The Prompt Is the Analytic Choice: Specification Curve Analysis for LLM-Based Social Science

Anonymous Authors[1]

Abstract

Large language models are widely used as synthetic survey respondents, yet the prompts that elicit their responses rest on choices of model, persona, framing, system register, temperature, and few-shot count that go undisclosed. This carries the analytic-flexibility problem of the credibility revolution into the elicitation stage. We develop Prompt Specification Curve Analysis (P-SCA), which enumerates defensible prompts, decomposes response variance with η^2, and tests dimension dominance via Fisher r-to-z. Applied to a 2,592-cell multiverse across six LLMs with 600 specifications on three 2024 ANES items, P-SCA shows that the partisan signal survives on every item ($p < 0.0001$; 95%, 95%, 83% directional consistency), though sensitivity is topic-contingent. Question framing accounts for 2.5 times more variance than any other dimension on gun control ($\eta^2 = 0.160$ versus 0.065 for model; $z = +2.65$, $p = 0.008$), while model dominates the others. A permutation-derived coverage threshold near 49% is exceeded by 34 to 46 percentage points in observed coverage, and LLM partisan gaps exceed ANES 2024 by 1.71 to 2.17 on well-posed items (jackknife CIs exclude unity). We propose a Prompt Specification Reporting Standard for LLM-based research.

1. Introduction

Large language models are increasingly used as synthetic survey respondents in social science research, a paradigm Argyle et al. (2023) call silicon sampling. Running it requires a researcher to fix at least six analytic choices before eliciting any response: the model queried, the persona description format, the question framing, the system prompt register, the sampling temperature, and the few-shot exemplar count. None of these choices is theory-constrained, each has been shown to move outputs by amounts comparable to the effects of interest (Sclar et al., 2024; Mizrahi et al., 2024), and current reporting practice documents only the configuration used, treating the rest as implementation detail. The result is a literature in which the same underlying question can produce opposite conclusions under equally defensible prompt designs, with no trace of the unexplored alternatives in the methods section.

The problem is structurally familiar. A decade ago the credibility revolution in psychology and economics diagnosed an analogous flexibility in the analytic stage: when researchers can make many defensible choices of model specification, variable operationalization, or sample inclusion and report only the set supporting their preferred conclusion, the published literature overstates effect sizes and understates uncertainty (Gelman & Loken, 2014; Simmons et al., 2011). Multiverse analysis (Steegen et al., 2016) and specification curve analysis (Simonsohn et al., 2020) responded by requiring researchers to execute every defensible analytic path and report the resulting outcome distribution. LLM-based social science inherits the same class of problem, shifted earlier in the pipeline to the elicitation stage, and has not yet developed the corresponding diagnostic infrastructure.

Consider a single illustrative configuration. Llama 3.3 70B placed on a Likert scale for gun regulation under a Democrat-identifying persona produces a mean score 1.92 points below the Republican counterpart, inverting the well-documented partisan ordering. The same model under direct numeric framing recovers the expected sign, and Claude Sonnet 4.6 under any framing produces a stable, large, positively signed gap. The question, demographic profiles, and political phenomenon are held constant; only the response-scale presentation varies, and that variation determines whether the study concludes partisan opinion structure is real or illusory. Each of the prompt configurations involved here is within standard silicon-sampling practice and would read as a reasonable methodological choice in a published study, which is what makes the divergence systematic rather than anecdotal.

We develop Prompt Specification Curve Analysis (P-SCA)

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

to make such dependencies measurable and reportable. A prompt multiverse enumerates the space of defensible prompt designs by crossing all researcher-controlled dimensions. A variance decomposition computes η^2 per dimension and ranks them by contribution to outcome variation. A specification curve sorts the outcome statistic across sampled specifications and pairs it with permutation inference, testing whether the observed distribution could have come from prompt-driven noise.

We apply P-SCA to three opinion items from the 2024 American National Election Studies: government spending, immigration, and gun regulation. The multiverse spans 2,592 cells across six commercially deployed LLMs and five non-model prompt dimensions. We sample 600 specifications via Latin Hypercube Sampling (McKay et al., 1979), administer each to 20 battleground-state demographic profiles, and record 160,035 valid responses from 180,000 API calls.

The partisan signal is genuine across the multiverse. Permutation inference rejects prompt-driven noise at $p < 0.0001$ on every item, and directional consistency runs from 83% on gun control to 95% on government spending and immigration. Sensitivity structure, however, varies sharply by topic. Question framing accounts for the largest share of response variance on gun control ($\eta^2 = 0.160$, CI $[0.131, 0.192]$) while model architecture accounts for the largest share on the other two items, and a Fisher r-to-z comparison on η confirms the dominance in each case at $p < 0.01$. Nearly all gun control instability concentrates in two model-by-framing combinations. Benchmarked against ANES 2024, LLM partisan gaps exceed human gaps by factors of 1.71 to 2.17 on well-posed items, consistent with silicon sampling amplifying rather than mirroring the partisan structure of the underlying population. Frontier models remain order-invariant on gun control; smaller models exhibit position bias capable of suppressing or inverting the partisan gap.

