The Journal of Navigation (2025), 1–16
doi:10.1017/S0373463325101148

RESEARCH ARTICLE

Nowcasting of typhoon tracks based on LLM and RAG

Baichuan Peng, Youchao Jiang and Li Yao

Merchant Marine College, Shanghai Maritime University, Shanghai, China
Corresponding author: Li Yao; Email: yaoli@shmtu.edu.cn

Received: 29 June 2024; **Revised:** 01 June 2025; **Accepted:** 23 July 2025

Keywords: LLM; nowcasting of typhoon tracks; RAG

Abstract

Accurate typhoon track nowcasting is vital for navigation and coastal disaster prevention. This research integrates a Large Language Model (LLM) with Retrieval-Augmented Generation (RAG) technology for typhoon path prediction. Leveraging LLMs as the predictive foundation, the approach tailors forecasts to individual typhoon characteristics. The methodology involves collecting satellite imagery, standardizing data, and employing optical flow methods to track typhoons and derive path coordinates. These coordinates are preprocessed and embedded into the LLM. RAG enhances the LLM's predictive performance, enabling effective forecasting. Increasing typhoon-specific embedded data further improves accuracy. Using the FY-4 dataset, the method achieved an average absolute error of 10.78 km in 12-hour predictions. The study demonstrates that LLM-RAG integration excels in nowcasting.

1. Introduction

Typhoons are a type of tropical cyclone, formed by tropical oceanic atmospheric circulation systems, characterised by intense rotating storms. These storms are often accompanied by hazardous weather conditions such as strong winds, heavy rain, floods and storm surges, causing significant damage to areas along their path, including destruction of buildings, infrastructure and crops, as well as loss of life. For instance, in 2023, Typhoon Doksuri struck Fujian Province in China, affecting 725,000 people, causing power outages for 1.21 million households and resulting in direct economic losses of approximately 177 million RMB. As the second most powerful typhoon to make landfall in Fujian since 1949, Doksuri's destructive force caused extensive damage to the communities along its path. This underscores the critical importance of accurate typhoon path prediction in safeguarding lives and property. Particularly for super typhoons, short-term forecasting is crucial for ships unable to return to port and for residents in affected areas to obtain timely information to avoid the typhoon and protect themselves. The nowcasting typhoon track prediction proposed in this study targets the prediction of typhoon movement within 0–12 h. This time frame is set based on relevant standards from the World Meteorological Organization. According to these standards, the 'nowcasting' of tropical cyclones typically refers to forecasts within 0–6 h. However, some studies extend this period to 12 h for the sake of operational continuity.

Traditional methods for predicting typhoon tracks primarily consist of numerical statistical methods and meteorological dynamics methods. Numerical statistics and meteorological physical models can also be combined. For instance, Chen and Duan (2018) employ a statistical-dynamic model to assess typhoon risk in southeastern China. This model comprises generation, movement and intensity models, improving traditional empirical footprint models through statistical-dynamic relationships. Hon (2020) uses the WRF model for typhoon path prediction. However, these prediction methods require extensive data support and high hardware demands. To address the shortcomings of traditional forecasting methods, many researchers have turned to machine learning technologies for predicting typhoon paths.

Chen et al. (2020) indicate that combining decision tree analysis, random forests and deep learning methods can enhance the accuracy of tropical cyclone path predictions. While these methods have improved prediction accuracy, challenges remain in feature extraction and response to external disturbances. Consequently, other researchers have opted for predictive models using recurrent neural networks (RNNs) and their advanced variant, long short-term memory (LSTM) networks (Pang et al., 2021; Gao et al., 2018; Suo et al., 2020; Jiang et al., 2024; Liang et al., 2023). Gao et al. (2018) trained an LSTM neural network using typhoon path observation data from 1949 to 2012, provided by the North China Sea Prediction Center. By comparing the predicted results with observed data, the feasibility of LSTM networks in typhoon path prediction was validated and the impact of different training dataset sizes on prediction accuracy was studied. Additionally, Song et al. (2022) used CNN layers to model and extract spatial relationships between meteorological variables and tropical cyclone positions, and GRU layers to mine deep features of time series, constructing an integrated deep learning model. Jabbar et al. (2021) mentions using GAN models to predict typhoon paths, addressing issues in GAN models such as Nash equilibrium, internal covariate shift, mode collapse, gradient vanishing and lack of proper evaluation metrics. These models can quickly and accurately perform future predictions. Beyond the aforementioned models, some researchers have enhanced prediction model performance by diversifying information feature extraction (Jiang et al., 2024; Liang et al., 2023; Qin et al., 2022; Fu et al., 2022; Ren et al., 2022; Lu et al., 2022a, 2022b, 2023, Li et al., 2024). Ren et al. (2022) combines convolutional neural networks (CNNs) and LSTM networks, proposing a deep learning model named DeepTyphoon for predicting typhoon paths. By labelling satellite images and combining CBAM's and LSTM's feature extraction and prediction methods, the model predicts typhoon paths using two datasets released by the Japan Meteorological Agency (JMA). Lu et al. (2022) processes typhoon path data through first-order differencing and cointegration tests to ensure data stationarity and cointegration relationships among variables. Subsequently, a C-LSTM model, combining CNN and LSTM networks, was constructed for typhoon path prediction. The study experimentally verifies the advantages of the C-LSTM model in typhoon path prediction and compares it with traditional LSTM models. Lu et al. (2022) employs a ConvLSTM model, comprising convolution operations and gating units, to better extract image features. By marking the typhoon centre's positions before and after on the generated heatmap of physical variables, the model learns these key features. Single-layer and multi-layer ConvLSTM experiments were conducted to validate the performance improvement of the multi-layer model. Additionally, some researchers have combined transformer and time-series analysis methods to enhance model performance (Zhan et al., 2023; Jiang et al., 2023; Jung et al., 2024). For instance, Zhan et al. (2023) construct a method based on TRAN, achieving prediction errors mostly below an average error of 40.93 km, with the error frequency below the average error being 62.74%. Although this method shows reliability, it still requires validation and improvement in practical applications. Issues such as model complexity, workload, forgetting problems and parameter adjustment complexity persist.

In recent years, the Large Language Model (LLM) has rapidly advanced, with explosive growth in various application modes across different fields (Chib, 2024). Peng et al. (2024) reported a novel interpretable lane change prediction model named LC-LLM, using LLM's powerful reasoning and self-explanation capabilities, which has shown significant improvements in prediction accuracy and interpretability. Lan et al. (2025) propose a dynamic traffic trajectory prediction method based on LLM, named Traj-LLM, envisioning its application in autonomous driving. Traj-LLM achieves high-precision prediction of complex trajectories and the embedment of physical rules by encoding multimodal trajectory data into language sequences, using the long-range dependency modelling and prompt-controlled generation capabilities of large language models. This approach allows for encoding trajectory points, time, and environmental factors (such as wind speed and pressure) into natural language sequences, breaking through the data fragmentation bottleneck of traditional models to form unified multimodal modelling. By dynamically controlling the output with prompts (such as 'predict the path under the westward extension of the subtropical high'), it supports real-time scenario simulations for interactive controlled generation. Chib (2024), Peng et al. (2024) and Lan et al. (2025) all use large language models as the foundational models for prediction, processing the data used in the research to

