# Project Report: Forecasting Hotel Ratings and Trump’s Post Frequency

**Author:** Your Name **McGill ID:** Your McGill ID **Date:** May 30, 2025

## 1. Methodology

This project aimed to forecast two distinct metrics for the period of June 2-6, 2025: (1) the average Google review star rating for the Montreal Marriott Château Champlain hotel, and (2) the average daily number of Truth Social posts by Donald J. Trump. A multi-phase methodology involving data acquisition, processing, prompt engineering, ensemble forecasting, and critical evaluation was employed for each target.

**Common Steps for Both Forecasts:**

1. **Environment Setup:** A Python 3.11 virtual environment was established, and necessary libraries (e.g., pandas, openai, anthropic, google-generativeai, python-dotenv, API-specific SDKs) were installed. API keys were managed using a .env file.
2. **Prompt Engineering & Selection:**
   * Multiple prompt variants were drafted for each forecast target: a base prompt, a Chain-of-Thought (CoT) prompt, and a context-aware/scenario-based prompt.
   * These prompts were back-tested against a historical holdout period (May 7, 2025, for hotel ratings; May 18-22, 2025, for Trump posts) using OpenAI’s gpt-3.5-turbo model.
   * The actual metrics for the holdout period were calculated from previously collected data to serve as ground truth.
   * The prompt variant yielding the lowest Mean Absolute Error (MAE) against the ground truth was selected for the main ensemble forecast. For the hotel forecast, hotel\_prompt\_cot.txt was chosen. For the Trump post forecast, trump\_prompt\_context.txt was selected.
3. **Ensemble Forecasting:**
   * The selected prompt for each target was run across three Large Language Models (LLMs): OpenAI’s gpt-4o, Anthropic’s claude-3-opus-20240229, and Google’s gemini-1.5-pro-latest.
   * Each LLM was queried twice, once with a low temperature (0.2) for more deterministic output, and once with a higher temperature (0.7) to encourage more diverse reasoning, resulting in 6 forecasts per target.
   * For the hotel forecast, which involved a prompt referencing external data (hotel\_daily\_metrics.csv), a modified version of the chosen prompt (hotel\_prompt\_cot\_with\_data.txt) was used for the Anthropic and Google models to directly embed the necessary daily metrics data, as these models did not have direct file access in the setup used.
4. **Aggregation & Finalization:**
   * The numerical forecasts extracted from the 6 LLM responses for each target were aggregated by calculating their mean and standard deviation. These formed the final predictions, stored in hotel\_final.json and trump\_final.json respectively.
5. **Critique and Revision:**
   * The final aggregated forecasts were subjected to a critique process. A separate LLM instance (gpt-4o) was prompted with a “critic” persona to evaluate the plausibility, methodology, potential biases, and missing considerations for each forecast.
   * Based on the critique of the hotel forecast, which highlighted a potential outlier (a 3.0 rating from one model run) significantly impacting the mean and standard deviation, the hotel forecast was revised by removing this outlier and recalculating the aggregate statistics. A rationale for this revision was documented.
   * The critique for the Trump post forecast did not lead to a numerical revision, as its main points concerned the inherent limitations of the “no major events” constraint in the chosen prompt, rather than an issue with the ensemble’s adherence to that prompt.

**Specific Data Acquisition & Processing:**

* **Hotel Ratings (Montreal Marriott Château Champlain):**
  + The Google Maps Place ID for the hotel was retrieved.
  + 200 recent Google reviews were fetched using the SerpAPI Google Local Services API, saving raw review data including ratings and dates to hotel\_reviews\_raw.csv.
  + A baseline overall mean rating and review count were calculated (hotel\_baseline.txt).
  + Daily new review counts and mean ratings for the last 30 days of activity were processed into hotel\_daily\_metrics.csv.
* **Trump’s Truth Social Posts:**
  + Truth Social posts by Donald J. Trump from the preceding 60 days were downloaded using an Apify actor (muhammetakkurtt/truth-social-scraper), saving to trump\_posts\_raw.json.
  + The raw JSON data was parsed to count daily posts, resulting in trump\_posts\_daily.csv.
  + A 30-day rolling mean and standard deviation of daily posts were calculated as a baseline (trump\_baseline.txt).
  + A visual plot of daily posts was generated for a sanity check.

