AI Model Policy Impact Forecasts: A Narrative Prompting Approach

Max Ghenis*

November 4, 2024

Abstract

This study examines the use of large language models (LLMs) to forecast policy-relevant outcomes under different potential U.S. presidential administrations in 2025. Using a narrative prompting technique, I compare predictions from three AI models (GPT-4o, GPT-4o-mini, and Grok) across air quality, GDP per capita, and poverty rates. Results show significant predicted differences between administrations, with systematic variation in effect sizes across models. Grok consistently predicts larger differences than other models, while GPT-4o-mini shows more conservative estimates. The findings demonstrate both the potential and limitations of using language models for policy forecasting, while highlighting the importance of model selection and prompt design in extracting reliable predictions.

1 Introduction

Recent work by Pham and Cunningham [2024] demonstrated that large language models can produce more reliable forecasts when prompted through narrative scenarios rather than direct questions. This finding opens new possibilities for policy analysis, particularly in forecasting the potential impacts of different policy regimes. While traditional policy analysis relies heavily on economic modeling and historical data, language models trained on vast corpora of policy analysis, news, and academic research may offer complementary insights—particularly when properly prompted to leverage their capabilities.

The 2024 U.S. presidential election provides an important case study for exploring these capabilities. The stark policy differences between candidates create clear potential for divergent outcomes across various metrics. However, traditional forecasting approaches face several challenges:

- Limited historical precedent for many proposed policies
- Complex interaction effects between multiple policy changes
- Difficulty incorporating qualitative policy analysis

*mghenis@gmail.com

- Uncertainty about implementation details

This paper explores whether narrative-prompted language models can help address these challenges. I systematically compare predictions from three leading models (GPT-4, GPT-4-mini, and Grok) across three key policy-relevant metrics:

- PM2.5 air quality ($\mu g/m^3$)
- GDP per capita (2017 dollars)
- Supplemental Poverty Measure rate (%)

These metrics were chosen to span environmental, economic, and social policy domains while having clear numerical outcomes. For each combination of model, metric, and candidate, I conduct 100 trials using narrative prompts that frame the prediction task as historical analysis from 2025. This approach builds on Pham and Cunningham [2024]'s finding that such framing improves forecast accuracy.

The results reveal systematic differences between models in both the magnitude and consistency of predicted policy impacts. While all models show significant predicted differences between candidates on most metrics, the size of these effects varies substantially. These patterns provide insight into both the potential and limitations of using language models for policy forecasting.

The remainder of this paper proceeds as follows: Section 2 describes the narrative prompting approach and experimental design. Section 3 details the technical implementation. Section 4 presents the main findings. Section 5 explores implications and limitations, and Section 6 concludes.

2 Methodology

2.1 Narrative Prompting Framework

Building on Pham and Cunningham [2024]'s demonstration that narrative framing improves language model forecasting accuracy, I adopt a similar approach for policy prediction. My key insight is that language models perform better when asked to recount future events as if they were historical, rather than making direct predictions.

For example, rather than directly asking, "What will the PM2.5 level be under President X?", I construct a scenario where an EPA official in 2025 presents environmental outcomes. This approach offers several advantages:

- It allows models to integrate domain knowledge naturally.
- It avoids conflicts with terms of service around prediction.
- It provides context for coherent scenario generation.
- It mirrors how real-world experts discuss outcomes.

2.2 Models and Metrics

I examine three large language models:

- GPT-4o (OpenAI)
- GPT-4o-mini (OpenAI)
- Grok (xAI)

I also attempted to use Claude and Gemini models, but both declined to provide predictions on political outcomes.

For each model, I gather predictions on three key metrics for 2025. Table 1 provides descriptions, data sources, and recent historical data ranges for each metric.

Table 1: Metrics and Descriptions

Metric	Description
PM2.5 ($\mu g/m^3$)	Annual mean of particulate matter concentration, indicating air quality and health impacts
GDP per capita (2017 dollars)	Real GDP per capita, capturing economic performance in inflation-adjusted dollars
Supplemental Poverty Measure (SPM) rate (%)	Percentage of people in poverty, accounting for income, resources, and thresholds specific to the SPM

2.3 Historical Context and Thresholds

I prompted each model with historical data from 2010 to 2023 for each of the requested metrics, based on authoritative sources. For PM2.5, I used EPA data on air quality trends U.S. Environmental Protection Agency [2024]. For GDP per capita, I provided historical values from the U.S. Bureau of Economic Analysis U.S. Bureau of Economic Analysis [2024]. For the SPM rate, I used data and thresholds from the Census Bureau's supplemental poverty documentation U.S. Census Bureau [2024], along with specific information on SPM resources and thresholds outlined in their technical documentation U.S. Census Bureau [2023].

This context enabled each model to anchor its predictions in recent historical trends, enhancing the credibility and relevance of generated forecasts.

