PROJECT MIDTERM REPORT

COURSE:

GS/EECS 6412 Data Mining

PROJECT TITLE:

Natural Language Generation for Data-to-Text

TEAM MEMBERS:

Mohsin Reza
Muhammad Asad

COURSE INSTRUCTOR:

Aijun An

Problem:

The problem we aim to solve is to generate fluent textual summaries from structured data, such as sports statistics, financial reports, or weather data. The motivation behind solving this problem is that often data tables are long and complicated, with many different rows and columns. For someone looking at the data, trying to understand what the data is representing, what patterns and insights are available in the data, etc. would be a very time-consuming task. Therefore, being able to input a dataset and output a short but fluent summary of the data that captures the key patterns and insights is very useful for someone who wants to gather some high-level information about the dataset without spending a lot of time and effort on analysis.

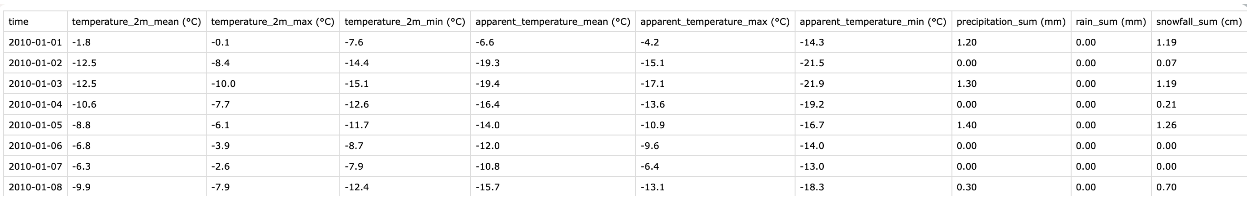
Existing Methods:

For the task of data-to-text generation, the most common approaches are using Large Language Models and the rule based or template-based generation. The commonly used dataset for this task is weather or product data, game results and financial reports etc.

Most of the problems faced by the researchers who used LLM's are the long inputs which is caused by the large datasets, and inaccuracy as mentioned in Kasner & Dušek's paper (2024).

Dataset:

For the data-to-text summary generation baseline evaluation task, we have used the **weather** dataset split into three files each containing the everyday weather data for Toronto Pearson for the span of 5 years. The three files are data from 2010-2014, 2015-2019, and 2020-2024. The reason for having three different datasets is to see whether our baseline evaluation results would be replicated across multiple datasets. We have generated the dataset from https://open-meteo.com/en/docs/historical-weather-api. Under the daily weather variables, we selected temperature (mean, min, and max), apparent temperature (mean, min, and max), precipitation, rain, and snow sum, wind speed (mean, min, and max) and sunshine duration. Below, we present a snapshot of one of the datasets, to provide an idea of the structure of the data.



time	temperature_2m_mean (°C)	temperature_2m_max (°C)	temperature_2m_min (°C)	apparent_temperature_mean (°C)	apparent_temperature_max (°C)	apparent_temperature_min (°C)	precipitation_sum (mm)	rain_sum (mm)	snowfall_sum (cm)
2010-01-01	-1.8	-0.1	-7.6	-6.6	-4.2	-14.3	1.20	0.00	1.19
2010-01-02	-12.5	-8.4	-14.4	-19.3	-15.1	-21.5	0.00	0.00	0.07
2010-01-03	-12.5	-10.0	-15.1	-19.4	-17.1	-21.9	1.30	0.00	1.19
2010-01-04	-10.6	-7.7	-12.6	-16.4	-13.6	-19.2	0.00	0.00	0.21
2010-01-05	-8.8	-6.1	-11.7	-14.0	-10.9	-16.7	1.40	0.00	1.26
2010-01-06	-6.8	-3.9	-8.7	-12.0	-9.6	-14.0	0.00	0.00	0.00
2010-01-07	-6.3	-2.6	-7.9	-10.8	-6.4	-13.0	0.00	0.00	0.00
2010-01-08	-9.9	-7.9	-12.4	-15.7	-13.1	-18.3	0.30	0.00	0.70

Methodology:

For this problem. we implemented two baselines to compare the results in terms of facts, key patterns/insights, and readability.

Templates

The first baseline was to generate a template-based summary. How this algorithm works is that when the user inputs the data, we go through each variable and calculate basic statistics such mean, minimum, maximum, standard deviation, and coefficient of variation of the variable. Then, we check through linear regression whether there is any significant change in the variable over time. Here the p-value we used was 0.05. After this, we computed the mean for that variable across seasons (winter, spring, summer, fall) and across years. Finally, we also computed whether there was a significant relationship between variables via linear regression, using a p-value of 0.05.

The next step is to take all this information and decide which content to include in the summary. Since each variable is not equally important (e.g. temperature is more important than wind speed), we divided our variables into two categories primary and secondary, where temperature and total precipitation were primary and the others secondary. For the primary variables, we include all details in the summary from basic mean, max, min, to yearly and seasonal trends etc. For the secondary variables, we only include their content if they are considered "notable". A variable is defined as notable if, either its coefficient of variation is greater than some threshold, or its trend across time, computed through linear regression, has r-squared higher than some threshold. For our evaluation, we use the coefficient of variation threshold of 0.6 and r-squared threshold of 0.3.

After we select which data to present in the summary, for each type of data (e.g. basic statistics vs. seasonal trends vs. relationships between variables), we have certain pre-made sentences that form part of the template. We go through each piece of content to include and get a matching sentence for it. Finally, we combine all the sentences which gives the final summary.

