Effectiveness of Large Language Models (LLMs) for Weather Forecasting in Arlington, VA

Aabha Belsare, Sara Pattanaik, Simran Ajwani, and Tanya Singh

ABSTRACT

We evaluated GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4 to predict future weather patterns based on past data. With controlling information fed to the Large Language Models (LLMs) and limiting it to a single dataset, we could evaluate the pattern recognition capabilities of each model. After careful prompt engineering, each model output the predicted high temperatures, low temperatures, and precipitation for Arlington, VA from June 1, 2025 to August 31, 2025. This predicted data was then compared to the actual high temperatures, low temperatures, and precipitation recorded by the Ronald Reagan Washington National Airport weather station. From the analysis, the predicted weather from June 1, 2025 to August 31, 2025 portrayed that GPT-5 was the best at predicting the low temperatures, while Gemini 2.5 Pro and Claude Sonnet 4 was the best at predicting the precipitation.

1 INTRODUCTION

The use of Artificial Intelligence, specifically LLMs, have recently been on the rise since the availability of large datasets and computer resources, as well as adaptation of deep learning techniques have been more prominent [4]. LLMs are models that use a neural-network to learn from large datasets in two main parts. The first being that they use a self-supervised learning approach where the data fed into these models do not need any external manipulation. Then, they then go through a fine-tuning stage where they use small datasets in order to use the knowledge they learned to complete tasks the user asks [5]. Based on multiple rounds of learning and training from this approach, LLMs have been launched for the public to be able to generate text, answer prompts, conduct analyses, translate text, solve math problems, and more. The use of LLMs have significantly increased, leading us to investigate the effectiveness of LLMs for future weather forecasting in Arlington, Virginia from June 1, 2025 to August 31, 2025.

Alongside the rise of LLMs, weather forecasting has also been an area of development since the year 1450, where Cardinal Nicholas de Cusa invented the first hygrometer. Weather forecasting is extremely important because the advancements in the ability to predict and model future weather conditions have allowed people to be able to plan activities for the future [3]. Currently, weather forecasting involves computers that run “models” in order to generate predictions regarding the temperature, air quality, and hydrological applications [3]. The most prominent model being used is the National Weather Prediction (NWP), which provides short-term forecasting predictions based on physics, information on initial conditions, and Earth system observations [1]. Though, it is important to

investigate if LLMs, especially GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4, could be used for weather predictions as well.

2 LITERATURE REVIEW

In recent years, LLMs have become a popular topic of research for their potential use in fields beyond regular tasks, including weather and climate research. As unpredictable weather patterns become more common, researchers have begun exploring if these models can help classify weather impacts, analyze disaster reports, and facilitate forecasting efforts. LLMs show that they are able to organize large amounts of information as well as derive patterns, but their accuracy and reliability is uncertain. So, research on models such as GPT-4o, Meta's Llama series, and other forecasting frameworks is helping to clarify the strengths and weaknesses of LLMs over time. Consequently, there have been a few studies in the past, mainly investigating the relationship between LLMs and weather predictions for models like GPT-4o, Meta AI (Llama-2, Llama-3, and Llama-3.1), and WindPromptCast.

2.1 Use of LLMs for Severe Weather Conditions

First, past research has been done regarding the use of LLMs to assess severe weather conditions and wind patterns [7]. A major use of LLMs has been for impact classification. LLMs have read weather reports and then can classify them based on their impacts on things like agriculture, infrastructure, and public health [7]. LLMs are also effective for ranking-based question answering. For example, in a study, researchers imputed questions based on the specific impact categories and to match this to the correct article from a pool of 100 candidates. Overall, from this GPT-4o scored the highest with a Hit@1 score of 91.94% meaning that 92/100 times GPT-4o was most accurate at extracting the correct article of disruptive weather events. This highlights how LLMs have been used successfully to organize the historical and modern records of weather events and classifying them successfully. Thus this translates into how LLMs can see past trends/impacts, organize them for us, and alert us on what to be cautious of [7].

Beyond impact classification, research has expanded to investigate whether LLMs like GPT-4o can be applied to climate forecasting. This includes predicting rainfall within short-term (~15 days) and long-term (~12 month) time periods [6]. In these studies, GPT-4o was evaluated with different conditions, including making predictions on its own, using expert models, using regional climate variables, and using teleconnection indices such as ENSO or PDO, which quantify the strength of geographically widespread, large-scale atmospheric and oceanic patterns [6]. When making its own predictions, GPT-4o came up with conservative forecasts, meaning that it relied on historical averages instead of considering extreme variabilities in rainfall [6]. However, when provided with expert models or teleconnection indices, GPT-4o improved long-term forecasts by a small degree, but it still underperformed compared to expert models and underestimated extreme events [6]. Overall, the studies that applied GPT-4o to climate forecasting concluded that the model is valuable for

synthesizing climate information, however traditional data-driven models are more accurate for severe weather conditions [6].

Additionally, another use of LLMs has been for disaster related analysis and classification. Meta's Llama models have been applied extensively to analyze and classify flood related data from the National Weather Service reports, which offers an innovative approach to disaster responses. By focusing on these official, structured reports instead of false, biased, unstructured data from social media that other models use, Llama models help to efficiently identify critical flood events [8]. This capability allows for a faster, and more accurate situational awareness, which is important in certain areas and situations, such as supporting emergency response efforts in frequently affected areas. The structured nature of the National Weather Service reports combined with Llama's advanced contextual understanding allows it to uncover complex patterns in flood events, contributing to more effective disaster management strategies [8].

Going beyond direct classification, Llama's outputs also have the potential to enhance the assessment of flood risk and forecasting when integrated with other environmental and hydrological data. For instance, insights gained from the flood report texts can be combined with physical features such as rainfall, elevation, and land cover, improving the prediction of flood prone areas [8]. This multi modal approach allows for more support in flood susceptibility mapping and resources allocation planning. Additionally, the adaptability of Llama to handle imbalanced and domain specific disaster data helps increase its value when it comes to more broader applications in environmental monitoring and emergency preparedness. As further research advances, Llama's interactions with time series and simulation based models could transform the way flood forecasting and disaster response are conducted in real world settings [8].

More recently, researchers developed WindPromptCast, an LLM-based framework tailored for multi-location, multi-step wind power forecasting. This model converts meteorological data such as wind speed and power into natural language–like inputs through a hard prompt generator, while a soft prompt adapter with gated attention fine-tunes a small set of parameters without altering the main LLM. A fusion mechanism then integrates these prompts, allowing the system to capture long-term patterns, short-term fluctuations, and sudden outliers that are common in severe weather conditions WindPromptCast has also demonstrated zero-shot adaptability, meaning it maintains accuracy when applied to unseen wind farms with limited data, which makes it valuable in real-world forecasting scenarios where weather patterns vary unpredictably [2]. As seen, GPT-4o, Meta AI's Llama, and WindPromptCast models have been used for the classification of severe weather conditions in the past.