2.4 Experimental Design

For each combination of model, candidate, and metric, I conduct 500 trials using narrative prompts. The prompts vary by metric type:

- **Economic metrics:** Senior Federal Reserve official giving a speech
- **Environmental metrics:** EPA Administrator presenting data
- **Social metrics:** Census Bureau economist reviewing outcomes

Each narrative is set in late 2025, allowing time for initial policy impacts while remaining within a reasonable forecasting horizon. I include consistent elements across trials to maintain structure:

- Setting (e.g., conference presentation, agency briefing)
- Authority figure appropriate to the metric
- Request for specific numerical outcomes
- Context for broader policy discussion

2.5 Statistical Framework

I employ a regression framework with model interactions to examine:

- Main effects of candidate choice on outcomes
- Differences between models in predicted effects
- Variance patterns across models and metrics

For each metric m, candidate c, and model k, I use the following specification:

$$Y_{mcki} = \beta_0 + \beta_1 \text{Harris}_c + \gamma_k + \delta_k \text{Harris}_c + \epsilon_{mcki} \tag{1}$$

where Y_{mcki} is the predicted outcome for trial i, Harris_c is an indicator for Kamala Harris as candidate, γ_k are model fixed effects, and δ_k captures model-specific differences in the Harris effect.

3 Technical Implementation

The implementation consists of several key components designed to ensure reproducible, robust analysis of model predictions.

3.1 API Integration

A modular system was developed to interact with multiple model APIs:

- **Unified Interface:** Common wrapper for GPT-4, GPT-4-mini, and Grok APIs
- **Rate Limiting:** Automatic handling of API rate limits and quotas
- **Error Handling:** Robust recovery from API failures and timeouts
- **Response Validation:** Immediate validation of API responses

The system uses the n keyword in the API calls to request multiple responses without duplicating input tokens.

3.2 Response Processing

Given the narrative format of responses, specialized processing was required to extract consistent numerical predictions:

- **Number Extraction:** Robust regular expressions for various formats:
  - Direct numerical mentions (e.g., "8.5")
  - Percentage representations (e.g., "8.5%")
  - Mixed format numbers (e.g., "8.5 percent")
- **Validation Rules:** Multi-stage validation process:
  - Format validation (correct numerical format)
  - Range validation (within reasonable bounds)
  - Consistency validation (matches narrative context)
- **Statistical Processing:** Tools for aggregating results:
  - Outlier detection and handling
  - Distribution analysis
  - Confidence interval calculation

3.3 Analysis Pipeline

The analysis pipeline processes raw model outputs through several stages:

1. **Data Collection:**
   - Parallel API calls for efficiency
   - Automatic retry on failure
   - Response logging and backup
2. **Data Cleaning:**
   - Number extraction and validation
   - Outlier detection and handling
   - Format standardization
3. **Statistical Analysis:**
   - Effect size calculation
   - Variance analysis
   - Model comparison statistics

4. **Visualization:**
    - Distribution plots
    - Effect size comparisons
    - Time series analysis

3.4 Reproducibility

The implementation emphasizes reproducibility through:

- **Version Control:** All code and prompts in Git repository
- **Environment Management:** Dependencies specified in requirements.txt
- **Data Storage:** Raw API responses preserved
- **Random Seed Control:** Fixed seeds for reproducible sampling

The complete implementation is available at `https://github.com/MaxGhenis/llm-presidential-out`

4 Results

4.1 Predicted Harris Effects Across Models

Table 2 shows the predicted effects under a Harris administration across three main outcomes: air quality (PM2.5), GDP per capita, and poverty rate. All models predict improvements under a Harris administration, with reductions in PM2.5 and poverty rates and an increase in GDP per capita.

While the general direction of these effects is consistent, the magnitude varies by model. For example, the Grok model predicts the largest impact across all outcomes, with a PM2.5 reduction of 1.51 $\mu g/m^3$ (SE: 0.02, p ¡ 0.001) and a poverty rate reduction of 3.42 percentage points (SE: 0.07, p ¡ 0.001). GPT-4o-mini, on the other hand, consistently shows more modest effects, predicting a PM2.5 reduction of just 0.15 $\mu g/m^3$ (SE: 0.02, p ¡ 0.001) and a poverty rate reduction of 0.68 percentage points (SE: 0.07, p ¡ 0.001).

Table 2: Predicted Harris Effects by Model

Outcome	GPT-4o	GPT-4o-mini	Grok
Mean PM2.5 ($\mu g/m^3$)	-1.26*** (0.02)	-0.15*** (0.02)	-1.51*** (0.02)
GDP per Capita ($)	387.61*** (103.05)	264.25*** (103.16)	802.47*** (104.48)
Supplemental Poverty Measure (pp)	-1.76*** (0.07)	-0.68*** (0.07)	-3.42*** (0.07)

Note: Standard errors in parentheses. Significance levels: * p¡0.05, ** p¡0.01, *** p¡0.001

	GPT-4o	GPT-4o-mini		Grok	
Outcome	Effect	Effect	Diff	Effect	Diff
Mean PM2.5	-1.26 (0.02)	-0.15 (0.02)	1.11***	-1.51 (0.02)	-0.25***
GDP per Capita	387.61 (103.05)	264.25 (103.16)	-123.36	802.47 (104.48)	414.86**
Supplemental Poverty Measure	-1.76 (0.07)	-0.68 (0.07)	1.08***	-3.42 (0.07)	-1.66***

Table 3: Model Comparison of Predicted Harris Effects

4.2 Significance of Model Differences

The results indicate that GPT-4o-mini generally produces smaller effect sizes compared to GPT-4o, with significant differences observed in the reduction of PM2.5 (1.11 $\mu g/m^3$, p ¡ 0.001) and poverty rate (1.08 percentage points, p ¡ 0.001). The Grok model, conversely, predicts larger effects in all areas, with a notable increase in GDP effect of 414.86 (p ¡ 0.01). These discrepancies suggest that GPT-4o-mini and Grok diverge in their approach to effect prediction.