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

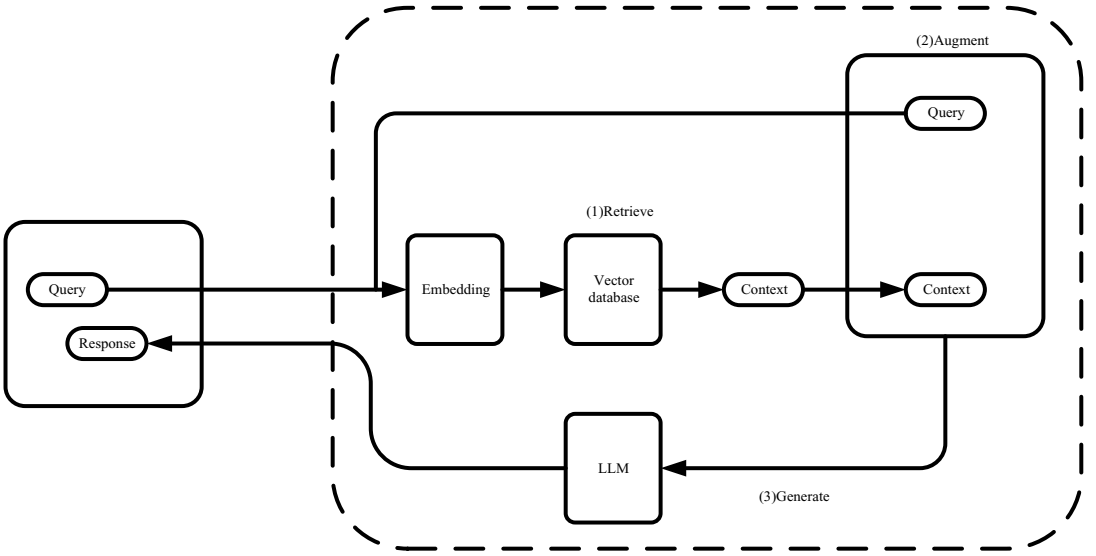

Figure 1. *Diagram of RAG structure principle.*

make it understandable by the large language models, and incorporating them into prediction scenarios across various research contexts. Human movement trajectories and vehicle movement trajectories provide insights into how large language models can be applied to typhoon trajectory prediction. Considering that typhoon movement is a manifestation of the complex interaction of weather systems, the study focuses on short-term intervals in near-term predictions. Additionally, given the characteristics of typhoon trajectory data, retrieval-enhanced generation technologies are employed to update the large language model in real-time, enabling interaction with the model to complete the prediction process. Based on the application experience of these large language models in other fields, the research on near-term prediction of typhoon trajectories using Retrieval-Augmented Generation (RAG) technology and LLM is feasible. Synthesising the latest advancements in trajectory prediction and previous work in typhoon path prediction, this study is the first to propose the application of combining LLM and RAG for typhoon path prediction.

2. Methodology

2.1. Related work

2.1.1. RAG

Retrieval-Augmented Generation combines retrieval and generation technologies to tackle complex natural language processing tasks (Munir and Sheraz Anjum, 2018). In this study, the Gemini large language model is used as the foundation, optimised for the task of typhoon trajectory prediction, with RAG introduced to enhance performance.

The basic principles of RAG can be summarised in the following steps (as shown in Figure 1).

1. Data Retrieval: based on the input.
2. Information Integration: the retrieved information is integrated with the input query.
3. Generation: the integrated data are used to generate the final output.

The formulation of the probability distribution for the generated output is given by

$$P_\eta(z|x) = Retriever(x, \eta) \tag{1}$$

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

$$P_\theta(y|x,z) = Generator(x,z,\theta) \tag{2}$$

where θ represents the parameters of the generator.

Additionally, RAG involves several supplementary technologies such as attention mechanisms and fusion strategies to better integrate the retrieved information with the input query and generate more accurate responses. While these technologies may introduce additional parameters and computational steps, the fundamental principles and equations remain based on the aforementioned retriever and generator.

In this study, the Retrieval-Augmented Generation question-answering (RAG QA) model was employed to enhance the performance of large language models. An augmented human–computer interaction interface was used to accomplish specific tasks. The detailed process is as follows.

The RAG QA approach was specifically selected, which combines retrieval and generation. This method retrieves information relevant to the questions from a large corpus and uses this information to assist in generating more accurate and targeted answers. The performance of the LLM, enhanced through RAG technology, was improved, enabling better understanding of user input and providing answers that meet expectations.

To implement this process, a human–computer interaction interface was constructed using the railway framework. Users can input questions through this interface and interact with the enhanced LLM. The interface transmits the user's question to the RAG system, which retrieves relevant information from the corpus and inputs it along with the question into the LLM for processing. The LLM generates answers based on the retrieved information and the question, and then returns the answers to the user. This approach resulted in an efficient, accurate and user-friendly human–computer interaction system.

The specific process is illustrated in Figure 2: (1) the user poses a question through the interface; (2) the server processes the question (e.g. rephrasing), followed by retrieval services; (3, 4) retrieval calls local documents and databases for more efficient search; (5) generating relevant contextual data; (6, 7) the integrated prompt template is used to call the LLM; (8, 9) the LLM generates the answer, which is parsed and post-processed through the prompt; and (10) the final answer is presented to the user. This process seamlessly integrates retrieval and generation, providing users with a more intelligent and convenient service experience.

2.1.2. Lucas–Kanade

The Lucas–Kanade method (Yang, 2018) is used to determine the motion vectors of each pixel in an image by minimising the sum of squared pixel residuals. For the specific task of image processing–analysing the relative motion between two input images, the following steps are proposed to estimate the motion of the camera or objects.

Selection of feature points within the typhoon region: initially, the typhoon region is identified from two satellite cloud images. Within this region, a series of feature points are selected, evenly distributed along the typhoon's contour. These points possess significant characteristics, including the typhoon centre, corners or edge points.

Calculation of local image gradients: for each selected feature point, its precise location in the first cloud image is determined. Based on the pixel information surrounding this point, a local image gradient is computed to capture the variation trends in the image structure near the point.

Feature point matching and iterative optimisation: in the second cloud image, the position of the feature point from the first image is used as the initial position. Using the local image gradient calculated in the first step, an iterative search is conducted along the gradient direction, as described by equations (4) and (5), to find the optimal matching position that minimises the pixel residuals between the two images at the corresponding feature point locations. This iterative process continuously optimises the feature point positions until a predefined convergence condition is met.

