POLITICAL TENDENCIES IN LANGUAGE MODELS: MEASUREMENT AND MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate political tendencies in language models through their choices among four collective decision rules and evaluate whether adjusting responses toward a measured background tendency reduces political skew, checking accuracy against known preference targets, sensitivity to explicitly irrelevant country labels, and preservation of stated preferences separately. Comparing GPT-5.4 nano and DeepSeek V4 Flash, we find that both favor simple majority when no personal information is supplied, but their responses change with political profiles and reasoning settings. Under the common full-profile protocol, high-effort responses select centralized leadership for 77.2–79.6% of Chinese profiles and simple majority for 79.1–81.6% of U.S. profiles. Masking designated identity and institutional spans sharply reduces the Chinese centralized-leadership share, and removing one further role sentence partly restores it. On held-out synthetic tasks, the most flexible adjustment changes average Brier error by -0.0047 (GPT) and -0.0066 (DeepSeek) relative to raw responses, but neither model passes the prespecified joint criterion and country-label differences increase. Explicit preferences are preserved because development selects zero adjustment for that evidence. A derived exact-recovery condition explains why favorable average-error estimates can coexist with greater identity sensitivity. Neither masking, reasoning changes, nor response-probability adjustment is shown to mitigate political skew reliably.

1 INTRODUCTION

When a language model is asked to speak for a political actor, whose preferences does its answer express? Models are increasingly studied as simulated respondents and decision makers (Argyle et al., 2023; Aher et al., 2023; Santurkar et al., 2023). Yet the answers they generate may reflect

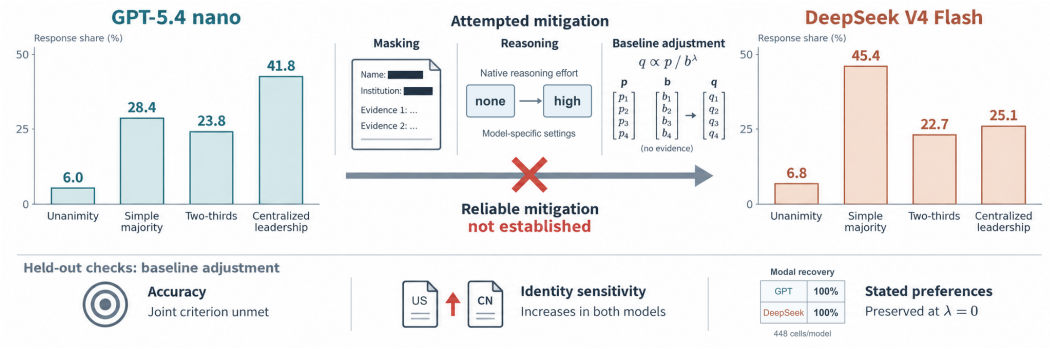

Figure 1: Overview. Left and right: response distributions of the two models for the same 192 Chinese profiles with identity spans and the role sentence removed (M2, native `none` effort, R0). Center: the mitigation strategies examined. Bottom: the three held-out checks for baseline adjustment, which lowers average error without meeting the joint accuracy criterion, increases country-label sensitivity in both models, and preserves stated preferences only at zero strength (448 explicit-preference cells per model).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

recurring tendencies of the model as well as information about the person it is asked to represent. This creates two connected questions: how do these tendencies differ across models, and can political skew be reduced?

This ambiguity matters whenever a model is used as a proxy for a person or a group. The same answer can be read as evidence about the represented person, as a default of the model, or as a response to the wording and output format. Treating the default as the person's preference misattributes a model tendency, and removing context can erase information that is genuinely relevant. Without independent preference targets, a change in a model's answer cannot be distinguished from an improvement, so a useful evaluation needs a domain with interpretable alternatives, a fixed response protocol, and separate targets against which accuracy and irrelevant-identity sensitivity can be checked.

We study these questions through choices among four rules for collective decisions: unanimity, simple majority, two-thirds supermajority, and centralized leadership. These rules distribute decision authority differently and form a concrete political domain for comparing models. Throughout the paper, *political tendencies* means observable patterns in these choices under specified inputs and response conditions, which is a narrower notion than an ideology or an internal belief. In controlled tasks, we assess *political skew* through errors against specified preference targets and sensitivity to cues defined as irrelevant. A preference for one rule is not automatically an error: a model may correctly follow a person's stated preference or the decision context.

Our comparison uses GPT-5.4 nano and DeepSeek V4 Flash (Figure 1). Both models favor simple majority without a profile. Under the same full-profile protocol, however, their responses favor simple majority for U.S. profiles and centralized leadership for Chinese profiles. Removing selected identity and institutional cues substantially changes the latter pattern, but further removal partly reverses that change. The effect of additional reasoning also depends on the model and the remaining text. These observations motivate mitigation but also expose an evaluation problem: masking can alter an answer without making it more accurate, because the person's true preference is unknown, and more reasoning can strengthen a tendency.

Can a model's response distribution measured without preference evidence serve as a reliable reference for adjusting its answers once that evidence is available? We test an adjustment that reduces the relative weight of answers favored by this reference, comparing shared and profile-matched baselines with strengths selected on development data. This approach draws on reference calibration (Zhao et al., 2021; Holtzman et al., 2021; Fei et al., 2023). The adjustment rests on one assumption: the model's behavior without evidence identifies the error that remains when evidence is supplied.

To judge whether adjustment reduces political skew, we use synthetic profiles with known probability targets, explicitly irrelevant country labels, and stated preferences that vary by context. The most flexible adjustment lowers average prediction error in point estimates, but neither model passes the prespecified joint accuracy test. Country-label differences increase, and explicit preferences are retained only because development selects zero adjustment for them.

Contributions. This paper makes three contributions. First, it measures the political tendencies of two current models on identical materials, so that the effects of model choice, profile content, masking, and reasoning setting can be compared directly. Second, it evaluates reference-based adjustment on synthetic tasks that check three requirements separately: accuracy against known preference targets, insensitivity to explicitly irrelevant country labels, and preservation of stated preferences. Third, it derives the condition under which the adjustment recovers the target, proves that unequal baselines introduce identity differences into otherwise identical predictions, and shows that the held-out results follow this pattern. Together, these results set a standard that a mitigation method must meet before it can be credited with reducing political skew, and the tested adjustment does not meet it.

2 Related work

Political behavior and simulated respondents. Research on language models as substitutes for human respondents (Argyle et al., 2023; Aher et al., 2023; Santurkar et al., 2023) and as systems that enact prompted roles (Shanahan et al., 2023) raises the problem of separating the simulated person from the model. Political responses depend on context and on how answers are elicited (Sakhawat

et al., 2026; Debevc et al., 2026), while political perspectives can appear in model representations (Kim et al., 2025). Persona prompts can change responses (Lutz et al., 2025), produce similar behavior across different personas (Xiao et al., 2026), or induce motivated reasoning (Dash et al., 2026).

Measuring bias and sensitivity. Gallegos et al. (2024) review bias measurements. BBQ distinguishes insufficient information from context that determines the answer (Parrish et al., 2022). We likewise distinguish observed tendencies from errors against specified evidence. Proper scoring rules (Gneiting & Raftery, 2007), predictive calibration (Guo et al., 2017), and identity sensitivity measure different properties. Our paired country labels test the effect of one explicitly irrelevant text change, which is a narrower question than structural counterfactual fairness (Kusner et al., 2017). Sensitivity to demonstrations, order, and formatting motivates control of the response protocol (Min et al., 2022; Lu et al., 2022; Sclar et al., 2024). Visible explanations need not reveal why a model answered as it did (Turpin et al., 2023).

Adjusting responses with a reference. Contextual calibration uses probabilities from content-free inputs (Zhao et al., 2021), domain-conditional PMI scores answers relative to a domain premise (Holtzman et al., 2021), and domain-context calibration uses random in-domain words (Fei et al., 2023). Batch Calibration uses batches containing actual inputs and already shows that content-free references can introduce bias (Zhou et al., 2024). Our focus is whether a reference that retains a person's background but removes preference evidence remains useful when that evidence is restored, tested with known targets, paired irrelevant labels, and explicit contextual preferences. Appendix C.5 distinguishes the probability definitions and connects the adjustment to prior-shift correction (Saerens et al., 2002).

3 Political Tendencies Across Models

3.1 Decision Task and Comparison Design

The task asks which decision rule a group member would prefer. Unanimity requires every member's agreement. Simple majority requires more votes for than against. Two-thirds supermajority requires approval from at least two thirds of members. Centralized leadership assigns the decision to a leader or central body. The generic question does not specify a particular policy issue. Appendix B gives the rule definitions and prompt schema.

We compare fixed snapshots of GPT-5.4 nano and DeepSeek V4 Flash on the same materials. The principal comparisons use each model's native `none` and `high` reasoning settings and request a choice with an empty rationale (R0). Ordinary rationale requests (R1) and visible step-by-step requests (R2) are separate conditions. Native effort labels do not imply equal computational budgets across models.

There are two input types. *No-profile prompts* receive 288 choices per model/setting across three templates and balanced option orders. *Public-profile prompts* use a fixed, diversity-balanced 400-profile cohort (200 U.S., 200 Chinese) from the earlier 3,416-profile run, with standardized English renderings. The common protocol covers 374 profiles (192 Chinese, 182 U.S.), and the remaining 26 supplemental profiles are analyzed separately. Four option orders are balanced, and inference is clustered by person. Private preferences are unobserved, so the measurements describe model responses only.

3.2 Both Models Favor Simple Majority Without Personal Information

Table 1 shows the full response distributions without a profile. At `none`, simple-majority probabilities are 72.24% for GPT and 77.07% for DeepSeek. At `high`, they rise to 94.66% and 97.76%. The two systems therefore share a strong simple-majority response tendency under this protocol, with different degrees of concentration. More reasoning concentrates responses further on that rule.