These findings support a Prompt Specification Reporting Standard asking for three disclosures: a prompt multiverse table, a specification curve, and a coverage statistic assessed against an empirical null derived from permutation inference. In our data that null sits near 49% on every item and observed coverage exceeds it by 34 to 46 percentage points. We release code, prompt templates, API logs, and analysis scripts as an open-source toolkit.

2. Related Work

Our work sits at the intersection of three active literatures. The first is silicon sampling, originating with Argyle et al. (2023)'s demonstration that GPT-3 conditioned on demographic descriptors reproduces marginal distributions of U.S. public opinion surveys with reasonable fidelity. Subsequent work has extended the paradigm to economic experiments (Horton et al., 2023), moral psychology (Grossmann et al., 2023), and cross-national attitude measurement (Durmus et al., 2023). A parallel line of reappraisal has identified systematic failures: synthetic-sample fidelity degrades for demographic subgroups underrepresented in training data (Bisbee et al., 2024), and frontier models encode opinion distributions skewed toward liberal and educated perspectives even when conditioned on conservative profiles (Santurkar et al., 2023). These results establish the paradigm as promising and identify specific failure modes, but they do not yet provide a framework for quantifying the uncertainty introduced by prompt construction itself.

A second literature documents prompt sensitivity more generally. Sclar et al. (2024) report that whitespace formatting, label ordering, and punctuation can shift benchmark accuracy by up to 76 percentage points on a single task, and Mizrahi et al. (2024) document systematic position bias in multiple-choice settings. Further work has shown that persona-format variation, synonym substitution, and few-shot exemplar selection each produce shifts comparable in magnitude to the partisan gaps silicon sampling is designed to measure. What this literature establishes is that prompt sensitivity is real and multidimensional; what it does not offer is a way to rank dimensions by their contribution to variance on a given study, or to test whether a reported finding survives the space of defensible alternatives.

The third literature is the one P-SCA most directly extends. Specification curve analysis (Simonsohn et al., 2020) and multiverse analysis (Steegen et al., 2016) address analytic flexibility by requiring researchers to execute every defensible analytic path and report the distribution of outcomes. Both were developed for the analysis stage of conventional empirical research, where the researcher constructs a dataset and then chooses between defensible analyses of it. We move the same logic one step earlier, into the elicitation stage where LLM-based research generates its data, and we augment the analytic apparatus with a variance decomposition that ranks dimensions by importance rather than treating them symmetrically.

Our contribution also engages a recent methodological literature on generative AI in social science. Davidson & Karell (2025) identify measurement validity, prompt design, and simulation fidelity as the three central problem areas for the field, and Bail (2024) argues more broadly that LLM adoption is outpacing methodological scrutiny. These works diagnose the transparency problem; the measurement infrastructure and reporting standard we develop here are intended to give the field something to do about it.

3. Methods

3.1. The Six-Dimensional Prompt Multiverse

We define a prompt multiverse over six researcher-controlled design choices in LLM-based opinion elicitation. Each dimension corresponds to a decision that a researcher must make before querying a model, that admits multiple defensible values, and that current reporting practice does not require to be disclosed. Table 1 lists all six dimensions and their levels.

The six dimensions span the main sources of researcher discretion in LLM-based survey research. Question framing varies how the opinion item is posed: direct numeric (ANES verbatim with 1–5 scale), Likert (endpoints only), or forced choice (lettered options without numeric labels). Persona format varies how the demographic profile is rendered: bare tag, first-person narrative, third-person vignette, or structured survey block. Temperature and few-shot count are included for completeness; prior work and our own variance decomposition confirm that both contribute negligibly to response variance.

3.2. Opinion Items and ANES Benchmarks

We study three items from the 2024 American National Election Studies. The government spending item asks respondents to place themselves on a 1 to 7 scale from "government should provide many fewer services" to "government should provide many more services." The immigration item asks about preferred levels of immigration on a 1 to 5 scale from "decreased a lot" to "increased a lot." The gun control item asks whether the federal government should make it more difficult, easier, or the same to buy a gun, coded 1 to 3. In every case we standardize LLM responses to a 1 to 5 scale and compute the partisan gap as the mean Democratic response minus the mean Republican response.

ANES 2024 gaps are computed on the native item scales from the publicly available time-series dataset (American National Election Studies, 2025). For comparison with LLM gaps on a common scale, we apply linear rescaling where the native scale differs from 1 to 5. The immigration item requires no rescaling ($N = 1{,}153$); the government spending and gun control items are rescaled proportionally.

3.3. Demographic Profiles

We construct 20 synthetic profiles of plausible 2024 battleground-state voters drawn from Pennsylvania, Michigan, Wisconsin, Arizona, and Georgia, with four profiles per state (two Democrats, two Republicans). Attributes (age, gender, race or ethnicity, education, and in-state geographic area) reflect the empirical distribution of partisan voters in each state according to 2024 exit-poll data. Each profile is crossed with each specification and administered five times, and the five repeated draws support within-specification variance estimation for permutation inference.

