefore, extracting both location and time descriptions from disaster-related messages presents a series of different challenges.

In response to the above issues, we have taken the following steps:

1. Collect and process high-quality typhoon disaster data from online texts for continuous pretraining. Continual pretraining involves continuously adding new data to the pretrained model and retraining it to further enhance its performance. The purpose of continual pretraining is to enable the model to adapt to evolving data and tasks, ensuring its ongoing updates and development. For data, our primary sources are Wikipedia, Baidu Encyclopedia, and various major news websites, which cover data on typhoon forecasts before landfall, disaster information after landfall, and typhoon prevention measures. We then process and categorize the raw data to form usable data, and subsequently construct the corpus needed for this study. To acquire the training set for the T5 model, we utilize the same training technique as the BERT model, which involves masked language modeling to process the data in the corpus. Additionally, we extract 2204 typhoon Q&A pairs from this corpus for subsequent fine-tuning.

2. Fine-tuning LLMs for typhoon disaster tasks. Fine-tuning is a technique in the field of deep learning that involves additional training of a pretrained model with a specific domain data set. This process enables it to adapt to tasks and information specific to that domain. In this experiment, we choose a high-quality pretrained language model (T5 model) as the base model. The initial phase involves continuous pretraining of the T5 model to encode information from the training set into the neural network's weights. Our goal is for the model to learn geographic information related to typhoon disasters, which aims to utilize the T5 model's vast parameters to store and process the expressions of geographic information in text data. Then, based on the typhoon Q&A pairs constructed in Step 1, we fine-tune the pretrained model. This phase involves parameter optimization to enhance the model's performance on typhoon-related Q&A tasks specifically. We refer to the model pretrained on the typhoon domain data set and subsequently fine-tuned on the Q&A data set as the Typhoon-T5 model.

3. Prompt argument based on RAG. The RAG technique enables LLMs to enhance their outputs with information from external knowledge sources. This process improves accuracy, ensures contextual relevance, and minimizes errors in generated responses. Specifically, we use the ColbertV2 (https://huggingface.co/colbert-ir/colbertv2.0 (accessed on 11 February 2024)) [55] model to convert the corpus knowledge into a vector database for storage. When a user makes a query, we embed the user's query into the same vector space as the vector database. Through similarity searches, we retrieve and return passages semantically related to the queries. These passages are then integrated into the prompt as contextual information. The enhanced prompt is then input into the Typhoon-T5 model, which has been fine-tuned in Step 2. This allows the model to generate detailed and targeted responses, providing users and emergency responders with relevant information. The specific implementation method is shown in Figure 4. It is noteworthy that the integration of typhoon disaster information significantly strengthens the T5 model's capability for comprehension of geographical spatial features. This advancement enables a more comprehensive and in-depth analysis of text information related to typhoon disasters.



Query: In what ways does the typhoon affect public mental health?

Context: The impact of typhoons on public mental health is reflected in anxiety, fear, post-traumatic stress disorder, etc.", "Typhoons may have a negative impact on people's mental health, including anxiety, panic and other emotions. They will also affect the mentality of the entire society and increase tension and instability.", " The impact of typhoons on people's mental health and social mentality is mainly reflected in aspects such as anxiety, panic, and increased social cohesion. Psychological assistance and community support are needed to alleviate these effects.

In what ways does the typhoon affect public mental health?

The impact of typhoons on public mental health is reflected in anxiety, fear, post-traumatic stress disorder, etc. Psychological assistance and community support are needed to alleviate these effects

Figure 4. Flowchart of the Typhoon-T5 method. The left half belongs to "RAG", and the right half belongs to "pretraining + fine-tuning". "R" stands for (1) retrieve. For user queries, first retrieve passages that are semantically similar to the question. "A" stands for (2) argument. The retrieved passage and prompt are combined to form "the enhanced prompt". The colored part highlights the key points, which indicates that the recalled text has the correct answer. "G" represents (3) generation, and the enhanced prompt is input into the fine-tuned Typhoon-T5 model. Finally, a response result is generated.

3.3. Model Evaluation

To evaluate the accuracy of the model's responses to user queries, this work employs a text similarity representation method. Text similarity measures the semantic or structural resemblance between two pieces of text, including both model-generated results and ideal answers. In this study, we employ multiple similarity measurement models.

From the perspective of semantic similarity of texts, we adopt the "all-MiniLM-L6-v2" (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (accessed on 11 February 2024)) model [56] as the representation model. "Cosine similarity" is utilized as the similarity measurement method, calculated as shown in Equation (1). In the case of information retrieval, the cosine similarity of two documents ranges from [0, 1], because term frequencies cannot be negative. A cosine similarity of 1 indicates complete similarity between two texts, while a cosine similarity of 0 signifies no correlation between them. The all-MiniLM-L6-v2 is a pretrained language model that utilizes large-scale word or sentence vector corpora for pretraining. It better captures the semantic correlations between words and is more suitable for fine-grained comparison and analysis of textual semantics.

$$Cosine\ similarity = Sc(A,B) := \cos(\theta) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{||\boldsymbol{A}||||\boldsymbol{B}||} = \frac{\sum_i^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2}\cdot\sqrt{\sum_{i=1}^n B_i^2}}, \tag{1}$$

where A_i and B_i are the ith components of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. For text matching, the attribute vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ are usually the term frequency vectors of the documents.

From the perspective of set-based similarity calculation, we adopt the Jaccard similarity coefficient method [57]. The Jaccard similarity coefficient is a set-based similarity calculation method that measures the similarity between two text word sets by assessing their overlap [58]. It offers a straightforward approach to measuring similarity between data samples. It is defined as the ratio of the size of the intersection of two data samples to the size of their union, as illustrated in Equation (2)

$$J(A,B) = \frac{|A\cap B|}{|A\cup B|} = \frac{|A\cap B|}{|A|+|B|-|A\cap B|}, \tag{2}$$

where, A and B represent two sets, $|A\cap B|$ denotes the intersection of sets A and B, and $|A\cup B|$ represents the union of sets A and B. If A and B are completely identical, then J (A, B) = 1. Therefore, J (A, B) takes values in the range [0, 1].