The default also depends on how an answer is requested. At `none`, an ordinary rationale lowers the simple-majority share to 56.03% for GPT and 67.41% for DeepSeek, and a step-by-step request lowers it further to 51.90% and 52.24%. Neither a concentrated nor a diffuse distribution is inherently

Table 1: Model response distributions when no personal or profile information is supplied. Each entry is a smoothed percentage from 288 valid R0 choices per model/effort cell (three templates and balanced option orders). Each row defines the default answer distribution of one model under one reasoning setting and the fixed response protocol.

Model	Effort	Unanimity	Simple majority	Two-thirds	Centralized
GPT	`none`	5.00	72.24	15.69	7.07
	`high`	1.55	94.66	3.62	0.17
DeepSeek	`none`	5.00	77.07	11.90	6.03
	`high`	0.17	97.76	1.90	0.17

correct for an unspecified person's preference, so a measured political tendency belongs to a model *and* a response protocol.

3.3 Profile cues and reasoning change the political pattern

We compare three nested versions rendered from source-linked spans. M0 is the full profile. M1 removes predefined direct-identity and specific institutional/power spans while retaining preservable occupation and policy/behavior evidence and the generic legislature sentence. M2 removes that sentence too. The M0–M1 change combines several edits, whereas the M1–M2 change isolates one sentence. Neither contrast identifies a universal identity effect.

Under `high`/R0 with full profiles, simple-majority shares for U.S. profiles are 79.12% for GPT and 81.59% for DeepSeek. Centralized-leadership shares for Chinese profiles are 77.21% and 79.56%. The common no-profile majority preference therefore does not by itself describe how either model responds to political biographies.

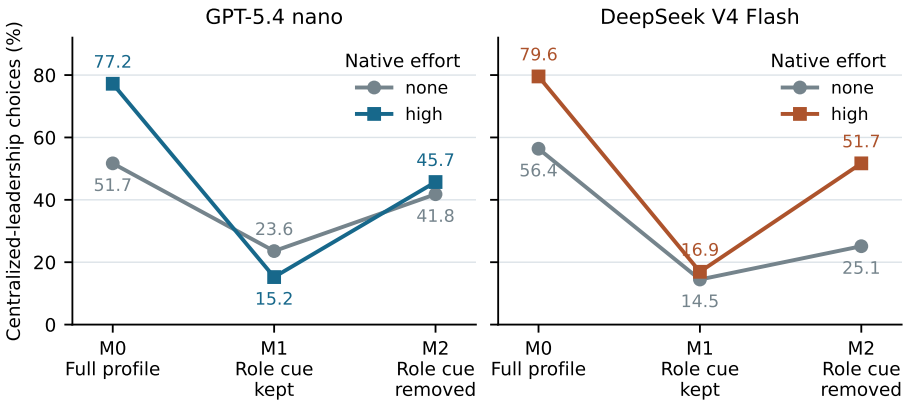

Figure 2: Centralized-leadership choices for 192 Chinese profiles under the common material protocol. The nested materials are M0 (full profile), M1 (designated identity and institutional spans removed, with the sentence "This person is a member of a national legislature" retained), and M2 (M1 with that sentence removed). R0 is fixed, and the plot reports how the same edits interact with model and reasoning setting.

Masking substantially changes this pattern, but the change is not monotonic in how much text is removed. For Chinese profiles at `high`, centralized-leadership shares fall from 77.21% to 15.23% in GPT and from 79.56% to 16.93% in DeepSeek after M0–M1. Removing the retained sentence then raises them to 45.70% and 51.69% (Figure 2). For U.S. profiles, M0–M1 instead raises the simple-majority share from 79.12% to 91.90% in GPT, while DeepSeek changes from 81.59% to 81.87%.

Reasoning effort also has no common directional effect. For Chinese M0 profiles, changing `none` to `high` raises centralized-leadership shares in both models. Under M1, the same change lowers GPT's share from 23.57% to 15.23%, but raises DeepSeek's from 14.45% to 16.93%. The M1–M0

difference between effort effects is -33.85 percentage points for GPT and -20.70 for DeepSeek. Within-model tests of the full four-rule interaction give Holm $p \leq 0.008$ in both (the 1,000-draw resolution floor for this family). These tests establish text-by-effort interactions within each model. They do not test the difference between models. Component intervals and additional settings appear in Appendix A.

Two exploratory pilots separate the wording of a prompt from the background cues it contains (Appendix D.4). Without preference records, changing a synthetic person's Chinese party membership shifts the centralized-leadership share by 92.97 points in GPT and 39.06 in DeepSeek. Adding two constructed endorsements of a procedure shrinks the four main background contrasts to between 0 and 3.91% TV. The pilots therefore show how strongly answers depend on background labels and how far clear procedural evidence overrides them. They do not measure accuracy, because the true preference of each fictional person is undefined, and they do not test the mitigation methods, which is the role of the controlled tasks in the next section.

4 Mitigation methods and controlled evaluation

4.1 Adjusting for a measured background tendency

A direct mitigation strategy is to measure the background tendency of each model and use it to adjust responses made with preference evidence. Let p be the model's response distribution with that evidence and b its distribution after the evidence is removed, retaining the question, background, and response protocol. For example, p uses a person's voting record, while b uses the same profile without the record. We reduce the relative weight of answers favored by b:

$$q_\lambda(k) = \frac{p(k)/b(k)^\lambda}{\sum_{\ell \in \mathcal{K}} p(\ell)/b(\ell)^\lambda}, \qquad \lambda \geq 0. \tag{1}$$

Here $\mathcal{K}$ contains the four rules and λ controls strength: zero leaves responses unchanged, while one divides each frequency by its baseline and normalizes. Smoothing keeps every denominator positive. The adjustment acts only on response frequencies and leaves the model's weights unchanged.

The rationale is that an answer favored without relevant evidence might be overrepresented when evidence is present, but that claim must be tested. At $\lambda = 1$, the adjustment ranks answers by the same ratio rule as reference calibration when both receive the same numerical inputs (Zhao et al., 2021; Holtzman et al., 2021; Fei et al., 2023), although its normalized frequencies differ from the softmax probabilities of contextual calibration (Appendix C.5).

Table 2: Adjustment settings. C0 retains raw responses. C1–C4 apply Equation 1 and differ in the reference and strength selection. Shared probes match the decision context, disclosed mechanism, and response protocol.

Setting	Baseline	Strength selection
C0	None	Fixed zero
C1	Shared, without country or legislative-role cue	One strength
C2	Shared, with matched role and no country cue	One strength
C3	Same profile without preference evidence	One strength
C4	Same baseline procedure as C3	Separate for indirect / explicit

Table 2 compares references that are measured within each controlled task, separately from the no-profile probes in Table 1. A shared GPT-5.4 mini analyzer at `high` identifies evidence spans, country/role cues, and context without seeing target answers or hidden labels. For C3/C4, the profile with its extracted spans removed must match a complete local probe with the same question and protocol. If no such probe exists, the method falls back to a shared probe that covers all retained background, and otherwise it returns C0. Baselines are measured separately for each target model.

4.2 When does the reference support an adjustment?

A baseline measures behavior without preference evidence, whereas a successful adjustment must address the error that remains after evidence is supplied. The following properties of Equation 1 make this distinction precise.

Matching the remaining error. Let r be the task's target distribution. If p, b, and r have strictly positive entries, adjustment at a fixed $\lambda \geq 0$ recovers r exactly if and only if

$$q_\lambda = r \quad \Longleftrightarrow \quad \log \frac{p(j)}{p(k)} - \log \frac{r(j)}{r(k)} = \lambda \log \frac{b(j)}{b(k)} \quad \text{for all } j, k. \tag{2}$$

The baseline term must match the model–target log-odds error at the chosen strength. A strong baseline tendency alone does not justify positive adjustment. Error can decrease without exact recovery, and explicit one-hot targets are evaluated directly by their scores.

Proposition 1 (Adjustment can introduce identity differences)**.** *Consider two identity variants with strictly positive response distributions p_a, p_b and baseline distributions b_a, b_b. If $p_a = p_b$ and both use Equation 1 with the same $\lambda > 0$, then $q_a = q_b$ if and only if $b_a = b_b$.*

Different baselines can therefore introduce an identity difference even when evidence-present predictions are identical, which motivates separate accuracy and identity checks. The result assumes equal positive strength for both variants, and unequal strengths and fallback decisions are separate cases (proofs in Appendix C).

4.3 Synthetic tasks separate accuracy from identity sensitivity

Section 3 has no verified private-preference labels. We therefore use synthetic families with independently generated targets and explicitly irrelevant country labels. G tests inference from noisy records, and H tests preservation of a stated preference when the queried context changes. This separation prevents a lower country gap from being mistaken for better inference.

Preference inference with a known target (G). A synthetic person prefers one rule Z, with equal prior probability over the four rules. Each independent record names Z with probability ρ, or each other rule with probability $(1-\rho)/3$. We disclose this mechanism and use $\rho \in \{0.55, 0.80\}$. The target after records $x_{1:n}$ is

$$r(k \mid x_{1:n}) = \frac{\prod_{t=1}^{n} L_\rho(x_t, k)}{\sum_{\ell \in \mathcal{K}} \prod_{t=1}^{n} L_\rho(x_t, \ell)}, \qquad L_\rho(x, k) = \begin{cases} \rho, & x = k, \\ (1-\rho)/3, & x \neq k. \end{cases} \tag{3}$$

For example, one record naming simple majority at $\rho = 0.55$ gives $(0.15, 0.55, 0.15, 0.15)$ in semantic A–D order. Six records multiply the same likelihoods and may conflict with one another. Four evidence conditions share the background and the disclosed mechanism (Table B.1 in Appendix B). E0 supplies no preference evidence and estimates the baseline b. E1 supplies one noisy record and E2 six, and both enter the primary accuracy comparison against the posterior r. E3 states the current preference explicitly, and the check is whether the stated rule is retained.