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

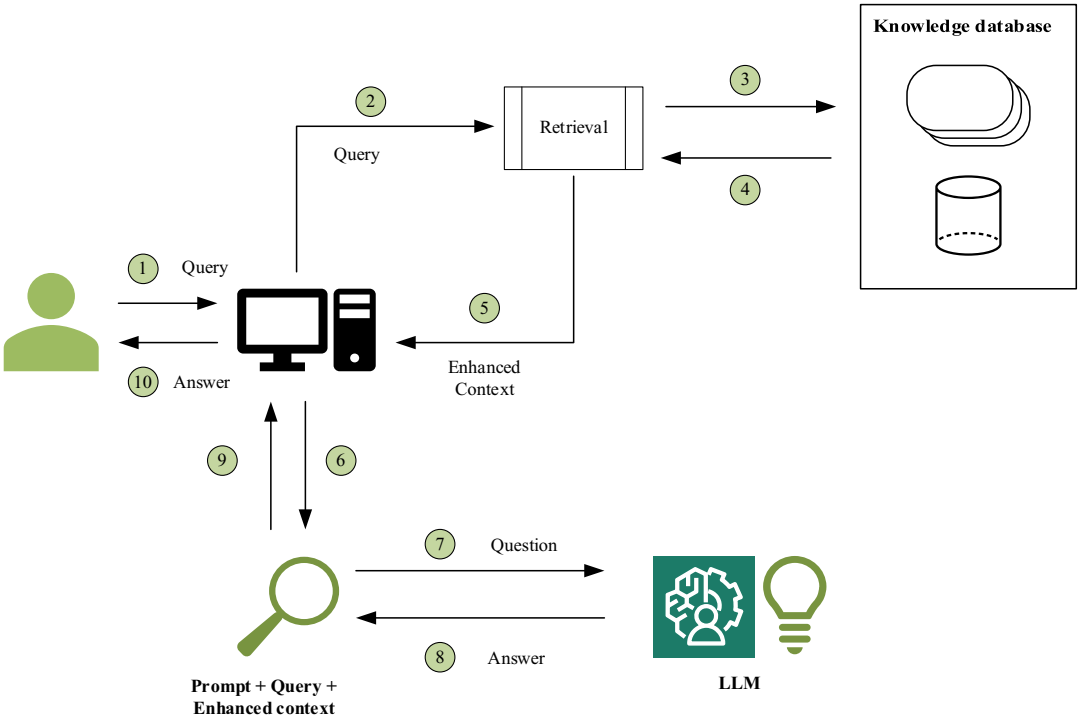

Figure 2. *Question and answering based on RAG and LLM.*

$$\sum_{\Delta p^{x,y}}^{\min} (I_1(x,y) - I_2(x+\Delta x, y+\Delta y))^2 \tag{3}$$

In this context, $\Delta p=[\Delta x,\Delta y]$ represents the displacement vector. By differentiating this equation, a system of linear equations can be obtained:

$$\begin{bmatrix} \sum_{x,y} I_x^2 & \sum_{x,y} I_x I_y \\ \sum_{x,y} I_x I_y & \sum_{x,y} I_y^2 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = -\begin{bmatrix} \sum_{x,y} I_x (I_2 - I_1) \\ \sum_{x,y} I_y (I_2 - I_1) \end{bmatrix} \tag{4}$$

where I_x and I_y denote the gradients in the x and y directions, respectively.

Global motion vector calculation: the process of feature point matching and iterative optimisation is repeated until the corresponding motion vectors for all selected feature points are computed. These motion vectors reflect the displacement of each feature point in the cloud images.

2.2. Datasets

2.2.1. Cloud data of the typhoon

This study used real-time infrared cloud image data provided by the FY-4 meteorological satellite, collecting over 8,000 images spanning from July 28, 2023, to September 17, 2023. These images cover the East Asia region from the coast of China to the Western Pacific. It is worth noting that over 8,000 images are sourced from multiple typhoons, including Typhoons Khanun, Saola and Haikui.

High-quality datasets are crucial for monitoring typhoon dynamics and issuing potential meteorological disaster alerts affecting oceanic regions. The FY-4A meteorological satellite, as a geostationary satellite, provides high-quality medium-resolution images for real-time weather

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

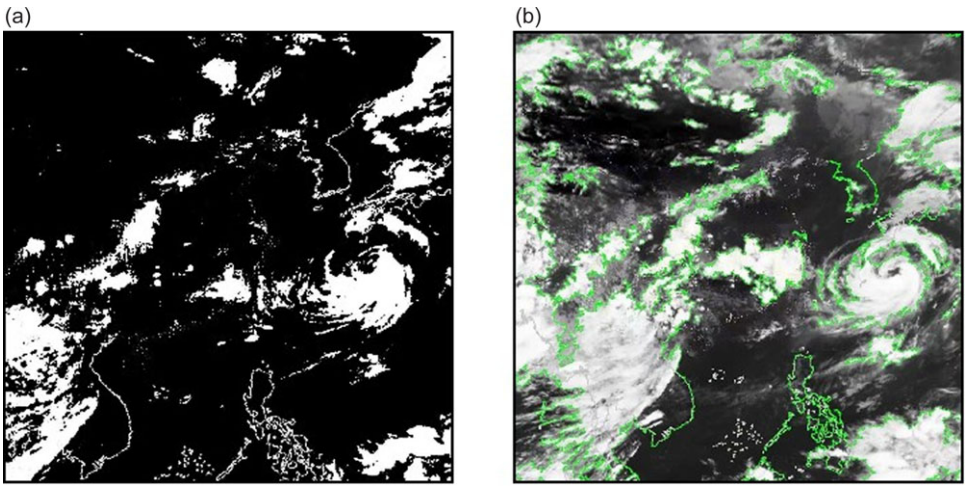

Figure 3. *(a) Binary image. (b) Image after boundary processing.*

monitoring, making it particularly suitable for typhoon observation and early warning. By combining automatic contour segmentation with manual key-point selection, complex typhoon image data are simplified into a series of typhoon particles, facilitating the analysis of typhoon dynamics using mathematical models and computational methods. Analysing Typhoon Kanu's data allows for the study of its movement trajectory and speed variation, which is essential for understanding and predicting typhoon behaviour. This analysis aids in better understanding typhoon development patterns, thereby improving the accuracy of typhoon path predictions.

The processed cloud image data include information such as the typhoon's latitude and longitude coordinates, direction, and speed. By automatically delineating typhoon contours in the images and manually selecting 35 tracking points, a series of typhoon particles are formed. The main focus of this study is analysing Typhoon Kanu's data, with each analysis conducted in 6-h time steps. Additionally, the FY-4A satellite, representing the top level of Chinese meteorological satellites, ensures the quality of cloud image data used in typhoon trajectory point tracking research through its high-resolution, high-timeliness, multi-spectral, stable and accurate images.

2.2.2. Data processing

In academic research, data acquisition and preprocessing are generally divided into two distinct yet interrelated processes. Initially, the collected images undergo a series of processing steps, which include reading the input images and converting them to the HSV colour space. This is followed by applying colour thresholding operations to create mask images for different colour regions and performing morphological operations (such as erosion and dilation) on the mask images to remove noise and fill gaps. The purpose of this step is to provide clear and denoised images for subsequent analysis.

The second step involves the selection of particles and the calculation of feature parameters from the preprocessed images. This includes finding contours in the processed mask images and generating the original images with boundaries by drawing these contours. Subsequently, optical flow analysis is conducted on consecutive image frames using the Lucas–Kanade method to track feature points. Finally, the displacement distance and direction of the feature points between consecutive frames are calculated. The aim of this step is to extract key information regarding cloud motion, including its position, speed and direction.

Overall, the combined application of technologies such as colour segmentation, morphological operations and optical flow analysis enables the detection and analysis of cloud motion. This process effectively visualises the speed and direction of clouds, as illustrated in Figure 3, which shows the