2.2 Accuracy of LLMs

Additionally, the accuracy of previous LLMs, such as GPT-4o, Meta AI (Llama-2, Llama-3, and

Llama-3.1), and WindPromptCast have been investigated when it comes to weather conditions as well as forecasting predictions. From one article, GPT-4o has been shown to be the strongest model when utilizing multi-label classification by testing each model for organizing information into categories of infrastructure, political, financial, agriculture, etc. From this, the models were given newspaper articles and had to assign them to those categories. The average score for GPT-4o was F1=66.9 which was the highest out of all the models with specific strengths in infrastructure with 81.25 and human health of 72.93. Thus, since LLMs have been shown to be very accurate at categorizing impacts from disasters, we can utilize models like GPT-4o to see these previous patterns from old newspaper sets to help predict any oncoming similar occurrences [7]. Additionally, as previously mentioned, for resource-based question answering, GPT-4o was most successful again for finding correct articles from a pool of 100 candidates, having a Hit@1 score of 91.94%. This is extremely helpful as LLMs have shown to take questions and then accurately find the correct corresponding in this case historical article of weather impacts [7]. From this, we are able to use GPT-4o to look at previous weather occurrences in specific areas (Arlington) to predict future trends and prevent any severe impacts.

When evaluating rainfall prediction at both short-term and long-term periods, researchers found that GPT-4o could generate reasonable average forecasts; however the model often reflected historical numbers and failed to consider extreme scenarios. When given expert models or teleconnection indices from ENSO and PDO, GPT-4o showed a slight improvement but continued to underperform compared to more specific models. This highlights that GPT-4o is slightly inaccurate in forecasting accuracy when considering extreme events [6].

Another research paper investigated the accuracy of Meta AI (Llama-2, Llama-3, and Llama-3.1) models. As previously mentioned, Meta's Llama models have shown to have strong performance when it comes to flood disaster response, especially when it comes to structured weather service reports rather than social media data. To make this happen, official reports from the National Weather Service provide consistent, well organized, and verified information which makes it a vital input for Llama models to achieve high accuracy in classification tasks [8]. For instance, the Llama 3 8B model achieved excellent results on flood event classification, with a precision of 0.989, a recall of 0.997, an F1 score of 0.992, and an accuracy of 0.97, which outperforms other transformer models. This ability to understand and analyze structured text is an important advantage for disaster response applications needing reliable and timely insights [8].

Moreover, Llama models extend their capabilities beyond just floods to multi label classification involving severe weather events, such as thunderstorms and cyclonic activity. Llama 3 models showed rapid convergence, reaching accuracy levels above 0.8 for thunderstorms by the second training epoch, whereas earlier model versions required more epochs to achieve similar performance. This efficiency comes partly from the architectural innovations such as the Grouped Query Attention (GQA), which optimizes the multi head attention mechanism allowing the models to handle large context windows in complex weather reports. These advancements help position Llama as the most effective tool for accurate, scalable disaster test classification in operational environments [8]. Similar to

GPT-4o, due to these high accuracy levels, Llama models could also be used to look at previous weather occurrences in specific areas such as Arlington to predict future trends and prevent any severe impacts.

In terms of accuracy, WindPromptCast outperformed traditional models like CNNs, RNNs, GRUs, Transformers, and BERT-based methods when evaluated across 14 wind farms in China spanning deserts, plains, mountains, and other terrains. It achieved the lowest mean absolute error (MAE = 11.67 MW) and root mean squared error (RMSE = 18.83 MW) among all tested models [2]. Notably, the model was also very good at handling sudden spikes or drops in wind power output, which typically trip up other models. Seasonal analysis showed the model performed best in autumn (MAE = 8.13 MW) and worst in winter (MAE = 13.3 MW), although it still surpassed baseline methods in every season [2]. These results highlight that WindPromptCast is one of the most accurate LLM-based forecasting approaches currently available, particularly for severe weather–driven variability.

In contrast to the two mentioned, there were other models that were not as effective. For example, some open source models for one-shot performance like DeepSeek-V3-671B was the strongest having a closer score to GPT-4o having a 62.90 F1, then Mistral-24B-IT with a score of 57.41 ,MIXTRAL-8X7B-IT scoring the worst of 40.53, etc. However, other GPT models like GPT-4 had a F1 score of 54.95 and GPT3.5-Turbo with 54.08. This emphasizes how usually larger LLMs will be more successful rather than smaller ones [7].

2.3 Limitations of LLMs

Lastly, there are a few limitations of LLMs that must be addressed. Since LLMs are a work-in-progress, with constant upgrades and changes, the accuracy of LLMs are not always as high as expected. For GPT-4o being the strongest model, there still wasn't a high F1 rating score. For zero shot performance this score was 65.93 and for one shot is 66.93 which is considerably low given how big the model is. This shows how GPT-4o still struggled with classifying certain impacts like politics with F1 scores of 58.46 and ecological scores of 46.81. This is because language wasn't very clear in these newspaper articles so these models were unable to understand how to successfully categorize them. Even when the models were provided with one example before having to classify categories (one shot-learning) GPT-4o still wasn't very successful. The F1 score only increased slightly from 65.93 to 66.93. This highlights that even when the models were given assistance regarding impact classification, it still did not help much, questioning the real accuracy of using LLMs to predict weather [7].

While GPT-4o had some strengths regarding impact classification, its performance in rainfall forecasting was much weaker. The model often relied on historical patterns, leading it to miss extreme rainfall events. Even when provided with regional climate data or teleconnection indices, GPT-4o’s predictions were still less accurate than expert models. This suggests that overall, GPT-4o struggles to capture variability that skews from historical patterns within climate systems [6].

On the other hand, a few limitations with Meta AI (Llama-2, Llama-3, and Llama-3.1) also need to be addressed. A key challenge is the imbalance in available training data, particularly for less frequent event types such as cyclonic events. Despite Llama models achieving high accuracy and F1 scores in flood and thunderstorm categories, the performance in the cyclonic category remains substantially lower, with F1 scores not exceeding 0.32 after fine tuning for ten epochs [8]. This shows us exactly the difficulty of training models effectively on underrepresented categories, which leads to bias toward majority classes and resulting in reduced performance for rare but critical disaster types. Although techniques like the Multi Label Synthetic Oversampling (MLSOL) help with mitigating this issue by generating synthetic data to balance categories, the lack of data continues to be a significant limiting factor for Llama full effectiveness in all weather classifications [8].