All relevant data, scripts, prompts, and outputs were version-controlled using Git.

## 2. Predictions, Rationales & Actionability

The following section details the final ensemble forecasts for both the hotel rating and Trump’s post frequency for the period of June 2-6, 2025.

### 2.1 Hotel Rating: Montreal Marriott Château Champlain

* **Final Forecast (Mean ± Std Dev):** 4.22 ± 0.20 stars
* **Individual Ensemble Forecasts (Original 6):** [4.0, 3.0, 4.4, 4.0, 4.5, 4.2] (Mean: 4.02, Std Dev: 0.49)
* **Individual Ensemble Forecasts (Revised 5, Outlier Removed):** [4.0, 4.4, 4.0, 4.5, 4.2]
* **Based on:** 5 aggregated LLM responses after outlier removal (GPT-4o, Claude 3 Opus, Gemini 1.5 Pro at varied temperatures).
* **Chosen Prompt Used:** hotel\_prompt\_cot.txt (with data embedded as hotel\_prompt\_cot\_with\_data.txt for Anthropic/Google).

**Rationale & Context:** The ensemble forecast for the hotel’s average Google review rating aims to balance the baseline mean rating of 4.12 (from 200 reviews) with recent, albeit sparse, daily review data from April-May 2025. This daily data showed some perfect 5.0 scores on days with review activity, but also a significant dip to a 1.0 mean on one day, indicating potential volatility or isolated incidents. The chosen Chain-of-Thought prompt (hotel\_prompt\_cot.txt) guided the LLMs to consider the baseline, recent trends, seasonality (early June in Montreal being a potentially busy period), typical hotel operational factors, and broader economic trends.

**Objective Outlier Rule Discussion & Revision:** An initial ensemble of 6 forecasts yielded a mean of 4.02 ± 0.49. The critique process (Task T50) highlighted that one forecast of 3.0 (from GPT-4o at temp 0.7) was a potential outlier. To formalize outlier handling, an objective rule could be: “Remove any forecast > 1.5 IQR (Interquartile Range) from the Q1 or Q3.” For the original 6 forecasts [4.0, 3.0, 4.4, 4.0, 4.5, 4.2]: Sorted: [3.0, 4.0, 4.0, 4.2, 4.4, 4.5] Q1 (median of lower half [3.0, 4.0, 4.0]): 4.0 Q3 (median of upper half [4.2, 4.4, 4.5]): 4.4 IQR = Q3 - Q1 = 4.4 - 4.0 = 0.4 Lower Bound for Outlier = Q1 - 1.5 \* IQR = 4.0 - 1.5 \* 0.4 = 4.0 - 0.6 = 3.4 Upper Bound for Outlier = Q3 + 1.5 \* IQR = 4.4 + 1.5 \* 0.4 = 4.4 + 0.6 = 5.0 The forecast of 3.0 is below the lower bound of 3.4, confirming it as an outlier by this rule. The outlier was removed (Task T52, rationale in hotel\_revision\_rationale.txt), resulting in the revised, more tightly clustered forecast of 4.22 ± 0.20. This revision leads to a mean closer to the baseline and a tighter standard deviation, addressing a key concern from the critique.

The critique also noted that the models might underweigh very negative single-day outliers if the overall trend is positive and that external factors like unannounced major city events or specific hotel issues (which LLMs have no direct knowledge of) remain a key uncertainty for a year-ahead forecast.

**Actionability:** \* The forecast of 4.22 ± 0.20 suggests a generally positive outlook, slightly above the historical baseline. \* The small standard deviation (after outlier removal) indicates good agreement among the majority of the ensemble models. \* **Measurable Follow-up:** If the actual average hotel rating for June 2-6, 2025, (once available from Google Reviews) deviates by more than 0.3 stars from the forecasted mean of 4.22 (i.e., below 3.92 or above 4.52), it would warrant a re-evaluation. This could involve: \* Re-running the forecast with an updated review corpus covering the period up to June 2025. \* Investigating if specific significant events (positive or negative) occurred at the hotel or in Montreal during the forecast period that were not anticipated. \* Reviewing the raw reviews for the forecast period to understand qualitative drivers of any deviation. \* The forecast can inform internal expectations but should be used cautiously for external commitments given the inherent uncertainties of long-range prediction and LLM-based methods.