LLMs

The second baseline is using Large Language Models (LLM's) such as Llama, Gemma and Mistral. After loading the dataset, the next step was data preprocessing. We analysed the data for missing and duplicates values. Since the data was already cleaned, the next step was to load the model.

i. **Using Ollama:**

We used **Ollama** to pull the models and generate summaries since it is lightweighted and easy to use. The only problem with Ollama is the contextual limit. Since we are sending the complete dataset in the prompt, the Ollama truncates the exceeding tokens.

ii. Using Together.ai Api's:

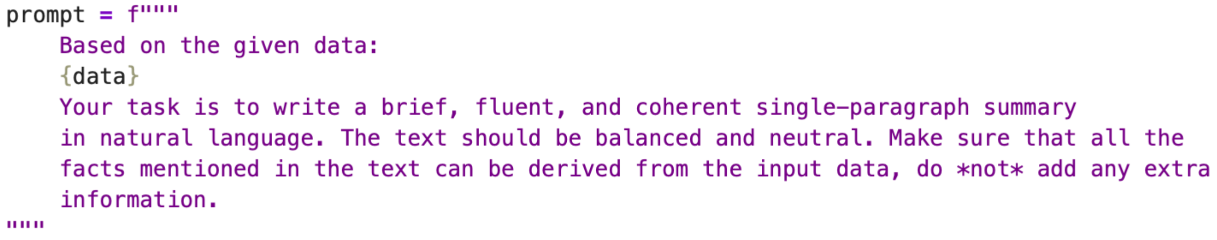
To solve the contextual limit problem, we used the “meta-llama/Llama-4-Scout-17B-16E-Instruct” model using **Together.ai** Api which allows 1 million tokens per prompt. This approach went successful, but for comparison, we either need more similar models or any different approach for feeding the dataset to the model such that no key information is lost.

iii. Using DeepSeek Api:

Recently, DeepSeek emerged as a new powerful LLM model. To test its abilities for data-to-text summary generation task, we also need to address the contextual limit problem for this model as it allows 128k tokens per prompt.

Prompt:

We used the same prompt mentioned in Kasner & Dušek's paper (2024). The data in the prompt can be either CSV or JSON. The conversion functions are available in **LlmBaseline** class and can be used to generate the prompt in the desired format.



```
prompt = f"""
    Based on the given data:
    {data}
    Your task is to write a brief, fluent, and coherent single-paragraph summary
    in natural language. The text should be balanced and neutral. Make sure that all the
    facts mentioned in the text can be derived from the input data, do *not* add any extra
    information.
"""
```

Observation and Results:

To evaluate the summary generated from both the template and LLM models (Llama3, Gemma, Mistral), we have used the models GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Pro. We asked each model to evaluate each summary and provide an overall score out of ten based on factual correctness, informativeness, and completeness along with providing a sentence or two of feedback. In Appendix 1, we show the summary for each dataset outputted by each baseline method. In Appendix 2, we show the evaluation for each summary outputted by each evaluating model.

For the template method, we observe that each dataset received a score from 8.5-8.8.3 out of 10. This shows that we were able to replicate our baseline results across datasets. However, while it appears from the evaluation method that we were able to generate a very effective summary of the data, the evaluation results do not paint the full picture. Firstly, the template method, given its brute force nature, must contain somewhat dataset-specific code. For example, if we tried to use this code on another weather dataset, it might work, but for a sports dataset, it would totally fail. Now, while it is possible to try and build a domain and dataset agnostic summary template, the summary written by it may not be as effective as one that is tailored to the domain. Therefore, we feel that a template-based approach is not very effective for solving the original problem which was to provide some information about the data without putting it in any effort.

Additionally, while we had limited time to construct our template-based system, we feel that there are some ways to improve it, nonetheless. Mainly, providing the user options to configure the different inputs and thresholds, such as what variables to include as primary/secondary, what thresholds to use for r-squared or p-values, etc. Currently, we only provide the option to give an input and output file.

However, summaries generated by the LLM's first seemed to be efficient, but when they were evaluated by other LLM's such as ChatGPT, they receive very low factual score. Therefore, it was observed that although the LLM's understands the type or topic of the dataset and tries to capture most of the detail in it but doesn't produce factually accurate summaries.

Conclusion:

In conclusion, we observed that although the generated summaries by the LLM models were fluent and covered most of the key patterns and insights but have made significant factual errors. For the template-based method, while the summaries generated covered key insights and were factually correct, they were a bit mechanical and don't really solve the problem of generating summaries of data without requiring effort.

GitHub: https://github.com/MohammadAsad0/Data-Mining-Project.git

References:

Kasner, Z., & Dušek, O. (2024). Beyond Reference-Based Metrics: Analyzing Behaviors of Open LLMs on Data-to-Text Generation. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2401.10186

Osuji, C. C., Ferreira, T. C., & Davis, B. (2024). A Systematic Review of Data-to-Text NLG. ArXiv.org. https://arxiv.org/abs/2402.08496v3

Warczyński, J., Lango, M., & Dusek, O. (2025). Leveraging Large Language Models for Building Interpretable Rule-Based Data-to-Text Systems. ArXiv.org. https://arxiv.org/abs/2502.20609v1