From the perspective of text generation quality, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) method is adopted. Additionally, two commonly used evaluation metrics, namely ROUGE-N and ROUGE-L, are selected. ROUGE-N measures the n-gram overlap rate between the reference answers and generated results to assess their similarity. ROUGE-L evaluates the fluency of the generated results by calculating the longest common subsequence between them and the reference answers [59], as shown in the following equations

(1) ROUGE-N (N = 1, 2)

$$\text{ROUGE-N} = \frac{\sum\limits_{S\in\{Ref\}}\ \sum\limits_{gram_n\in S} Count_{\text{match}}(gram_n)}{\sum\limits_{S\in\{Ref\}}\ \sum\limits_{gram_n\in S} Count(gram_n)}, \tag{3}$$

where, $gram_n$ represents the n-grams, which indicates n consecutive words; $S\in\{Ref\}$ denotes the reference answers, i.e., the standard answers obtained in advance; $Count_{\text{match}}(gram_n)$ denotes the number of n-grams matched between the generated answers and the reference answers; and $Count(gram_n)$ represents the number of n-grams in the reference answers.

(2) ROUGE-L

$$R_{lcs} = \frac{LCS(X,Y)}{m_1}, \quad (4)$$

$$P_{lcs} = \frac{LCS(X,Y)}{m_2},$$

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}},$$

where, $LCS(X, Y)$ represents the length of the longest common subsequence of X and Y; m_1 and m_2 denote the lengths of the reference answers and the generated results, respectively; and R_{lcs} and P_{lcs} represent recall and precision, respectively. Generally, β is set to a large number, so ROUGE-L almost only considers recall.

4. Experimental Results and Analysis

This section mainly describes the classification and production of data and applies it to the proposed method. Additionally, we compare the proposed method with three other scenarios, the T5 model without typhoon knowledge guidance, the T5 model with typhoon knowledge guidance, and the Typhoon-T5 model without typhoon knowledge guidance. Finally. Our method is implemented using T5 models of varying parameter sizes, specifically T5-large and T5-base. Detailed configuration information related to the experiments is presented below.

4.1. Experimental Environment

The experimental setup in this work utilizes the Ubuntu 20.04 operating system. The system is equipped with an NVIDIA GeForce RTX 4090 graphics card boasting 24 GB of video memory. A GPU was utilized for both training and testing, with PyCharm software version 2021.3, which served as the training platform. During the training process of the T5 model, we conducted 20 epochs of training with a batch size of 4. As the optimizer, we employed AdaFactor [60] with specific settings, including a learning rate of 1×10^{-3}, scale_parameter as False, relative_step as False, and warmup_init as False.

4.2. Experimental Procedure

To construct the pretraining and fine-tuning data sets for the T5 model, the experimental procedure constructs a knowledge system on typhoon disaster events focused around the following three aspects: typhoon meteorological knowledge, typhoon disaster case studies, and typhoon disaster management. Specifically, typhoon meteorological knowledge entails basic concepts such as typhoon definition, naming, and classification from a meteorological perspective. It's important to note that due to practical application needs, the general public primarily seeks fundamental concepts and definitions in the field of typhoon disasters. They often overlook underlying mechanisms and processes. Therefore, this knowledge framework does not delve into such mechanistic knowledge. Typhoon disaster case studies, which involve historical cases of typhoon disasters, detail the disasters and their impacts from a disaster management perspective. Typhoon disaster management knowledge pertains to measures taken by humans to prevent and mitigate typhoon disasters, including typhoon forecasting and warning, emergency response measures, and disaster recovery. Specifically, we utilized data from encyclopedias and news websites as our primary sources. We collected 59 encyclopedia articles from Baidu Encyclopedia, which resulted in 432 Q&A pairs. Additionally, we gathered 72 encyclopedia articles from Wikipedia, which generated 690 Q&A pairs. These pairs mainly cover factual knowledge about typhoons. They contribute to the dissemination and popularization of typhoon disaster knowledge. Furthermore, to effectively address typhoon disasters and engage in disaster prevention, mitigation, and relief efforts, we also selected reports about Typhoon "In-Fa" from various major typhoon news websites at the time of its landfall in 2021 (e.g.,http://typhoon.nmc.cn/web.html, http://typhoon.org.cn (accessed on 11 February 2024)). These reports served as the training data set for this experiment, amounting to 150 reports. This selection resulted in 1082 Q&A pairs. Table 1 details the types, descriptions, and examples of these data.

Table 1. Experimental data classification.

Type of Data	Data Description	Demonstration	
Typhoon Meteorological Data	From a meteorological perspective, describe fundamental knowledge related to typhoon concepts.	Typhoon Definition	A typhoon is a tropical cyclone that develops between 180° and 100° E in the Northern Hemisphere.
		Typhoon Naming	Since 2000, the tropical cyclone naming list in the northwest Pacific has been developed by the WMO Typhoon Committee. There are five naming lists, each consisting of two names provided by 14 members.
		Typhoon Classification	A tropical depression is upgraded to a tropical storm should its sustained wind speeds exceed 34 knots. Should the storm intensify further and reach sustained wind speeds of 48 knots then it will be classified as a severe tropical storm.
		……	
Typhoon Disaster Case Data	From a disaster studies perspective, select relevant information about Typhoon “In-Fa” from historical occurrences of typhoons, and describe the disasters it caused and their associated impacts.	Evolution Mechanism	“In-Fa” has a structurally complete and symmetrical form, with a clear eye of the typhoon and a vast expanse of cloud cover. True to its name, it is a “beautiful” typhoon.
		Characteristics and Attributes	On 25 July, the Typhoon “In-Fa” made landfall along the coast of Putuo District, Zhoushan City, Zhejiang Province, around 12:30 p.m. The maximum wind force near the center reached 13 on the Beaufort scale (38 m per second), with the minimum central pressure of 965 hPa.
		Disaster Situation Information	Before making landfall, Typhoon “In-Fa” had already impacted the climate on the Chinese mainland. On 20 July, Henan Province experienced catastrophic extreme precipitation, which results in the deaths of 302 people.
		……	
Typhoon Disaster Management Data	From a disaster management perspective, describe the relevant knowledge generated by humans to prevent and mitigate typhoon disasters.	Typhoon Forecast and Warning	Typhoon warnings are issued by specialized agencies in various regions during the period when the storm may strike, providing forecasts and alerts.
		Emergency Response Measures	When a typhoon approaches, it is necessary to secure doors and windows tightly, remove all kinds of hanging objects indoors and outdoors, close doors and windows, and if necessary, reinforce them with nailed wooden boards.
		Disaster Recovery	All levels of government departments mobilize the entire population and achieve full coverage, carrying out emergency rescue, garbage cleaning, sludge and pollution removal, and disinfection and sterilization, to ensure no major epidemic outbreaks after major disasters.
		……	

During the data collection phase, the acquired data are rigorously screened to remove duplicate, inaccurate, and irrelevant information to ensure data quality. To obtain the pretraining data set, we employed masked language modeling to format the data according to the training needs of the T5 model. For the construction of the fine-tuning data set (typhoon Q&A pairs), we designed a series of questions related to typhoon disasters from different levels and perspectives. Subsequently, we used open-source projects on GitHub (https://github.com/SupritYoung/free-self-instruct (accessed on 11 February 2024)) to extract answers matching the questions, which built a data set of 2204 typhoon-related Q&A pairs. The data set is divided into training, testing, and validation sets in a ratio of 70:20:10 for model fine-tuning purposes. Moreover, to further enhance model response quality, we employed RAG technology, which retrieves passages from the typhoon corpus that are semantically similar to user queries, as context to enhance prompts. Finally, to assess the response quality of the Typhoon-T5 model, we randomly selected questions from the test set and generated answers through the Typhoon-T5 model. The experimental flowchart is shown in Figure 5.