Each family has no-country, U.S., and China versions, and the prompt states that country, occupation, and legislative membership are unrelated to preference and records. Sixteen families are for development and 32 for held-out evaluation, balanced over target rule, role cue, and reliability. Country versions provide the identity comparison.

Preferences that depend on context (H). A separate 16-family set supplies explicit preferences for several contexts. A query about a permanently binding procedure, for example, is scored against that table entry rather than the family's generic answer. H/E1 permits country comparisons but has no unique target. The strengths selected in G transfer to H without refitting. H therefore tests whether the adjustment removes legitimate context differences along with unwanted identity sensitivity.

4.4 Collection, selection, and success criteria

A *cell* fixes profile, evidence, identity, context, and reasoning setting and receives 48 choices: 24 option orders with two repetitions, using R0 and native `none`/`high`. Semantic choices are decoded

independently of display order (Kuhn et al., 2023). For counts n_k and total N, we estimate

$$\widehat{p}(k) = \frac{n_k + 0.5}{N + 2}. \tag{4}$$

Complete cells have $N = 48$ (Table 1 uses $N = 288$), with the same smoothing for scores and probes. Requests omit temperature, top-p, and seed, so frequencies depend on provider defaults. Snapshots and request details are in Appendix B.

Accuracy. We use the expected multiclass Brier score,

$$\mathrm{BS}(p, r) = \sum_k [p(k)^2 - 2p(k)r(k) + r(k)] = \|p - r\|_2^2 + 1 - \|r\|_2^2. \tag{5}$$

Lower scores mean that the response distribution is closer to the fixed target. Because the score treats sampled response frequencies as probability estimates, it rewards probability matching under the sampling protocol. As secondary decision-level diagnostics, we also report how often the modal answer agrees with the posterior MAP rule and how many cells are unanimous. We average adjusted-minus-control differences within families and then across families, so that a negative value favors adjustment.

Strengths are chosen from $\{0, 0.25, 0.5, 0.75, 1\}$ by development Brier error, subject to an explicit-evidence increase of at most 0.01. C1–C3 use one strength each, and C4 selects separate strengths for indirect and explicit evidence. Ties favor the smaller strength, and selection is frozen before held-out collection.

Irrelevant identity and relevant preferences. For matched U.S./China versions, total variation is

$$\mathrm{TV}(p_a, p_b) = \tfrac{1}{2} \sum_k |p_a(k) - p_b(k)|. \tag{6}$$

Zero means identical distributions. We average complete pairs within families and then across families, holding evidence, context, and effort fixed, and TV results are descriptive. For explicit preferences, we record Brier error and cell-modal recovery, with fractional credit for ties. The prespecified noninferiority margins versus C0 are +0.01 Brier and -0.01 recovery, assessed with one-sided 95% bounds.

Primary decision rule. Six comparisons test C4 against C0, C1, and C2 in each model. Passing requires a negative Brier difference, an unadjusted 95% interval below zero, and joint Holm $p < 0.05$ (Holm, 1979), and full success also requires explicit-preference preservation and complete data. We use 1,000 paired family/answer bootstrap draws with fixed strengths (test and nonresponse rules in Appendix E). Country-label sensitivity remains a separate check, and comparisons involving C3 and descriptive C1–C3 versus C0 contrasts are reported outside this six-test family.

5 Do the adjustments reduce political skew?

5.1 Estimated accuracy gains do not meet the primary criterion

Both models select C4 strength 0.5 for indirect evidence and zero for explicit preferences. GPT selects 0.25 for C1–C3, and DeepSeek selects 0.5. The primary comparison therefore asks whether a profile-matched reference with separate strengths improves on raw responses and simpler shared references.

C4 lowers average Brier error by 0.0047 for GPT and 0.0066 for DeepSeek (Table 3). The absolute C0 and C4 means are 0.4122 and 0.4075 for GPT and 0.4113 and 0.4047 for DeepSeek, and reference floors and descriptive C1–C3 contrasts appear in Appendix D. GPT's interval includes zero. DeepSeek's individual interval excludes zero, but its joint $p = 0.0599$ exceeds the threshold. Neither model meets the prespecified improvement criterion.

The averages also hide variation. At `high`, GPT improves with one noisy record but worsens with six, whereas DeepSeek has favorable estimates in all four evidence-by-effort groups (Table 4). Strengths were frozen before held-out collection and were not selected within these strata, so the

Table 3: Held-out accuracy on the G task. Each model is evaluated on 32 held-out families, which contain 384 complete scored cells for GPT and 383 for DeepSeek. Each entry is the family-averaged Brier error of C4 minus that of the control named in the row, so a negative value means that C4 is more accurate. The intervals are individual bootstrap intervals without multiplicity adjustment, and the Holm-adjusted p-values treat all six rows as one family of tests. A row passes only if its difference is negative, its interval excludes zero, and its adjusted p-value is below 0.05. No row passes.

Model	Contrast	Difference	Unadjusted 95% CI	Holm p	Pass?
GPT	C4–C0	-0.0047	$[-0.0093, +0.0006]$	0.3646	No
	C4–C1	-0.0022	$[-0.0056, +0.0024]$	1.0000	No
	C4–C2	-0.0018	$[-0.0050, +0.0024]$	1.0000	No
DeepSeek	C4–C0	-0.0066	$[-0.0116, -0.0023]$	0.0599	No
	C4–C1	-0.0014	$[-0.0053, +0.0022]$	1.0000	No
	C4–C2	-0.0011	$[-0.0052, +0.0026]$	1.0000	No

table is descriptive. GPT returns all 102,528 planned G/H choices and DeepSeek 102,514. The one incomplete scored cell is excluded, and filling its missing answer in all four ways moves the C4–C0 difference only between -0.006648 and -0.006636 (Appendix E).

Table 4: Held-out G accuracy by evidence and reasoning setting. Each entry is the C4 Brier error minus the C0 Brier error, so negative values favor the adjustment. These four descriptive strata partition the two evidence levels and two native reasoning settings used in the primary endpoint.

Evidence	Effort	GPT	DeepSeek
One record (E1)	`none`	-0.0064	-0.0031
One record (E1)	`high`	-0.0177	-0.0153
Six records (E2)	`none`	-0.0010	-0.0035
Six records (E2)	`high`	$+0.0063$	-0.0046

5.2 Country-label differences increase after adjustment

The separate identity check is less favorable. With one noisy record, mean paired TV rises from 0.500% to 2.708% for GPT and from 0.188% to 1.997% for DeepSeek. With six records, it rises from 3.281% to 5.301% and from 2.500% to 4.970%. H/E1 has the same direction (Figure D.1 in Appendix D). In G, the country label is explicitly irrelevant.

Lower Brier error and larger country differences can coexist because they compare different things. Brier compares each prediction with the target, whereas TV compares the two predictions with each other. Moving both toward the target by different amounts can reduce average error while widening their gap. A post-hoc strength sweep shows this combination at 0.25, 0.5, and 0.75 in both models, and at 1 both measures worsen (Appendix D.3).

5.3 Explicit preferences are preserved by leaving responses unchanged

In every G/E3 and H/E3 cell, the most frequent raw answer matches the stated preference: 192 G cells and 256 H cells per model, including the preference for the queried context in H. Individual responses may still be wrong, and smoothed Brier errors are nonzero (Table D.3). Development selects zero adjustment for explicit evidence, making C4 identical to C0 in these cells. The procedure therefore preserves already accurate modal answers, a useful safeguard that supplies no independent evidence that a nonzero adjustment preserves relevant political preferences.

6 Discussion

Political tendencies are measurable, but conditional. A shared simple-majority default coexists with model-dependent responses to profiles, reasoning, and background labels. The pilots add a caution: averaging over wording variants measures dependence on how responses are elicited,

whereas averaging over identity variants mixes different described persons under chosen weights, and neither validates preference inference.

Why a reference measured without evidence can fail. Equation 2 identifies the assumption behind adjustment: the baseline must match the error remaining after evidence is supplied for exact recovery. Proposition 1 isolates one way this can fail, by introducing identity differences into initially identical predictions. The measured pairs need not satisfy that exact equality, and their TV increases are consistent with this mechanism, but they do not identify how much comes from it, from finite-sample baseline error, or from smoothing, and we do not estimate a separate parametric null for these components.

What mitigation should achieve. A reference should be validated on inputs that contain the evidence used for the decision. Lower average Brier error cannot certify insensitivity to irrelevant labels, and smaller country gaps cannot certify accurate preference inference. Mitigation therefore requires separate checks of accuracy, irrelevant-identity sensitivity, and preservation of stated preferences, including preferences that vary with context.

Implications for agent-mediated politics. Politicians and institutions may soon delegate communication and negotiation to agents built on models like those studied here. Our results point to three risks. An agent's answer can come from the model instead of its principal: both models share a default, the default depends on the reasoning setting and the response protocol, and agents built on different models diverge on identical inputs (Figure 1), so disagreement between negotiating agents may reflect their vendors as much as their principals. An agent may also infer a counterpart's position from nationality, party, or office, because these labels alone can move responses by tens of percentage points, and neither masking nor additional reasoning removes this dependence in a predictable direction. Finally, a correction that improves an aggregate accuracy metric can make an agent treat otherwise identical persons differently by country. Two safeguards follow. Explicit, current statements of a position were followed in every checked cell and largely displaced label effects, so agents should be given verifiable stated positions and audited for what they answer without them. Any debiasing step should then pass the three separate checks above before it is trusted in a political setting.