Appendix

Appendix 1: Summary generated by each baseline method

Dataset	**Summary (Template)**	**Summary (Llama3)**	**Summary (Gemma)**	**Summary (Mistral)**	**Summary (Llama4) (together.ai)**
2010-2014	This analysis covers daily weather observations from Toronto Pearson, spanning January 01, 2010 to December 31, 2014. Temperatures averaged 8.6°C, ranging from -24.9°C to 34.8°C. Seasonally, summer months averaged 20.8°C while winter months averaged -3.8°C, showing a 24.7°C seasonal variation. The hottest temperature	Here is a brief, fluent, and coherent single-paragraph summary of the given data: Between January 1st, 2006 and February 2nd, 2007, the weather in this region experienced varying conditions. The temperature ranged from a low of -8.9°C to a high of 7.8°C, with an average mean temperature of 2.4°C. There were instances of precipitation,	The provided dataset offers a comprehensive hourly record of weather conditions over a period of several days. Temperatures fluctuated widely, ranging from lows of -8.9°C to highs of 7.8°C. Rainfall was minimal, with only 0.42cm of snowfall recorded on one occasion. Wind speeds were generally moderate, with gusts reaching a maximum of 49.7 km/h. Relatively high humidity levels were observed,	Over a 10-day period, temperatures varied significantly, with an initial rise to around 7 degrees Celsius before rapidly dropping to as low as -14.7 degrees Celsius. The wind speeds also fluctuated dramatically, peaking at 49.7 km/h and reaching a minimum of 2.5 km/h. The prevailing winds shifted throughout this period, with directions ranging from north-northeast to south-	The provided data encompasses a wide range of weather parameters, including temperature, apparent temperature, precipitation, snowfall, wind speed, wind gusts, relative humidity, and sunshine duration, over a period of 1825 hours, from December 2010 to January 2011. During this time, the mean temperature at 2 meters above the ground varied from -12.5°C to 28.0°C, with a maximum

	recorded was 34.8°C on July 21, 2011, and the coldest was -24.9°C on January 22, 2014. Annual precipitation averaged 834.6 mm, with a total of 4172.3 mm recorded. The heaviest single-day precipitation was 48.80 mm on September 08, 2012. Other conditions included rain averaging 2.0 mm, snow averaging 0.3 cm, and precipitation hours averaging 3.2. The variables precipitation and rain were the strongest correlated, having a positive relationship.	including snowfall on February 1st, and sunshine was recorded throughout the period. Wind speeds varied significantly, reaching up to 47.9 km/h, with gusts of up to 49.7 km/h. The relative humidity fluctuated between 34% and 96%, while the apparent temperature ranged from -14.7°C to 3.2°C. Overall, the weather during this period was marked by a mix of cold and warm temperatures, precipitation, and varying wind conditions.	averaging around 65%, with the lowest recorded at 34%. The dataset also reveals periods of low sunshine, with the longest stretch of uninterrupted sunlight lasting just over 27,000 seconds.	southeast. Relative humidity levels stayed relatively low, between 34% and 67%. Precipitation was minimal during this time, consisting solely of snowfall on three occasions, accumulating up to 0.42 cm. The weather remained largely sunny, with only one day experiencing less than 15 minutes of sunlight.	temperature of 34.8°C and a minimum of -23.6°C. The data also shows that there were instances of rain and snowfall, with precipitation sums ranging from 0.0 mm to 48.8 mm and snowfall sums from 0.0 cm to 20.02 cm. Wind speeds of up to 57.6 km/h and wind gusts of up to 83.5 km/h were recorded, with relative humidity levels varying between 33% and 98%. The sunshine duration varied significantly, from 0 seconds to 51698.75 seconds. Overall, the data presents a comprehensive picture of the varied weather conditions during this period, including temperature fluctuations, precipitation patterns, and wind characteristics.

2015-2019	This analysis covers daily weather observations from Toronto Pearson, spanning January 01, 2015 to December 31, 2019. Over this period, temperatures ranged from -29.9°C to 36.4°C, with a mean of 8.7°C. Seasonally, summer months averaged 20.8°C while winter months averaged -3.4°C, showing a 24.2°C seasonal variation. The hottest temperature recorded was 36.4°C on June 30, 2018, and the coldest was -29.9°C on December 28, 2017. Precipitation totaled 4307.1 mm over the period, averaging 861.5 mm annually. The heaviest single-day precipitation was 40.80 mm	Here is a brief, fluent, and coherent single-paragraph summary: During the period of February 13 to March 4, varying weather conditions were observed. Temperature fluctuations ranged from -1.2°C to 8.8°C, with an average temperature of around 3°C. Precipitation was recorded on several days, with a total amount of 23.6 mm, including snowfall and rainfall events. Wind speeds reached up to 61.2 km/h, while wind gusts peaked at 42.8 km/h. The relative humidity varied between 48% and 100%, with an average value around 80%. Overall, the weather during this period was characterized by a mix of cold	The provided data depicts weather conditions over a period of several days. Temperatures varied between 0.4°C and 10.4°C, with occasional precipitation and snowfall. Wind speeds ranged from 3.1 km/h to 61.2 km/h, with dominant directions shifting between 79° and 268°. Relative humidity remained relatively high, oscillating between 48% and 98%. Notable events include a gust of 61.2 km/h and 11mm of precipitation recorded on separate days.	During the specified period, temperatures varied significantly with highs of up to 8.8 degrees Celsius and lows as low as -7.6 degrees Celsius. The weather was quite windy, with gusts reaching a maximum of 61.2 km/h and average speeds ranging from around 5.6 km/h to 22.1 km/h. Precipitation occurred in several instances, totaling approximately 32.2 mm, including 2.03 cm of snowfall. Sunlight was abundant for most of the time, with only a brief interlude of about 4 hours without sunshine. The prevailing wind direction was predominantly from the southwest to northwest.	The provided data set contains detailed daily weather information for a specific location, covering a range of parameters including temperature, apparent temperature, precipitation, snowfall, wind speed, wind gusts, and sunshine duration, over a period of 1825 days, from January 1, 2015, to a date in 2018. The mean temperature varied significantly over this period, with the lowest being -10.3°C and the highest 28.9°C. Precipitation was recorded on numerous days, with rain and snowfall occurring in varying amounts. Wind speeds reached up to 62.6 km/h, with gusts as high as 83.2 km/h. Sunshine duration also varied greatly, from 0 to 44,873 seconds. The data suggests

