Project Report: Forecasting Hotel Ratings and Trump's Post Frequency

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

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:** [4.0, 4.4, 4.0, 4.5, 4.2]
- **Based on:** 5 aggregated LLM responses (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.

An initial ensemble of 6 forecasts yielded a mean of 4.02 ± 0.49. However, a critique process (Task T50) highlighted that one forecast of 3.0 (from GPT-4o at temp 0.7) was a potential outlier significantly widening the standard deviation and pulling down the mean. This outlier was removed (Task T52, rationale in `hotel_revision_rationale.txt`), resulting in the revised, more tightly clustered forecast above. The revised mean of 4.22 is slightly above the historical baseline, suggesting an expectation of good performance, while the reduced standard deviation (0.20) indicates higher agreement among the remaining models after outlier removal.

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.

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.

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.