Additionally, computational demands represent another limitation of Llama models. While Llama 3 exhibits superior efficiency, achieving high performance with fewer training epochs compared to Llama 2, the fine tuning process still requires considerable computational resources, especially for larger model variants with billions of parameters. For instance, fine tuning Llama 3.1 8B requires approximately 16 GB of memory even with parameter efficient techniques such as Low Rank Adaptation (LoRA), which shows this trade off between model size, accuracy, and resource consumption [8]. Furthermore, few learning methods, although promising for rapid deployment, currently show unstable and lower performance when it comes to more complex categories which shows that Llama and other LLMs still require significant data and training to achieve reliable results in specialized tasks. These limitations suggest that while Llama models are powerful tools for disaster response and weather event classification, challenges with data imbalance and computational cost also must be addressed as the technology evolves [8]. From these limitations, it can be seen that LLMs demand a lot of resources and data in order to achieve the highest accuracy, which shows that they are not extremely efficient when it comes to weather forecasting and predictions.

Similarly, WindPromptCast comes with important limitations despite its strong performance. Training is computationally expensive, requiring about 143,000 seconds compared to only 3,000 seconds for simpler models [2]. Its accuracy also varies seasonally, as mentioned earlier, with the best results in autumn and weaker forecasts in winter (MAE rising to 13.3 MW). Additionally, unlike GPT-4o or Llama, which process text directly, WindPromptCast relies on converting numerical data into text-like prompts. This translation step may introduce complexity or noise, especially when dealing with out-of-distribution data [2]. Finally, as with many LLM-based systems, WindPromptCast predictions lack interpretability, making it harder for meteorologists to fully trust and explain its outputs in real contexts [2].

Based on our findings, we determined that LLMs have both strengths and weaknesses when applied to weather-related disasters. Among the specific models, GPT-4o has been the most successful compared to Meta Llama models in areas such as classification accuracy, while newer frameworks like WindPromptCast are beginning to show promise in forecasting applications. Overall, all of the AI

models we reviewed were able to perform well in tasks such as disaster-related analysis and impact classification, revealing the potential of LLMs to contribute to predicting weather events. However, despite their size and success, these models often did not achieve the level of accuracy that might be expected given their extensive resources. This raises important questions not only about their reliability but also about their sustainability, since training and running large-scale LLMs consumes significant energy and contributes to carbon emissions. Therefore, as we move forward with programming our chosen LLMs, it will be important to remain mindful of both the technical limitations of these systems and their environmental impact, ensuring that future applications balance predictive power with sustainable energy practices.

3 METHODOLOGY

Since previous research has shown some accuracy in weather forecasting from LLMs, it was necessary to investigate how each model interprets past weather patterns and how accurately they can use those patterns to generate future predictions. In order to do this, an extensive methodology was created to interpret the generated data. This methodology is organized into two major components. The first is data collection, which describes our procedure to gather and compile historical weather data for Arlington, VA into a standardized dataset for input. The second is data analysis which describes our procedure to visualize results and apply statistical tests to compare predicted and actual weather predictions. Through this organization, our study aims to quantify differences in performance among the 3 LLMs to evaluate the potential of using them in real world predictive modeling scenarios.

3.1 Data Collection

Weather data on the high temperatures, low temperatures, and precipitation was collected initially from the source ‘The Weather Channel’ and ‘Weather Underground’ in the timeframe of June 1, 2024 to May 31, 2025, which would be fed into the LLMs as a reference for past weather for Arlington, VA. This location was chosen due to the accessibility of weather data from the weather station at the Ronald Reagan Airport, which is located in Arlington, VA.

This data was then downloaded as a .pdf file and uploaded into LLMs of GPT-5 from Microsoft Copilot, Gemini 2.5 Pro from Gemini by Google, and Claude Sonnet 4 from Claude AI with the settings of having web searches turned off. Then, the following prompt was added with the uploaded data:

“Solely using the data linked in the .pdf file, which portrays the High (in degrees Fahrenheit), Low (in degrees Fahrenheit), and Precipitation (in inches) for the weather in Arlington, Virginia from June 1, 2024 to May 31, 2035, predict the High (in degrees Fahrenheit), Low (in degrees Fahrenheit), and Precipitation (in inches) for each day in Arlington, Virginia from June 1, 2025 to August 31, 2025. Do not just map the data from 2024 to 2025. Use the data given to provide a prediction for the weather in 2025. Only utilize the .pdf file provided. Do not fetch any resources or data from outside the .pdf file.”

After data was gathered and condensed into 92 rows from June 1, 2025 to August 31, 2025, it was then separated according to each LLM model used into a finalized google spreadsheet. This was done by formatting each LLM result into tables to compare the predicted high temperature (in degrees Fahrenheit), low temperature (in degrees Fahrenheit), and precipitation (in inches) to the actual weather within the time frame. Several visualizations and statistical tests were also performed to evaluate whether the LLMs' predictions aligned with the observed weather data. Then, a normality check was performed by generating a histogram of the high temperatures, low temperatures, and precipitation per LLM. From the histogram, it was determined that the temperatures were normally distributed, but the precipitation was skewed right. Thus, Table 1 represents summary statistics of Mean and Standard Deviation found in accordance with each LLM and the actual temperature, and Table 2 represents the summary statistics of Median and Interquartile Range (IQR) found in accordance with each LLM and the actual precipitation.

Table 1: Mean and Standard Deviation of the High and Low Temperatures in Arlington, VA

		Summary Statistics for Weather Forecast in Arlington, VA	
		Mean	SD
High Temperatures (°F)	Actual	86.01	6.31
	GPT-5	89.91	4.11
	Gemini 2.5 Pro	89.9	5.01
	Claude Sonnet 4	90	5.84
Low Temperatures (°F)	Actual	70.57	6.06
	GPT-5	70.77	3.81
	Gemini 2.5 Pro	72.99	4.21
	Claude Sonnet 4	73.48	4.96

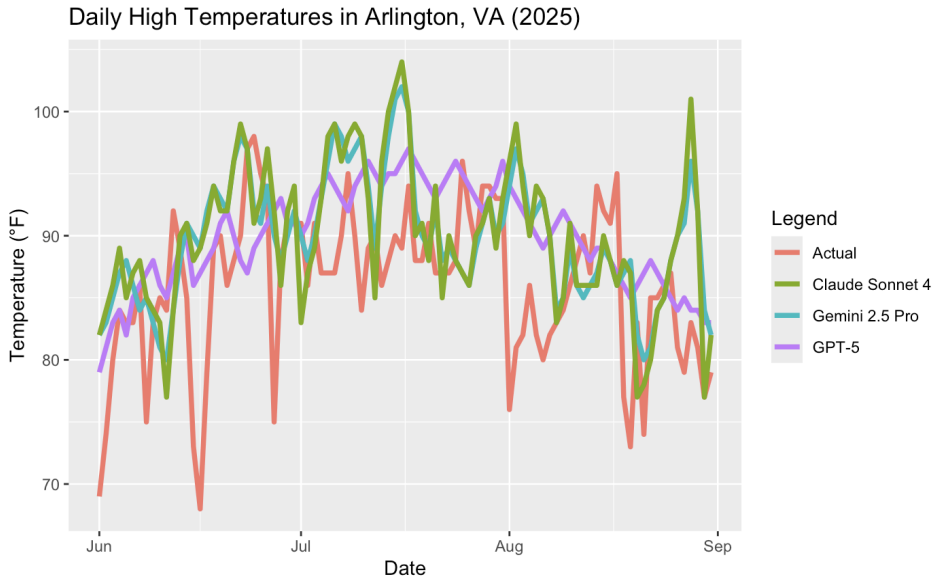
Table 2: Median and IQR of the Precipitation in Arlington, VA