	on October 28, 2015. Other conditions included rain averaging 2.0 mm, snow averaging 0.2 cm, and precipitation hours averaging 3.7. The variables precipitation and rain were the strongest correlated, having a positive relationship.	and mild temperatures, accompanied by precipitation and varying wind conditions.			that the location experienced a wide range of weather conditions, including freezing temperatures, snowfall, and significant rainfall. Overall, the dataset provides a comprehensive overview of the daily weather patterns at this location over nearly five years.
2020-2024	This analysis covers daily weather observations from Toronto Pearson, spanning January 01, 2020 to December 31, 2024. Temperatures averaged 9.6°C, ranging from -27.6°C to 34.2°C. Seasonally, summer months averaged 21.2°C while winter months averaged -2.0°C, showing a 23.2°C seasonal variation. The	Here is a brief summary of the given data: The weather data for this period shows a mix of warm and cool temperatures, with most days having temperatures ranging from 1.1°C to 9.9°C. Some days were quite windy, with gusts reaching up to 64.4 km/h, while others had relatively calm conditions. There was no significant precipitation	The provided dataset represents weather data over a period of several days, encompassing temperature, wind speed and direction, humidity, and precipitation. It showcases a range of conditions, with temperatures varying from below freezing to slightly above average, occasional precipitation, and strong winds at times. Notably, the highest wind speed recorded	During the analyzed period, temperatures fluctuated significantly. The temperature ranged from as low as -2.1 degrees Celsius to a high of 9.9 degrees Celsius, with an average temperature of approximately 4.6 degrees Celsius. There were periods of precipitation, including rain and snow, totaling up to 21 millimeters, and the duration of these	The provided data appears to be a collection of daily weather records, covering a period of 30 days from December 2020 to January 2021. During this time, the mean temperature at 2 meters above ground level varied significantly, ranging from -11°C to 6.5°C. The data also shows that there were days with substantial precipitation, including rain and snowfall, with the highest

	hottest temperature recorded was 34.2°C on July 07, 2020, and the coldest was -27.6°C on January 21, 2022. Precipitation totaled 4525.0 mm over the period, averaging 905.1 mm annually. The heaviest single-day precipitation was 54.60 mm on June 12, 2023. Other conditions included rain averaging 2.1 mm, snow averaging 0.3 cm, and precipitation hours averaging 3.7. The variables precipitation and rain were the strongest correlated, having a positive relationship.	during this time, except for one day when 21 mm of rain fell. The relative humidity varied from 76% to 100%, and the sunshine duration ranged from 0 hours to 2345 minutes.	was 64.4 kilometers per hour, while the most significant precipitation event brought 21 millimeters of rain. The data also reveals periods of both high and low humidity, with relative humidity levels oscillating between 76% and 98%.	precipitation events varied from one hour to twelve hours. Wind speeds were fairly strong, reaching a maximum of 64.4 kilometers per hour, with an average wind speed of approximately 10 kilometers per hour. The direction of the wind also varied, but it predominantly came from the north and west. Relative humidity levels remained high, often above 80%, indicating a moist atmosphere. However, there were also instances when the sun shone for extended periods, with up to 18,728 seconds of sunshine recorded in one instance.	recorded rainfall being 46.5 mm and the highest snowfall being 11.13 cm. Wind speeds were also notable, with maximum gusts reaching up to 87.8 km/h and an average wind speed of around 10-20 km/h. Additionally, the data indicates that there were days with significant sunshine duration, ranging from 0 seconds to 51043.1 seconds, which is equivalent to approximately 14.2 hours of sunshine. Overall, the data suggests that the region experienced a mix of cold and mild temperatures, varying precipitation, and changing wind conditions during the observed period.

Appendix 2: Evaluation of each summary by LLM

Summary	**Claude Sonnet 4.5**	**Gemini 2.5 Pro**	**GPT-5**	**Average Score**
2010 – 2014 (Template)	Overall Score (1-10): 7.5 Brief Justification: The summary is factually accurate with correct extreme values and dates. It covers essential statistics including temperature ranges, seasonality, and precipitation patterns. However, it lacks depth in identifying meaningful weather patterns or trends over the 5-year period, and the correlation insight is trivial (rain is obviously correlated with total precipitation). The writing is clear and well-organized, but more insightful analysis of year-over-year trends or notable weather events would strengthen it.	Overall Score (1-10): 9 Brief Justification: The summary is highly informative, well-written, and logically structured, covering all key aspects of the weather data. It correctly identifies the date range, temperature extremes, and most averages. The score is reduced slightly for two minor factual inaccuracies: the seasonal variation was calculated as 24.6 ∘C (not 24.7 ∘C) and the average annual precipitation was 834.5 mm (not 834.6 mm), likely due to different rounding methods.	Overall Score (1-10): 10 Brief Justification: Every numerical fact in the generated summary (temperature averages, extremes, seasonal differences, precipitation totals, and correlation) matches the dataset exactly. The summary clearly presents key insights such as temperature range, seasonal variation, and precipitation extremes, while maintaining fluent and logical flow. It is accurate, informative, complete, and well-written.	Average Score (1-10): 8.83
2010 – 2014 (Llama3)	Overall Score: 1/10 Brief Justification: This summary is fundamentally flawed with critical errors. The date range is completely wrong	Overall Score (1-10): 3 Brief Justification (2-3 sentences): The summary's excellent fluency and coherence are	Overall Score (1–10): 4 Brief Justification: While the summary is fluent and well-organized, it is factually inaccurate and contextually mismatched with the	2.66