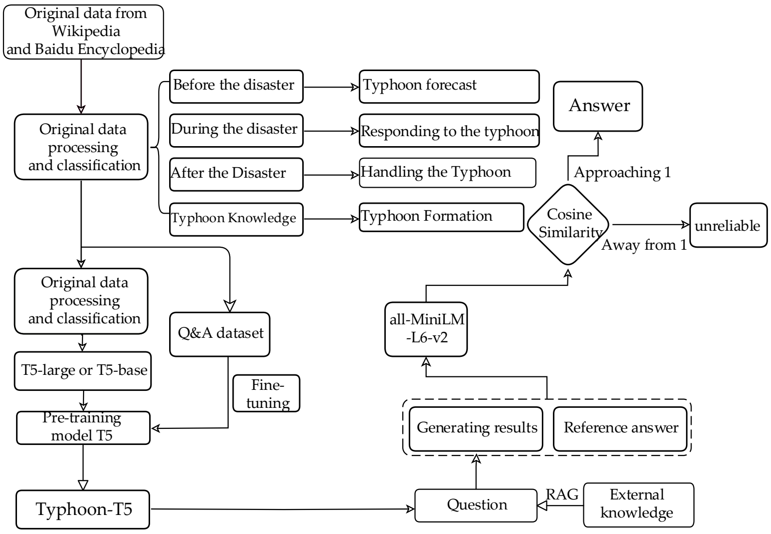


Figure 5. Experimental workflow diagram.

4.3. Experimental Rusults

In this work, we conducted text similarity measurements between the outputs of the T5 model and the reference answers, which specifically included three distinct evaluation methods as follows: model-based evaluation, intelligent evaluation, and manual evaluation. Subsequently, we further discussed the performance of T5 models with varying parameter sizes.

4.3.1. Model Metric Evaluation

Model metric evaluation refers to a method of assessing model performance or predicting results using quantitative metrics. In the field of natural language processing, model metric evaluation typically involves using various metrics to quantify model performance, which aims to provide quantitative measures of model performance on specific tasks. From the perspective of semantic similarity, we adopted the 'all-MiniLM-L6-v2' model. Through several sets of comparative experiments (Group 1: typhoon-T5-large_no_context, Group 2: typhoon-T5-large_with_context, Group 3: T5-large_with_context, and Group 4: T5-large_no_context), we conducted an in-depth investigation into the performance of the proposed method (Group 2 experiments) in the 'all-MiniLM-L6-v2'. As shown in Figure 6, the experiments in Group 3 and Group 4 showcase the performance of the T5-large model with and without typhoon knowledge guidance. It is evident that when guided by typhoon knowledge, the model's responses have a higher cosine similarity with reference answers in the high-score range. Conversely, without typhoon knowledge guidance, the model's responses display a higher cosine similarity with reference answers in the low-score range. This is because the model lacks sufficient information for guidance, which results in the inability to generate answers that match the reference answers. Under the influence of the RAG task, the model retrieves passages with higher semantic similarity to the user queries as context prompts, akin to an "open-book exam". The model can fully utilize these contextual cues to generate more accurate and reliable answers. Therefore, a retrieval-based Q&A method demonstrates superior quality and reliability. In the experiments of Group 1 and Group 3, the Group 1 experiment integrates typhoon knowledge into the T5-large model through continuous pretraining. It is observed that, in the trustworthy range of 0.8 to 1.0, the model with integrated typhoon knowledge performs more effectively. This observation may be caused from various factors. Primarily, through fine-tuning, typhoon-related knowledge is integrated into the model and stored implicitly within its weights. Consequently, when users make queries, the model can effectively utilize all available information to respond to questions. The results are more comprehensive and accurate outcomes. Additionally, the RAG task involves integrating retrieved knowledge explicitly into prompts, which enables the model to respond to user queries based on this information. However, due to the inherent limitations of the RAG task, there may be poor quality of the recalled text, potentially leading to misleading model responses. This indicates that incorporating typhoon disaster knowledge through fine-tuning results in superior response outcomes compared to retrieval-based methods, which enhances its effectiveness in addressing unknown questions. The comparison results between Group 1 and Group 2 experiments reveal that the Group 2 experiment incorporates domain knowledge into the model and retrieves passages semantically similar to the question as context to guide Typhoon-T5 in generating responses. It is apparent that responses generated in this manner predominantly fall in the high-score range and less so in the low-score range. Consequently, the incorporation of typhoon disaster knowledge into LLMs, along with the inclusion of contextual information into the prompt, enhances the credibility, flexibility, and interactivity of the generated responses. However, this approach is more complex, which necessitates higher hardware configuration and greater computational resources for model training and fine-tuning.

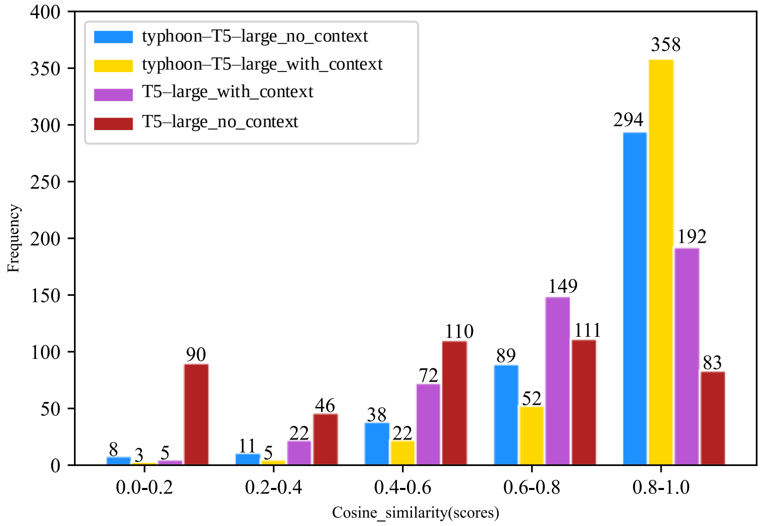


Figure 6. Experimental results comparison cased on cosine similarity.