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

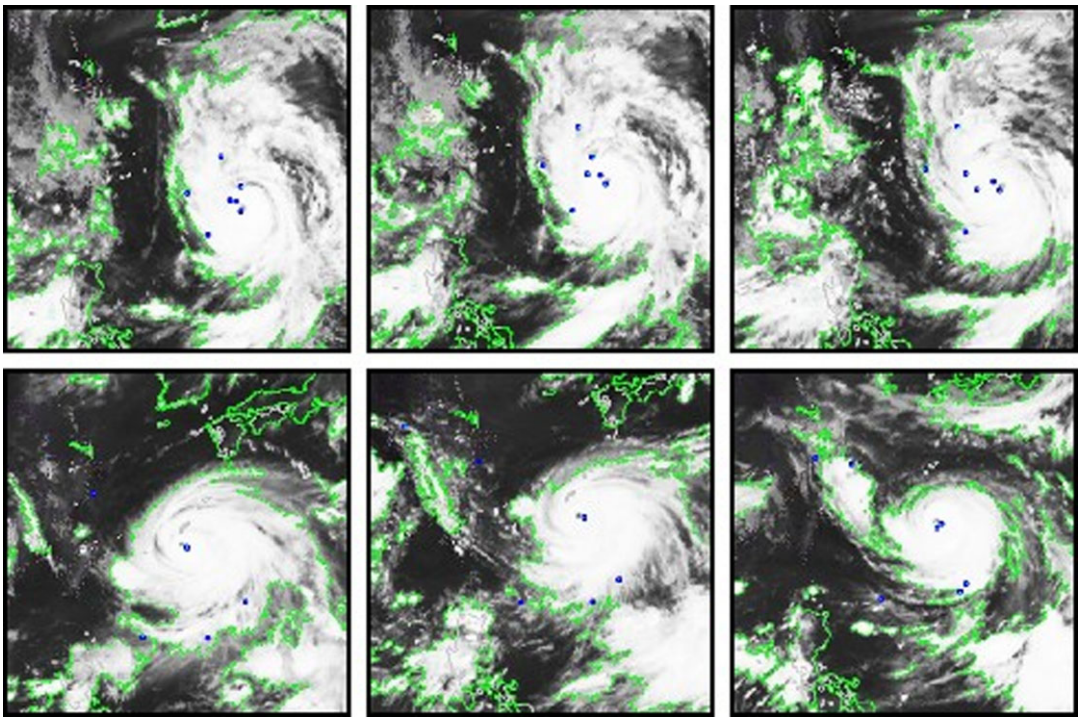

Figure 4. *Changes in the tracked points marked during the typhoon within 36 h.*

processed cloud images. This methodological approach reveals the dynamic characteristics of cloud movement, providing crucial data support for further meteorological research and decision-making.

2.2.3. Haversine

In cloud image analysis, not all initially selected points can maintain accuracy over extended tracking periods. As shown in Figure 4, which illustrates the initial annotation and subsequent changes of typhoon feature points over a 36-h period, some feature points deviate from the typhoon's contour during continuous tracking. Therefore, this study aims to identify and select points that accurately represent the typhoon's central movement trajectory, and to collect their latitude and longitude coordinates. To quantify the horizontal displacement of these selected feature points over a specific period, the study employs the Haversine formula to calculate the distance between the tracked points at different times. The distance between tracked points at different times is calculated using the Haversine formula (Mahmoud and Akkari, 2016), where (R) is the Earth's radius, valued at 6,371 km:

$$D = 2R\sin^{-1}\sqrt{\sin\left(\frac{lat_{\text{real}} - lat_{\text{pred}}}{2} \times \frac{\pi}{180}\right)^2 + \cos\left(lat_{\text{pred}} \times \frac{\pi}{180}\right)\cos\left(lat_{\text{real}} \times \frac{\pi}{180}\right)\sin\left(\frac{lon_{\text{real}} - lon_{\text{pred}}}{2} \times \frac{\pi}{180}\right)^2} \quad (5)$$

2.3. RAG establishment and prediction

2.3.1. Model establishment and improvement

In this study, the model leverages RAG technology to enhance the performance of the LLM in understanding and answering complex questions. By integrating RAG, the LLM can retrieve and incorporate extensive historical typhoon data, significantly improving its contextual understanding and accuracy in responses. This approach allows the model to combine retrieved information with the

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

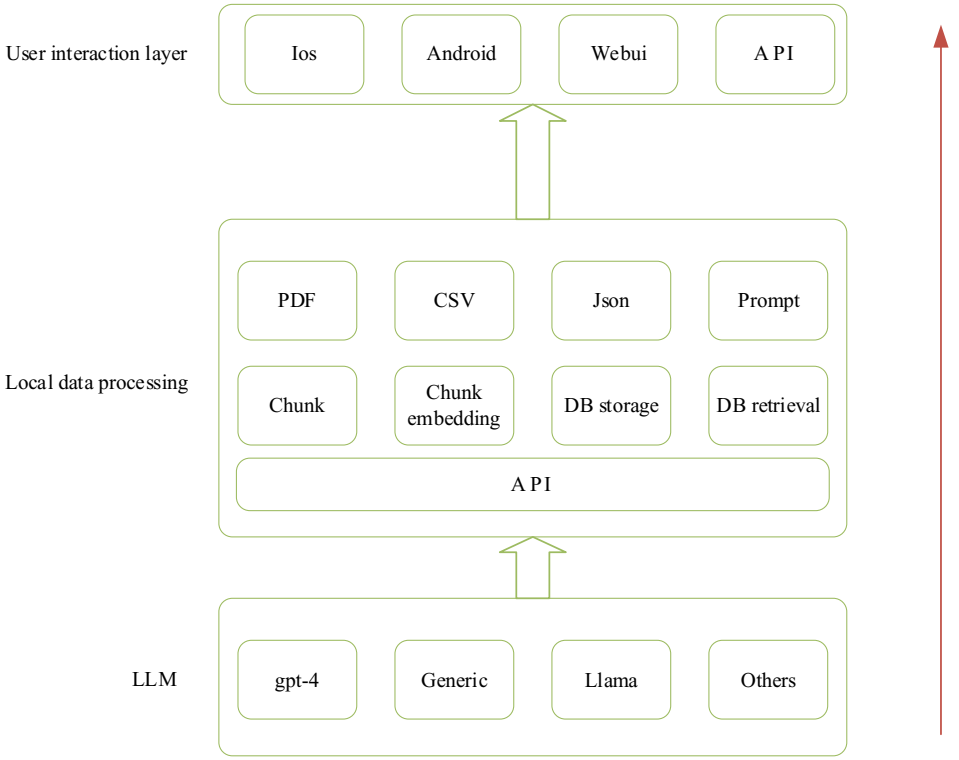

Figure 5. *Deployment architecture of the prediction model.*

current context, resulting in deeper comprehension of the questions and the generation of more precise, detailed, and relevant texts and answers. The synergy of a broad knowledge base and targeted data retrieval ensures the model's superior performance in handling complex queries, thereby improving the quality and relevance of its responses.

In a multi-layered system, the overall framework is illustrated in Figure 5. The first layer serves as the user interaction layer, primarily responsible for deploying user interfaces on Web UI or iOS and Android platforms. This layer aims to provide intuitive user experience and also offers API call services to meet specific prediction requirements and cater to user needs.

The second layer, functioning as the local knowledge processing and vector retrieval layer, is primarily responsible for efficiently handling external input documents. This includes, but is not limited to, document parsing, information extraction, and necessary preprocessing tasks to ensure the accuracy and efficiency of subsequent analyses. The third layer, positioned at the system's foundation, serves as the large language model layer. This layer provides robust language processing capabilities, enabling the invocation of third-party large language models or the use of locally deployed models as a basis for RAG services. Given that running large language models locally typically requires extremely high hardware configurations, which exceed the capabilities of most computers, API calls are employed when interacting with the aforementioned layers. This method ensures service reliability while reducing dependency on hardware resources.

This layered architecture design not only offers flexible user interaction methods but also ensures efficient data processing and accurate language comprehension, thereby providing robust support for various prediction scenarios.

The specific dialogue process for deploying local data into a large language model is illustrated in Figure 6. The local data are processed into JSON file format and embedded in this format, then saved into the OpenAI GPT-4o model via API calls. Subsequently, the data are loaded into the specific predictive dialogue using the pin function. This approach enables the application of local data in predictive methods.