		Summary Statistics for Weather Forecast in Arlington, VA	
		Median	IQR
Precipitation (in)	Actual	0	0.023
	GPT-5	0.05	0.1
	Gemini 2.5 Pro	0	0.05
	Claude Sonnet 4	0	0.088

3.2 Data Analysis

In order to test the accuracy and effectiveness of the various LLMs in weather forecasting, it is important to check if there is any similarity between the predicted high and low temperatures, as well as the precipitation and the actual weather data recorded from the weather station in Ronald Reagan Airport in Arlington, VA from June 1, 2025 to August 31, 2025. As seen in Table 1, the mean high temperatures across all models were relatively similar to the actual mean high temperature, which led to visually representing the high temperatures across the 92 days in order to check for any overlapping patterns. While visual trends provided initial insights into predictive similarities, further statistical validation was necessary to confirm whether these patterns were significant. Overlapping patterns between the actual and predicted weather from either one of the LLMs could portray a potential accurate prediction from the LLM when it comes to the specific time range for the weather. Thus, the high temperatures across all the LLMs were graphed with the actual high temperature from June 1, 2025 to August 31, 2025, as seen below in Figure 1.

Figure 1: High Temperatures in Arlington, VA from June 1, 2025 to August 31, 2025



Similarly, based on Table 1, the mean low temperatures across all the models were relatively similar to the actual mean low temperatures. This also led to visually representing the low temperatures across the 92 days in order to check for any overlapping patterns. Additionally, from Table 2, the median precipitation predictions across all the models were similar, if not the same, as the actual median precipitation in Arlington, VA, which also facilitated the observation of overlapping patterns. Therefore, the low temperature predictions from the LLMs in comparison to the actual low temperatures from June 1, 2025 to August 31, 2025 was graphed and represented in Figure 2 , and the

precipitation predictions from the LLMs and actual precipitation from June 1, 2025 to August 31, 2025 was graphed and represented in Figure 3 below.

Figure 2: Low Temperatures in Arlington, VA from June 1, 2025 to August 31, 2025

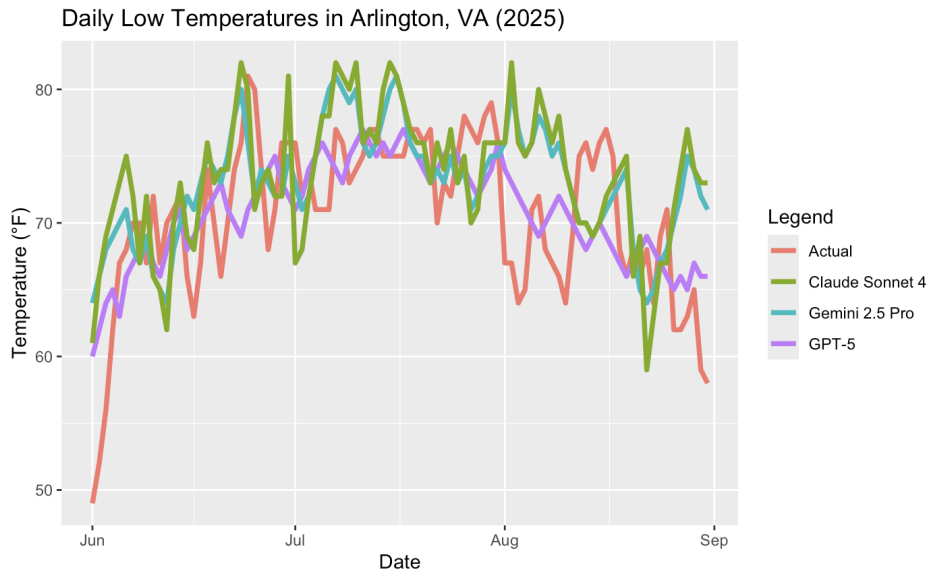
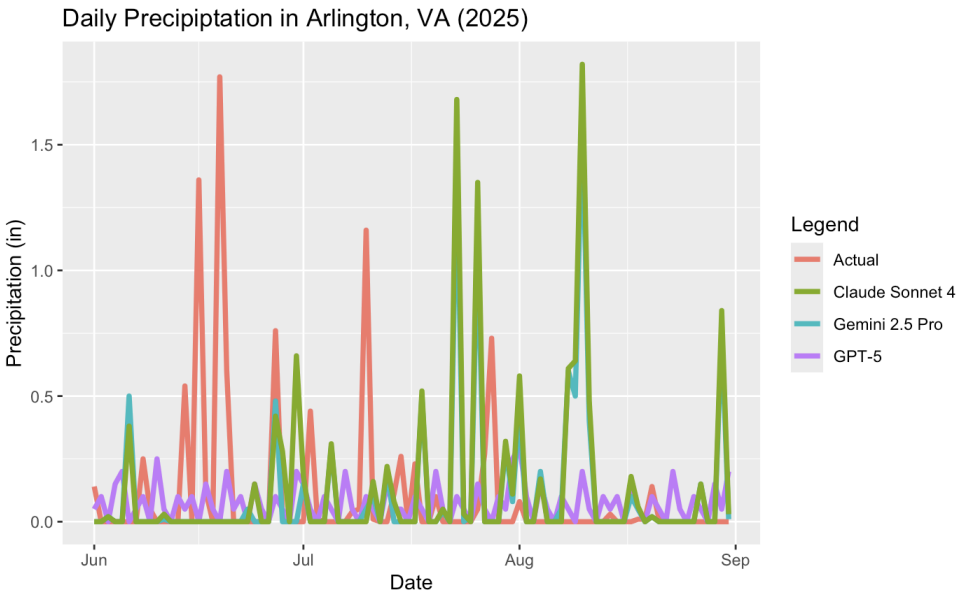


Figure 3: Precipitation in Arlington, VA from June 1, 2025 to August 31, 2025



Based on Figures 1 and 2 above, there were a few trends that were noticed. Firstly, the predictions by GPT-5 had the lowest standard deviation, which aligns with the data portrayed in Table 1. Additionally, predictions by Gemini 2.5 Pro and Claude Sonnet 4 regarding the high temperatures, low temperatures, and precipitation, portrayed in Figures 1, 2, and 3 were relatively similar. Though, when comparing the predictions from the LLMs with the actual weather data, there were a few days where

the weather data overlapped/intersected. The similarities across the predictions from the LLMs and the actual weather from June 1, 2025 to August 31, 2025 shows that there might be some accuracy in the LLM prediction when comparing each LLM to the actual weather data. Thus, six paired T-tests, and three paired Wilcoxon tests were conducted in R to compare the data, mainly because at one time, two groups were compared. Paired tests were selected because each model's prediction corresponded directly to the same actual daily measurement. All statistical tests were conducted at a significance level of $\alpha = 0.05$.