### 2.2 Trump’s Truth Social Post Frequency

* **Final Forecast (Mean ± Std Dev):** 16.2 ± 0.59 posts per day
* **Individual Ensemble Forecasts:** [16.3, 16.3, 16.3, 16.3, 15.0, 17.0]
* **Based on:** 6 aggregated LLM responses (GPT-4o, Claude 3 Opus, Gemini 1.5 Pro at varied temperatures).
* **Chosen Prompt Used:** trump\_prompt\_context.txt.

**Rationale & Context:** The ensemble forecast for Donald J. Trump’s average daily post frequency on Truth Social was generated using the trump\_prompt\_context.txt. This prompt specifically instructed the LLMs to assume a “business-as-usual” week for June 2-6, 2025, with **no major pre-scheduled political events, significant anniversaries, major court dates, or national holidays** that would unusually inflate or deflate posting activity.

The LLMs were provided with baseline statistics from late May 2025: a mean of 16.31 daily posts and a standard deviation of 7.84.

The resulting ensemble forecast (16.2 ± 0.59 posts) is very close to this baseline mean, with a remarkably small standard deviation among the 6 model runs. This suggests the LLMs, when constrained by the “no major events” context, converged strongly on the idea that activity would likely mirror the recent past under such conditions.

The critique of this forecast (Task T51, trump\_critique.txt) highlighted several points: \* The forecast is plausible *given the specific “no major events” constraint*. \* The low standard deviation is expected under such a narrowly defined scenario. \* There is a strong likelihood of **anchoring bias**, with models heavily relying on the provided baseline when other significant drivers are explicitly excluded by the prompt. \* The main limitation is the artificiality of the “no major events” constraint itself. Trump’s posting is known to be highly reactive to minor daily news cycles or spontaneous thoughts, which this prompt intentionally downplayed. Therefore, while the forecast adheres to the prompt, its real-world applicability is contingent on the actual period being devoid of even minor stimuli that typically drive his activity.

No numerical revision was made to this forecast based on the critique, as the critique primarily addressed the inherent limitations of the chosen prompt’s context rather than a flaw in the models’ interpretation of that context.

**Actionability:** \* The forecast of 16.2 ± 0.59 posts per day suggests a continuation of recent baseline activity, assuming a period devoid of major unscheduled catalysts. \* The very low standard deviation indicates strong agreement among the models under the specified “no major events” constraint. \* **Measurable Follow-up:** If the actual average daily post count for June 2-6, 2025, deviates by more than 3 posts from the forecasted mean of 16.2 (i.e., below 13.2 or above 19.2), this would indicate that either the “no major events” assumption was incorrect or other factors significantly influenced posting behavior. A review would involve: \* Analyzing news and events during June 2-6, 2025, to identify any potential drivers of deviation. \* Comparing the forecast period’s activity to a more extended historical dataset if available, to check for broader shifts in posting patterns not captured by the 60-day baseline. \* This forecast is highly conditional on the “no major events” premise. Its utility lies in establishing a baseline expectation for a “quiet” period, against which actual activity (and the impact of real-world events) can be measured.

## 3. Analysis of Limitations & Mitigation Strategies

This forecasting project, while systematic, is subject to several limitations inherent in using Large Language Models (LLMs) for predictive tasks and in the nature of forecasting itself. Efforts were made to mitigate these, particularly concerning potential LLM “hallucinations” or ungrounded outputs.

