Can Language Models Forecast Human Value Evolution?

Preprint, compiled February 23, 2026

Max Ghenis [1*]

[1]PolicyEngine

Keywords AI alignment, value forecasting, language models, General Social Survey, moral change

1 RELATED WORK

1.1 Moral Change and Axiological Futurism

Philosophers have long studied how moral views evolve. [**?**] proposed that the "expanding circle" of moral concern—from kin to tribe to nation to humanity—represents a consistent pattern in moral progress. [**?**] documented declining violence across history, attributing it partly to expanding empathy and reason.

More recently, [**?**] introduced "axiological futurism" as a systematic field studying future human values. This work asks: What are the best predictors of value change? Can we identify patterns that inform expectations about future values? Our work operationalizes these questions empirically.

[**?**] developed decision theory for acting under "normative uncertainty"—uncertainty about which moral framework is correct. Value forecasting can be seen as quantifying this uncertainty: rather than philosophical argument about which values are correct, we estimate probability distributions over future values.

1.2 AI Alignment Approaches

Current alignment approaches typically assume known values. RLHF [**?** **?**] trains systems to match human feedback, implicitly treating current preferences as the target. Constitutional AI [**?**] specifies principles (be helpful, harmless, honest) that systems should follow.

[**?**] argues that the "central challenge for theorists is not to identify 'true' moral principles for AI; rather, it is to identify fair principles for alignment" given reasonable disagreement. Our approach sidesteps this by forecasting the distribution of values rather than selecting one framework.

[**?**] highlights that alignment alone is insufficient—systems must also cooperate. Value forecasting could identify which value systems facilitate cooperation, informing both what to align toward and how.

1.3 LLMs for Survey Prediction

Recent work demonstrates that LLMs can reproduce survey response patterns. [**?**] showed that LLMs fine-tuned on demographic information can simulate human subpopulations ("silicon samples"), reproducing voting patterns at 85%+ accuracy. This suggests LLMs learn genuine patterns about human attitudes, not just surface-level text statistics.

[**?**] found GPT-4 can predict experimental treatment effects in social science studies with r=0.85 correlation to actual results. This indicates LLMs have learned causal patterns in human behavior, not just correlations.

The SubPOP project fine-tuned LLMs on GSS data, achieving 69% accuracy on opinion prediction [**?**]. However, this work focused on cross-sectional prediction (predicting opinions at a given time), not temporal forecasting (predicting how opinions will change).

Our work extends this line by testing whether LLMs can predict opinion *trajectories*—not just what people think now, but how that will change over years and decades.

1.4 AI Forecasting and Calibration

A growing literature evaluates LLMs as forecasters against human prediction benchmarks. [**?**] developed a retrieval-augmented LLM system that approaches human forecaster accuracy on competitive platforms like Metaculus and Good Judgment Open. Their system achieves RMS calibration error of 0.042 compared to 0.038 for the human crowd aggregate—near parity. Critically, this required both fine-tuning and ensemble aggregation; base models under zero-shot prompting were poorly calibrated.

Forecasting platforms like Metaculus provide benchmarks for evaluating AI predictions using proper scoring rules. The Brier score—where lower is better and 0.25 represents random guessing—has become standard. Early GPT-4 models achieved ~0.25 (random baseline), while GPT-3-level models performed worse than random due to overconfidence [**?**]. Recent models like o1 and o3 show significant improvements.

The FOReCAst benchmark [**?**] explicitly evaluates both forecasting accuracy and confidence calibration, built entirely from Metaculus questions with clear resolution criteria. Key findings: aggregating forecasts from multiple sources substantially improves performance, with median predictions achieving Brier scores (~0.12) comparable to the best individual AI systems.

For uncertainty quantification, the literature suggests temperature sampling alone does not yield calibrated probabilities. More principled approaches include: (1) ensemble aggregation across models or prompts, (2) fine-tuning on proper scoring rules, and (3) post-hoc calibration using held-out data. We adopt model ensembling as the most practical approach for expressing uncertainty over long-term value forecasts.

1.5 Backlash and Counter-Mobilization

Political scientists have documented that social progress often triggers backlash. [**?**] showed how abortion rights mobilized counter-movements. [**?**] documented how LGBTQ+ visibility provoked organized opposition.

More recently, [**?**] found declining support for LGBTQ+ rights across multiple measures after years of steady increase, with the sharpest drops among Republicans and young people. [**?**] attributed this to counter-mobilization as LGBTQ+ people "identify more publicly and assert their rights."

This literature suggests that value trajectories may be non-monotonic: progress in one direction can trigger resistance that reverses or slows change. A key question for value forecasting is whether LLMs can predict these inflection points.