The six paired T-tests were between the following groups, which were normally distributed, in order to see if the LLM prediction for the temperatures were similar to the actual temperatures from June 1, 2025 to August 31, 2025: Actual High Temperatures vs. GPT-5 High Temperatures, Actual High Temperatures vs. Gemini 2.5 Pro High Temperatures, Actual High Temperatures vs. Claude Sonnet 4 High Temperatures, Actual Low Temperatures vs. GPT-5 Low Temperatures, Actual Low Temperatures vs. Gemini 2.5 Pro Low Temperatures, and Actual Low Temperatures vs. Claude Sonnet 4 Low Temperatures. The three paired Wilcoxon tests were between the rest of the groups, which were not normally distributed, in order to see if the LLM prediction for the precipitation were similar to the actual precipitation as well: Actual Precipitation vs. GPT-5 Precipitation, Actual Precipitation vs. Gemini 2.5 Pro Precipitation, and Actual Precipitation vs. Claude Sonnet 4 Precipitation.

Based on the results generated from these statistical tests, it would be determined if there is a statistically significant difference between the predictions and the actual weather, which could determine the accuracy of each of the three LLMs.

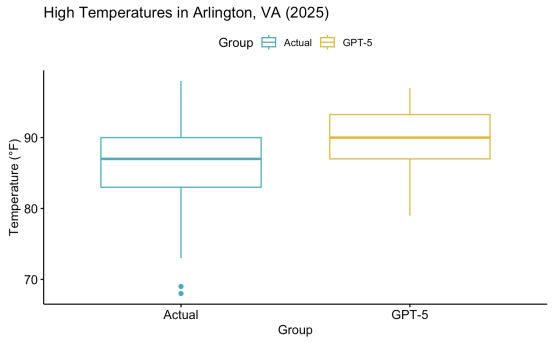
Overall, this methodology established a controlled and replicable experiment for evaluating how effectively LLMs can forecast weather based on historical data. By standardizing the dataset, keeping inputs consistent, and applying statistical analyses, the study provides a foundation for comparing model performance. These methods allow us to assess whether GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4 demonstrate meaningful and relevant predictive capabilities in real-world scenarios.

4 RESULTS

This analysis focuses on identifying whether the predicted values differed from the actual values, both visually through plotted distributions and statistically though the appropriate tests. By examining the temperature and precipitation results in a paired manner to the actual weather data, an accurate analysis of the difference between the two could be calculated. To do this, daily high temperatures, low temperatures, and precipitation totals from June 1 to August 31, 2025, were compared to predictions generated by GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4. Paired t-tests were used to assess differences in temperature predictions, while Wilcoxon signed-rank tests were used to evaluate precipitation predictions. The following results present the deduced statistical findings accompanied by visual comparisons.

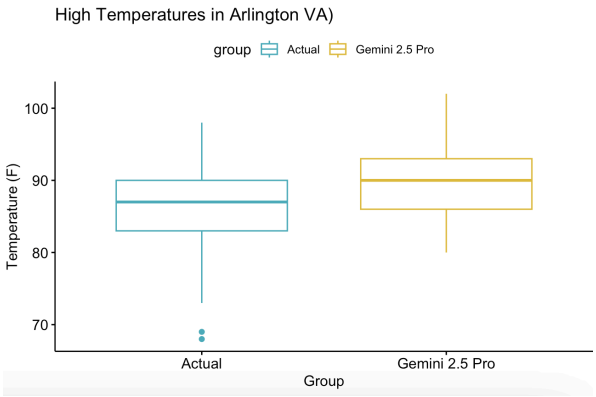
Initially, three paired t-tests were performed in order to emphasize the accuracy for LLMs to predict high temperatures from June 1, 2025 to August 31, 2025 in Arlington Virginia. In Table 3, the p-value, comparing the means of high temperature for actual to the GPT-5 model, was 0.000000001007, suggesting that there was a statistically significant difference between both. This reinforces how the difference between GPT-5 and the actual high temperature were not random, with consistently different results throughout the sample, thus producing such a significant difference. Consequently as seen in Figure 4, there is a clear distributional difference between GPT-5 vs the actual high temperatures, with GPT-5 having a greater deviation. Thus, showing a higher mean prediction and lower variation. From this, a medium effect size was found with -0.73, highlighting how GPT-5 predicted lower temperatures than what was recorded, supporting the large gap between both GPT-5 and actual high temperatures in Arlington, VA. This supports how GPT-5 is unlikely to be used in real-world events as it ultimately failed at producing accurate weather predictions.

Figure 4: Actual vs. GPT-5 High Temperatures in Arlington, VA



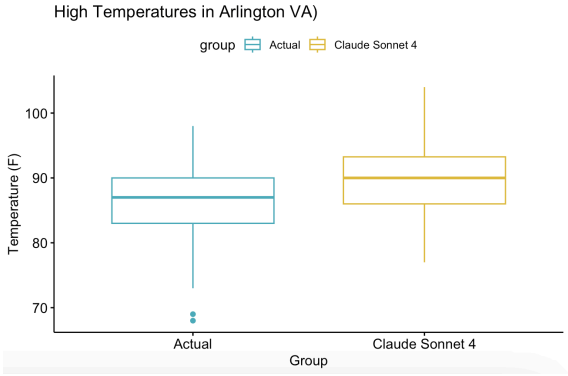
Likewise, another paired p-test was performed to compare the means of the actual high temperatures to the predictions of Gemini 2.5 Pro's high temperatures. A p-value of 0.000000272 was found, further suggesting a statistically significant difference between actual data to Gemini 2.5 pro, thus not random. This proves that Gemini 2.5 Pro was unsuccessful at accurately predicting weather trends for high temperatures by having such a significant difference in means. Moreover, as seen in Figure 5, there is a noticeable distributional difference between the Actual High temperatures to Gemini 2.5 Pro, with higher mean prediction and low variation. From this, a medium effect size was shown, with -0.68, suggesting that Gemini 2.5 Pro is unlikely to be used in real-world application, providing that it is inaccurate at producing weather forecasts.