### 3.1 Limitations

1. **Data Limitations:**
   * *Hotel Reviews:* The recent daily hotel review data (hotel\_daily\_metrics.csv) was sparse, with activity only on a few days within the 30-day window. This makes it difficult to establish a strong recent trend. The overall baseline relied on 200 reviews, which is a reasonable number, but older reviews may not reflect current quality.
   * *Trump Posts:* The 60-day window for Trump’s posts (trump\_posts\_raw.json) provides a recent snapshot but might not capture longer-term cyclical patterns or shifts in communication strategy.
2. **LLM Characteristics & Prompt Sensitivity:**
   * *Black-Box Nature:* LLMs operate as complex, non-transparent models. While their reasoning can be prompted (e.g., via CoT), the exact internal mechanisms leading to a specific forecast number remain opaque, making it hard to debug unexpected deviations beyond a certain point.
   * *Sensitivity to Prompting:* LLM outputs are highly sensitive to the phrasing and structure of prompts. Minor changes can lead to different results. While prompt selection via back-testing aimed to find a robust option, the chosen prompt still shapes the outcome significantly.
   * *Consistency Issues:* As seen with the trump\_prompt\_cot.txt failing during back-testing with gpt-3.5-turbo, or the gpt-4o hotel forecast producing a 3.0 outlier at a higher temperature, LLMs can sometimes produce unexpected or non-compliant outputs even with established prompts.
3. **Artificial Forecasting Contexts:**
   * The “no major events” constraint in the selected Trump forecast prompt (trump\_prompt\_context.txt) creates an artificial scenario. While useful for isolating baseline behavior, real-world post frequency is almost always influenced by an unpredictable mix of major and minor events, and spontaneous reactions.
4. **Inability to Predict Unforeseen Events:**
   * The forecasts inherently cannot account for sudden, unannounced major events (e.g., for the hotel: unexpected closures, major new competitor openings, city-wide emergencies; for Trump: significant personal news, major geopolitical crises) occurring between the forecast generation and the forecast period.
5. **Generalization from Training Data:**
   * LLMs generate forecasts based on patterns learned from their vast training data up to their knowledge cut-off date. While this provides a broad understanding, specific future contexts might diverge from these learned patterns in unpredictable ways.

### 3.2 Mitigation Strategies for Ungrounded Outputs

Several strategies were employed to ground the LLM outputs and mitigate the risk of uncontextualized “hallucinations” or irrelevant forecasts:

1. **Providing Numerical Baselines:** All prompts included key historical data (e.g., mean hotel rating, mean daily Trump posts, and associated standard deviations) to anchor the LLMs in relevant factual context.
2. **Structured & Contextual Prompts:**
   * Chain-of-Thought (CoT) prompts were used to encourage step-by-step reasoning, making the forecast derivation process more transparent and allowing for evaluation of the reasoning path.
   * Context-specific instructions, such as the “no major events” clause for the Trump forecast or embedding daily hotel metrics directly into the hotel prompt for some models, aimed to focus the LLMs on the desired scenario.
3. **Empirical Prompt Selection via Back-testing:** Rather than relying on a single assumed best prompt, multiple prompt variants were drafted and quantitatively evaluated against a historical holdout period using gpt-3.5-turbo. The prompt with the lowest Mean Absolute Error (MAE) was chosen, providing an empirical basis for prompt selection.
4. **Ensemble Forecasting:** Using multiple LLMs (GPT-4o, Claude 3 Opus, Gemini 1.5 Pro) and varied temperatures (0.2 and 0.7) provided a range of outputs. Convergence among these models (as seen in parts of the Trump forecast) can increase confidence, while divergence (as initially seen in the hotel forecast) can highlight uncertainty or potential issues with specific model responses.
5. **Output Format Specification:** Prompts included explicit instructions for the final forecast output (e.g., “Final Forecast: X.X”), which aids in consistent parsing and reduces the chance of narrative-only responses without a clear numerical prediction.
6. **Self-Critique Process:** The final aggregated forecasts were fed to a separate LLM instance (GPT-4o) tasked with a “critic” role. This provided an independent (albeit still LLM-based) check on the plausibility, methodology, and potential biases of the forecasts. This critique directly led to the revision of the hotel forecast by identifying and justifying the removal of an outlier.
7. **Iterative Refinement:** The process involved several iterative steps, such as realizing the need for hotel\_prompt\_cot\_with\_data.txt when models couldn’t access local files, and addressing API key issues, which are part of a practical LLM application workflow.

While these strategies do not eliminate all risks associated with LLM-based forecasting, they provide a framework for producing more grounded, explainable, and critically evaluated predictions.