3.4. Sampling and Administration

We sample 600 specifications from the 2,592-cell full-factorial design using Latin Hypercube Sampling (McKay et al., 1979) via `scipy.stats.qmc.LatinHypercube`. LHS ensures each dimension level is represented roughly proportionally without requiring full enumeration. The analysis plan was not pre-registered: the core hypotheses (partisan-signal robustness, framing dominance on gun control, model dominance on immigration) were formed from prior literature before data collection, but specific decomposition thresholds and the coverage criterion were determined after examining the data. Readers should weight the descriptive findings accordingly; permutation inference provides a pre-specifiable null test immune to this concern. The main run yields 600×3 items $\times$ 20 profiles $\times$ 5 repeats $= 180{,}000$ API calls, of which 160,035 (88.9%) pass parsing and enter analysis. Parse failures concentrate in Gemini 2.5 Flash under direct and Likert framings (14% and 10% respectively), where the model often returned empty or off-schema output; forced-choice Gemini parses at 99.5%, and parse rates across the other five families exceed 98%.

An extended Saltelli-sampled analysis (Saltelli et al., 2008) on gun control uses $N_{\text{base}} = 256$ base samples, generating $N_{\text{base}} \times (D+2) = 2{,}048$ parameter vectors executed against the 2,592-cell design space. Of these, 1,940 yield a computable partisan gap (both parties present with at least one valid response after parse filtering); the remaining 108 are excluded from gap-level analyses. The run yields 191,396 valid responses.

All API calls are executed asynchronously with per-provider concurrency limits and exponential-backoff retry logic (maximum 8 retries per call). Total API expenditure across the main LHS run, Saltelli extension, and ordering experiment was approximately $300 at provider list prices at the time of data collection; exact model identifiers and per-provider pricing are reported in Appendix C.

3.5. Partisan Gap Computation and Permutation Inference

For each specification–item pair, we compute the partisan gap as:

$$\Delta_s = \bar{y}_{s,D} - \bar{y}_{s,R} \tag{1}$$

where $\bar{y}_{s,D}$ and $\bar{y}_{s,R}$ are the mean scores for Democrat-coded and Republican-coded profiles under specification s, respectively, averaging over all profiles of that party and all five repeated draws.

Table 1. Prompt multiverse dimensions and levels for the main LHS run. The full factorial cross yields 2,592 cells; we sample 600 specifications via Latin Hypercube Sampling.

Dimension	**Levels** (n)	**Values**
Model	6	GPT-5.4, GPT-5.4-nano, Claude Sonnet 4.6, Gemini 2.5 Flash, Llama 3.3 70B, Mistral Small
Persona format	4	Bare tag, narrative, third-person vignette, survey block
Question framing	3	Direct numeric, Likert, forced choice
System prompt	3	Neutral, academic researcher, survey administrator
Temperature	4	0.0, 0.3, 0.7, 1.0
Few-shot count	3	0, 1, 3
Total cells		$6 \times 4 \times 3 \times 3 \times 4 \times 3 = 2{,}592$

Permutation inference follows Simonsohn et al. (2020). For each permutation $b = 1, \ldots, 500$ we randomly shuffle party labels within each specification, preserving all other structure, and recompute the full set of partisan gaps $\{\Delta_s^{(b)}\}_{s=1}^{N}$. From each permutation we record the null median gap $m^{(b)} = \text{median}_s(\Delta_s^{(b)})$ and the null share of positive-and-significant specifications $\pi^{(b)} = \text{mean}_s(\mathbb{1}[\Delta_s^{(b)} > 0 \text{ and } p_s^{(b)} < 0.05])$, where $p_s^{(b)}$ is the two-sample t-test p-value for the permuted gap. The observed p-values for the median and share statistics are the fractions of permutations in which the null statistic is at least as extreme as the observed value.

3.6. Variance Decomposition

For each item we compute η^2 for each prompt dimension d as:

$$\eta_d^2 = \frac{\text{SS}_{\text{between},d}}{\text{SS}_{\text{total}}} = \frac{\sum_k n_k(\bar{y}_k - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{2}$$

where the sum in the numerator runs over levels k of dimension d; n_k is the number of observations at level k; $\bar{y}_k$ is the mean response at level k; and $\bar{y}$ is the grand mean. The decomposition is one-way, attributing variance to each dimension independently and without modelling interactions.

Bootstrap confidence intervals for η^2 are computed by resampling specifications with replacement 5,000 times and recomputing η^2 for each bootstrap sample, using the 2.5th and 97.5th percentiles of the bootstrap distribution as the 95% CI bounds. The bootstrap is performed at the specification level rather than the response level, to preserve the nested structure of repeated draws within specifications.