Figure 5: Actual vs. Gemini 2.5 Pro High Temperatures in Arlington, VA



Moreover, a last paired t-test was performed in order to compare the means of the actual high temperatures to Claude Sonnet 4's High temperatures prediction. Hence, a p-value of 0.0000009365 was found, indicating a statistically significant difference between Actual and Claude Sonnet 4's results, which was also not random. This demonstrates that Claude Sonnet 4 was similarly unsuccessful at accurately predicting a weather forecast for high temperatures this past year; however, had the largest p-value of all three models. In addition, as seen in Figure 6, the distributional difference is evident, as Claude Sonnet 4 was not similar to the actual high temperatures having a higher mean prediction and lower variation with the data. Therefore, a medium effect size was found again, with -0.65, establishing that Claude Sonnet 4 is overall inadequate at predicting accurate weather trends.

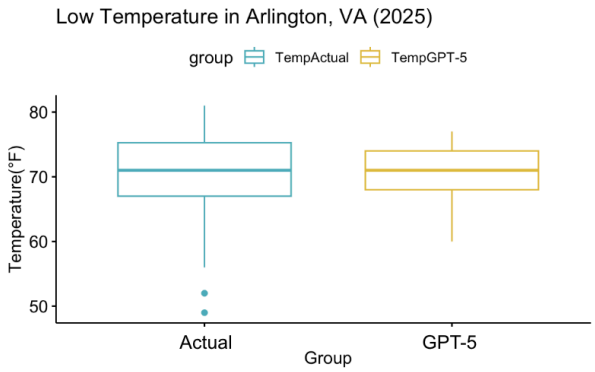
Figure 6: Actual vs. Claude Sonnet 4 High Temperatures in Arlington, VA



Then, the results for the three paired T-tests that pertained to the accuracy for LLMs to predict the low temperatures from June 1, 2025 to August 31, 2025 in Arlington, VA were analyzed. As seen in Table 3 below, the p-value for the first group, which compared the means of the actual low temperature compared to the low temperature predicted by the GPT-5 model, was 0.66, suggesting that there was no statistically significant difference between the two groups, and any difference that occurred in the prediction of the weather from the LLMs may have been due to random chance. Even though this doesn't suggest that the groups were extremely identical, there is a possibility that future GPT models could generate accurate predictions if fed data containing weather for more than the past year.

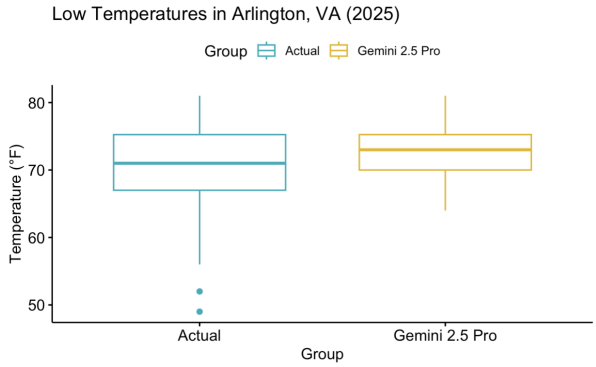
Furthermore, as seen in Figure 7, this derivation is supported by the distribution of the low temperatures predicted from GPT-5, as they were similar to the actual low temperatures, with an extremely accurate mean prediction and variation in the data. However, the effect size for this comparison was negligible, amounting to -0.041, which suggests that these results would need to be further studied in order to be eventually applied in the real world.

Figure 7: Actual vs. GPT-5 Low Temperatures in Arlington, VA



For the second tests, which compared the means of the actual low temperature compared to the low temperature predicted by the Gemini 2.5 Pro model, the p-value was 0.00012, suggesting that there was a statistically significant difference between the two groups, which likely did not occur at random. This portrays that Gemini 2.5 Pro was unable to accurately generate a weather forecast for the low temperatures based on the weather data for the past year. Furthermore, as seen in Figure 8, this derivation is supported by the distribution of the low temperatures predicted from Gemini 2.5 Pro, as they were not similar to the actual low temperatures, with a higher mean prediction and lower variation in the data. However, the effect size for this comparison was small, amounting to -0.465, which suggests that these results would also need to be further studied in order to be applied to the real world, which means that Gemini models should be further tested with larger amounts of data to confirm that they could not be used in weather forecasting for low temperatures.

Figure 8: Actual vs. Gemini 2.5 Pro Low Temperatures in Arlington, VA



As seen in Table 3, the last test, which compared the means of the actual low temperature compared to the low temperature predicted by the Claude Sonnet 4 mode, had a p-value of 0.000018, suggesting that there was a statistically significant difference between the two groups, which likely did not occur at random as well. This portrays that Claude Sonnet 4 was unable to accurately generate a weather forecast for the low temperatures based on the weather data for the past year, and it performed worse compared to the previous two models. Furthermore, as seen in Figure 9, this derivation is supported by the distribution of the low temperatures predicted from Claude Sonnet 4, as they were not similar to the actual low temperatures, with a higher mean prediction and lower variation in the data. However, Claude Sonnet 4 was the only model that generated outliers in the weather predictions, mimicking natural weather patterns. Regardless, the effect size for this comparison was medium, amounting to -0.526, suggesting that these results do have significance in the real world, which means that Claude models are likely to not be used in weather forecasting for low temperatures.

Figure 9: Actual vs. Claude Sonnet 4 Low Temperatures in Arlington, VA

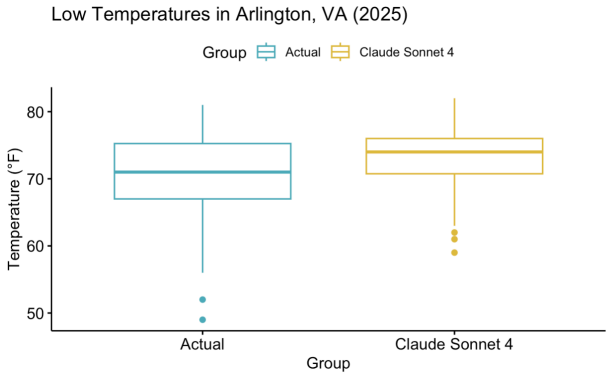
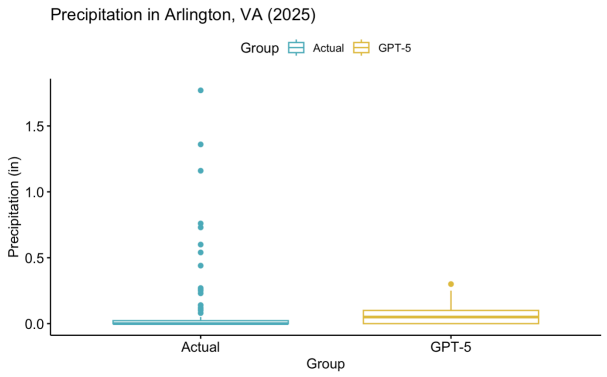