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

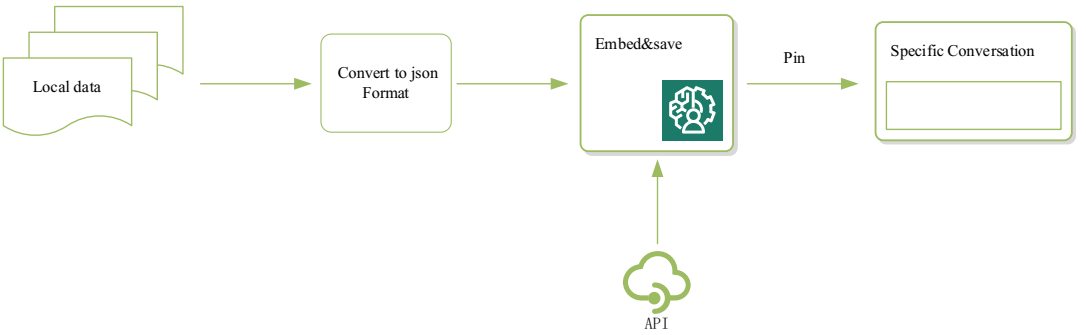

Figure 6. *Specific conversation process for embedding local files into an LLM.*

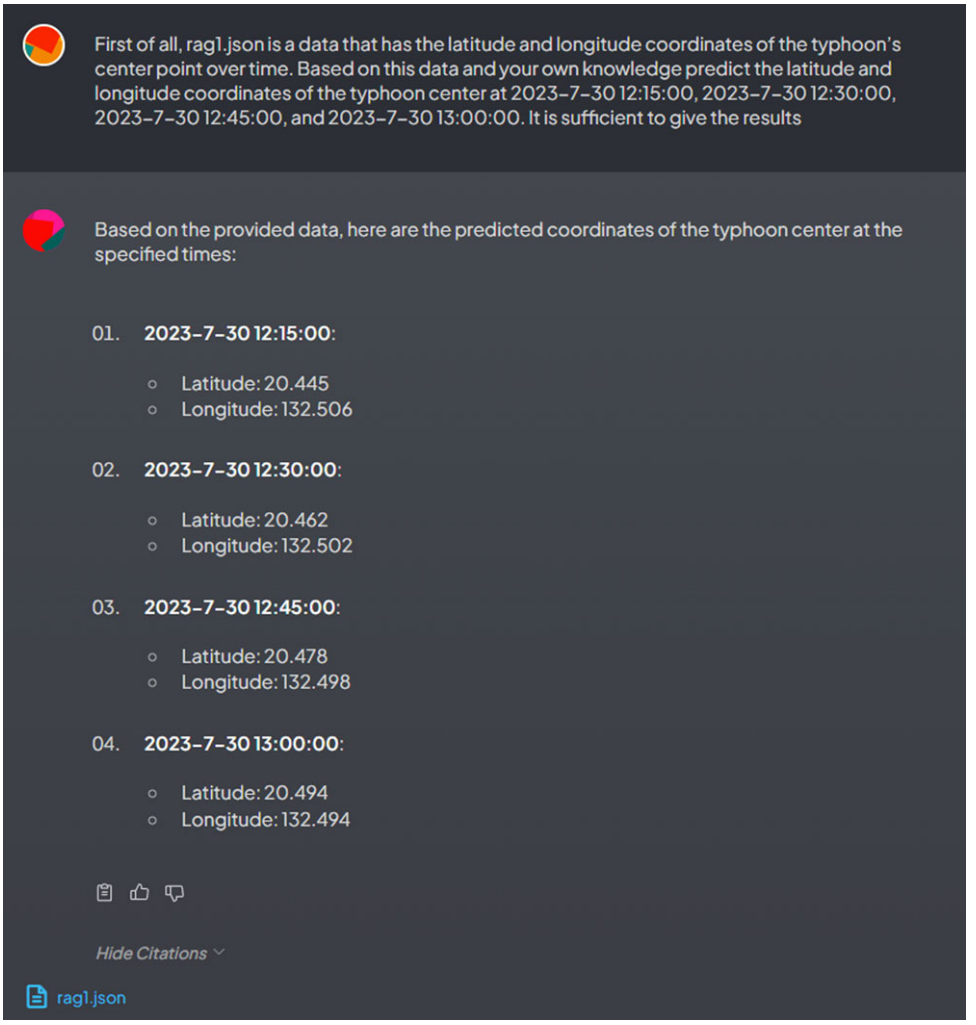

Figure 7. *Prediction results with direct querying method.*

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

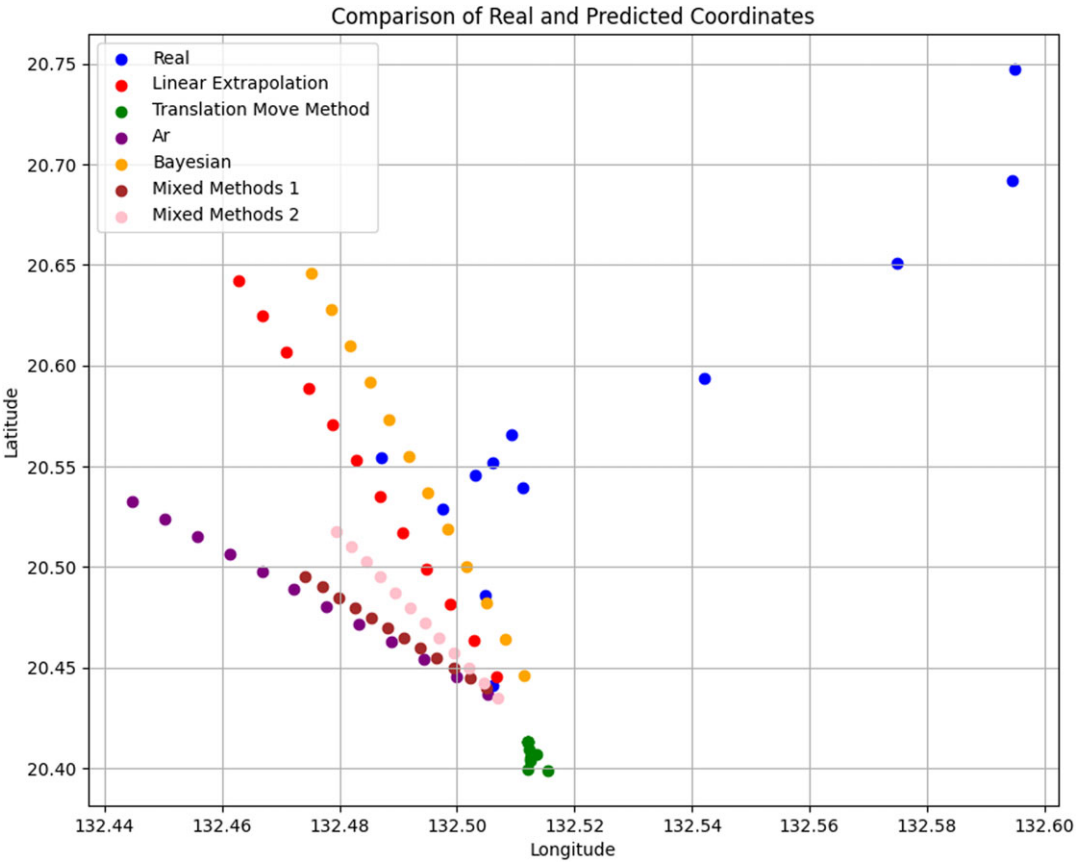

Figure 8. *Comparison of real and predicted coordinates using various methods.*

Through the user interface, a prediction command is issued to the locally deployed LLM, as illustrated in Figure 7. Initially, the deployed local data are described. Subsequently, the specific time points for prediction are provided, instructing the LLM to use the relevant knowledge and the uploaded local data to forecast the future trajectory of the typhoon. The LLM then generates the prediction results.