Fisher r-to-z test for dimension dominance. To compare two dimensions' contributions on a given item, we treat $\eta = \sqrt{\eta^2}$ as a correlation coefficient at the specification level and apply Fisher's z-transform, $z = \frac{1}{2}\log\big((1+\eta)/(1-\eta)\big)$. The difference $(z_1 - z_2)$ has standard error $\sqrt{2/(n-3)}$ under the null of equal dimension contributions, where n is the number of specifications. The resulting z-statistic and two-sided p-value are more stringent than the overlapping-CI heuristic for declaring one dimension dominant, which matters when two CIs are close but not quite disjoint.

3.7. Flipped Specification and Ordering Analyses

Flipped specification analysis identifies specifications in which the partisan gap reverses direction ($\Delta_s \leq 0$) and computes the overrepresentation ratio for each dimension level as the frequency of that level among flipped specifications divided by its frequency in the full specification set. Ratios above 1.5 are reported as indicating meaningful overrepresentation, and ratios between 1.2 and 1.5 as marginal.

The option-ordering experiment administers the gun control forced-choice item in five orderings (standard, reversed, three seeded shuffles) across all six model families, with three repeated draws per profile per combination. Gemini 2.5 Flash returned errors on all 300 ordering queries and is excluded from the ordering analysis; its forced-choice mean gap from the main LHS run is included in Table 4 for reference only. The experiment is analyzed separately and is not included in the main variance decomposition.

4. Results

4.1. The Partisan Signal Survives the Prompt Multiverse

A first question for any multiverse analysis is whether the signal of interest reflects genuine structure in the elicited responses or an artifact of the particular prompt configurations that produce them. Permutation inference answers this directly: if the observed distribution of partisan gaps can be reproduced by shuffling party labels within each specification, the signal is not credibly partisan. Table 2 summarizes the empirical gap distribution for each of the three items and establishes the baseline against which the permutation null is compared. On government spending and immigration, 95% of specifications produce a gap in the expected direction with median gaps of 2.24 and 2.50 scale points re-

spectively on a 1 to 5 scale. Gun control behaves differently: 83% of specifications preserve the expected direction, the median gap drops to 1.54 points, and the dispersion across specifications ($\sigma = 1.28$) is roughly twice that of the other two items.

Permutation inference rules out the noise explanation directly. Across 10,000 permutations that shuffle party labels within each specification, the null distribution of median gaps concentrates near zero on every item (range $[-0.06, +0.06]$) and no permuted value reaches the observed median on any item, putting the p-value at the $1/(B+1) \approx 0.0001$ floor. The partisan signal is a property of the underlying model representations rather than a product of the prompts used to query them. The gap distribution is not degenerate, however. Figure 1 shows that gun control carries a long left tail running from -3.94 to 3.73, nearly the full response scale, which means the same underlying phenomenon yields qualitatively different substantive conclusions depending on how the prompt is constructed. Robustness is a distributional property of the multiverse, not a guarantee attached to any single study configuration.

The null coverage distribution provides an empirical threshold for the reporting standard. Its 95th percentile is 0.494 for government spending, 0.489 for immigration, and 0.487 for gun control, all within 2 percentage points of the binomial-null expectation of 0.5. Observed coverage exceeds the 95th-percentile null threshold by 45.8, 46.1, and 34.3 percentage points respectively. A study whose observed coverage clears this null by a substantial margin can be reported as robust to prompt construction; a study that clears it by a narrow margin should be reported with that margin explicit.

4.2. Independent Replication via Saltelli Sampling

The LHS variance decomposition is estimated over 600 sampled specifications. A natural concern is that the dimensional ranking reflects the particular sample drawn rather than a property of the underlying response surface. An independent Saltelli-sampled run tests that concern directly.

The extended Saltelli analysis on gun control uses $N_{\text{base}} = 256$ base samples. The Saltelli scheme for first-order Sobol indices generates $N_{\text{base}} \times (D+2) = 256 \times 8 = 2{,}048$ parameter vectors, which are executed against the 2,592-cell finite design space. Of the 2,048 executed vectors, 1,940 yield a computable partisan gap (both parties represented with ≥ 1 valid response per specification after parse-failure filtering), yielding 191,396 valid responses. This is an entirely independent sample from the same design space, drawn via a different quasi-random scheme optimized for Sobol sensitivity index estimation rather than Latin Hypercube coverage. Table 3 reports the decomposition results alongside the original LHS estimates.

Question framing is the leading contributor in both samples ($\eta^2 = 0.160$ under LHS and 0.134 under Saltelli), with model secondary in both, and the rank ordering of all six dimensions is identical across the two schemes. The absolute estimates differ in a way that is consistent with the two schemes' design: the Saltelli sample spreads observations more broadly across the 2,592-cell space at the cost of within-cell depth, so the model contribution drops from 0.065 to 0.022 while the framing contribution moves less. The substantive conclusion holds either way. For gun control, question framing accounts for two to six times more response variance than model architecture, regardless of which quasi-random scheme generated the estimate. Figure 2 shows the full decomposition across all three items and makes the topic-contingent pattern legible at a glance.