From the perspective of set-based similarity calculation, we employed the Jaccard similarity coefficient method. To standardize the representation of text, we converted all text to lowercase. This helps reduce calculation errors caused by differences in character casing and punctuation. The experimental results are shown in Figure 7. It can be observed that the number of high-score intervals in the Jaccard index is significantly lower than that of the 'all-MiniLM-L6-v2' model. Through specific case analysis, it was found that the Jaccard index only considers the cooccurrence of words. This neglects semantic and word order information, which leads to the inability to capture deep semantic correlations in the text. For example, for the query: "'In-Fa' caused Heavy downpour in Henan, and how many people died?", the model generated the result: "The 'In-Fa' brought Heavy downpour to Henan before it made landfall, which causes the deaths of 302 people". The reference answer is "The 'In-Fa' caused 302 deaths". From this example, it can be seen that both the generated result and the reference answer can accurately answer the user query, with differences only in semantic richness. Therefore, the Jaccard index generally remains low. Nevertheless, our proposed method still exhibits a higher frequency in the high-score intervals compared to other models, with a lower frequency in the low-score intervals.

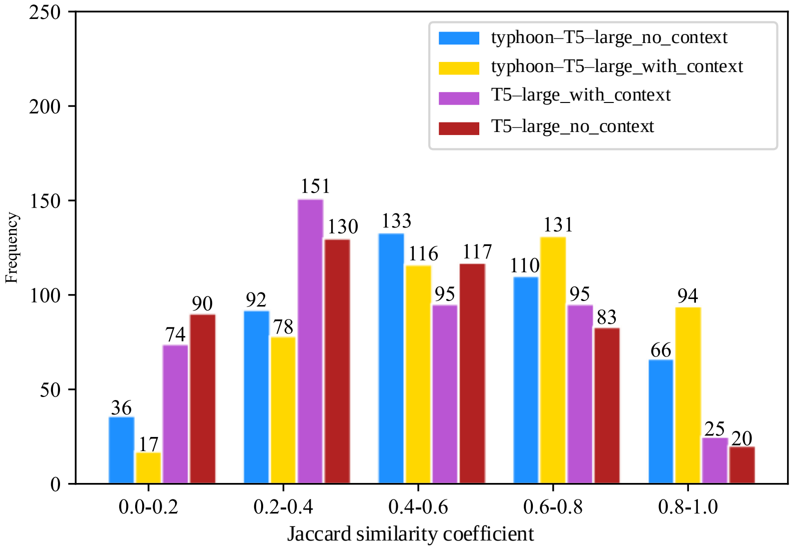


Figure 7. Comparison of experimental results based on the set similarity calculation.

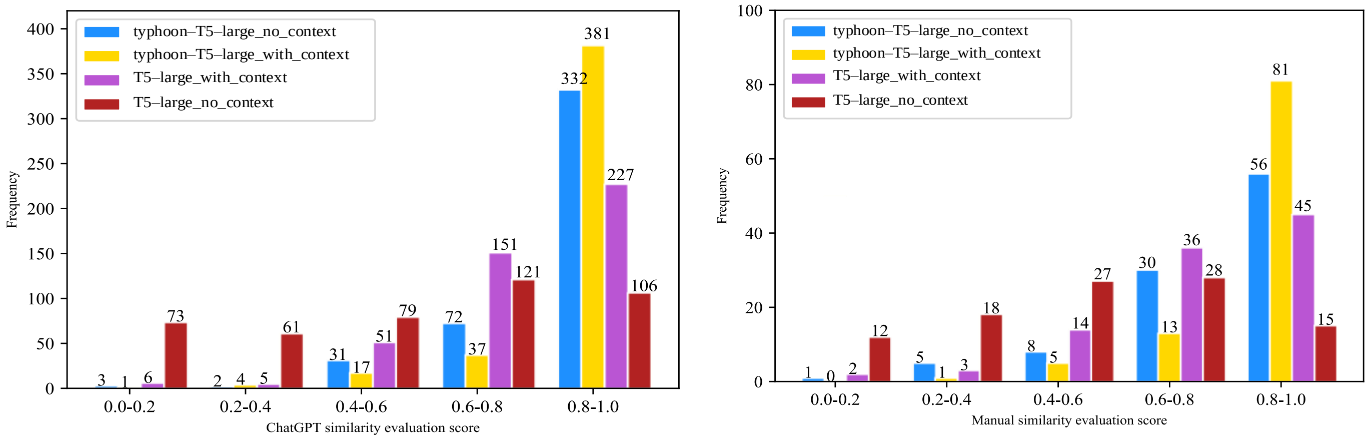
For the ROUGE method (ROUGE-N and ROUGE-L), we used the number of Q&A pairs with scores between 0.8–1.0 divided by the total number of test sets. The experimental results are shown in Table 2. It can be observed that typhoon-T5-large_with_context outperforms the other three cases in ROUGE-1, ROUGE-2, and ROUGE-L performance. Additionally, ROUGE-1 calculates similarity using unigrams, while ROUGE-2 uses bigrams. Due to the larger number of word combinations in bigrams compared to unigrams, there are fewer matching opportunities. This results in lower ROUGE-2 scores.

Table 2. Comparison of experimental results under ROUGE metrics.

Model	**T5-Large no_Context**	**T5-Large with_Context**	**Typhoon-T5-Large no_Context**	**Typhoon-T5-Large with_Context**
ROUGE-1	11.36%	25.45%	30.45%	40.82%
ROUGE-2	6.23%	10.57%	12.15%	21.72%
ROUGE-L	16.81%	26.36%	29.09%	37.27%

4.3.2. Intelligent Evaluation

Recent studies [61,62] have utilized LLMs for vertical domain evaluation tasks. These studies have empirically demonstrated that the evaluation of ChatGPT is consistent with manual evaluation. In this work, we used ChatGPT to avoid intensive labor. Specifically, given standard answers and answers generated by the T5 model, ChatGPT is prompted to return similarity scores. Finally, the frequencies of all scores in each interval are aggregated, as shown in Figure 8a.