Table 3: Paired T-test Results

	Paired T-test Results				
	Groups	p-value	Significant Difference (α = 0.05)	Effect Size	Interpretation of Effect Size
High Temperatures (°F)	Actual	1.01E-09	Yes	-7.33E-01	Medium
	GPT-5				
	Actual	2.72E-07	Yes	-6.81E-01	Medium
	Gemini 2.5 Pro				
	Actual	9.37E-07	Yes	-6.53E-01	Medium
	Claude Sonnet 4				
Low Temperatures (°F)	Actual	0.66	No	-0.04083666	Negligible
	GPT-5				
	Actual	0.00012	Yes	-0.4647079	Small
	Gemini 2.5 Pro				
	Actual	1.81E-05	Yes	-5.26E-01	Medium
	Claude Sonnet 4				

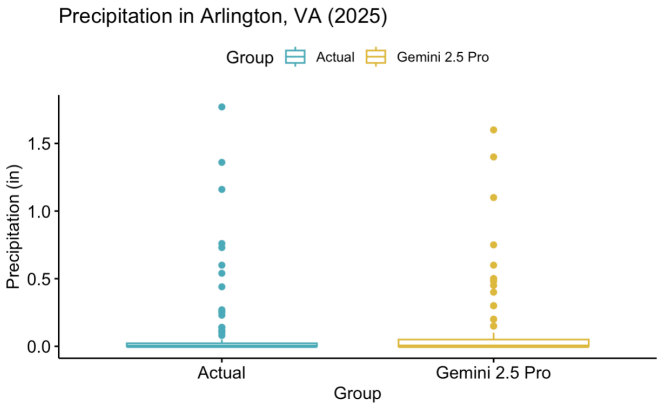
Later, the results for the three Wilcoxon signed-rank tests that evaluated the accuracy of each LLM in predicting daily precipitation from June 1, 2025 to August 31, 2025 in Arlington, VA were analyzed. As seen in Table 4 below, the p-value for the first group, which compared the actual precipitation values to the precipitation predicted by the GPT-5 model, was 0.01382, indicating a statistically significant difference between the two groups. This suggests that GPT-5's precipitation predictions differed meaningfully from actual precipitation totals, and the discrepancy was unlikely to have occurred by random chance. As reflected visually in Figure 10, GPT-5 tended to produce precipitation estimates that did not closely match the observed data, with noticeable differences in magnitude across multiple days. Although this does not imply that GPT-5 is entirely incapable of improving with additional or more specialized meteorological training data, these findings demonstrate that its current predictive capabilities for precipitation are limited.

Figure 10: Actual vs. GPT-5 Precipitation in Arlington, VA



For the second test, which compared the actual precipitation values to those predicted by the Gemini 2.5 Pro model, the p-value was 0.7165, indicating no statistically significant difference between the two groups. This suggests that any deviations between Gemini 2.5 Pro's predictions and the real precipitation values were likely due to random variation rather than systemic inaccuracy. As illustrated in Figure 11, the distribution of Gemini's precipitation predictions more closely resembled the pattern of the actual data compared to the other two models, although some inconsistencies in magnitude and range were still apparent. Despite the lack of statistical significance, this result implies that Gemini 2.5 Pro may have comparatively stronger potential for accurate precipitation forecasting, though further research using larger datasets would be necessary to confirm its reliability.

Figure 11: Actual vs. Gemini 2.5 Pro Precipitation in Arlington, VA



As shown in Table 4, the final test compared the actual precipitation values to the precipitation predicted by the Claude Sonnet 4 model. The resulting p-value of 0.3457 indicated no statistically significant difference between the two groups. This suggests that Claude Sonnet 4's predictions did not systematically deviate from actual precipitation measurements. However, as seen in Figure 12, Claude's distribution exhibited several inconsistencies in day-to-day variability, including certain overestimations and underestimations that reduced its predictive alignment with real observations. While this model did not generate large outliers, its moderate deviation patterns suggest that its precipitation forecasting ability remains limited. Overall, although Claude Sonnet 4 did not differ significantly from the actual data in a statistical sense, its performance indicates that further refinement and larger training samples would be necessary before applying this model to real-world precipitation forecasting.

Figure 12: Actual vs. Claude Sonnet 4 Precipitation in Arlington, VA

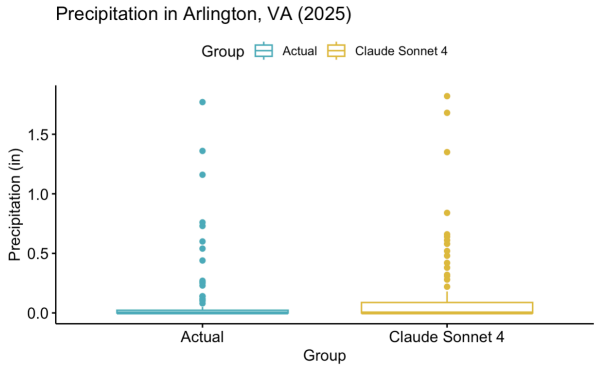


Table 4: Paired Wilcoxon Test Results

	Paired Wilcoxon Test Results				
	Groups	p-value	Significant Difference (α = 0.05)	Effect Size	Interpretation for Effect Size
Precipitation (in)	Actual	0.01382	Yes	0.3317089	Small
	GPT-5				
	Actual	0.7165	No	0.01323087	Negligible
	Gemini 2.5 Pro				
	Actual	0.3457	No	0.03923768	Negligible
	Claude Sonnet 4				

Across all weather variables analyzed, high temperatures, low temperatures, and precipitation, the predictive accuracy of the evaluated LLMs varied substantially. All three models demonstrated statistically significant differences from actual high temperatures. For low temperatures, GPT-5 was the only model that did not significantly differ from actual observations. For precipitation, only GPT-5 showed a statistically significant difference from observed values. Although some models aligned more closely with specific weather variables, none demonstrated consistent accuracy across all measures. These findings suggest that further refinement is necessary before these LLMs can be considered reliable for weather forecasting.

5 DISCUSSIONS

The results from this study demonstrate that while LLMs show some ability to recognize general weather trends, their effectiveness in accurately forecasting future weather conditions is limited. Across high temperatures, low temperatures, and precipitation, none of the models consistently outputted predictions that matched the actual observed data. Although comparisons showed some overlapping patterns between the predicted and actual values, the statistical tests revealed that these similarities were often not strong enough to conclude that the models were reliably accurate. This suggests that LLMs may be better suited for identifying broader patterns rather than producing precise daily weather forecasts.

One of the most notable findings from this study was the performance of GPT-5 in predicting low temperatures. The paired t-test comparing the actual low temperatures to GPT-5's predicted low temperatures resulted in a non-significant p-value, indicating that there was no statistically significant difference between the two groups. This suggests that GPT-5 was able to capture the general trend of low temperatures relatively well. Low temperatures tend to vary more gradually than high temperatures, which may explain why GPT-5 was more successful in modeling them. However, despite this interpretation, the effect size was negligible, indicating that while the results were not significantly different, they were also not strong enough to confidently support the use of GPT-5 for real-world low temperature forecasting.