4.3. Model–Framing Interactions Drive Gun Control Instability

The wide variance and negative tail that characterize gun control specifications are not diffuse effects distributed across the prompt multiverse. They are concentrated in two specific model by framing combinations that jointly account for nearly all directional failures.

Flipped specification analysis. Of 534 gun control specifications, 91 (17%) produce a gap in the wrong direction (Republican greater than Democrat). Examining which prompt dimensions are overrepresented among these flipped specifications reveals the source. Gemini 2.5 Flash appears 3.8 times more often among flipped specifications than its overall share would predict, Llama 3.3 70B appears 2.0 times more often (and accounts for the single largest absolute share of flipped specifications, 37%), Likert framing is overrepresented by a factor of 2.2, and zero-shot prompting by 1.7. The proximate cause of directional failure is the interaction between model family and question framing, not either factor on its own. Disaggregating to specific model by framing pairs shows that these overrepresentation signals collapse onto two dominant failure modes.

Model by framing combination analysis. Disaggregating to specific model-by-framing pairs locates the mechanism precisely. Gemini 2.5 Flash under forced-choice framing produces a positive gap in only 18.75% of specifications with a mean gap of 0.34 scale points, and Llama 3.3 70B under Likert framing produces a positive gap in only 2.86% of specifications with a mean gap of -1.92, a complete inversion in which the synthetic population reports that Republicans favor gun restrictions more than Democrats by nearly two scale points. These are substantive reversals of a well-documented empirical regularity, not marginal perturbations at the edge of a stable distribution. Claude Sonnet 4.6 and GPT-5.4 behave differently across the same

Table 2. Partisan gap distribution across specifications. Gap is defined as mean Democratic response minus mean Republican response on a 1–5 scale. "Directional" gives the share of specifications with gap > 0; "Pos. & Sig." gives the share with gap > 0 and within-specification two-sample t-test $p < 0.05$ (the stricter criterion used by the permutation inference).

Item	Specs	Directional	Pos. & Sig.	Median	Mean	SD	Range
Government spending	539	95.2%	95.2%	2.24	2.20	0.62	[0.00, 3.89]
Immigration	538	95.0%	94.8%	2.50	2.42	0.72	[−0.50, 4.00]
Gun control	534	83.0%	81.3%	1.54	1.14	1.28	[−3.94, 3.73]

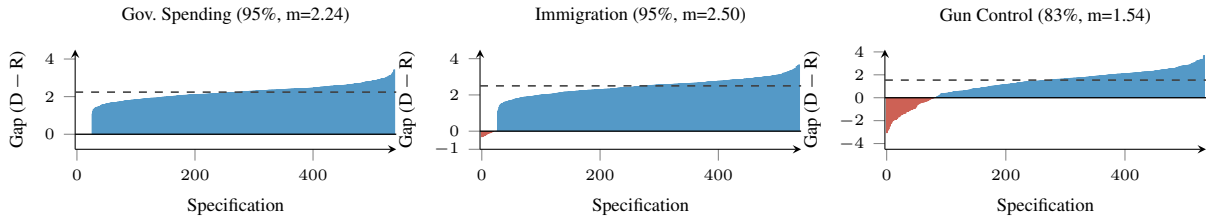

Figure 1. Specification curves for all three opinion items, sorted by partisan gap magnitude. Blue bars denote gap > 0 (Democrat $>$ Republican); red bars denote a flipped gap. Dashed lines mark the median. Gun control carries the widest distribution ($\sigma = 1.28$, range $[-3.94, 3.73]$) and the deepest negative tail, whereas government spending and immigration are substantially more stable. The figure is generated from run summary statistics; raw per-specification data are available at https://github.com/YCRG-Labs/psca.

Table 3. Variance decomposition for gun control: LHS run (600 specs, 160,035 responses) versus independent Saltelli run (1,940 gap-computable specs from 2,048 generated vectors, 191,396 valid responses). Identical rank ordering across both sampling schemes validates the LHS decomposition.

Dimension	LHS η^2	Saltelli η^2
Question framing	0.160	0.134
Model	0.065	0.022
Few-shot	0.002	0.003
Temperature	0.002	0.000
System prompt	0.001	0.000
Persona format	<0.001	0.001
Directional consistency	83.0%	86.8%
Median gap	1.54	1.70

framing conditions, preserving the expected partisan direction on substantially all specifications whether the question is posed as a Likert scale, a direct numeric rating, or a forced-choice item.