(**a**) (**b**)

Figure 8. Using ChatGPT and human workers to assess the model outputs and standard answers. (**a**) Recording the frequency distribution of ChatGPT's evaluation scores across different intervals. (**b**) Recording the frequency distribution of evaluation scores from 12 human reviewers across various intervals.

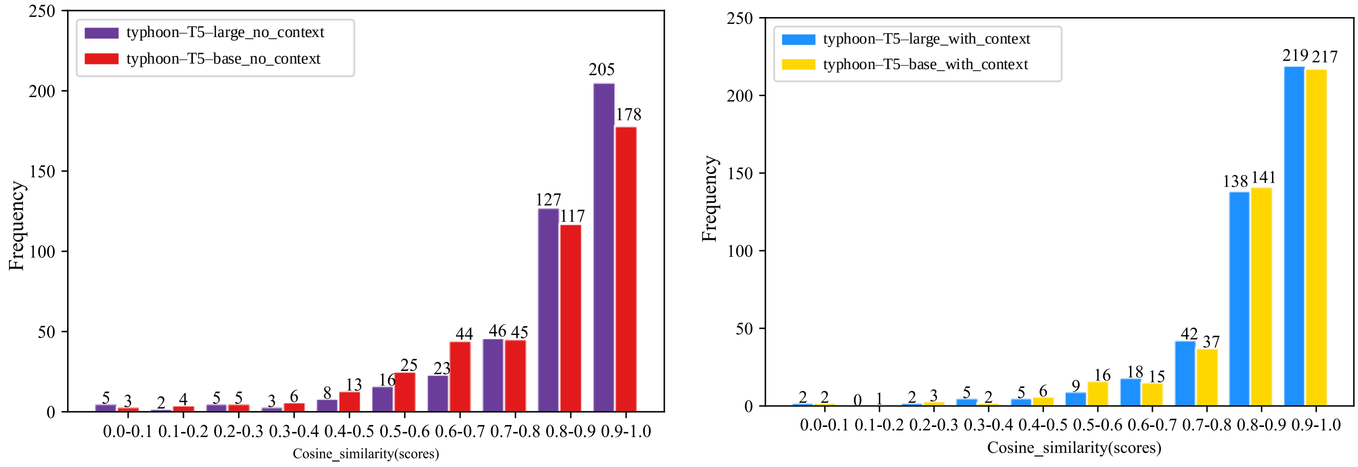
4.3.3. Manual Evaluation

Manual evaluation is crucial for assessing model performance. It particularly addresses issues such as data ambiguity and context dependency, which often require human judgment. This ensures the accuracy and robustness of model performance. At the same time, to mitigate the subjectivity and bias of individual evaluations, we employed 12 workers to randomly evaluate 100 results from the test data. The main purpose is to judge whether the output of the T5 model meets user queries and the similarity to standard answers. The experimental results are shown in Figure 8b.

From different perspectives such as index evaluation, intelligent evaluation, and manual evaluation, the performance of the model is optimal. This reflects that incorporating typhoon disaster knowledge into LLMs and adding context information to prompts can improve the credibility, flexibility, and interactivity of generated responses. However, it can also be observed from the experimental process that with the increase in external knowledge bases, this method becomes more complex. Through training and fine-tuning, knowledge is implicitly stored in model weights. This will require higher hardware configurations and greater consumption of computing resources. At the same time, the quality of RAG-recalled text will decrease. It may even recall text sentences unrelated to user queries, which misleads LLMs into generating incorrect results. Therefore, the correct selection of models and methods is crucial for experimental results across different tasks.

4.3.4. Comparison between T5-Base and T5-Large

To further validate the performance of our method across the T5 models with varying parameter sizes, we specifically focused on the models' effectiveness in parsing user questions and generating responses related to typhoons. For instance, we consider questions like "When did Typhoon 'In-Fa' make landfall?" and "How much economic loss did Typhoon 'In-Fa' cause?" To explore this, we conducted experiments using the T5-large model, which has a parameter size of 770 M, and the T5-base model, which has a parameter size of 248 M. Evaluated using cosine similarity. In Figure 9a, "typhoon-T5-large_no_context" demonstrates a performance improvement over "typhoon-T5-base_no_context". This suggests that a larger parameter size enables the model to retain more knowledge details, which results in superior responses to user questions. On the other hand, as shown in Figure 9b, while "typhoon-T5-large_with_context" shows improved performance relative to "typhoon-T5-base_with_context", the degree of improvement is not as significant. This may be attributed to the relatively limited size of the typhoon disaster knowledge data set that we studied. While T5-large captures more fine-grained details during training, T5-base successfully compensates for deficiencies in response generation by retrieving contextual information. Therefore, the results are not significantly different from T5-large in terms of performance. This also indicates that when optimizing the application of LLMs in the geoscientific domain, selecting an appropriate parameter size can not only generate excellent results but also reduce hardware requirements and save computational resources. This also underscores that optimizing the application of LLMs in the geoscientific domain, which involves selecting an appropriate parameter size, can not only generate excellent results but also reduce hardware requirements and save computational resources.



(**a**) (**b**)

Figure 9. Performance differences of T5 models with different parameter sizes: (**a**) performance of the Typhoon-T5 model with context guidance; (**b**) performance of the Typhoon-T5 model without context guidance.

5. Discussion

5.1. Scalability

The proposed method not only improves the processing efficiency of specific typhoon information, its high flexibility also makes it suitable for scenarios involving detailed information about various typhoon disasters. Meanwhile, it offers a viable approach for application to other disaster events. By integrating additional disaster knowledge into the ongoing pretraining of the T5 model and reindexing the collected data, it becomes possible to swiftly retrieve passages semantically similar to user queries. This approach facilitates the generation of disaster-related information. Key elements of this method include the selection of LLMs, pretraining, and the adoption of RAG technology. LLMs selection may consider options like LLaMA (Large Language Model Meta AI) and PaLM (Pathways Language Model). They have stronger text generation capabilities, but are also accompanied by significant resource consumption. Furthermore, training data can cover other types of disaster events such as earthquakes, floods, and fires. Through fine-tuning and customization in diverse disaster scenarios, this method can be adapted to various disaster events and provide tailored knowledge Q&A services for different disaster scenarios. It is crucial to ensure the authenticity and reliability of the collected data. LLMs inherently suffer from hallucination phenomena. Therefore, the introduction of authentic and reliable data is essential, especially in the collection of data in vertical domains. Such data not only helps alleviate model hallucinations but also enhances the model's performance and accuracy. Additionally, for RAG tasks, the authenticity and effectiveness of recalled text are crucial for generating correct results.