Evidently, the prediction results provided by the LLM in Figure 8 appear to follow only some basic mathematical patterns. As shown in Figure 8, various methods were employed in the study to query the LLM, and the prediction results from these different methods were plotted for comparison. The outcomes were far from satisfactory, with the trajectories primarily derived using simple linear extrapolation technologies. These suboptimal results were attributed to issues such as hallucinations in the large model, which led to errors in file loading and inadequate prompt phrasing during questioning.

To address this issue and enhance the final prediction outcomes, one solution is to limit the amount of local data embedded within the platform. Additionally, prior to making predictions, the LLM can be guided to display the already uploaded local datasets, as shown in Figure 9.

Incorporating more specialised terminology, such as ‘long-term patterns’ and ‘short-term trends’, and considering factors like volatility and atmospheric characteristics, can help mitigate the issue of insufficient prompt phrases. Using highly specialised vocabulary to constrain and direct the LLM can result in more accurate and effective predictions.

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

Table 1. *Configuration of experimental platform*

Parameters and Device Name	Values and Version
LLM Selection	Generic OpenAI compatible service
Embedding Preference	AnythingLLM Embedder
Vector Database	LanceDB
Chat Model Name	GPT-4o

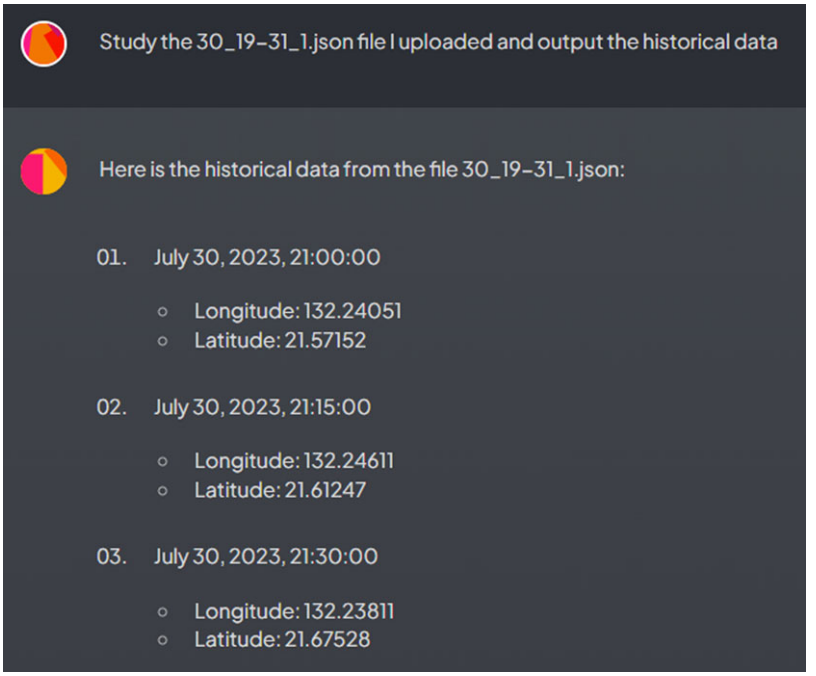

Figure 9. *Prediction results with optimised querying method.*

3. Result and analysis

3.1. Experimental environment

The experimental environment is shown in Table 1.

3.2. Evaluation indicators of method

The accuracy of the predicted location of the typhoon's centre is evaluated by calculating the Mean Absolute Error (MAE) for distance, longitude and latitude, as well as the Root Mean Squared Error (RMSE) for distance. The absolute error is determined using the Haversine formula, and the Mean Absolute Error is then computed based on the sequence number (n) of the test set. Additionally, the Root Mean Squared Error is calculated based on the absolute error. The calculation formulae are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{6}$$

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

Table 2. Comparison of results for different prediction time intervals

Data	Time Range (h)	MAE_{lat}	MAE_{lon}	*MAE* (km)	*RMSE* (km)
2023.8.7	2(10:00–12:00)	0.047	0.051	7.45	7.83
2023.8.8	4(7:00–12:00)	0.100	0.066	13.19	16.10
2023.8.7	7(11:00–18:00)	0.149	0.078	19.92	22.56
2023.8.8	12(8:00–20:00)	0.063	0.067	10.78	15.20

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} D_i^2} \tag{7}$$

$$\mathrm{MAE}_{lat} = \frac{1}{n}\sum_{i=1}^{n} \left|\mathrm{lat}_{real} - \mathrm{lat}_{pred}\right| \tag{8}$$

$$\mathrm{MAE}_{lon} = \frac{1}{n}\sum_{i=1}^{n} \left|\mathrm{lon}_{real} - \mathrm{lon}_{pred}\right| \tag{9}$$

where lat_{real} represents the actual latitudinal coordinate of the current typhoon centre, lat_{pred} represents the predicted latitudinal coordinate of the current typhoon centre, lon_{real} represents the actual longitudinal coordinate of the current typhoon centre and lon_{pred} represents the predicted longitudinal coordinate of the current typhoon center.

3.3. Experimental result

In the experiment, the prediction of typhoon centre trajectories was conducted by embedding a small amount of pre-prediction trajectory coordinate data into the LLM. These data, consisting of approximately 100 trajectory points from the 12 h prior to the prediction time, had time intervals ranging from 4 to 15 min.

Table 2 provides the prediction error results for four time periods and indicates the specific times of the prediction trajectories. The predicted trajectories for four-time spans (2 h, 4 h, 7 h and 12 h) were consistent with the actual typhoon trajectory trends. Figure 10 compares the predicted and actual trajectories of Typhoon Khanun from July to August 2023, with the red dashed line representing the predicted path and the blue solid line representing the actual path. The predicted typhoon centre coordinates demonstrated excellent performance in terms of both latitude/longitude errors and spherical distance errors.

However, the proposed method has certain limitations. As the prediction time span increases, the errors tend to grow, as shown in Table 2. The prediction errors for 2-h, 4-h and 7-h spans increase with longer time spans. This indicates that for long-term predictions, the performance of the proposed method gradually deteriorates. Interestingly, at a 12-h time span, the error decreases and the prediction performance is even better than that of the 4-h span. This anomaly is likely related to the complexity of typhoon trajectory changes over certain time periods.

3.4. Comparison with other methods

Table 3 provides a detailed comparison of the prediction errors between the proposed method in this study and other existing methods. Our approach employs LLM in conjunction with RAG technology to predict typhoon trajectories, demonstrating superior accuracy in error metrics. The comparative data for our method are based on a 12-h prediction horizon. It is noteworthy that the prediction horizons of the