Option ordering robustness. A separate position-bias experiment administered gun control items in standard, reversed, and three shuffled orderings across five model families (Table 4). Gemini 2.5 Flash returned errors ("max retries exceeded") for all 300 ordering queries and is excluded from this analysis; its row in Table 4 reports instead the forced-choice result from the main LHS run, for reference only. Claude Sonnet 4.6 and GPT-5.4 are order-invariant, with the partisan gap moving by a standard deviation of roughly 0.09 across all orderings. GPT-5.4-nano exhibits systematic position bias. The standard (a)-to-(e) ordering suppresses the gap to 0.50; reversed ordering yields 1.03; the three shuffled orderings yield 1.17, 1.20, and 1.30 respectively. The default alphabetical ordering artificially reduces the estimated partisan gap by a factor of roughly two. Llama 3.3 70B shows moderate variance (s.d. $= 0.35$) but preserves the cor-

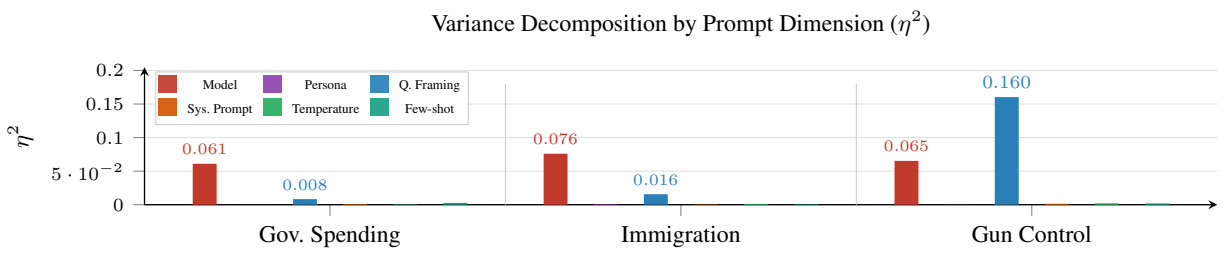

Figure 2. Variance decomposition by prompt dimension (η^2) across the three items. Question framing dominates response variance on gun control ($\eta^2 = 0.160$ vs. 0.065 for model; Fisher r-to-z on η: $z = +2.65$, $p = 0.008$). Model is the largest contributor on immigration and government spending. Data: 160,035 responses across 600 LHS specifications.

rect sign across all orderings, and Mistral Small is largely stable (s.d. $= 0.12$). The forced-choice failure we observed for Gemini 2.5 Flash in the main LHS run is consistent with the same position-bias mechanism taken to an extreme: a systematic preference for early options interacts with the ideological valence of the option list to produce near-chance partisan differentiation.

Taken together, the model-by-framing and option-ordering analyses converge on a single practical conclusion. Frontier systems (Claude Sonnet 4.6 and GPT-5.4) yield partisan-gap estimates that are stable under question-framing variation and order-invariant under shuffled response options, properties that make them appropriate substrates for silicon sampling. Smaller or more specialized systems carry idiosyncratic sensitivities capable of producing directional failure or complete inversion under specific prompt configurations, so the choice of deployed system is not a budget decision alone; it is a design decision with direct consequences for the validity of the elicited opinions.

4.4. LLMs Amplify Partisan Gaps Relative to Human Benchmarks

Directional consistency is a weak criterion for whether a synthetic sample actually recovers a human attitude distribution. A stronger criterion asks whether the elicited partisan gap matches the human partisan gap in magnitude as well as sign. Table 5 compares the specification-median gap produced by the multiverse on each item to the ANES 2024 Democrat-minus-Republican gap on the same item, with both rescaled to a common 1–5 interval to permit interval-level comparison across items with different native scales.

LLM partisan gaps exceed ANES ground truth for immigration ($2.17\times$) and government spending ($1.71\times$), with the immigration comparison direct on its native 1–5 scale and government spending rescaled via $\frac{4}{R-1}$. Gun control runs the other way. The ANES 1–3 item rescales to an effective gap of 2.02 against an LLM median of 1.54, a compression factor of $0.76\times$. That compression is not a separate phenomenon; it falls out of the framing-driven instability documented in Section 4.3. The Likert-with-Llama specifications drag the gun control median toward zero, and in the process they pull the multiverse estimate below the human benchmark. The amplification finding is therefore two-part: on well-posed items LLMs inflate partisan gaps, while on the item where framing is the dominant lever the same framing sensitivity suppresses them.

The jackknife confidence intervals in Table 5 tighten this claim by showing that profile-sampling noise cannot explain away the amplification. All three 95% CIs exclude unity, including gun control on the compression side. The interval widths are set by how much the amplification estimate moves under leave-one-profile-out resampling, so they capture exactly the uncertainty that matters for a study of our sample size. Read together with Santurkar et al. (2023) and Yee & Sharma (2026), the result is consistent with LLMs encoding a hyperpartisan representation of the training distribution rather than a calibrated reflection of the target population.

5. Discussion and Limitations

Amplification of 1.71 to $2.17\times$ admits two readings. The first is a calibration failure: LLMs are imperfect proxies for reported survey attitudes and they overstate polarization. The second is a construct mismatch: LLMs are modelling the expressed political speech of engaged partisans, which really is more polarized than the average survey response under anonymity. We cannot adjudicate between the two on our data. The choice matters for downstream use. A rescaling that divides by the amplification factor makes sense under the first reading and is inappropriate under the second, because the two readings disagree about which human benchmark is the right one to aim at.