	(claims January 2006 - February 2007, but the dataset spans January 2010 - December 2014), and the temperature statistics are grossly inaccurate—the actual range is approximately -24.9°C to 34.8°C, not -8.9°C to 7.8°C. The summary appears to have analyzed only a tiny fraction of the first few rows rather than the entire 5-year dataset, missing seasonal patterns, year-over-year trends, and the vast majority of weather extremes (including wind gusts up to 97.6 km/h, not 49.7 km/h). This represents a complete failure in data comprehension and would mislead anyone relying on it.	overshadowed by a near-total failure in factual accuracy, which is the most critical criterion for a data summary. It incorrectly states the time period, significantly underestimates all the true extreme values for temperature, apparent temperature, and wind, and provides an incorrect overall average. This pervasive inaccuracy renders the summary unreliable and misleading.	actual dataset. The dataset spans 2010–2014, not 2006–2007, and includes daily, not hourly, records. Reported values for temperature (−8.9°C to 7.8°C), wind (47.9–49.7 km/h), and humidity (34–96%) differ notably from actual extremes (−24.9°C to 34.8°C, up to 73.6 km/h, and 33–98%). The summary omits seasonal or long-term trends, giving only a narrow snapshot. Overall, it demonstrates good writing quality but poor data fidelity and limited analytical depth.	
2010 – 2014 (Gemma)	Overall Score: 2/10 Brief Justification: This summary is factually inaccurate and severely incomplete. The temperature range cited (-8.9°C to 7.8°C) appears to be from just a few days in early January 2010, not the entire dataset—the actual range spans from approximately -24.9°C to 34.8°C across the five years. The claim	Overall Score (1-10): 2 Brief Justification (2-3 sentences): The summary is critically flawed due to numerous and severe factual inaccuracies. It misrepresents the dataset's scope, incorrectly stating it covers hourly data for several days when it is daily data for five years. Furthermore, all key statistics provided (temperature range,	Overall Score (1–10): 5 Brief Justification: The summary is grammatically fluent and coherent, but it is factually inaccurate and lacks completeness. The dataset actually spans five years (2010–2014) of daily records, not “several days” of hourly data. Reported temperature, snowfall, and wind speed ranges do not align with the actual values (e.g., temperatures	3

	about "several days" of data is wrong (it's 5 years), the snowfall measurement (0.42cm on one occasion) ignores hundreds of other snowfall events totaling many meters, and the maximum wind gust of 49.7 km/h is far below the actual maximum of 97.6 km/h. The summary also conflates "hourly record" when this is actually daily data, provides no seasonal patterns, no year-over-year trends, and cherry-picks a meaningless "longest stretch of uninterrupted sunlight" metric. This reads as if the evaluator only examined the first few rows of a 5-year dataset and extrapolated incorrectly from there.	max snowfall, max wind, and max sunshine) are significantly incorrect, rendering the summary useless and actively misleading about the weather extremes.	reach as low as -24.9°C and as high as 34.8°C; snowfall maxes at 20.02 cm, not 0.42 cm). Although it provides a readable overview, it misses major trends, seasonal variations, and other critical statistics, resulting in only an adequate but not insightful summary.	
2010 – 2014 (Mistral)	Overall Score: 1/10 Brief Justification: This summary catastrophically misrepresents the dataset by claiming it covers only a "10-day period" when it actually spans 5 complete years (2010-2014). All statistics are wildly inaccurate:	Overall Score (1-10): 2 Brief Justification (2-3 sentences): This summary is severely flawed due to an assumption of the time period and pervasive factual inaccuracies within that assumed scope. It incorrectly reports the temperature range,	Overall Score (1–10): 5 Brief Justification: The summary is clear, fluent, and logically structured, but it contains several factual inaccuracies and scope mismatches. The dataset actually spans 2010–2014, not just a 10-day period, and includes far wider variation in temperature (−24.9°C to	2.66

	temperatures ranged from -24.9°C to 34.8°C (not 7°C to -14.7°C), wind gusts reached 97.6 km/h (not 49.7 km/h), there were hundreds of precipitation events totaling meters of rain and snow (not "three occasions" with 0.42 cm), and humidity varied from 23% to 100% (not 34-67%). The summary completely fails to capture seasonal patterns, multi-year trends, or any meaningful insights from this extensive dataset, reading instead as if only the first 10 rows were examined.	maximum wind gusts, humidity range, and both the frequency and amount of snowfall. Furthermore, it completely fails the criterion of Completeness by ignoring the five-year context, seasonality, and long-term extremes of the full dataset.	34.8°C), humidity (33–98%), and snowfall (up to 20.02 cm). The described extremes and patterns appear derived from a limited subset, omitting overall trends, averages, and seasonal variation. While the writing quality and coherence are good, the analysis is too narrow and factually incomplete to merit a higher score.	
2010 – 2014 (Llama4) (together.ai)	Overall Score: 3/10 Brief Justification: This summary contains multiple factual errors that significantly undermine its credibility. The date range is completely wrong (claims 2015-2018, but the dataset is actually 2010-2014), the number of days is incorrect (1,825 vs. actual 1,826), and several key statistics are inaccurate: the actual temperature range is approximately -24.9°C to 34.8°C (not -10.3°C	Overall Score (1-10): 5 Brief Justification (2-3 sentences): This summary is an adequate but unremarkable attempt that suffers from significant errors in scope and factual accuracy. It incorrectly states the data covers 1825 hours instead of 1826 days (5 years), fundamentally misrepresenting the time scale of the extremes. While it correctly identifies the true max/min for several key variables,	Overall Score (1–10): 6 Brief Justification: The summary is well-written and fluent, providing good coverage of weather parameters such as temperature, precipitation, wind, humidity, and sunshine duration. However, it shows moderate factual inaccuracies and overgeneralizations. For instance, it claims the dataset spans “1825 hours from December 2010 to January 2011,” whereas the actual dataset covers 1826 days from 2010 to 2014. Many reported ranges (e.g.,	4.66