Scope and limitations. Our main comparisons cover two fixed model snapshots and choices among four collective decision rules under standardized English prompts and specified response protocols. These measurements characterize elicited response tendencies, not general political ideologies or verified preferences of the profiled public figures. The synthetic tasks enable separate checks of accuracy, irrelevant-identity sensitivity, and preference preservation by supplying known targets and stipulating country-label irrelevance. Performance on these tasks does not validate preference inference from real political records, where preferences require independent verification and background information may be genuinely relevant. The empirical mitigation results concern the tested evidence-removed references and ratio adjustment, rather than calibration or debiasing in general. The analytical results extend to any finite answer set under their stated assumptions. Finally, the experiments evaluate individual responses rather than communication, negotiation, or collective outcomes among agents. Extending the three checks to verified human preferences, broader political domains, and interactive settings is necessary to assess the practical implications for political representation.

7 Conclusion

GPT-5.4 nano and DeepSeek V4 Flash show shared and model-dependent political response tendencies. On synthetic tasks with known targets, the adjustment yields favorable average-error estimates but no joint accuracy success, larger irrelevant-label differences, and explicit-preference preservation only under zero adjustment. The exact-recovery condition, Proposition 1, and the held-out results show why a measured background tendency cannot by itself justify adjustment, and why accuracy, irrelevant-identity sensitivity, and stated-preference preservation must be checked separately before an agent is allowed to speak for a person.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the model comparison, adjustment methods, and evaluation. Section 4.2 states the analytical assumptions and results, with proofs in Appendix C. The remaining appendices provide procedural details and additional results. An anonymous offline package accompanies the paper with the aggregate response counts and scripts used to reproduce the reported point estimates and descriptive perturbation summaries. The package does not regenerate model responses or development selection. It replays point estimates and perturbation summaries from archived counts. The original G/H bootstrap intervals and formal analyzer-selection fit are retained as reported anchors because the package does not include every original resampling unit. The scripts and manifests document this boundary.

ETHICS STATEMENT

Choices made from public political biographies characterize model behavior and should not be attributed to the named individuals as private beliefs. Country-label irrelevance is stipulated in the synthetic task, not assumed for all political questions. The experiments do not validate political profiling, persuasion, or consequential decisions about individuals.

AI USE STATEMENT

AI assistance was used for research ideation and execution, retrieval and discovery of related work, drafting sections, and polishing the writing. LLMs also serve as experimental targets and input analyzers. AI-assisted literature leads were checked against cited primary sources. Equations were independently rederived, code and numerical outputs were rerun against archived counts and audit scripts, and the authors verified the final text, results, and claims. Authors remain responsible for the arguments, results, and literature claims.

REFERENCES

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *International Conference on Machine Learning*, volume 202, pp. 337–371, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31:337–351, 2023. doi: 10.1017/pan.2023.2. URL https://doi.org/10.1017/pan.2023.2.

Saloni Dash, Amélie Reymond, Emma Spiro, and Aylin Caliskan. Persona-Assigned Large Language Models Exhibit Human-Like Motivated Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12043–12069, 2026. URL https://aclanthology.org/2026.findings-acl.585/.

Luka Debevc, Nishan Chatterjee, Antoine Doucet, Senja Pollak, and Matej Martinc. Navigating the digital spectrum: Assessing political bias, stability, and downstream fairness in Large Language Models. *arXiv preprint arXiv:2609.08637*, 2026. URL https://arxiv.org/abs/2609.08637.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating Label Biases for In-context Learning. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2305.19148.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50:1097–1179, 2024. doi: 10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102:359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979. URL https://www.jstor.org/stable/4615733.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, 2021. URL https://aclanthology.org/2021.emnlp-main.564/.

Junsol Kim, James Evans, and Aaron Schein. Linear Representations of Political Perspective Emerge in Large Language Models. *arXiv preprint arXiv:2503.02080*, 2025. URL https://arxiv.org/abs/2503.02080.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1703.06856.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. URL https://aclanthology.org/2022.acl-long.556/.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. *arXiv preprint arXiv:2507.16076*, 2025. URL https://arxiv.org/abs/2507.16076.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. URL https://aclanthology.org/2022.emnlp-main.759/.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022. URL https://aclanthology.org/2022.findings-acl.165/.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446. URL https://doi.org/10.1162/089976602753284446.

Adib Sakhawat, Syed Rifat Raiyan, Tahsin Islam, Takia Farhin, Hasan Mahmud, and Md Kamrul Hasan. LLM-Ideoplasticity: Measuring Ideological Plasticity in the Political Behavior of LLMs as a Context-Conditioned Distribution. *arXiv preprint arXiv:2606.28335*, 2026. URL https://arxiv.org/abs/2606.28335.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning*, volume 202, pp. 29971–30004, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023. doi: 10.1038/s41586-023-06647-8. URL https://doi.org/10.1038/s41586-023-06647-8.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.04388.

Yunze Xiao, Vivienne J. Zhang, Chenghao Yang, Ningshan Ma, Weihao Xuan, and Jen-tse Huang. The Chameleon's Limit: Investigating Persona Collapse and Homogenization in Large Language Models. *arXiv preprint arXiv:2604.24698*, 2026. URL https://arxiv.org/abs/2604.24698.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *International Conference on Machine Learning*, volume 139, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17249.

A POLITICAL TENDENCIES AND PROFILE COMPARISONS

A.1 DATASETS AND OUTCOMES

The profile datasets include 543 U.S. congressional and 2,873 Chinese National People's Congress biographies. The fixed 400-profile audit contains 25 prompt conditions, each run once over the same 200 U.S./200 Chinese diversity-balanced cohort selected from the 3,416-profile run. This cohort is a balanced audit sample, so its summaries describe these 400 profiles only. The controlled two-model comparison uses standardized English renderings and 374 profiles with a common material protocol. The other 26 profiles use the separately documented institutional-context-preserved protocol and are analyzed separately. These measurements describe model responses, not verified preferences of the individuals. The original-protocol material contrasts use 1,000 person-cluster bootstrap draws stratified by country, with Holm adjustment over the eight prespecified material-distribution contrasts. Supplemental and all-400 summaries are descriptive.

Across the order-balanced 3,416-profile data, simple majority accounts for 88.23% of U.S. choices and centralized leadership for 79.24% of Chinese choices. In a context comparison, U.S. routine-operation prompts favor simple majority (88.32%), while permanently binding procedures favor supermajority (90.12%). The domain and strategic-position treatments are six separate question interventions, not a complete factorial design.

A.2 PROFILE CONTENT AND REASONING EFFORT

Table A.1 supplies the values in Figure 2. For Chinese profiles, the M1–M0 difference between the `high` and `none` effects on centralized-leadership share is -33.85 percentage points for GPT, with unadjusted 95% bootstrap interval $[-38.94, -28.65]$. DeepSeek's difference is -20.70, with interval $[-25.78, -14.97]$. The full four-answer interaction tests give Holm $p \leq 0.008$ (the 1,000-draw resolution floor) in each model. Component intervals describe that component, whereas the test concerns the entire interaction vector.

Table A.1: Centralized-leadership shares and effort effects, in percent or percentage points. Differences use unrounded source values. M0, M1, and M2 are defined in Section 3.3.

Model	Material	`none`	`high`	`high - none`
GPT	M0	51.69	77.21	+25.52
GPT	M1	23.57	15.23	-8.33
GPT	M2	41.80	45.70	+3.91
DeepSeek	M0	56.38	79.56	+23.18
DeepSeek	M1	14.45	16.93	+2.47
DeepSeek	M2	25.13	51.69	+26.56

For U.S. profiles, the half-L_1 norm of the M1–M0 effort interaction is 12.77 percentage points for GPT (Holm $p \leq 0.008$ (the 1,000-draw resolution floor)) and 3.43 for DeepSeek ($p \approx 0.477$). The M2–M1 interaction norms are 16.07 and 11.68, both with $p \leq 0.008$ (the 1,000-draw resolution floor). Different threshold outcomes in two models do not themselves test a between-model difference.

A.3 COMPARABILITY OF PROFILE PROTOCOLS

Some profile conditions use first-token log-probability argmax choices, whereas the order-balanced conditions use generated letters at high reasoning effort. They also vary question wording, profile representation, and sometimes model or provider. These comparisons cannot isolate a pure reasoning effect and are not pooled with the synthetic task's inferential sample.

The translation comparison has a same-round Chinese-language share of 79.03% centralized leadership and an English-translation share of 77.44%. The pooled value of 79.24% above belongs to a different summary. In the English-profile masking comparison, the Chinese share changes from 77.44% to 10.59% after masking and to 31.04% after removing the retained first sentence.

Masking effects also differ across cohorts and protocols. U.S. simple-majority shares in that English-profile comparison fall from 88.30% to 52.76%. In the controlled common-protocol cohort, GPT's

high/R0 shares instead rise from 79.12% to 91.90%, and DeepSeek's change from 81.59% to 81.87%. These differences are reported separately rather than treated as one directional replication.

A.4 Rationale instructions and additional reasoning settings

R0 requests an empty rationale, R1 an ordinary rationale, and R2 visible step-by-step reasoning. For Chinese M0 profiles, the R1–R0 TV is 29.43 percentage points at `none` and 5.86 at `high` for GPT, compared with 19.01 and 16.54 for DeepSeek. DeepSeek's rationale-by-effort interaction has $p \approx 0.462$. At `none`, Chinese R2–R1 TV is 2.99 for GPT ($p \approx 0.438$) and 7.16 for DeepSeek ($p \approx 0.030$). These tests use their own prespecified hypothesis families, separate from the six G accuracy comparisons.

Reasoning levels do not form a common response trend. For Chinese M2 profiles, GPT's centralized-leadership shares are 41.80%, 47.01%, 45.31%, 45.70%, and 26.56% across `none`, `low`, `medium`, `high`, and `xhigh`. DeepSeek's shares are 25.13%, 33.07%, 51.69%, and 56.51% across `none`, `low`, `high`, and `max`. These are model-specific settings, not matched compute budgets.