In contrast, all three LLMs performed poorly when predicting high temperatures. The paired t-tests for high temperature predictions all resulted in statistically significant differences between the predicted and actual values. This suggests that the discrepancies observed were unlikely to be due to random chance. The distribution plots further supported this conclusion, as the predicted high temperatures from all models showed higher mean values and lower variability compared to the actual data. High temperatures are influenced by several atmospheric factors, such as cloud cover and radiation, which LLMs are unable to account for since they do not incorporate physical weather modeling. As a result, the models appeared to rely heavily on historical averages, leading to less accurate predictions.

The results for precipitation forecasting were more mixed. GPT-5 showed a statistically significant difference between its predicted precipitation values and the actual observations, suggesting that it struggled to accurately model precipitation patterns. In contrast, Gemini 2.5 Pro and Claude Sonnet 4 did not show statistically significant differences when compared to actual precipitation values. This indicates that these two models may have been more effective at capturing the overall distribution of precipitation. However, further analysis revealed inconsistencies in daily precipitation amounts, suggesting that the lack of statistical significance does not necessarily imply high predictive accuracy. Since precipitation is highly variable and often skewed, it is one of the most challenging weather variables to predict accurately.

These findings align with previous research that has shown LLMs to be more effective at organizing, summarizing, and classifying weather related information than producing precise forecasts. Prior studies have demonstrated that models such as GPT-4o and Meta's Llama series perform well in tasks related to disaster classification and impact analysis but underperform when applied to forecasting tasks that require modeling complex physical processes. The results of this study apply their conclusions to newer LLMs, including GPT-5, indicating that improvements in model size and architecture are not sufficient to overcome these fundamental limitations.

Overall, the results suggest that LLMs are not yet suitable for use as tools for weather forecasting yet. While they are capable of identifying general trends and patterns, their inability to accurately predict daily weather conditions limits their practical application in real-world settings. However, LLMs still have potential as supplementing tools, especially when used alongside traditional forecasting models. By assisting with data interpretation, trend analysis, or summarization of forecast information, LLMs could contribute meaningfully to weather predictions as the technology continues to evolve.

6 LIMITATIONS

There were several limitations that influenced the scope of this study, particularly regarding the constraints imposed by the LLM interfaces that were used for forecasting. Because the free versions of GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4 restrict both the number and size of allowable file uploads, only a single PDF with one year of historical weather data could be given to each model.

Although the dataset fit within the attachment size limit, these platform restrictions prevented the ability to include more extensive or multi year datasets that may have improved the accuracy of the resulting predictions. Additionally, the free tier prompt limits reduced the number of opportunities there were to refine the instructions given, making it difficult to improve the model's performance iteratively. For GPT-5 specifically, the model failed to generate a complete set of predictions in the first two attempts. This required the full prompt to be resubmitted multiple times before the output including all 92 days was obtained. These barriers may have caused inconsistency across models and limited the strength of the predictions.

There were also methodological constraints related to the behavior of the models. During dearly testing, several models had a tendency to replicate the 2024 values directly into the 2025 predictions. This required having to include an extra, explicit instruction within every prompt to not map the data directly from 2024 to 2025. Even with this clarification, there is still the possibility that some residual pattern copying influenced the forecasts. This limitation reflects a broader challenge of prompting LLMs for quantitative predictions, which is that their outputs can be sensitive to subtle variations in instruction wording, while the opportunities to refine the prompts were limited due to usage caps. Future tests using paid-tier versions of the LLMs would allow for more extensive experimentation and the use of more informative datasets.

Lastly, the study was limited to focus on a single location, Arlington, VA, using data from the Ronald Reagan Washington Airport weather station. While this allowed conduction of accurate, local predictions, it meant that the study did not explore other surrounding regions. So, it cannot be generalized to other areas that could have potentially different weather patterns. The choice of using Arlington, VA was intentional based on the availability of a continuous dataset, rather than because of a constraint in the methodology.

7 CONCLUSION

In conclusion, this study evaluated the effectiveness of three LLMs, GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4 in order to predict weather temperature from June 1, 2025 to August 31, 2025 in Arlington, Virginia. When comparing each model's performance at generating predictions to observed precipitation and temperature, there were clear differences. GPT-5 had more accurate forecasts with a lower variability, and closer to the observed weather trends. On the other hand, Gemini 2.5 Pro and Claude Sonnet 4 had a greater deviation to the actual data. From this, LLLMs, specifically GPT-5, are shown to be capable at generating general weather trends; however, overall findings show that the models are not precise enough to be used in a real-world setting.

LLMs are shown to be a growing source and do have the potential to eventually contribute to practical forecasting, through more research. Each model, GPT-5, Claude Sonnet 4, and Gemini 2.5 Pro continue to advance and receive more focused model developments day by day. Thus, they gain more improved

tools allowing them to analyze data, identify specific trends, and locate connections that arise when predicting weather trends. While, in the current day, the discrepancies between observed and predicted data emphasize how LLMs are not yet ready to supersede transitional weather projections, more research is being gathered as each model continues to grow. Therefore, by reducing limitations and providing models with increased methods to evaluate weather patterns, LLM's do face the possibility of being integrated in meteorological applications.

8 ACKNOWLEDGEMENTS

The feedback and assistance from Dr. Kristina Kramarczuk and Valentina Hong on this research project was greatly appreciated.

9 REFERENCES

1. Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., & Pappenberger, F. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883. https://doi.org/10.1175/BAMS-D-23-0162.1
2. Duan, Z., Bian, C., Yang, S., & Li, C. (2025). Prompting a large language model for multi-location multi-step zero-shot wind power forecasting. *Expert Systems with Applications, 280,* 127436. https://doi.org/10.1016/j.eswa.2025.127436
3. Nabi, Z., Kumar, D., & Central University of Jammu. (2019). The Evolution and Revolution of Weather Forecasting: A review. *International Journal of Research and Analytical Reviews*, *6–6*, 505–506. https://www.ijrar.org
4. Raiaan, M. A. K., Mukta, Md. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742
5. Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, *307*(2), e230163. https://doi.org/10.1148/radiol.230163
6. Wang, Y., & Karimi, H. A. (2025). Exploring large language models for climate forecasting. *AIMS Artificial Intelligence and Computation, 3*(1), 1–24. https://doi.org/10.3934/aci.2025001
7. Yu, Yongan, et al. "WXImpactBench: A Disruptive Weather Impact Understanding Benchmark for Evaluating Large Language Models." arXiv, 2025. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2505.20249.
8. Zafarmomen, N., & Samadi, V. (2025). Can large language models effectively reason about adverse weather conditions? *Environmental Modelling & Software*, *188*, 106421.

https://doi.org/10.1016/j.envsoft.2025.106421