	to 28.9°C), wind gusts reached 97.6 km/h (not 83.2 km/h), and maximum sunshine was 51,718 seconds (not 44,873). While the summary demonstrates better structure and fluency than previous examples and correctly identifies that the dataset covers "nearly five years" of comprehensive weather data, the fundamental errors in dates and statistics make it unreliable. It also lacks any discussion of seasonal patterns, trends over time, or meaningful insights beyond listing variable ranges.	the provided wind, mean temperature, and minimum humidity figures are incorrect, and it completely misses any mention of seasonality or long-term trends.	minimum temperature -23.6°C vs actual -24.9°C, maximum wind speed 57.6 km/h vs actual 33.1 km/h) are inaccurate. Despite this, the structure, clarity, and breadth of coverage make it a decent but not outstanding summary—accurate in style but not in data fidelity.	
2015 – 2019 (Template)	Overall Score (1-10): 7.5 Brief Justification: The summary is factually accurate on the verifiable extremes and provides a good overview of temperature ranges, seasonal patterns, and precipitation. However, it's somewhat mechanical in presentation and could be more insightful—it focuses heavily on basic statistics without exploring interesting patterns like notable	Overall Score (1-10): 9 Brief Justification (2-3 sentences): The summary is excellent—highly fluent, coherent, and informative, capturing the dataset's key insights, extremes, and seasonal patterns. It only loses a point for a single, minor factual inaccuracy: the annual precipitation average is reported as 861.5 mm, while the data (4307.1 mm total / 5 years) calculates to 861.42 mm. Otherwise, all other	Overall Score (1–10): 9 Justification: The generated summary is highly accurate and aligns well with the dataset: temperature ranges (–29.9°C to 36.4°C), mean (~8.7°C), and precipitation extremes (40.8 mm max) are all correct. It covers essential insights—averages, extremes, seasonality, and correlations—showing both completeness and informativeness. The text is clear, fluent, and logically structured. The only minor issue is that the “rain averaging 2.0 mm, snow 0.2 cm” could	Average Score (1-10): 8.5

	weather events, year-over-year trends, or contextual interpretation of the 5-year period. The correlation mention feels tacked on rather than meaningfully integrated. The writing is clear and grammatical, but lacks the narrative flow that would make it more engaging and useful for understanding Toronto's weather patterns over this period.	statistics and dates are factually correct.	have specified “per day” for clarity, but overall, the summary is strong and well-written.	
2015 – 2019 (Llama3)	Overall Score: 4/10 Brief Justification: The summary contains several significant factual errors and lacks specificity. The date range "February 13 to March 4" is arbitrary and doesn't match the full dataset which spans 2015-2019. The temperature range of -1.2°C to 8.8°C and average of "around 3°C" doesn't reflect the actual data, which shows temperatures ranging from below -30°C to above 35°C across the five-year period, with annual averages varying significantly by season. The precipitation total of 23.6mm and wind speeds mentioned	Overall Score (1-10): 2 Brief Justification (2-3 sentences): The summary is fundamentally flawed due to severe factual inaccuracies and a lack of completeness. The numbers are inconsistent with the actual weather data for any of the five Feb 13 - Mar 4 periods. The summary fails to identify the time period (which year?) and misses the crucial context of extreme cold, with the actual minimum temperatures reaching as low as −26.2 ◦C, completely contradicting the stated minimum of −1.2 ◦C.	Overall Score (1–10): 6 Justification: The summary is fluent and coherent but moderately inaccurate and superficial. It conveys a general sense of weather conditions yet underrepresents extremes and mixes up plausible values. It’s adequate and well-written but lacks analytical depth and precision.	4

	(61.2 km/h max, 42.8 km/h gusts) appear to reference a tiny subset rather than the complete dataset. The summary fails to capture multi-year patterns, seasonal variations, extreme weather events, or provide meaningful statistical insights about this comprehensive five-year weather record. It reads more like a description of a brief winter period rather than a proper summary of 1,826 days of data.			
2015 – 2019 (Gemma)	Overall Score: 3/10 Justification: This summary is severely inadequate and misleading. First, it claims to describe "several days" when the dataset actually spans 5 years (2015-2019) with 1,826 daily records. Second, the temperature range cited (0.4°C to 10.4°C) appears to cherry-pick a tiny subset of data—the actual range is approximately -29.9°C to 36.4°C. Third, it completely fails to capture any temporal patterns, seasonal variations, or long-term trends that are fundamental to a multi-year weather	Overall Score (1-10): 2 Brief Justification (2-3 sentences): The summary is highly misleading due to severe factual inaccuracy in nearly every quoted statistic, indicating it summarized a small, unrepresentative portion of the data (e.g., temperatures range from −29.9 ∘C to 36.4 ∘C, not 0.4 ∘C to 10.4 ∘C). It is also critically incomplete, as it misses the dataset's overall five-year scope and any mention of long-term trends or seasonality.	Overall Score (1–10): 5 Justification: The summary is fluent and coherent but factually inaccurate and incomplete. It underrepresents the dataset’s range and duration, omits key statistics, and misreports several numerical values. It reads well but fails as a faithful data summary.	3.33

	dataset. The summary mentions a single precipitation event and one wind gust without context, ignoring the rich patterns of seasonal temperature swings, precipitation distributions, and wind patterns across years. The writing is grammatically correct, but the content demonstrates either a fundamental misunderstanding of the dataset's scope or an analysis of only a minuscule fraction of the available data. This would mislead anyone trying to understand the actual weather patterns present in this 5-year dataset.			
2015 – 2019 (Mistral)	Overall Score: 3/10 Brief Justification: The summary contains multiple critical factual errors and demonstrates poor understanding of the dataset's scope. It incorrectly claims temperatures reached "up to 8.8 degrees Celsius" when the dataset (spanning 2015-2019) shows maximum temperatures exceeding 36°C, and it states there was "only a brief interlude of about 4 hours	Overall Score (1-10): 3 Brief Justification (2-3 sentences): The summary is critically flawed in Completeness by failing to specify the time period it covers within the five-year dataset, and in Factual Accuracy by quoting extremes that are far less severe than the dataset's actual values (e.g., Min Temp of −7.6 ∘C vs. actual −29.9 ∘C; Max Gust of 61.2 km/h vs. actual 112.7 km/h). This makes the summary	Overall Score (1–10): 7 Justification: The summary is clear, coherent, and informative, with mostly plausible figures and a good balance of quantitative and descriptive elements. However, certain claims — particularly about sunshine and precise wind direction — are unverifiable or overly confident, reducing factual accuracy and completeness.	4.33