5.2. Exploration of Application Scenarios

To provide factual typhoon knowledge, during the model's pretraining phase, we gathered information about typhoons from authoritative sources. This process helped us establish a repository of factual knowledge. This fact-based knowledge is essential for the model to provide accurate and reliable information when responding to user queries. However, factual knowledge accumulates gradually over time. Despite collecting a considerable amount of typhoon-related information, there may still be potential omissions or oversights. Utilizing the scalability of our proposed method, we can continually pretrain T5 model and incorporate new information to gradually fill knowledge gaps.

It is rather regrettable that our trained T5 model, like most LLMs, cannot respond to ongoing typhoon events. However, owing to the relatively smaller parameter size of T5, even the T5-large model is only 770 M. This limitation can be addressed by gathering the latest information from sources such as news websites and social media. These data can be utilized for model training, which enables it to cope with ongoing typhoon events. The implementation of this entire process can be achieved with an NVIDIA GeForce 4090. This is advantageous compared to larger LLMs (e.g., ChatGPT (gpt-3.5-turbo-0613)). They consume significant resources, which makes practical implementation more challenging.

5.3. Analysis of Limitation

Table 1 demonstrates that model-generated results form complete expressions that address questions on the psychological impact of typhoons on the public. However, the responses appear overly concise and lack sufficient detail. This may be attributed to the inherent lack of more detailed information in the collected typhoon knowledge base. Moreover, thanks to the limitations of the T5 model, it cannot automatically adjust the length of generated responses. As shown in Table 3, when responding to the question "What is the initial stage of typhoon formation?" with min_length = 128, the generated response includes the key phrase "a tropical depression" but introduces additional information. In contrast, with min_length = 64, which avoids introducing extraneous information, a complete expression is formed. Furthermore, when generating a response to the question "How are medical services and health facilities guaranteed during emergency response?" the response with min_length = 128 appears to be slightly more favorable compared to min_length = 64. It is anticipated that in the future, the T5 model can be improved to automatically adjust the length of generated responses, which enhances the accuracy of generating appropriate results. This is significant for leveraging the T5 model as a baseline for the geographic information field.

Table 3. Response results with different parameter settings.

Len(Min_Length)	Question	Answer
128	What is the initial stage of typhoon formation?	A tropical depression. At this stage, the initial low-pressure circulation needs to evolve until forming a complete typhoon structure.
128	How are medical services and health facilities guaranteed during emergency response?	Emergency response methods for ensuring medical services and health facilities include preparing drugs and medical equipment in advance, dispatching medical personnel, setting up emergency medical stations, etc., to deal with possible public health problems.
64	How are medical services and health facilities guaranteed during emergency response?	Preparations involve securing drugs and medical equipment, dispatching personnel, and establishing emergency medical stations.
64	What is the initial stage of typhoon formation?	A tropical depression is the initial stage of typhoon formation

Furthermore, the current data representation is a series of questions and answers. The content mostly focuses on the spatiotemporal description of typhoon text information. However, this representation has its limitations because it fails to fully consider key factors such as geographic knowledge and spatiotemporal features. In reality, disaster prevention and mitigation related to typhoons involve various aspects, including but not limited to meteorological data, political decisions, and socioeconomic data, which are not adequately covered by the current data representation. Therefore, despite providing spatiotemporal descriptions of typhoon text information, the current data still has limitations in achieving comprehensive, multidimensional analysis, and application of typhoon disaster data. It is precisely because the T5 model is trained on a large amount of text covering geographic knowledge. It focuses on the spatiotemporal features of geographic information and stores a considerable amount of geographic spatiotemporal information in its parameters. As a result, the model possesses strong spatiotemporal awareness and reasoning capabilities when responding to user queries. This enables it to represent factual, specific geographic knowledge and guide the public and emergency relief personnel in disaster prevention, mitigation, and relief efforts.

6. Conclusions

Typhoon disasters, as a seasonally significant meteorological disaster, occur primarily in the summer and autumn seasons, which pose substantial challenges to efforts in disaster prevention and mitigation. In the era of deepening digital and internet technology development, a wealth of popular materials on typhoon disasters has been accumulated. However, the precise and rapid acquisition of targeted information has not been effectively resolved. Although search engines have made efforts to optimize the information retrieval experience, users still depend on keyword searches, which face the tough task of filtering through extensive search results. In the context of today's information-overloaded society, such retrieval methods struggle to meet users' growing pursuit of information. The emergence of intelligent Q&A systems has made up for these shortcomings. These systems organize unstructured corpus information in an orderly and scientific manner and construct knowledge-based classification models. This approach precisely addresses users' queries, offering personalized information services. This work proposes a novel approach that integrates typhoon knowledge with the T5 model, which uses a Q&A format to help the public better understand typhoon-related knowledge. The required data, implementation details, and conducted experiments for this method are thoroughly detailed. We compared it with other viable alternatives, including the T5 model without typhoon knowledge guidance, the T5 model with typhoon knowledge guidance, and the Typhoon-T5 model without typhoon knowledge guidance. Experimental results show that the proposed method, which leverages RAG technology to retrieve passages semantically similar to the question from the knowledge base, significantly outperforms the alternatives in guiding the fine-tuned Typhoon-T5 model to generate responses. Meanwhile, this study found that when using the same typhoon knowledge base, the larger-parameter T5-large model does not perform much better than the smaller-parameter T5-base model. The difference in their ability to parse user questions is minimal, which might be attributed to the limited typhoon disaster data used. Finally, the analysis pointed out that due to the inherent limitations of the T5 model. It is necessary to adjust the T5 model's parameter settings to achieve variable-length response outputs. Fixed parameter settings in the T5 model can lead to redundancy in longer responses and incompleteness in shorter ones. In future planning, we intend to expand the dimensions of corpus retrieval to include diverse natural disasters information such as typhoons, earthquakes, and torrential rains, which enhances the model's cognition and analysis capabilities of disaster knowledge. Furthermore, we plan to develop a multilevel and structured disaster knowledge system incorporating knowledge representation mechanisms from knowledge graphs in fine-tuning and RAG to build a disaster information Q&A model that collaborates LLMs with knowledge graphs, further improving the understanding of disaster events in complex contexts. Meanwhile, we will use more interactive and not merely text-limited multimodal large models. In the case of processing disaster information, realizing the analysis of multimodal data (such as traffic flow data and remote sensing imagery) opens up new avenues for the comprehensive utilization of disaster knowledge.