## 4. Accuracy Evaluation Plan

To formally evaluate the accuracy of the forecasts post-factum, the following plan will be used once actual data for June 2-6, 2025, becomes available:

1. **Data Collection:**
   * **Hotel Rating:** Collect all Google Reviews for the Montreal Marriott Château Champlain posted during or reflecting stays between June 2, 2025, and June 6, 2025. Calculate the actual mean star rating from these reviews.
   * **Trump Posts:** Collect all Truth Social posts by Donald J. Trump for each day from June 2, 2025, to June 6, 2025. Calculate the actual average number of posts per day over this 5-day period.
2. **Error Metric Function:**
   * The primary error metric will be the **Mean Absolute Error (MAE)**, calculated as: MAE = | Actual Value - Forecasted Value |
   * For the hotel rating, the “Value” is the average star rating.
   * For Trump posts, the “Value” is the average number of daily posts.
3. **Denominator for Relative Error (Contextual):**
   * While MAE is the primary metric, for contextual understanding, a relative error can be considered against the **forecasted value** or the **baseline value** used in the prompt. For instance: Relative Error = MAE / Forecasted Value
   * Alternatively, against the Standard Deviation (SD) of the baseline data: Error in terms of Baseline SD = MAE / Baseline\_SD (Baseline SD for hotel was not explicitly used in prompts but was implicitly part of daily metrics; Baseline SD for Trump posts was 7.84).
4. **Acceptance Threshold (Qualitative Alignment with Rubric Principles):**
   * A primary goal is that the **Actual Value should fall within a reasonable range of the Forecasted Value ± Ensemble\_Standard\_Deviation**.
   * **Hotel Rating (Forecast: 4.22 ± 0.20):**
     + Highly Acceptable: Actual rating falls within 4.22 ± 0.20 (i.e., 4.02 to 4.42).
     + Acceptable: Actual rating falls within 4.22 ± 0.40 (i.e., 3.82 to 4.62, or within twice the ensemble SD).
   * **Trump Posts (Forecast: 16.2 ± 0.59):**
     + Highly Acceptable: Actual posts fall within 16.2 ± 0.59 (i.e., 15.61 to 16.79).
     + Acceptable: Actual posts fall within 16.2 ± 1.18 (i.e., 15.02 to 17.38, or within twice the ensemble SD).
   * If the MAE is significantly larger than the ensemble standard deviation, the forecast’s practical utility would be considered low for that specific target. The “SD rule” from a rubric often implies that a forecast is good if the error is within one or two standard deviations (either of the forecast error if multiple forecasts were made over time, or of the ensemble itself as a proxy for confidence).

This plan provides a clear method for quantitatively assessing the forecast accuracy once the actual data is available.

## Appendix A – Prompt Listings

### A.1 Final Hotel Rating Prompt (hotel\_prompt\_cot\_with\_data.txt)

(Used for Anthropic Claude 3 Opus and Google Gemini 1.5 Pro. hotel\_prompt\_cot.txt used for OpenAI GPT-4o was similar but referred to hotel\_daily\_metrics.csv externally.)

You are a forecasting expert. Your task is to predict the average Google review star rating for the Montreal Marriott Château Champlain hotel for the period of June 2, 2025, to June 6, 2025.  
  
Current baseline information:  
- The current overall mean Google review star rating for this hotel is 4.12 (based on 200 reviews).  
  
Recent daily review trends for the last 30 days of observed activity (April 10, 2025 - May 09, 2025) are as follows:  
- April 10-18: No new reviews.  
- April 19: 25 new reviews, mean rating 5.0.  
- April 20: No new reviews.  
- April 21: 25 new reviews, mean rating 5.0.  
- April 22-May 1: No new reviews for most days.  
- May 02: 25 new reviews, mean rating 4.0.  
- May 03: 25 new reviews, mean rating 1.0 (Note: this is a significant dip).  
- May 04-06: No new reviews.  
- May 07: 25 new reviews, mean rating 5.0.  
- May 08: No new reviews.  
- May 09: 25 new reviews, mean rating 5.0.  
- On days with no new reviews, the daily mean rating is 0.0 and count is 0.  
- Consider the general trends from this data (e.g., if ratings are generally stable, increasing, or decreasing; if review volume is high or low, and the impact of outlier days).  
  
Before providing your final numerical forecast, please provide a step-by-step reasoning process. Consider the following:  
1. Current baseline rating and its stability/trend based on the recent daily metrics provided above.  
2. Potential impact of seasonality (early June in Montreal).  
3. Typical hotel operational factors over a 1-year horizon (e.g., renovations, staff changes - assume no specific news unless provided).  
4. Broader economic or travel industry trends that might influence guest experiences or review scores by mid-2025.  
5. Any other factors you deem relevant to this forecast.  
  
After your step-by-step reasoning, conclude with your forecast.  
  