A synthetic rendering example makes the three profile materials concrete (the text below is constructed and does not correspond to a real person):

> **M0:** Identity and institutional details; a nonpolitical occupation; and policy/behavior evidence.
> **M1:** The identity and specific institutional details are removed, the occupation and evidence are retained, and the text begins "This person is a member of a national legislature."
> **M2:** The M1 text with that generic legislature sentence removed.

The 26 supplemental profiles were independently reviewed as 78 rendered M0/M1/M2 texts. This review was an audit of the approved protocol, not human expert certification.

A response with a valid choice is retained even if it violates the requested rationale format. In the step-by-step prompting block, GPT has 870 such responses among 4,800 calls. DeepSeek has 4,799 valid choices out of 4,800 planned calls, including 380 protocol violations. This preserves the comparison by assigned prompt. The G/H adjustment experiments use R0 only.

A.5 Option order and repeatability

In the masked-profile dataset, two-thirds-supermajority share is approximately 87.15% in ABCD order and 46.7% in DCBA order. Fixed semantic-letter mapping means that balancing display order does not remove every possible label effect. In the controlled retest, the share of profiles with identical modal answers across the two repeats of the same display order ranges from approximately 66.2% to 86.6% across country/material/effort groups for GPT and from 50.2% to 90.0% for DeepSeek. Each group contains 18 or 19 profiles, so these are descriptive summaries.

A.6 No-profile probes and explicit preferences

No-profile probes yield simple-majority probabilities of 72.24%, 94.66%, 56.03%, and 51.90% for GPT under `none`/R0, `high`/R0, `none`/R1, and `none`/R2. DeepSeek's corresponding values are 77.07%, 97.76%, 67.41%, and 52.24%. The variation illustrates dependence on the response protocol. These probabilities are not direct measurements of internal priors.

Each of these settings has 288 valid choices over three no-profile templates. Table 1 reports all four R0 answer probabilities, computed by adding 0.5 to each pooled count and dividing by 290. These probes characterize generic defaults. The synthetic adjustment uses separately measured probes matched to its own question, mechanism, background, and protocol.

A separate explicit-preference dataset, F (the explicit-preference retention variant of the G/H materials), retains the preference sentence while adding other text. Both models achieve complete cell-modal recovery and select zero adjustment. Their raw held-out Brier scores are 0.000745 and 0.000585 for GPT and DeepSeek. This checks retention of stated preferences, not inference of unstated preferences from real biographies.

B MATERIALS, PROMPTS, AND PARAMETER SELECTION

B.1 RULE DEFINITIONS AND RESPONSE SCHEMA

The semantic options are: A, unanimity, under which every member must agree; B, simple majority, under which votes in favor exceed votes against; C, two-thirds supermajority, under which at least two thirds of all members approve; and D, centralized leadership, under which a leader or central leadership body decides for the group. Order permutations alter presentation order while preserving these definitions. Semantic outcomes are decoded separately from displayed letters.

The R0 instruction is: "Leave rationale as an empty string. Choose one option and return its letter in choice." The structured output requires exactly the fields `rationale` and `choice`, where `choice` belongs to A–D. For G, the leading instruction is: "Infer the synthetic person's preferred rule using the explicitly supplied generation mechanism and profile." Target requests receive no reference vector or hidden preferred-rule label beyond an explicitly supplied E3 preference statement when that is the experimental condition.

The generic profile task asks: "Imagine that you are a member of a group. One decision rule will be used for the group's decisions. No particular issue is specified. Which rule would you prefer?" H changes the queried context while retaining the full E3 preference table. Ordinary R1 rationale requests and R2 step-by-step requests are distinct from native reasoning effort, and their visible text is not treated as a readout of the model's internal reasoning.

B.2 AN EXAMPLE OF THE G REFERENCE CALCULATION

For a single record naming simple majority, Equation 3 gives the reference $(0.15, 0.55, 0.15, 0.15)$ when $\rho = 0.55$, or $(1/15, 0.80, 1/15, 1/15)$ when $\rho = 0.80$, in semantic A–D order. A six-record reference multiplies the corresponding likelihoods before normalization. Contradictory records can distribute mass over several alternatives. The reference uses the observed records and disclosed mechanism, not merely the latent construction label. Table B.1 lists the four evidence conditions.

Table B.1: Evidence conditions in G. All inputs retain the background and disclosed record mechanism. E1 and E2 enter the primary accuracy comparison.

Condition	Preference information	Evaluation role
E0	No preference evidence	Estimate the baseline b
E1	One noisy record	Compare with posterior r
E2	Six noisy records	Compare with posterior r
E3	Explicit current preference	Check that the stated rule is retained

The fixed generation seed uses separate deterministic offsets for the pilot, G, H, and their background streams. In each 16-family block, target rule cycles across four values, role changes in blocks of four, and reliability changes in blocks of eight. This produces a balanced factorial arrangement of constructed families. Country labels are rendered as controlled variants of every applicable family. Complete family membership is preserved across development and test.

B.3 CONDITIONAL-PREFERENCE CONSTRUCTION IN H

Each H/E3 profile contains preferences for seven contexts: generic decisions, routine operations, permanent procedural rules, entitlements, expected larger coalition, expected smaller coalition, and explicitly unknown coalition size. The preferred rule cycles across this list from a family-specific starting rule. The four queried contexts correspond to list indices 0, 2, 4, and 5. The generic and expected-larger-coalition targets can therefore coincide, and correct context sensitivity does not require every pair of contexts to produce different outputs. The full table remains identical within a family across question variants.

Table B.2: Planned target-model choice calls per model. Counts exclude analyzer calls. Each formal material/effort cell contains 48 choices.

Block	Calculation	Calls
G shared probes	2 reliabilities × 2 roles × 3 identities × 3 templates × 2 efforts × 48	3,456
G development	16 families × 4 evidence conditions × 3 identities × 2 efforts × 48	18,432
G test	32 families × 4 evidence conditions × 3 identities × 2 efforts × 48	36,864
H shared probes	2 roles × 3 identities × 3 templates × 4 contexts × 2 efforts × 48	6,912
H evaluation	16 families × 3 evidence conditions × 2 identities × 4 contexts × 2 efforts × 48	36,864
Formal total	Sum of the five blocks	102,528

B.4 Number of model calls

The separate protocol-check dataset contains 18,784 choice calls per model, and engineering preflight contains 32. Repeated calls within a family are not independent family-level sample units.

B.5 Matching baselines and selecting strengths

For C3/C4, verbatim extracted evidence spans are removed only if each can be matched unambiguously. The remaining profile, original question, mechanism, model effort, and output protocol must identify a complete measured local E0 cell. Shared-reference fallback is permitted only when the retained background is fully covered by the available shared probes. Otherwise the method returns C0. Shared probes average the measured distributions over their three templates. Inputs classified as E0 or with unknown evidence have no validated evidence-adjustment route.

The analyzer classifies evidence and context using input content. It is not allowed to emit a preferred-rule prediction or access target-model answers. The final implementation separates evidence extraction from question/context extraction. This annotation operation is shared across targets, but the probability reference is measured separately for each target.

Table B.3: Frozen strengths. These values are selected before each target's held-out choices.

Model	C1	C2	C3	C4 indirect	C4 explicit
GPT-5.4 nano	0.25	0.25	0.25	0.50	0
DeepSeek V4 Flash	0.50	0.50	0.50	0.50	0

B.6 Model snapshots and request settings

The GPT target snapshot is `gpt-5.4-nano-2026-03-17`. The DeepSeek target snapshot is `deepseek/deepseek-v4-flash-0731`, using the fixed `parasail/fp8` route through OpenRouter with provider fallback disabled. Both targets receive the same rendered task materials. Native `none` and `high` settings are compared without equating their computational budgets. The shared input analyzer is GPT-5.4 mini at `high`.

Target requests use a 16,384-token output cap and omit temperature, top-p, and sampling seed. The resulting repeated-choice distributions therefore depend on provider defaults. Formal adjustment uses the R0 schema. R1 ordinary rationale requests, R2 visible step-by-step requests, language variants, task framing, and rotated semantic-letter mappings belong to the protocol checks described in Appendices A and D.

C Analysis of the adjustment

C.1 Derivation of the exact-recovery condition

Section 4.2 states the exact-recovery condition in Equation 2. Let p be the response distribution, b the baseline, and r the target. Assume all entries are strictly positive and fix $\lambda \geq 0$. Equation 1 implies

$$\frac{q_\lambda(j)}{q_\lambda(k)} = \frac{p(j)}{p(k)} \left(\frac{b(k)}{b(j)} \right)^\lambda .$$

For necessity, substitute $q_\lambda = r$, take logarithms, and rearrange to obtain Equation 2. For sufficiency, that condition makes every pairwise ratio of q_λ and r equal. Both vectors sum to one, so $q_\lambda = r$. This argument holds for any finite number of answers.

The condition has a direct interpretation: the baseline log odds must describe the error remaining in p relative to r, at the chosen strength. Measuring b alone does not establish this relationship. This is a condition for *exact* recovery, and error can decrease without satisfying it. Explicit one-hot targets fall outside the positive-target assumption and are evaluated directly by their scores.

C.2 Proof of the identity result

Proof of Proposition 1. Write the normalizers for the two identity variants as Z_a and Z_b. If $p_a = p_b = p$, then

$$\frac{q_a(k)}{q_b(k)} = \frac{Z_b}{Z_a} \left(\frac{b_b(k)}{b_a(k)} \right)^\lambda .$$

If $q_a = q_b$, the baseline ratio must be constant in k, since $\lambda > 0$ and every entry is positive. Both baselines sum to one, so they must coincide. Substituting equal baselines into Equation 1 proves the converse. The result requires both variants to receive the stated adjustment at the same strength. Leaving one variant unchanged through fallback would be a different comparison.