Author Contributions: Conceptualization, Yongqi Xia, Yi Huang and Xueying Zhang; methodology, Yongqi Xia, Yi Huang and Qianqian Qiu; validation, Yongqi Xia and Qianqian Qiu; formal analysis, Yongqi Xia and Lizhi Miao; data curation, Yongqi Xia; writing—original draft preparation, Yongqi Xia; writing—review and editing, Yi Huang and Yixiang Chen; funding acquisition, Yi Huang. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China, grant number KLSMNR-G202214; NUPTSF, grant number NY220165; the Open foundation of Key Lab of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, grant number 2022VGE01.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors express their gratitude towards the journal editors and the reviewers, whose thoughtful suggestions played a significant role in improving the quality of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Knutson, T.R.; McBride, J.L.; Chan, J.; Emanuel, K.; Holland, G.; Landsea, C.; Held, I.; Kossin, J.P.; Srivastava, A.; Sugi, M. Tropical cyclones and climate change. *Nat. Geosci.* **2010**, *3*, 157–163. [CrossRef]
2. Elsner, J.B.; Elsner, S.C.; Jagger, T.H. The increasing efficiency of tornado days in the United States. *Clim. Dyn.* **2015**, *45*, 651–659. [CrossRef]
3. Murakami, H.; Wang, B. Patterns and frequency of projected future tropical cyclone genesis are governed by dynamic effects. *Commun. Earth Environ.* **2022**, *3*, 77. [CrossRef]
4. Sarker, M.N.I.; Peng, Y.; Yiran, C.; Shouse, R.C. Disaster resilience through big data: Way to environmental sustainability. *Int. J. Disaster. Risk Reduct.* **2020**, *51*, 101769. [CrossRef]
5. Yang, J.; Li, Y.; Liu, Q.; Li, L.; Feng, A.; Wang, T.; Zheng, S.; Xu, A.; Lyu, J. Brief introduction of medical database and data mining technology in big data era. *J. Evid. Based Med.* **2020**, *13*, 57–69. [CrossRef] [PubMed]
6. Zhou, C.; Su, F.; Pei, T.; Zhang, A.; Du, Y.; Luo, B.; Cao, Z.; Wang, J.; Yuan, W.; Zhu, Y. COVID-19: Challenges to GIS with big data. *Geogr. Sustain.* **2020**, *1*, 77–87. [CrossRef]
7. Naeem, M.; Jamal, T.; Diaz-Martinez, J.; Butt, S.A.; Montesano, N.; Tariq, M.I.; De-la-Hoz-Franco, E.; De-La-Hoz-Valdiris, E. Trends and future perspective challenges in big data. In *Advances in Intelligent Data Analysis and Applications, Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, Arad, Romania, 15–18 October 2019*; Springer: Singapore, 2022; pp. 309–325.
8. Liu, N.F.; Zhang, T.; Liang, P. Evaluating verifiability in generative search engines. *arXiv* **2023**, arXiv:2304.09848.
9. Shams, A.B.; Hoque Apu, E.; Rahman, A.; Sarker Raihan, M.M.; Siddika, N.; Preo, R.B.; Hussein, M.R.; Mostari, S.; Kabir, R. Web search engine misinformation notifier extension (SEMiNExt): A machine learning based approach during COVID-19 Pandemic. *Healthcare* **2021**, *9*, 156. [CrossRef] [PubMed]
10. Zaib, M.; Zhang, W.E.; Sheng, Q.Z.; Mahmood, A.; Zhang, Y. Conversational question answering: A survey. *Knowl. Inf. Syst.* **2022**, *64*, 3151–3195. [CrossRef]
11. Martinez-Gil, J. A survey on legal question–answering systems. *Comput. Sci. Rev.* **2023**, *48*, 100552. [CrossRef]
12. Huang, D.; Wei, Z.; Yue, A.; Zhao, X.; Chen, Z.; Li, R.; Jiang, K.; Chang, B.; Zhang, Q.; Zhang, S. DSQA-LLM: Domain-Specific Intelligent Question Answering Based on Large Language Model. In Proceedings of the International Conference on AI-Generated Content, Shanghai, China, 25–26 August 2023; pp. 170–180.
13. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
14. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 11809–11822.