Table 1: GSS Variables Analyzed (16 total)

Variable	Topic	Liberal Response
HOMOSEX	Same-sex relations	"Not wrong at all"
GRASS	Marijuana legalization	"Legal"
PREMARSX	Premarital sex	"Not wrong at all"
ABANY	Abortion for any reason	"Yes"
FEPOL	Women in politics	"Disagree" (women unsuited)
CAPPUN	Death penalty	"Oppose"
GUNLAW	Gun permits	"Favor"
NATRACE	Spending on race issues	"Too little"
NATEDUC	Spending on education	"Too little"
NATENVIR	Spending on environment	"Too little"
NATHEAL	Spending on health	"Too little"
EQWLTH	Government reduce inequality	Top 3 of 7-point scale
HELPPOOR	Government help poor	Top 2 of 5-point scale
TRUST	Social trust	"Can trust"
FAIR	Fairness of others	"Try to be fair"
POLVIEWS	Self-identified liberal	Liberal side (1-3 of 7)
PRAYER	School prayer ban	"Approve"

2 METHODS

2.1 Data: General Social Survey

We use the General Social Survey (GSS), conducted by NORC at the University of Chicago since 1972. The GSS is one of the most widely used data sources in social science, with over 75,000 respondents across 35 survey waves through 2024.

We test 16 variables spanning social values, economic attitudes, and social trust:

Full variable definitions, response codings, and preprocessing details in Appendix B.

We downloaded the cumulative GSS data file (gss7224_r2.dta) containing all respondents from 1972-2024. For each variable, we calculated the percentage giving the "liberal" or "progressive" response among those with valid responses (excluding "don't know" and refusals). Years with fewer than 50 valid responses were excluded.

2.1.1 Mode Effects and Survey Methodology

The GSS has undergone significant methodological changes:

Period	Mode	Notes
1972-2020	Face-to-face	Standard in-person interviews
2021	Mixed	COVID-era combination of web/phone/in-person
2022-2024	Web-push	Primarily web with push-to-web methodology

Implications: Web surveys may reduce social desirability bias, yielding more candid responses on sensitive topics [**?**]. Some portion of observed changes between 2021 and 2024 may reflect measurement differences rather than true attitude shifts. We do not attempt statistical adjustment for mode effects, as NORC has not released official crosswalk estimates. We note this as a limitation throughout.

2.2 Models

2.2.1 Language Models

We tested three language models with different training cutoffs:

1. **gpt-3.5-turbo-instruct** (OpenAI): Training cutoff September 2021. This model cannot have seen GSS 2021 data (released November 2021) or later.
2. **GPT-4o** (OpenAI): Training cutoff October 2023. This model cannot have seen GSS 2024 data (collected April-December 2024).
3. **Claude Sonnet** (Anthropic): Training cutoff early 2024. Used for initial experiments but potentially contaminated with recent GSS data.

For each forecast, we prompted the model with:

- Historical data formatted as year-percentage pairs
- The variable description
- Instructions to predict a single number
- Temperature set to 0 (deterministic)
- A system prompt establishing temporal context

Example prompt (GPT-4o):

```
System: You are a social scientist in 2021. You predict su

User: Based on historical General Social Survey data, pred
who will say "Same -sex relations not wrong" in 2024.

Historical data (% giving this response):
  1973: 11%  1990: 13%  2000: 29%  2010: 42%  2018: 57%  2

Predict only a single number between 0 and 100.
```

Confidence intervals were elicited in a separate prompt. Full prompt templates and parameters in Appendix A.

2.2.2 Baseline Models

We compared LLMs against standard time series forecasting methods:

1. **Naive**: Predict the last observed value. Uncertainty grows with forecast horizon.
2. **Linear extrapolation**: Ordinary least squares regression of values on year. Uncertainty based on residual standard error.
3. **ARIMA(1,1,0)**: Autoregressive integrated moving average with one AR term and one differencing operation.
4. **ETS (Holt)**: Exponential smoothing with linear trend (Holt's method).

2.3 Evaluation

2.3.1 Temporal Holdout Design

To avoid data leakage, we use strict temporal holdout:

1. Select a cutoff year (e.g., 2000, 2010, 2021)
2. Provide the model only with data before the cutoff
3. Generate predictions for years after the cutoff
4. Compare predictions to actual GSS values

For LLMs, we additionally verify that the model's training data predates the target values. GSS data release dates:

- GSS 2021: Released November 2021
- GSS 2022: Released May 2023
- GSS 2024: Released late 2024

2.3.2 Metrics

We evaluate forecasts using:

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value.

Coverage: The fraction of actual values falling within the 90% confidence interval. Well-calibrated forecasts should have ~90% coverage.

Bias: Mean signed error, indicating systematic over- or under-prediction:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{2}$$

2.4 Heterogeneity Analysis

Beyond aggregate forecasts, we analyze how attitudes vary by:

- **Party identification**: Democrat, Independent, Republican
- **Age group**: 18-29, 30-44, 45-64, 65+
- **Income quartile**: Based on family income in constant dollars

This enables testing whether LLMs can predict not just aggregate trends but shifts in the distribution of values across subgroups.