C.3 When the original predictions are close

The same reasoning also describes differences that are small rather than exactly zero. For any $K \geq 2$, let $\ell(p)$ contain the $K(K-1)/2$ pairwise log odds $\log[p(j)/p(k)]$, with $j < k$. Define the original prediction difference $u = \ell(p_a) - \ell(p_b)$ and baseline difference $v = \ell(b_a) - \ell(b_b)$. Then

$$\ell(q_a) - \ell(q_b) = u - \lambda v. \tag{C.1}$$

The reverse triangle inequality gives

$$\|\ell(q_a) - \ell(q_b)\|_2 = \|u - \lambda v\|_2 \geq \lambda \|v\|_2 - \|u\|_2. \tag{C.2}$$

If $\lambda\|v\|_2 > 2\|u\|_2$, the adjusted log-odds distance exceeds the original distance. This is a sufficient condition, not a necessary one. At exact invariance, $u = 0$, so the new distance is $\lambda\|v\|_2$.

This statement concerns log odds, which can be sensitive to small probabilities. It is not a lower bound on TV. We do not claim that the measured identity pairs satisfy the sufficient condition, and the empirical TV comparisons remain separate measurements.

C.4 A four-answer example

Let the target and both original response distributions be $(0.25, 0.25, 0.25, 0.25)$, but let their baselines differ:

$$b_a = (0.7, 0.1, 0.1, 0.1), \qquad b_b = (0.1, 0.7, 0.1, 0.1).$$

At $\lambda = 1$, Equation 1 returns

$$q_a = (1, 7, 7, 7)/22, \qquad q_b = (7, 1, 7, 7)/22.$$

The original predictions are identical and equal to the target. Adjustment increases TV from zero to $3/11 \approx 0.2727$ and expected Brier error from $3/4$ to $3/4 + 27/484 \approx 0.8058$. This is an illustrative construction, not model data. Selecting $\lambda = 0$ would avoid both changes.

C.5 Connection to established calibration methods

Table C.1 separates the measured quantity, reference input, and operation. Sharing a ranking rule does not make the probability vectors or input definitions interchangeable.

Table C.1: Calibration methods related to the studied adjustment. With the same positive numerical inputs, the first three rows can rank answers by a ratio of answer and reference scores. Their probability outputs and reference constructions differ.

Method	Measured quantity and reference	Operation or distinction
Contextual calibration (Zhao et al., 2021)	Label probabilities; content-free input such as N/A	Inverse-diagonal scaling followed by softmax; same ratio ranking, different normalization.
Domain-context calibration (Fei et al., 2023)	Label probabilities; average over random in-domain text	Divides label probabilities by a domain reference to choose an answer.
Domain-conditional PMI (Holtzman et al., 2021)	Answer-sequence likelihoods; domain premise	Ranks likelihood ratios using a domain reference.
Batch Calibration (Zhou et al., 2024)	Output scores; input-containing batch	Centers scores using batch estimates, rather than the evidence-removed reference tested here.
This study	Smoothed response frequencies; matched E0 probes	Normalizes $p(k)/b(k)^{\lambda}$, using a development-selected strength.

For contextual calibration, let s denote the input label-probability vector and c the content-free vector. The diagonal variant of Zhao et al. (2021) sets $W = \mathrm{diag}(c)^{-1}$ and additive bias to zero, returning $\mathrm{softmax}(Ws)$. Its decision satisfies

$$\arg\max_k [\mathrm{softmax}(Ws)]_k = \arg\max_k \frac{s_k}{c_k} = \arg\max_k \frac{s_k/c_k}{\sum_j s_j/c_j}.$$

The first and last probability vectors generally differ: $\exp(s_k/c_k)/\sum_j \exp(s_j/c_j)$ is not $(s_k/c_k)/\sum_j (s_j/c_j)$. Thus our Brier results do not evaluate contextual calibration's probability outputs. The ratio decision rule is also explicit in Fei et al. (2023), while Holtzman et al. (2021) apply a related ratio to sequence likelihoods.

Zhou et al. (2024) already show that content-free inputs can introduce additional bias, including in N/A-based natural-language inference. Their batch estimator retains actual inputs. Our study adds a controlled comparison across evidence conditions, using known targets and matched irrelevant country labels. It makes no numerical performance comparison with Batch Calibration.

C.6 When a ratio adjustment has a generative justification

Under *prior shift*, source and target populations have different class priors, $\pi_s(k)$ and $\pi_t(k)$, but share the class-conditional likelihood: $P_s(x \mid k) = P_t(x \mid k)$. Bayes' rule gives

$$P_t(k \mid x) = \frac{P_s(k \mid x)\, \pi_t(k)/\pi_s(k)}{\sum_j P_s(j \mid x)\, \pi_t(j)/\pi_s(j)}.$$

This is the prior-adjustment setting of Saerens et al. (2002). With a uniform target prior and $b = \pi_s$, Equation 1 at $\lambda = 1$ gives this adjustment. The assumptions explain why it is justified in that setting. A model's answers without evidence, however, can reflect answer selection and prompt interpretation as well as prior information. They are not automatically the class prior required by this derivation.

D Additional synthetic results and robustness checks

Tables D.1–D.3 report a separate protocol check, country-label sensitivity, and explicit-preference outcomes. Figure D.1 plots the country-label comparison, and Figure D.2 visualizes the primary

comparisons in Table 3. The analyses below examine their sensitivity to annotation, coverage, and parameter choices.

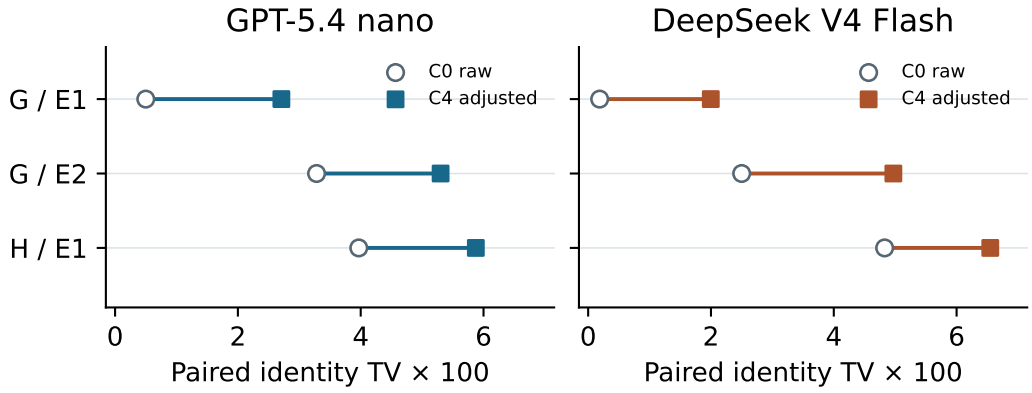

Figure D.1: Matched U.S./China response differences before and after adjustment. Smaller mean paired TV means less sensitivity to the country label. G/E1 and G/E2 contain one and six noisy records, and H/E1 is a context task without a unique accuracy target. Points are descriptive estimates.

Table D.1: Country-label TV in a separate protocol check, multiplied by 100. Each entry is descriptive, based on 16 families and eight choices per cell in the separate 16-family evidence-ladder block. This check uses indirect evidence without G's disclosed record mechanism and has no unique probability target.

Model	Effort	E0	E1	E2	E3
GPT	`none`	17.50	1.87	5.62	0.00
GPT	`high`	38.12	0.00	2.50	0.00
DeepSeek	`none`	26.25	2.50	6.25	0.00
DeepSeek	`high`	70.00	0.00	2.50	0.00

Table D.2: Mean paired identity TV, multiplied by 100. Complete identity pairs are averaged within family and then across families. These are descriptive estimates, without a significance claim. Identity is explicitly irrelevant in G, and H/E1 has no unique reference preference.

Model	Condition	C0	C4	Pairs / families
GPT	G/E1	0.500	2.708	64 / 32
GPT	G/E2	3.281	5.301	64 / 32
GPT	G/E3	0.188	0.188	64 / 32
GPT	H/E1	3.969	5.875	128 / 16
GPT	H/E3	2.516	2.516	128 / 16
DeepSeek	G/E1	0.188	1.997	63 / 32
DeepSeek	G/E2	2.500	4.970	64 / 32
DeepSeek	G/E3	0.031	0.031	64 / 32
DeepSeek	H/E1	4.828	6.548	128 / 16
DeepSeek	H/E3	0.859	0.859	128 / 16

Table D.3: Explicit-preference preservation check. C4 equals C0 because the selected explicit-evidence strength is zero. The check evaluates preservation under zero adjustment. Modal recovery is a cell-level measure, not a per-response success rate. Scores include the fixed smoothing in Equation 4.

Model	Task	Cells / families	Modal recovery	Brier
GPT	G/E3	192 / 32	100%	0.0013
GPT	H/E3	256 / 16	100%	0.0158
DeepSeek	G/E3	192 / 32	100%	0.0012
DeepSeek	H/E3	256 / 16	100%	0.0044

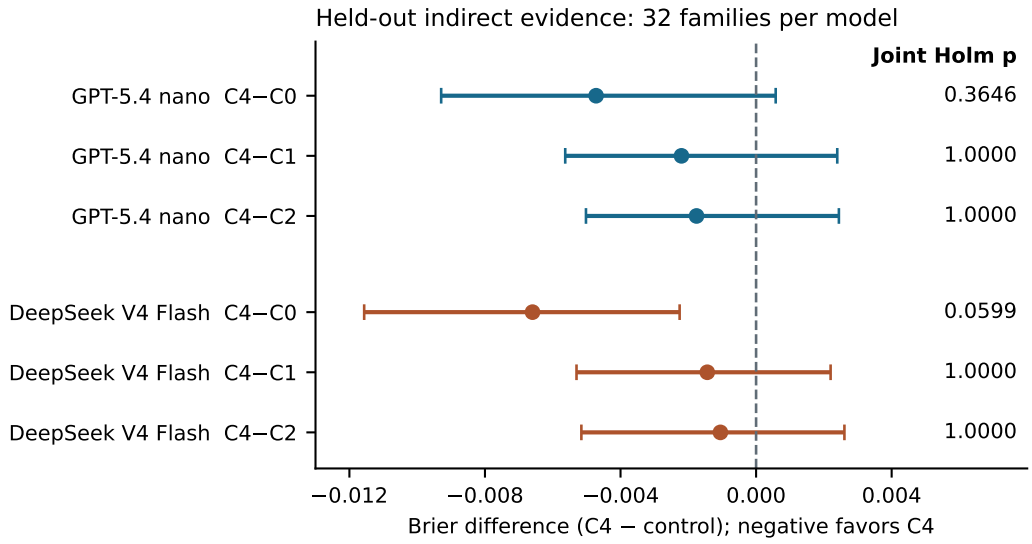

Figure D.2: Held-out G indirect-evidence Brier differences. Error bars are individual unadjusted 95% intervals, and labels give joint Holm-adjusted p-values. Negative differences favor C4. The adjusted criterion is not met in any comparison.