15. Lyu, Y.; Li, Z.; Niu, S.; Xiong, F.; Tang, B.; Wang, W.; Wu, H.; Liu, H.; Xu, T.; Chen, E. CRUD-RAG: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *arXiv* **2024**, arXiv:2401.17043.
16. Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; Nanayakkara, S. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1–17. [CrossRef]
17. Tang, Y.; Yang, Y. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv* **2024**, arXiv:2401.15391.
18. Krause, A.; Cohen, S. Geographic Information Retrieval Using Wikipedia Articles. In Proceedings of the ACM Web Conference, Austin, TX, USA, 30 April–4 May 2023; pp. 3331–3341.
19. Witmer, J.T. Mining Wikipedia for Geospatial Entities and Relationships. Doctoral Dissertation, University of Colorado at Colorado Springs, Colorado Springs, CO, USA, 2009.
20. Choukolaei, H.A.; Ghasemi, P.; Goodarzian, F. Evaluating the efficiency of relief centers in disaster and epidemic conditions using multi-criteria decision-making methods and GIS: A case study. *Int. J. Disaster Risk Reduct.* **2023**, *85*, 103512. [CrossRef] [PubMed]
21. Clemente-Suárez, V.J.; Navarro-Jiménez, E.; Ruisoto, P.; Dalamitros, A.A.; Beltran-Velasco, A.I.; Hormeño-Holgado, A.; Laborde-Cárdenas, C.C.; Tornero-Aguilera, J.F. Performance of fuzzy multi-criteria decision analysis of emergency system in COVID-19 pandemic. An extensive narrative review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5208. [CrossRef] [PubMed]
22. Esmaelian, M.; Tavana, M.; Santos Arteaga, F.J.; Mohammadi, S. A multicriteria spatial decision support system for solving emergency service station location problems. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 1187–1213. [CrossRef]
23. Saha, A.K.; Agrawal, S. Mapping and assessment of flood risk in Prayagraj district, India: A GIS and remote sensing study. *Nanotechnol. Environ. Eng.* **2020**, *5*, 11. [CrossRef]
24. Yang, W.; Zhang, L.; Liang, C. Agricultural drought disaster risk assessment in Shandong Province, China. *Nat. Hazards* **2023**, *118*, 1515–1534. [CrossRef]
25. Shao, Y.; Wang, Z.; Feng, Z.; Sun, L.; Yang, X.; Zheng, J.; Ma, T. Assessment of China's forest fire occurrence with deep learning, geographic information and multisource data. *J. For. Res.* **2023**, *34*, 963–976. [CrossRef]
26. Jena, R.; Pradhan, B.; Beydoun, G. Earthquake vulnerability assessment in Northern Sumatra province by using a multi-criteria decision-making model. *Int. J. Disaster Risk Reduct.* **2020**, *46*, 101518. [CrossRef]
27. Fang, G.; Pang, W.; Zhao, L.; Cui, W.; Zhu, L.; Cao, S.; Ge, Y. Extreme typhoon wind speed mapping for coastal region of China: Geographically weighted regression–based circular subregion algorithm. *J. Struct. Eng.* **2021**, *147*, 04021146. [CrossRef]
28. Wang, S.; Mu, L.; Yao, Z.; Gao, J.; Zhao, E.; Wang, L. Assessing and zoning of typhoon storm surge risk with a geographic information system (GIS) technique: A case study of the coastal area of Huizhou. *Nat. Hazards Earth Syst. Sci.* **2021**, *21*, 439–462. [CrossRef]
29. Wu, K.; Wu, J.; Ding, W.; Tang, R. Extracting disaster information based on Sina Weibo in China: A case study of the 2019 Typhoon Lekima. *Int. J. Disaster Risk Reduct.* **2021**, *60*, 102304. [CrossRef]
30. Zhang, T.; Cheng, C. Temporal and spatial evolution and influencing factors of public sentiment in natural disasters—A case study of typhoon haiyan. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 299. [CrossRef]
31. Sufi, F.K.; Khalil, I. Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis. *IEEE Trans. Comput. Soc. Syst.* **2022**, 1–11. [CrossRef]
32. Rao, P.R.; Jhawar, T.N.; Kachave, Y.A.; Hirlekar, V. Generating QA from Rule-based Algorithms. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 16–18 March 2022; pp. 1697–1703.
33. Thorat, S.A.; Jadhav, V. A review on implementation issues of rule-based chatbot systems. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC), Delhi, India, 21–23 February 2020.
34. Jin, S.; Lian, X.; Jung, H.; Park, J.; Suh, J. Building a deep learning-based QA system from a CQA dataset. *Decis. Support Syst.* **2023**, *175*, 114038. [CrossRef]
35. Abdel-Nabi, H.; Awajan, A.; Ali, M.Z. Deep learning-based question answering: A survey. *Knowl. Inf. Syst.* **2023**, *65*, 1399–1485. [CrossRef]
36. Huang, X.; Zhang, J.; Li, D.; Li, P. Knowledge graph embedding based question answering. In Proceedings of the twelfth ACM international conference on web search and data mining, Melbourne VIC, Australia, 11–15 January 2019; pp. 105–113.
37. Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.H.; Riedel, S. Language models as knowledge bases? *arXiv* **2019**, arXiv:1909.01066.
38. Da, J.; Bras, R.L.; Lu, X.; Choi, Y.; Bosselut, A. Analyzing commonsense emergence in few-shot knowledge models. *arXiv* **2021**, arXiv:2101.00297.
39. Safavi, T.; Koutra, D. Relational world knowledge representation in contextual language models: A review. *arXiv* **2021**, arXiv:2104.05837.
40. Hu, Y.; Mai, G.; Cundy, C.; Choi, K.; Lao, N.; Liu, W.; Lakhanpal, G.; Zhou, R.Z.; Joseph, K. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *Int. J. Geogr. Inf. Sci.* **2023**, *37*, 2289–2318. [CrossRef]
41. Bhandari, P.; Anastasopoulos, A.; Pfoser, D. Are large language models geospatially knowledgeable? In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, Hamburg, Germany, 13–16 November 2023; pp. 1–4.

42. Jiang, Z.; Araki, J.; Ding, H.; Neubig, G. How can we know when language models know? on the calibration of language models for question answering. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 962–977. [CrossRef]
43. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef] [PubMed]
44. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]
45. Cui, J.; Li, Z.; Yan, Y.; Chen, B.; Yuan, L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv* **2023**, arXiv:2306.16092.
46. Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; Mann, G. Bloomberggpt: A large language model for finance. *arXiv* **2023**, arXiv:2303.17564.
47. Yang, H.; Liu, X.-Y.; Wang, C.D. FinGPT: Open-Source Financial Large Language Models. *arXiv* **2023**, arXiv:2306.06031. [CrossRef]
48. Huang, J.; Wang, H.; Sun, Y.; Shi, Y.; Huang, Z.; Zhuo, A.; Feng, S. ERNIE-GeoL: A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 3029–3039.
49. Gao, Y.; Xiong, Y.; Wang, S.; Wang, H. GeoBERT: Pre-Training Geospatial Representation Learning on Point-of-Interest. *Appl. Sci.* **2022**, *12*, 12942. [CrossRef]
50. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv* **2024**, arXiv:2401.16822.
51. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv* **2023**, arXiv:2312.10115.
52. Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; Xiao, P. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. *arXiv* **2024**, arXiv:2402.02544.
53. Ni, J.; Ábrego, G.H.; Constant, N.; Ma, J.; Hall, K.B.; Cer, D.; Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv* **2021**, arXiv:2108.08877.
54. Karimzadeh, M.; Pezanowski, S.; MacEachren, A.M.; Wallgrün, J.O. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Trans. GIS* **2019**, *23*, 118–136. [CrossRef]
55. Khattab, O.; Zaharia, M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China, 25–30 July 2020; pp. 39–48.
56. Zhou, Y.; Li, C.; Huang, G.; Guo, Q.; Li, H.; Wei, X. A Short-Text Similarity Model Combining Semantic and Syntactic Information. *Electronics* **2023**, *12*, 3126. [CrossRef]
57. Bag, S.; Kumar, S.K.; Tiwari, M.K. An efficient recommendation generation using relevant Jaccard similarity. *Inf. Sci.* **2019**, *483*, 53–64. [CrossRef]
58. Verma, V.; Aggarwal, R.K. A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: Empirical and theoretical perspective. *Soc. Netw. Anal. Min.* **2020**, *10*, 43. [CrossRef]
59. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
60. Shazeer, N.; Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4596–4604.
61. Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685.
62. Wang, C.; Cheng, S.; Xu, Z.; Ding, B.; Wang, Y.; Zhang, Y. Evaluating open question answering evaluation. *arXiv* **2023**, arXiv:2305.12421.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.