Output your final forecast as a single number between 1.0 and 5.0, rounded to one decimal place, on a new line after your reasoning, prefixed with "Final Forecast:". For example:  
Reasoning step 1...  
Reasoning step 2...  
Final Forecast: 4.3

### A.2 Final Trump Post Frequency Prompt (trump\_prompt\_context.txt)

(Used for all models in the Trump post ensemble forecast.)

You are a forecasting expert specializing in social media trends. Your task is to predict the average daily number of Truth Social posts by Donald J. Trump for the period of June 2, 2025, to June 6, 2025 (a 5-day period).  
  
Current baseline information (based on activity from May 4, 2025, to May 29, 2025, covering 26 days with posts):  
- Mean Daily Posts: 16.31  
- Standard Deviation of Daily Posts: 7.84  
  
\*\*Specific Context for this Forecast:\*\*  
For the purpose of this forecast (June 2-6, 2025), assume a period of \*\*no major pre-scheduled political events, significant anniversaries, major court dates, or national holidays\*\* that would unusually inflate or deflate posting activity. Consider it a typical, "business-as-usual" week in early June, unless your general knowledge strongly indicates specific, regularly occurring minor events for that week that might have a subtle influence.  
  
Before providing your final numerical forecast, please provide a step-by-step reasoning process. Consider the following:  
1. \*\*Baseline Analysis under Assumed Context:\*\* Given the baseline mean/std dev and the assumption of a "normal" week, what is your initial expectation?  
2. \*\*Calendar & Seasonality (No Major Events):\*\* Reiterate the assumption of no major events. Does early June, in a typical year without extraordinary circumstances, have any subtle seasonal patterns for political discourse or Trump's activity?  
3. \*\*Behavioral Consistency in a "Quiet" Period:\*\* How does Trump's communication style manifest during periods without major external news drivers directly involving him?  
4. \*\*Trend Stability:\*\* In the absence of major catalysts, would you expect his posting frequency to remain close to the recent baseline, drift, or show other patterns?  
5. \*\*Synthesis & Uncertainty (Given Context):\*\* Synthesize these points under the "no major events" context and comment on your confidence.  
  
After your step-by-step reasoning, conclude with your forecast.  
  
Output your final forecast as a single number (which can be a non-integer, e.g., 15.5), rounded to one decimal place, on a new line after your reasoning, prefixed with "Final Forecast:". For example:  
Reasoning step 1...  
Reasoning step 2...  
Final Forecast: 14.7

## Appendix B – Back-test Metrics

The following tables summarize the Mean Absolute Error (MAE) for each prompt variant when back-tested against a historical holdout period using OpenAI’s gpt-3.5-turbo model. The chosen prompt for the final ensemble forecast is highlighted.

### B.1 Hotel Rating Prompts Back-test

* **Holdout Period:** May 7, 2025
* **Ground Truth Rating for Holdout Period:** 5.0 (based on 25 reviews on that day)
* **Baseline Rating (prior to holdout):** 4.12 (from 200 reviews)

| Prompt Variant File | LLM Forecast | MAE | Notes |
| --- | --- | --- | --- |
| hotel\_prompt\_base.txt | 4.1 | 0.90 |  |
| **hotel\_prompt\_cot.txt** | **4.2** | **0.80** | **Chosen for ensemble** |
| hotel\_prompt\_scenario.txt | 4.1 | 0.90 |  |

*(Data from hotel\_prompt\_eval.csv)*

### B.2 Trump Post Frequency Prompts Back-test

* **Holdout Period:** May 18-22, 2025 (5 days)
* **Ground Truth Average Daily Posts for Holdout Period:** 13.2
* **Baseline (prior to holdout, from May 4-29):** Mean 16.31 posts/day, Std Dev 7.84

| Prompt Variant File | LLM Forecast | MAE | Notes |
| --- | --- | --- | --- |
| trump\_prompt\_base.txt | 16.3 | 3.1 |  |
| trump\_prompt\_cot.txt | N/A | N/A | Failed to parse forecast during back-test |
| **trump\_prompt\_context.txt** | **16.0** | **2.8** | **Chosen for ensemble** |

*(Data from trump\_prompt\_eval.csv)*

**Report Footer** GitHub Repository: https://github.com/Julian-Oppedisano/insy697\_individual\_assignment