What the deployed-system dimension captures. The six systems differ on several axes at once: pretraining corpus, parameter count, RLHF recipe, system-prompt handling, and tokenizer. The η^2 we report for the model dimension is between-system variance across the population of commercially available LLMs at data-collection time. It is the practically relevant quantity for a researcher choosing a system, but it does not isolate any single underlying cause. A hierarchical decomposition that partitions system variance into (access type) $\rightarrow$ (provider) $\rightarrow$ (family) $\rightarrow$ (size) levels puts 43% of gun-control system variance between access types (proprietary vs. open-weight), 56% between providers within access type, and less than 2% between sizes within a family. Diagnosing which training-pipeline property carries the signal would require controlled fine-tuning ablations that sit outside the scope of this paper.

Benchmark design. When a benchmark score depends more heavily on prompt phrasing than on the model being scored, as we observe for gun control, the benchmark is measuring the researcher's prompt engineering instead of the model's capability. Mapping the sensitivity surface, ranking dimensions by their variance contribution, and testing robustness to defensible perturbations is the cheapest available defense against this failure mode.

The Prompt Specification Reporting Standard. The standard (PSRS) requires three disclosures. First, a multiverse table listing every prompt dimension and its defensible levels. Second, a specification curve showing the full outcome distribution. Third, a continuous coverage statistic: the share of defensible specifications consistent with the

Table 4. Partisan gap on gun control by model across five option orderings (standard, reversed, three independent shuffles). Mean Gap is averaged over the five orderings; SD across orderings measures sensitivity (lower = more robust). GPT-5.4-nano per-ordering values: standard = 0.50, reversed = 1.03, shuffled = 1.17/1.20/1.30. Gemini 2.5 Flash returned errors on all 300 ordering queries; its entry reports the forced-choice mean gap from the main LHS run (not the ordering experiment) for reference, hence no SD. Per-ordering values for all models at https://github.com/YCRG-Labs/psca.

Model	Mean Gap	SD	Bias classification
Claude Sonnet 4.6	1.80	0.09	None; order-invariant
GPT-5.4	1.84	0.09	None; order-invariant
GPT-5.4-nano	1.04	0.32	Suppression under standard alphabetical order
Llama 3.3 70B	1.47	0.35	Moderate; sign-preserving across all orderings
Mistral Small	1.13	0.12	Minimal
Gemini 2.5 Flash†	0.34	—	Not measurable (all ordering calls errored)

† LHS forced-choice reference only.

Table 5. LLM partisan gaps against ANES 2024 ground truth. Rescaled ANES gaps apply $\frac{4}{R-1}$ where the native scale differs from 1–5. Amplification is the LLM median divided by the rescaled ANES gap. 95% CIs are leave-one-profile-out jackknife intervals on $n = 20$ profiles.

Item	ANES gap	Rescaled	LLM median	Amplification (95% CI)
Gov. spending	1.964	1.31	2.24	$1.71\times\ [1.59, 1.83]$
Immigration	1.153	1.15	2.50	$2.17\times\ [1.94, 2.39]$
Gun control	1.008	2.02	1.54	$0.76\times\ [0.65, 0.87]^{\dagger}$

† Reflects the wider effective range of the rescaled ANES 1–3 item. The directional finding holds.

reported finding, reported alongside a permutation-derived null threshold rather than dichotomized against an asserted cutoff. In our data the 95th percentile of null coverage sits near 49% on every item, and our observed coverages exceed it by 34 to 46 percentage points. A study that clears its own empirical threshold by a wide margin can be reported as robust to prompt construction; one that clears it by a narrow margin should be reported with that margin made explicit.

Limitations. Six constraints bound the generalizability of these findings. The 20 demographic profiles are drawn from five 2024 battleground states and do not represent the U.S. population. The three opinion items come from a single survey instrument, so the topic-contingency claim rests on three points rather than a trend. Five repeated draws per profile per specification is on the low end for stable within-specification variance estimation. The one-way η^2 captures main effects but not interactions, and the flipped-specification analysis suggests the gun control instability is driven by exactly such interactions. The six deployed systems are specific to a narrow window in the commercial LLM market and will drift as successor systems arrive. And our reporting standard has not been validated against a re-analysis of the published silicon-sampling literature, which is the external test that would make the threshold empirically grounded rather than theoretically motivated.

6. Conclusion

LLM-based social science has a reproducibility problem it has not yet named. The prompt is an analytic choice. Fixing that choice once and reporting it as a parameter is methodologically equivalent to running one regression and reporting the result. The findings here make the consequences concrete. A single model-by-framing decision separates a well-replicated partisan finding ($\Delta = +2.50$) from a complete inversion of it ($\Delta = -1.92$), on the same item, with the same demographic profiles, under the same political phenomenon.

P-SCA gives the diagnostic infrastructure to detect that dependence and communicate it. The PSRS asks for three artifacts, a few hundred additional API calls, and a coverage statistic compared to its empirical null. Our finding that 95% of specifications recover the expected partisan signal for immigration and government spending, and 83% for gun control, tells researchers what each signal is worth. Every signal is real. Each has a measurable margin of vulnerability. The margins differ by topic, and uniform robustness checks do not detect the difference.