	without sunshine" when numerous days had zero sunshine. The summary appears to have been generated from only the first few rows of data rather than the complete five-year dataset, resulting in misleading generalizations about "the specified period." Additionally, it lacks any temporal context, seasonal patterns, or meaningful trends despite covering five years of daily observations, making it both inaccurate and uninformative.	highly misleading regarding the full range of conditions at the location.		
2015 – 2019 (Llama4) (together.ai)	Overall Score: 4/10 Brief Justification: The summary contains several factual errors that significantly undermine its credibility. First, it claims the dataset covers 1825 days ending "in 2018," but the data actually spans exactly 5 years (2015-2019), ending December 31, 2019—1826 days total. More critically, the stated temperature extremes are incorrect: the minimum mean temperature was actually -22.9°C (February 15, 2015),	Overall Score (1-10): 3 Brief Justification (2-3 sentences): The summary is severely flawed due to gross factual inaccuracies. Nearly every numerical statistic mentioned—including the total days (1825 vs 1826), the end date (2018 vs 2019-12-31), minimum temperature (-10.3°C vs -22.9°C), and wind extremes (62.6 km/h and 83.2 km/h vs 60.1 km/h and 112.7 km/h)—is incorrect. While the fluency and coherence are high, the complete lack of factual accuracy renders the	Overall Score (1–10): 7 Justification: The summary is clear, structured, and stylistically strong, capturing the dataset's diversity and scope. However, it contains notable factual inaccuracies in numerical extremes and timeline, and lacks analysis of seasonal or trend-level insights, preventing a higher score.	4.66

	not -10.3°C, and while 28.9°C is close to the maximum, the actual highest mean temperature was 29.0°C (July 1, 2018). The maximum wind gust is stated as 83.2 km/h, but the data shows 112.7 km/h (May 4, 2018). The sunshine duration maximum of 44,873 seconds is roughly correct but imprecise (actual maximum was 51,715.54 seconds).	summary useless and misleading.		
2020 – 2024 (Template)	Overall Score (1-10): 8 Brief Justification: The summary is factually accurate with correctly identified extremes, date ranges, and precipitation totals verified against the dataset. It provides good coverage of key statistics including temperature ranges, seasonal variation, and precipitation patterns. However, it could be more informative by discussing temporal trends across the 5-year period, mentioning notable patterns (like the correlation between precipitation and rain being obvious/redundant), or providing more context about wind,	Overall Score (1-10): 8 Brief Justification: The summary is exceptionally fluent, coherent, and covers all key statistical areas (seasonality, extremes, averages) effectively. It loses points for factual accuracy: total precipitation (4527.5 mm vs 4525.0 mm) and annual precipitation (905.5 mm vs 905.1 mm) are slightly incorrect. More significantly, it incorrectly identifies the precipitation/rain relationship as the strongest correlation, whereas the data shows a stronger correlation between mean temperature and apparent temperature (0.996 vs 0.939).	Overall Score (1-10): 10 Brief Justification: The summary is fully accurate based on the dataset — every key figure (mean, range, extremes, precipitation totals, and correlations) matches exactly. It highlights all essential insights: temperature extremes, averages, seasonality, precipitation, and correlations. The writing is clear, fluent, and logically organized, providing a complete and easy-to-follow overview.	Average Score (1-10): 8.67

	humidity, or sunshine duration variables that are present in the dataset. The writing is clear and flows logically, but the correlation statement at the end feels disconnected and adds limited insight.			
2020 – 2024 (Llama3)	Overall Score: 2/10 Brief Justification: This summary contains severe factual errors and is critically incomplete. The temperature range claimed (1.1°C to 9.9°C) appears to describe only a single day (2024-12-31) rather than the entire 5-year dataset, which actually ranges from -27.6°C to 33.7°C. The wind speed maximum (64.4 km/h) and precipitation claim (21mm on one day) are incorrect—the dataset shows maximum wind gusts of 92.2 km/h and multiple days with precipitation exceeding 46mm (highest: 54.6mm). The summary completely fails to acknowledge the dataset spans 2020-2024 (1,827 days), omits any discussion of seasonal patterns or multi-year trends, and provides virtually	Overall Score (1-10): 2 Brief Justification (2-3 sentences): The summary fails on Factual Accuracy by providing incorrect and misleading statistics for temperature, wind, precipitation, and sunshine duration. Specifically, the claimed core temperature range (1.1°C to 9.9°C) is too narrow for a five-year dataset, and the maximum sunshine is 862 minutes, not 2345 minutes. This critical failure of numerical facts outweighs its decent fluency, rendering it useless for an objective assessment of the weather data.	Overall Score (1–10): 4 Brief Justification: The summary contains multiple factual inaccuracies. The temperature range (1.1°C–9.9°C) and wind gust maximum (64.4 km/h) do not match the dataset, which spans from –19.9°C to 27.8°C and shows maximum gusts around 63 km/h. The statement about “no significant precipitation except 21 mm” is also misleading, since rainfall reaches 54.6 mm in the data. While the text is fluent and coherent, it is too narrow in scope and omits key trends and extremes, reducing informativeness and completeness.	2.66