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

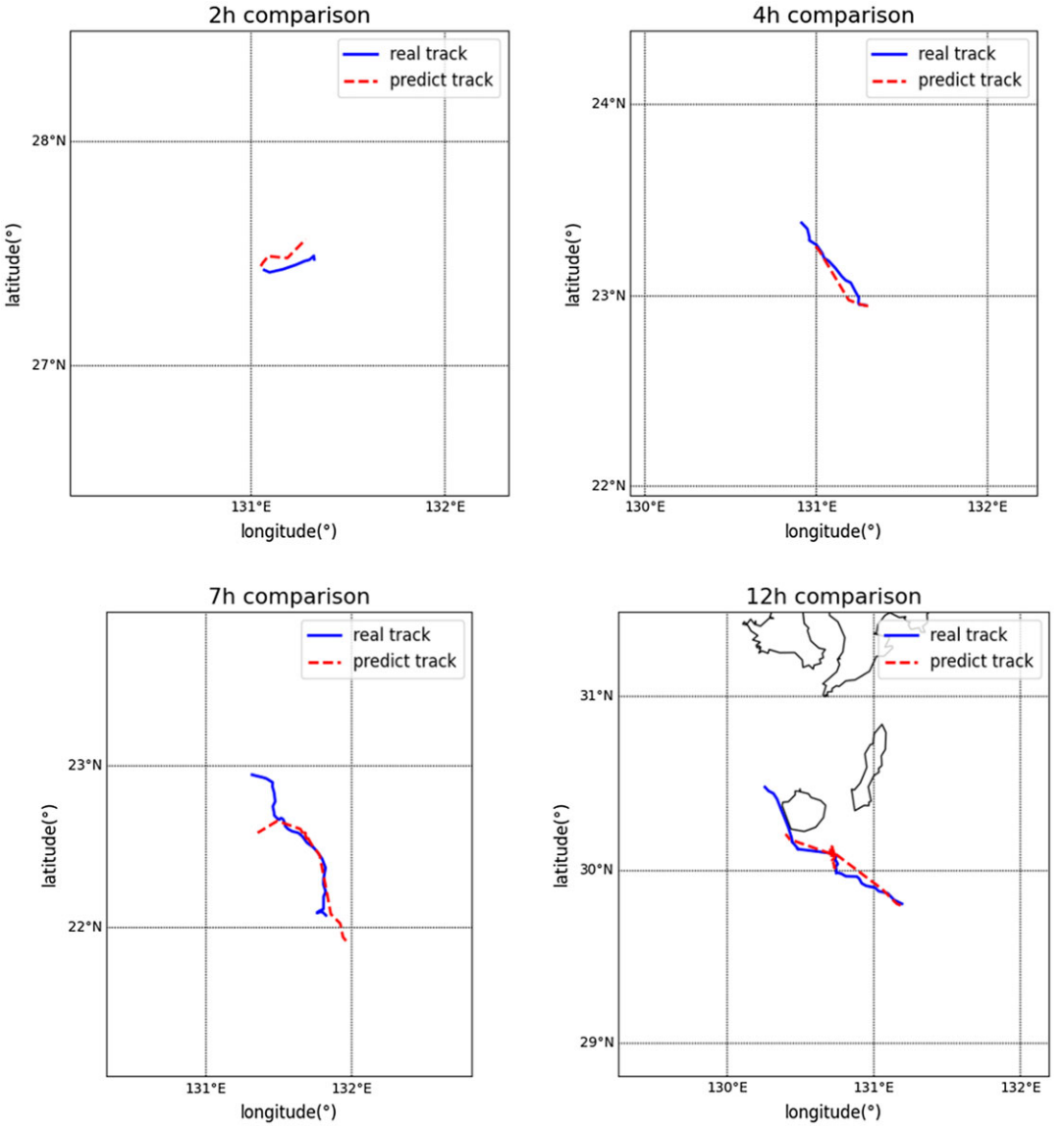

Figure 10. *Comparison of typhoon trajectory predictions and actual tracks across 2-h, 4-h, 7-h and 12-h intervals.*

Table 3. Comparison of the results with other methods

Methods	MAE_{lat}	MAE_{lon}	*MAE* (km)	*RMSE* (km)
Proposed method	0.063	0.067	10.78	15.20
LSTM (Gao et al., 2018)	–	–	63.367	–
DeepFR (Qin et al., 2022)	1.065	1.353	–	–
ConvLSTM + Reanalyse data (Lu et al., 2022)	0.303	0.393	54.69	71.27
Trj-DMFMG (Qin et al., 2022)	0.92748	1.13288	–	–
GAN + related data (Ruttgers et al., 2022)	–	–	68.7	–
Ensemble (Hao and Jin, 2022)	0.4712	0.8298	104.97	–
Multi+CNN-GRU (Lian et al., 2020)	–	–	102.32	–

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

models listed in the table are not uniform; for instance, the model of Lu et al. (2022) is designed for longer-term forecasts.

Maintaining consistency in prediction intervals, our method shows a reduction in MAE_{lat} and MAE_{lon} by an order of magnitude compared with those of Qin et al. (2022). Additionally, the *MAE* of our method is 58 km less than that reported by Ruttgers et al. (2022). Compared with Hao and Jin (2022) and Lian et al. (2020), the reduction in error is even more significant. These results indicate that even with limited data support, our method effectively leverages LLM and RAG technologies to achieve more precise predictions.

This demonstrates that our method can provide high-quality predictions within shorter forecast intervals. Not only does this corroborate the efficiency of our approach in handling complex prediction problems, but it also underscores its reliability in practical applications. These findings further highlight the potential of LLM and RAG technologies in enhancing the accuracy of typhoon trajectory predictions, offering valuable insights for future research in related fields.

The main advantages and innovations of the LLM + RAG prediction method proposed in this paper, compared with various other methods, are as follows.

1. The most notable advantage is the significant reduction in dependency on data volume. For example, the model of Qin et al. (2022) applied data covering 76 typhoons affecting or about to affect the Korean Peninsula from 1993 to 2017. To overcome the lack of satellite image data, reanalysis data were introduced, expanding the dataset to 757 typhoons.
2. The most distinctive feature is the use of a large language model as the basis for prediction. From a deep learning perspective, the emergence of large language models represents profound exploration and application in this field. Large language models, with their rich data reserves and generalisation capabilities, are suitable for various scenarios. This means that large language models are deep learning models that have been extensively trained and possess a vast knowledge reserve. For instance, previous deep learning training methods achieved a functional leap from 0 to 1, and while various methods and explorations can make predictions more accurate and enhance the model's generalisation ability, directly using large language models as the prediction foundation is like pursuing prediction accuracy from a height of 100.

3.5. Discussion and prospects

The experimental results indicate that employing LLM and RAG technologies for prediction can yield accurate results with minimal data and without the need for extensive modelling and parameter tuning. The LLM effectively learns the temporal coordinate data of typhoons, producing reliable predictions. However, despite its significant advantages, LLM may still encounter some unknown errors during the prediction process, such as hallucinations and minor information processing biases.

To enhance the accuracy and generalisability of predictions, future work should focus on fine-tuning the LLM. Fine-tuning typically involves making minor adjustments to the network architecture, hyperparameters or training process to optimise model performance. For the specific challenge of typhoon trajectory prediction, fine-tuning could improve the LLM's ability to capture the patterns of typhoon movements, thereby enhancing prediction accuracy. This may involve adjusting network weights, optimising learning rates or introducing new data augmentation strategies.

Though accurately predicting typhoon trajectories remains a complex challenge, we remain optimistic about the potential of LLM technology as it continues to mature and become more tailored to address specific problems in this domain.

https://doi.org/10.1017/S0373463325101148 Published online by Cambridge University Press

4. Conclusion

This paper introduces an innovative approach to typhoon path near-term prediction by integrating an LLM with RAG technology, marking the first application of this combination in the field. The method leverages the adaptability of LLMs and the retrieval capabilities of RAG to significantly enhance the accuracy and efficiency of predictions. The research uses FY-4 satellite imagery as the primary data source, which, after preprocessing and tracking with the Lucas–Kanade method, yields key dynamic change data of typhoon feature points. These data are meticulously organised into a dataset and embedded into the LLM as a vector database, providing essential input for the RAG process.

The RAG process plays a pivotal role in this study. It enhances the generative capabilities of the LLM by efficiently retrieving and using specific information from the vector database, leading to more precise predictions. Furthermore, an interactive question-and-answer format via the User Interface (UI) facilitates user interaction with the model to obtain immediate prediction results. During model deployment and optimisation, the parameters of the LLM were finely tuned and effective prompting words were introduced to further refine the prediction process.

The experimental results demonstrate the high accuracy of this method in short-term typhoon path prediction. This innovative approach not only offers a new perspective for typhoon path near-term prediction but also paves the way for new technological avenues in meteorological warning and disaster prevention, holding significant academic value and application prospects.