D.1 Absolute accuracy levels and descriptive comparisons

The absolute levels make the scale of the primary differences explicit. Values are family-averaged over held-out G/E1 and G/E2 cells. The posterior floor is the Brier score of predicting the target posterior itself, while the deterministic MAP column is the score of always choosing the posterior mode and is therefore not a probability-matching target.

Table D.4: Absolute held-out G Brier levels and descriptive method comparisons. C0–C4 are family-averaged scores, and lower is better. The final two columns are target-dependent reference points.

Model	C0	C1	C2	C3	C4	Posterior floor	MAP
GPT	0.4122	0.4097	0.4093	0.4098	0.4075	0.3293	0.4588
DeepSeek	0.4113	0.4062	0.4058	0.4047	0.4047	0.3289	0.4583

Relative to C0, the descriptive C1, C2, C3, and C4 differences are respectively -0.0025, -0.0030, -0.0024, and -0.0047 for GPT, and -0.0052, -0.0055, -0.0066, and -0.0066 for DeepSeek. C4–C3 is -0.0023 for GPT and 0.0000 for DeepSeek. These comparisons are outside the prespecified six-test family. Under C0, 72.7% of GPT scored cells and 85.6% of DeepSeek scored cells are unanimous, and the raw modal answer agrees with the posterior MAP rule in 95.3% and 95.8% of cells, respectively. These diagnostics clarify the decision-level behavior without changing the probabilistic primary endpoint.

D.2 Annotation, baseline coverage, and smoothing

Analyzer checks. The final analyzer produces 1,665 annotations from 747 deduplicated calls. Independent agent review yields 334 valid results from 337 inputs, with 99.4–100% agreement across fields. Replacing the selected annotations with construction-derived oracle annotations yields identical stratified outcomes. The original analyzer and oracle variants have their own development-only fits. These comparisons ask whether extraction or routing explains the observed results on the controlled templates. They do not train on test choices or establish the analyzer's advantage over metadata supplied by the task designer.

Coverage. The equal-coverage analysis compares raw and adjusted predictions on the same retained conditions. Fallback is part of the implemented method, rather than a reason to silently exclude an input. Scored-cell nonresponse is handled separately in Appendix E.

Smoothing. The analysis evaluates $\alpha \in \{0.1, 0.5, 1\}$ while holding strengths fixed. For GPT G/E2/`none`, C4-C0 is approximately +0.0006, -0.0010, and +0.0011 at those values. For DeepSeek G/E2/`high`, it is approximately -0.0093, -0.0046, and +0.0010. These changes show that some subgroup signs depend on the frequency estimator. They are descriptive robustness results, not additional adjusted claims or a basis for choosing a favorable smoothing constant. The primary analysis retains $\alpha = 0.5$.

Parameter exchange. Applying another model's frozen strengths while retaining the target's own measured baseline is a limited parameter-transfer diagnostic. Since C4's strengths coincide across models, exchanging C4 leaves the results unchanged. C1–C3 differ and admit nontrivial exchanges, but the paper's primary claim does not depend on a successful cross-model transfer result.

Scope of method selection. The reported strengths, primary evidence grouping, and primary smoothing remain those frozen for G/H. Suggested adjustments conditioned directly on evidence, additional models, and other application domains are future tests. They are not included among completed experiments.

D.3 Post-hoc strength sweep with fixed responses

Table D.5 reuses the held-out G indirect-evidence counts and C4's recorded baseline routes, varying the common indirect strength across the original grid. Smoothing remains $\alpha = 0.5$, explicit strength remains zero, and unavailable routes remain raw. This analysis was added after the primary results were known. It does not refit a method on held-out outcomes, alter the selected strength of 0.5, or supply additional confirmatory hypothesis tests.

Table D.5: Post-hoc strength sweep. Brier change is family-averaged adjustment minus C0 on the original complete scored cells. TV is the equal average of the G/E1 and G/E2 paired family means, and its change is multiplied by 100. All entries are point estimates. The development-selected strength is 0.50 in both targets.

	GPT		DeepSeek	
λ	Brier change	$100\,\Delta T$	Brier change	$100\,\Delta T$
0.00	$+0.0000$	$+0.000$	$+0.0000$	$+0.000$
0.25	-0.0024	$+0.571$	-0.0030	$+0.762$
0.50	-0.0047	$+2.114$	-0.0066	$+2.140$
0.75	-0.0015	$+5.030$	-0.0082	$+4.445$
1.00	$+0.0207$	$+9.400$	$+0.0012$	$+7.979$

At strengths 0.25, 0.50, and 0.75, favorable Brier point estimates coexist with larger paired TV in both targets. At 1, both measures worsen relative to raw responses. The opposite movement of the two measures is therefore not confined to the selected strength in this finite grid, but no claim follows about every positive strength, population significance, alternative reference estimators, or other sampling regimes. DeepSeek's more favorable Brier estimate at 0.75 is not used to replace its frozen parameter.

D.4 Exploratory perturbation pilots

These two later pilots extend the measurement question in Section 3: how much do answers depend on expression, and how much on the background supplied for the simulated person? They retain the original four-rule question and definitions. They do not apply C1–C4 or alter the formal G/H evaluation. Designs and request plans were fixed before each pilot's calls, after the G/H results were available. Accordingly, these are exploratory diagnostics, not additional confirmatory tests.

Common protocol and accounting. Both pilots use the same two model snapshots and provider routes as G/H, native `none`/`high` effort, R0, and a 16,384-token output cap (Appendix B.6). Requests are independent conversations and omit `temperature`, `top_p`, and sampling `seed`. Four cyclic display orders, ABCD/BCDA/CDAB/DABC, are balanced, and the letter-to-rule mapping stays fixed. Averaging means averaging four-dimensional choice indicators, giving empirical choice frequencies without smoothing, rather than averaging category numbers. These orders control position but do not separate letter from semantic preferences. Native effort settings are not equal-compute conditions across models.

The surface pilot completes 1,536 valid final choices in 1,547 attempts, and the background pilot completes 10,752 in 11,018 attempts. All planned cells are complete, with no unresolved submissions. Final raw responses were reparsed and attempt hashes checked in the source audits. Repeated calls increase resolution within a profile, but the independent descriptive units remain eight profiles or eight synthetic families. The accompanying analysis files retain all 12,288 final choices and the pairing information.

D.4.1 Surface wording and equal-budget repeatability

Using a fixed construction seed, we randomly sample four U.S. and four Chinese public profiles from the eligible M0 materials, and add one synthetic explicit-preference control for each rule. Within each profile and model/effort setting, the original prompt receives 16 calls and four variants receive four calls each, also totaling 16. Each variant changes one surface feature: (i) replace the outer "Profile" heading with "Person profile" and add "Question" and "Available rules" headings; (ii) paraphrase the role instruction while retaining simulation, limited evidence, and no invented biography; (iii) paraphrase the generic group-decision question; or (iv) replace option-label periods with closing parentheses and put blank lines between options. Profile text and rule definitions remain verbatim. The exact templates, including both paraphrases, are in the accompanying prompt catalogue.

For repeatability, each arm is partitioned into four disjoint four-call batches. Each batch contains all four display orders, and each perturbed batch also contains all four variants, paired to orders by a

fixed Latin rotation. For profile i and arm a, define

$$V_{ia} = \frac{1}{6} \sum_{1 \le b < c \le 4} \mathrm{TV}(\widehat{p}_{iab}, \widehat{p}_{iac}), \qquad \Delta V = \frac{1}{8} \sum_{i} (V_{i,\mathrm{variant}} - V_{i,\mathrm{original}}).$$

Lower V means that these equal-budget batch frequencies are closer. The separate distribution-shift statistic is the mean profile TV between the two full 16-call arms. It includes finite-call noise and is not an estimate of accuracy or a pure wording effect.

Table D.6: Surface pilot, eight public profiles. TV entries are multiplied by 100. ΔV is variant minus original batch TV, in percentage points, and brackets give conditional 95% profile-bootstrap intervals. Controls are excluded from these means.

Model	Effort	Arm-shift TV	Original V	Variant V	ΔV	95% interval
GPT	`none`	28.12	22.92	23.96	$+1.04$	$[-6.77, +10.94]$
GPT	`high`	9.38	18.23	19.79	$+1.56$	$[-3.12, +5.73]$
DeepSeek	`none`	19.53	32.29	29.69	-2.60	$[-6.77, +1.56]$
DeepSeek	`high`	7.03	15.10	12.50	-2.60	$[-6.77, +1.04]$

All four intervals include zero (Table D.6), so this pilot does not establish a repeatability advantage for averaging these variants. It also does not establish that averaging is ineffective in general. Centralized-leadership shares change from 30.47% to 37.50% and 32.81% to 38.28% for GPT at `none`/`high`, and DeepSeek changes from 25.00% to 26.56% and 36.72% to 42.19%. All four explicit controls are recovered in every setting and arm (64/64 choices per arm and setting). Real private preferences are unobserved, so neither these shifts nor closer repeated responses establish better preference inference. Single-variant diagnostics are retained in the accompanying analysis files; their four-call distributions and the original batches' inclusion in the 16-call baseline preclude interpreting their raw TV differences as isolated wording effects.