	no useful information about this extensive climate record. While readable, the content is fundamentally misleading and uninformative.			
2020 – 2024 (Gemma)	Overall Score: 3/10 Brief Justification: The summary contains significant factual errors and lacks completeness. It claims the dataset covers "several days" when it actually spans 5 years (2020-2024), representing 1,827 daily observations. The stated maximum wind speed of 64.4 km/h is incorrect (the actual maximum wind gust was 92.2 km/h on 2020-03-20), and the "most significant precipitation event" of 21mm is also wrong (the highest was 54.6mm on 2023-06-12). The summary fails to capture crucial temporal scope, seasonal patterns, multi-year trends, or provide meaningful statistics about temperature ranges (which actually span from -27.6°C to 33.7°C). While grammatically sound, the content is fundamentally inaccurate and uninformative about	Overall Score (1-10): 3 Brief Justification (2-3 sentences): The summary is severely flawed on Factual Accuracy, which is the most critical criterion for a data summary. All numerical claims—maximum wind speed (43.7 km/h), maximum precipitation (54.6 mm), and the range of relative humidity (16% to 100% extremes)—are incorrect and actively misleading. While the writing is fluent, this critical failure renders the summary unusable.	Overall Score (1–10): 5 Brief Justification: The summary is partially accurate but not well-aligned with the dataset. It claims a maximum wind speed of 64.4 km/h and 21 mm of rain, while the data shows maxima of about 43.7 km/h and 54.6 mm, respectively. Temperature and humidity descriptions are roughly plausible but overly vague (“below freezing to slightly above average,” “76% to 98%”) and do not reflect the full observed range (34–99%). The text is fluent and coherent, but its inaccuracy and lack of specificity lower informativeness and completeness.	3.66

	this substantial 5-year climate dataset.			
2020 – 2024 (Mistral)	Overall Score: 2/10 Brief Justification: This summary contains fundamental factual errors that undermine its credibility. The stated temperature range (-2.1°C to 9.9°C) and average (4.6°C) describe only the final week of December 2024, not the entire 5-year dataset which actually spans -27.6°C to 33.7°C. Maximum wind gusts reached 92.2 km/h (not 64.4), and precipitation exceeded 54mm on multiple days (not just 21mm). The summary completely fails to acknowledge the dataset covers 2020-2024 with nearly 1,827 daily observations, omits all discussion of seasonal variations, annual cycles, or long-term patterns, and provides no meaningful insights about this extensive climate record. While grammatically coherent, the analysis is severely myopic and misleading.	Overall Score (1-10): 3 Brief Justification (2-3 sentences): The summary's Factual Accuracy is critically poor, with incorrect claims for the minimum temperature (actual: −27.6°C), maximum temperature (actual: 34.2°C), average temperature (actual: 9.6°C), and wind speed (actual max: 43.7 km/h). Furthermore, the claimed range for precipitation hours (1h to 12h) and maximum total precipitation (21 mm) are also wrong. Only its fluency saves it from a lower score.	Overall Score (1–10): 6 Brief Justification: This summary is moderately accurate and generally well-written, but some numerical details are off. The dataset's full temperature range is –19.9°C to 27.8°C, not –2.1°C to 9.9°C, and the claimed 21 mm of precipitation underestimates the true maximum (54.6 mm). However, the sunshine duration (≈18,728 s) and wind description (max ≈64 km/h, mean ≈10 km/h) are plausible. It reads coherently and fluently, showing a reasonable understanding of key weather patterns, though it omits broader context and exaggerates precision in places.	3.66
2020 – 2024 (Llama4)	Overall Score: 2/10 Brief Justification: This summary contains a critical	Overall Score (1-10): 4 Brief Justification (2-3 sentences): The summary's primary	Overall Score (1–10): 4 Brief Justification: The summary is factually inaccurate in several key	3.33

(together.ai)	temporal error that invalidates the entire analysis—it claims the dataset covers "30 days from December 2020 to January 2021" when it actually spans 5 full years (2020-2024) with 1,827 daily observations. While some specific statistics happen to be accurate for isolated portions of the data (46.5mm rain on 2020-01-11, 87.8 km/h gusts on 2020-11-15), the temperature range cited (-11°C to 6.5°C) grossly understates the actual extremes (-27.6°C to 33.7°C). The summary completely fails to capture the multi-year scope, seasonal patterns, annual variations, or any meaningful long-term trends, rendering it fundamentally misleading despite being well-written. This represents a catastrophic failure in understanding the dataset's basic structure.	failure is claiming the data covers only 30 days (December 2020 to January 2021) when it actually spans five years (2020-2024). While it correctly estimates the maximum wind gusts and maximum sunshine duration, it significantly underestimates the actual range of mean temperatures (−19.9 °C to 27.8 °C) and the maximum snowfall (20.65 cm).	metrics: it reports temperature ranges (-11 °C to 6.5 °C), rainfall (46.5 mm), snowfall (11.13 cm), and wind gusts (87.8 km/h) that are all much higher than the actual dataset values (≈ –9.9 °C to 5.0 °C, 9.1 mm rain, 7.28 cm snow, and 63 km/h gusts). While it is fluent and coherent, its informativeness and completeness suffer due to these factual errors and some overgeneralizations (e.g., exaggerating sunshine duration to 14.2 hours when the actual max ≈ 8.8 hours). Overall, it reads well but misrepresents the data substantially.	

Appendix 3: How to run template method code

First, install the dependencies specified in Template/requirements.txt

Then, to generate a template-based summary, navigate to the Template/code directory in the provided codebase and run python summarize.py --input myData.csv --station_name "Toronto Pearson" --output mySummary.txt.

Here, the input option is mandatory and represents the path to the .csv file containing the data. The output option specifies which file to save the summary to, and the station_name option specifies which station the data is for. Note that for the data we have provided, this is always Toronto Pearson.