Funding statement. The project was supported by the National Key Research and Development Program of China (2022YFB4301400).

References

Chen, R., Zhang, W. and Wang, X. (2020). Machine Learning in Tropical Cyclone Forecast Modeling: A Review. *Atmosphere*, **11**, 676–701.

Chen, Y. and Duan, Z. (2018). A Statistical Dynamics Track Model of Tropical Cyclones for Assessing Typhoon Wind Hazard in the Coast of Southeast China. *Journal of Wind Engineering and Industrial Aerodynamics*, **172**, 325–340.

Chib, P.S.S. (2024). Pravendra, LG-Traj: LLM Guided Pedestrian Trajectory Prediction. *Computer Vision and Pattern Recognition*.

Fu, H., Gu, Z. and Wang, Y. (2022). Ship Pitch Prediction Based on Bi-ConvLSTM-CA Model. *Journal of Marine Science and Engineering*, **10**, 840–859.

Gao, S., Zhao, P., Pan, B., Li, Y., Zhou, M., Xu, J., Zhong, S. and Shi, Z. (2018). A Nowcasting Model for the Prediction of Typhoon Tracks Based on a Long Short Term Memory Neural Network. *Acta Oceanologica Sinica*, **37**, 8–12.

Hao, F. and Jin, J. (2022). A Broad Learning Ensemble System Using Bagging for Typhoon Trajectory Forecasting. In: *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 729–733.

Hon, K.-K. (2020). Tropical Cyclone Track Prediction Using a Large-Area Wrf Model at the Hong Kong Observatory, *Tropical Cyclone Research and Review*, **9**, 67–74.

Jabbar, A., Li, X. and Omar, B. (2021). A Survey on Generative Adversarial Networks: Variants, Applications, and Training. *ACM Computing Surveys*, **54**, 1–49.

Jiang, D., Shi, G., Li, N., Ma, L., Li, W. and Shi, J. (2023). TRFM-LS: Transformer-Based Deep Learning Method for Vessel Trajectory Prediction. *Journal of Marine Science and Engineering*, **11**, 880–901.

Jiang, Y., Yu, Y., Ma, Y., Zhuang, J., Ye, X., Yuan, Y. and Su, G.(2024). Typhoon Track Image Prediction Using a Comprehensive LSTM Attention Model in Deep Learning, In: *2024 4th International Conference on Neural Networks, Information and Communication (NNICE)*, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 537–541.

Jung, H., Baek, Y.-H., Moon, I.-J., Lee, J. and Sohn, E.-H. (2024). Tropical Cyclone Intensity Estimation Through Convolutional Neural Network Transfer Learning using Two Geostationary Satellite Datasets. *Frontiers in Earth Science*, **11**, 104798.

Lan, Z.X., Liu, L.S., Fan, B., Lv, Y.S., Ren, Y.L. and Cui, Z.Y. (2025). Traj-LLM: A New Exploration for Empowering Trajectory Prediction with Pre-trained Large Language Models. *IEEE Transactions on Intelligent Vehicles*, **10**, 794–807.

Li, T., Lai, M., Nie, S., Liu, H., Liang, Z. and Lv, W. (2024) Tropical Cyclone Trajectory Based on Satellite Remote Sensing Prediction and Time Attention Mechanism ConvLSTM Model. *Big Data Research*, **36**, 100439-100447.

Lian, J., Dong, P., Zhang, Y., Pan, J. and Liu, K. (2020). A Novel Data-Driven Tropical Cyclone Track Prediction Model Based on CNN and GRU With Multi-Dimensional Feature Selection. *IEEE Access*, **8**, 97114–97128.

Liang, Z., Yang, Y., Sun, C., He, W., Dai, Z., Su, G. and Yu, Y. (2023). Typhoon Track Prediction Based on Dual Attention LSTM Model. In: *2023 11th International Conference on Information Systems and Computing Technology (ISCTech)*, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA,519–525.

Lu, P., Sun, A., Xu, M., Wang, Z., Zheng, Z., Xie, Y. and Wang, W. (2022a). A Time Series Image Prediction Method Combining a Cnn and Lstm and its Application in Typhoon Track Prediction. *Mathematical Biosciences and Engineering*, **19**, 12260–12278.

Lu, P., Xu, M., Chen, M., Wang, Z., Zheng, Z. and Yin, Y. (2023). Multi-Step Prediction of Typhoon Tracks Combining Reanalysis Image Fusion Using Laplacian Pyramid and Discrete Wavelet Transform with ConvLSTM. *Axioms*, **12,** 107036.

Lu, P., Xu, M., Sun, A., Wang, Z. and Zheng, Z. (2022b). Typhoon Tracks Prediction with ConvLSTM Fused Reanalysis Data. *Electronics*, **11,** 3279.

Mahmoud, H. and Akkari, N. (2016). Shortest Path Calculation: A Comparative Study for Location-Based Recommender System. In: *2016 World Symposium on Computer Applications & Research (WSCAR)*, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 1–5.

Munir, K. and Sheraz Anjum, M. (2018). The Use of Ontologies for Effective Knowledge Modelling and Information Retrieval. *Applied Computing and Informatics*, **14**, 116–126.

Pang, Y., Zhao, X., Yan, H. and Liu, Y. (2021). Data-Driven Trajectory Prediction with Weather Uncertainties: A Bayesian Deep Learning Approach. *Transportation Research Part C: Emerging Technologies*, **130,** 103326.

Peng, M.G., Chen, X., Zhu, X., Chen, M., Yang, K., Wang, H., Wang, Y.X. (2024). LC-LLM explainable lan-change intention and trajectory predictions with large language models.

Qin, W., Tang, J. and Lao, S. (2022). DeepFR: A Trajectory Prediction Model Based on Deep Feature Representation. *Information Sciences*, **604**, 226–248.

Qin, W., Tang, J., Lu, C. and Lao, S. (2022). A Typhoon Trajectory Prediction Model Based on Multimodal and Multitask Learning. *Applied Soft Computing*, **122,** 108804.

Ren, J., Xu, N. and Cui, Y. (2022). Typhoon Track Prediction Based on Deep Learning. *Applied Sciences*, **12**, 80–107.

Ruttgers, M., Jeon, S., Lee, S. and You, D. (2022). Prediction of Typhoon Track and Intensity Using a Generative Adversarial Network With Observational and Meteorological Data. *IEEE Access*, **10**, 48434–48446.

Song, T., Li, Y., Meng, F., Xie, P. and Xu, D. (2022). A Novel Deep Learning Model by BiGRU with Attention Mechanism for Tropical Cyclone Track Prediction in the Northwest Pacific, *Journal of Applied Meteorology and Climatology*, **61**, 3–12.

Suo, Y., Chen, W., Claramunt, C. and Yang, S. (2020). A Ship Trajectory Prediction Framework Based on a Recurrent Neural Network. *Sensors (Basel)*, **20,** 5133.

Yang, Z.W.X. (2018). Moving Target Detection and Tracking Based on Pyramid Lucas-Kanade Optical Flow. In: *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC),* Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA.

Zhan, R., Lu, J., Jiang, Y., Li, T., Zhong, G. and Lv, W. (2023). A Multimodal Deep Learning Approach for Typhoon Track Forecast by Fusing CNN and Transformer Structures. In: *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 391–394.

Cite this article: Peng, B., Jiang, Y., and Yao, L. (2025). Nowcasting of typhoon tracks based on LLM and RAG. *The Journal of Navigation* 1–16. https://doi.org/10.1017/S0373463325101148