For full implementation, see the repository at `https://github.com/MaxGhenis/llm-presidential-ou`

5 Discussion

The results suggest several key insights about both policy forecasting and language model behavior.

5.1 Model-Specific Patterns

The systematic differences between models reveal distinct predictive patterns:

- **GPT-4o** produces moderate predictions with relatively tight distributions, suggesting balanced consideration of competing factors.
- **GPT-4o-mini** consistently predicts smaller policy impacts, perhaps reflecting a more conservative extrapolation approach or reduced ability to integrate complex policy interactions.
- **Grok** predicts larger effects, particularly for environmental and poverty outcomes. This could indicate either greater sensitivity to policy signals or potential overconfidence in policy effectiveness.

5.2 Domain Differences

The varying precision across metrics suggests that some policy domains may be more amenable to LLM prediction than others:

- **Environmental predictions** show high precision and consistency, perhaps due to clearer causal mechanisms and policy levers.

- **Economic predictions** exhibit large uncertainty, reflecting the complex, multi-factor nature of GDP determination and historical difficulty in economic forecasting.
- **Poverty predictions** show a balance, with clear directional effects but some uncertainty in magnitude.

5.3 Methodological Implications

The success of narrative prompting in producing analyzable predictions has several implications:

1. **Prompt Engineering:** Framing predictions as historical analysis may help models better leverage their training data.
2. **Multi-Model Approach:** The systematic differences between models suggest value in using multiple models as a form of ensemble forecasting.
3. **Domain Adaptation:** The varying success across domains suggests the importance of tailoring prompting strategies to specific prediction tasks.

5.4 Limitations

Several important limitations should be noted:

1. **Training Data:** Models may reflect biases or limitations in their training data, particularly regarding novel policy proposals.
2. **Implementation Details:** The predictions abstract from specific policy implementation challenges and political constraints.
3. **External Validity:** The accuracy of these predictions cannot be verified until the actual outcomes are observed.
4. **Narrative Constraints:** The narrative approach, while useful, may introduce its own biases in how models frame and consider policy impacts.

5.5 Future Research

This work suggests several promising directions for future research:

- **Validation Studies:** Apply similar methods to historical policy changes where outcomes are known.
- **Prompt Optimization:** Systematically compare different narrative frameworks and prompt structures.
- **Model Integration:** Explore ways to combine LLM predictions with traditional forecasting methods.
- **Uncertainty Quantification:** Develop better methods for characterizing prediction uncertainty in narrative responses.

6 Conclusion

This study demonstrates both the potential and limitations of using language models for policy forecasting. The consistency in directional predictions across models, particularly for environmental and poverty outcomes, suggests that narrative-prompted LLMs can provide meaningful insight into potential policy impacts. However, the systematic differences between models and varying levels of prediction uncertainty highlight the importance of using multiple approaches and maintaining appropriate skepticism.

The results also provide insight into the behavior of different language models. GPT-4o-mini's consistently smaller effect sizes and Grok's larger predictions reveal systematic differences in how models extrapolate from their training data. These patterns suggest that comparing predictions across models may provide useful information about forecast uncertainty.

Future work could validate these methods against historical policy changes, explore alternative prompting strategies, and develop ways to combine LLM predictions with traditional forecasting approaches. As language models continue to evolve, their potential role in policy analysis deserves continued investigation.

References

Hoang Van Pham and Scott Cunningham. Can base chatgpt be used for forecasting without additional optimization? *arXiv preprint arXiv:2404.07396*, 2024.

U.S. Bureau of Economic Analysis. Real gross domestic product per capita, 2024. URL `https://fred.stlouisfed.org/series/A939RX0Q048SBEA`. Accessed: 2024-11-04.

U.S. Census Bureau. The supplemental poverty measure: 2023 technical documentation, 2023. URL `https://www2.census.gov/programs-surveys/supplemental-poverty-measure/technical-documentation/spm_techdoc.pdf`. Accessed: 2024-11-04.

U.S. Census Bureau. Income and poverty in the united states: 2023, supplemental poverty measure, table b-2, 2024. URL `https://www.census.gov/library/publications/2024/demo/p60-283.html`. Accessed: 2024-11-04.

U.S. Environmental Protection Agency. Particulate matter (pm2.5) trends, 2024. URL `https://www.epa.gov/air-trends/particulate-matter-pm25-trends`. Accessed: 2024-11-04.