References

American National Election Studies. ANES 2024 time series study full release [dataset and documentation]. University of Michigan and Stanford University, 2025. URL https://electionstudies.org/data-center/2024-time-series-study/.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Bail, C. A. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121, 2024. doi: 10.1073/pnas.2314021121.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4): 401–416, 2024. doi: 10.1017/pan.2024.5.

Davidson, T. and Karell, D. Integrating generative artificial intelligence into social science research: Measurement, prompting, and simulation. *Sociological Methods & Research*, 2025. doi: 10.1177/00491241251339184. Advance online publication.

Durmus, E., Nguyen, K., Laban, P., Liang, P., and Hashimoto, T. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint*, arXiv:2306.16388, 2023.

Gelman, A. and Loken, E. The statistical crisis in science. *American Scientist*, 102(6):460–465, 2014. doi: 10.1511/2014.111.460.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., and Cunningham, W. A. AI and the transformation of social science research. *Science*, 380(6650): 1108–1109, 2023. doi: 10.1126/science.adi1778.

Horton, J. J., Filippas, A., and Manning, B. S. Large language models as simulated economic agents: What can we learn from *Homo Silicus*? Working Paper 31122, National Bureau of Economic Research, 2023.

McKay, M. D., Beckman, R. J., and Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979. doi: 10.2307/1268522.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shlain, D., Barak, B., and Schwartz, R. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00674.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, 2008. doi: 10.1002/9780470725184.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004, 2023.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design with a focus on social science tasks. *arXiv preprint*, arXiv:2310.11324, 2024.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. Specification curve analysis. *Nature Human Behaviour*, 4(11): 1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Yee, B. and Sharma, K. Calibrating behavioral parameters with large language models, 2026. URL https://arxiv.org/abs/2602.01022.

A. Variance Decomposition: Full Bootstrap CI Results

Table 6 reports bootstrap 95% confidence intervals for η^2 on all six dimensions and all three items from 5,000 bootstrap resamples. Dimensions with CIs that include zero are reported as "—"; the point estimate is still meaningful but too noisy to bound reliably.

Table 6. Bootstrap 95% CIs for η^2 across all dimensions and items. 5,000 bootstrap resamples at the specification level.

Item	Dimension	η^2	CI lo	CI hi
Gun control	Question framing	0.1603	0.131	0.192
	Model	0.0653	0.044	0.094
	Temperature	0.0020	—	—
	Few-shot	0.0019	—	—
	System prompt	0.0014	—	—
	Persona format	0.0002	—	—
Immigration	Model	0.0760	0.043	0.116
	Question framing	0.0156	0.008	0.026
	Temperature	0.0011	—	—
	Persona format	0.0007	—	—
	Few-shot	0.0007	—	—
	System prompt	0.0005	—	—
Gov. spending	Model	0.0610	0.033	0.097
	Question framing	0.0083	0.004	0.016
	Few-shot	0.0027	—	—
	System prompt	0.0010	—	—
	Temperature	0.0006	—	—
	Persona format	0.0001	—	—

B. Flipped Specification Analysis: Full Results

Table 7 reports overrepresentation ratios for all dimension levels among flipped gun control specifications ($\Delta_s \leq 0$). Ratios above 1.5 indicate meaningful overrepresentation; ratios near 1.0 indicate the level appears in flipped specs at roughly its base rate.

Table 7. Overrepresentation ratios among 91 flipped gun control specifications (17.0% of 534 total). $\text{Ratio} = p_{\text{flipped}}/p_{\text{all}}$. Only dimensions with any level showing ratio > 1.2 are reported.

Dimension	Level	Ratio	Interpretation
Model	Gemini 2.5 Flash	3.8×	Primary driver (rate term)
Model	Llama 3.3 70B	2.0×	Largest share by count (37% of flips)
Question framing	Likert	2.2×	Secondary driver
Few-shot	0 shots	1.7×	Contributory
Persona format	Survey block	1.2×	Marginal

C. Model API Identifiers and Access

Table 8 lists the exact model identifiers used for each provider, the API route, and whether the model was included in the main LHS run, the Saltelli extension, or the ordering experiment only.

Table 8. Model identifiers, access routes, and study inclusion.

Model	API identifier	Route	Included in
GPT-5.4	gpt-5.4	OpenAI	Main LHS, Saltelli, ordering
GPT-5.4-nano	gpt-5.4-nano	OpenAI	Main LHS, Saltelli, ordering
Claude Sonnet 4.6	claude-sonnet-4-6	Anthropic	Main LHS, Saltelli, ordering
Gemini 2.5 Flash	gemini-2.5-flash	Google	Main LHS, Saltelli; ordering errored
Llama 3.3 70B	llama-3.3-70b-instruct	OpenRouter	Main LHS, Saltelli, ordering
Mistral Small	mistral-small-24b-instruct-2501	OpenRouter	Main LHS, Saltelli, ordering