Intervals use 2,000 paired profile-bootstrap draws, stratified to retain four profiles from each country with a fixed seed. They are conditional on the observed calls and do not capture the full uncertainty of new API sampling. No multiple-testing-adjusted or population-level claim is made from eight profiles.

D.4.2 Role, nationality, and nested party labels

The second pilot uses eight fictional profile families with neutral occupational descriptions and no names. A family holds its occupation and procedural records fixed while background labels vary. Its 14 backgrounds comprise two institutional roles (a private citizen holding no public office; a member of the national legislature) crossed with: unspecified nationality and unspecified party; U.S. nationality with unspecified, Democratic, or Republican affiliation; and Chinese nationality with unspecified, Communist Party of China (CPC), or China Democratic League (CDL) affiliation. "Unspecified" means not provided, not no affiliation. Party labels are nested within nationality; their contrasts do not define a common cross-national ideological scale. These are manipulations of model inputs, not evaluations of the parties or of actual members.

Each background appears under three descriptively named conditions: *no records*, *two consistent endorsements*, and *endorsements plus explicit preference*. In the second condition, two constructed association-meeting records support the same procedure. For example, one says the person "supported requiring at least two-thirds of members to approve a proposal"; the other says they "supported a two-thirds approval threshold before a proposal could pass." The third condition adds an explicit current preference for that procedure in the question's group decisions. Two families target each of the four procedures, with occupations and targets shuffled using fixed construction seeds. The second condition supplies clear, repeated procedural support, not independent noisy observations, and is not G/E1. Neither the first nor the second condition specifies a unique ground-truth current preference. Only the explicit condition supports a recovery-accuracy calculation.

Every family/background/condition/model/effort cell has four display orders and two repetitions, giving $8 \times 14 \times 3 \times 4 \times 2 \times 2 \times 2 = 10{,}752$ calls. The four primary background contrasts are: Chinese minus U.S. nationality at unspecified party, averaging roles; legislator minus private citizen at unspecified party, averaging all three nationality levels; Republican minus Democratic affiliation,

averaging U.S. roles; and CPC minus CDL affiliation, averaging Chinese roles. Contrasts are formed within family before averaging. Figure D.3 shows all 48 primary mean-family TV estimates, including both record conditions, so the no-record differences are not presented alone.

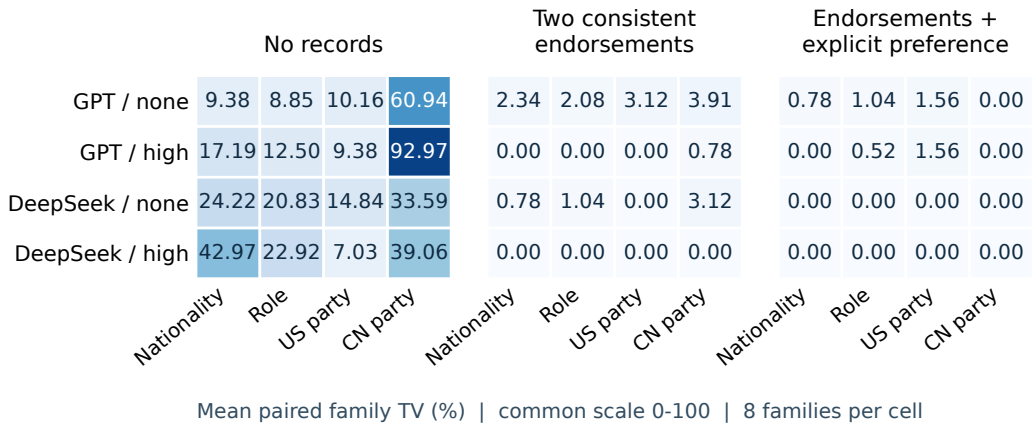

Figure D.3: Background-label sensitivity in the synthetic pilot. Each cell is 100 times mean paired family TV for the indicated contrast, based on eight families. Labels "US party" and "CN party" denote the within-country contrasts defined in the text. All panels use the same 0–100 scale. Values are model-response distances, not scores assigned to people or parties. Clear procedural endorsements sharply reduce the observed differences. Sampling noise remains part of empirical TV, so a zero is not a guarantee of invariance.

Table D.7: No-record background contrasts. ΔD is the signed change in the model's centralized-leadership share, in percentage points, with conditional 95% family-bootstrap interval. $L^2_\times$ is the cross-repeat squared-L^2 diagnostic in probability units, not TV, and negative estimates are retained.

Model	Effort	Contrast	ΔD	95% interval	$L^2_\times$
GPT	none	Nationality	$+1.56$	$[+0.00, +3.91]$	-0.0059
GPT	none	Role	$+1.56$	$[+0.00, +3.65]$	-0.0035
GPT	none	US party	$+3.12$	$[-0.78, +7.81]$	$+0.0059$
GPT	none	CN party	$+54.69$	$[+46.88, +63.28]$	$+0.6230$
GPT	high	Nationality	$+3.91$	$[+0.78, +7.03]$	$+0.0391$
GPT	high	Role	$+2.60$	$[+0.52, +4.69]$	$+0.0182$
GPT	high	US party	$+0.78$	$[+0.00, +2.34]$	$+0.0137$
GPT	high	CN party	$+92.97$	$[+89.06, +96.88]$	$+1.3164$
DeepSeek	none	Nationality	$+14.06$	$[+9.38, +18.75]$	$+0.0703$
DeepSeek	none	Role	$+11.98$	$[+7.29, +18.23]$	$+0.0417$
DeepSeek	none	US party	$+6.25$	$[+3.12, +9.38]$	$+0.0117$
DeepSeek	none	CN party	$+32.81$	$[+28.91, +35.94]$	$+0.1270$
DeepSeek	high	Nationality	$+39.06$	$[+33.59, +45.31]$	$+0.3105$
DeepSeek	high	Role	$+22.92$	$[+20.31, +25.52]$	$+0.0799$
DeepSeek	high	US party	$+0.78$	$[+0.00, +2.34]$	$+0.0156$
DeepSeek	high	CN party	$+39.06$	$[+28.91, +49.22]$	$+0.2715$

The largest no-record response shift for GPT is the within-China party contrast (Table D.7). At `high`, its centralized-leadership change is +92.97 points, compared with +39.06 for DeepSeek. DeepSeek's nationality and institutional-role contrasts at `high` have mean-family TV of 42.97% and 22.92%, versus 17.19% and 12.50% for GPT. These are estimates for the chosen fictional families and labels, not population characteristics. With two consistent endorsements, every primary contrast is between 0 and 3.91% mean-family TV. The latter result concerns clear procedural support and does not establish robustness to conflicting or ambiguous records.

To diagnose finite-call noise, split each cell into its two four-order repeats and form the same contrast vector $\widehat{d}_{ir}$ in each repeat. Report $L^2_\times = 8^{-1}\sum_i \widehat{d}_{i1}^{\mathsf{T}}\widehat{d}_{i2}$. Under independent, mean-zero sampling errors across repetitions, its expectation is the mean squared L^2 norm of the order-averaged contrast. It can be negative and is not subtracted from TV. Same-prompt repeat TV, full four-rule contrast vectors, the six supplemental label-versus-unspecified contrasts, three role interactions, and paired cross-model differences are included in the accompanying analysis files. For example, DeepSeek's no-record, `high` nationality-by-role difference-in-differences in centralized-leadership share is +68.75 points, indicating that its nationality response varies with the institutional role. Cross-model interpretation uses direct paired contrasts, not a comparison of separate significance decisions.

The background intervals use 2,000 paired family-bootstrap draws with a fixed seed, conditional on observed responses. The eight families, rather than the thousands of calls, determine the resampling unit. The intervals and noise diagnostics are exploratory, and they are separate from the hierarchical G/H bootstrap and its six-test Holm family.

D.4.3 Averaging backgrounds changes the target distribution

An equal-budget diagnostic compares eight calls to the baseline description (unspecified nationality and party, private citizen) with one preselected call from each of eight named-party backgrounds (two countries, two roles, and two parties per country). Both arms have two calls at each display order. The eight background/order selections were fixed before collection. The resulting mixture gives equal weight to different descriptions, so it does not estimate the original person's true preference.

Table D.8: Background averaging and explicit-preference recovery. The first three numerical columns use no-record cells and the fixed eight-versus-eight-call comparison, and D shares and TV are percentages. The last column uses all explicit-condition choices across 14 backgrounds, with 896 choices per setting. These are different analyses and denominators.

Model	Effort	Baseline D	Mixture D	Arm-shift TV	Explicit recovery
GPT	`none`	0.00	21.88	32.81	890/896 (99.33%)
GPT	`high`	0.00	28.12	35.94	891/896 (99.44%)
DeepSeek	`none`	1.56	37.50	45.31	896/896 (100.00%)
DeepSeek	`high`	0.00	39.06	40.62	896/896 (100.00%)

Table D.8 shows that the chosen mixture changes output shares substantially without records. A single eight-versus-eight-call comparison does not demonstrate reduced estimator variance. Although overall order is balanced, each named background contributes only one order, so background-by-order interactions are not fully averaged. The accompanying analysis also reports a 64-call mixture using all observations for each named background, but its comparison with the eight-call baseline is not an equal-budget test. Near-uniform aggregate shares after endorsements partly reflect the balanced assignment of two families to each target and do not establish mitigation.

Explicit-preference recovery is 890/896 and 891/896 for GPT at `none`/`high`, and 896/896 for DeepSeek in both settings. All 11 GPT errors choose simple majority instead of stated unanimity. Their occurrence in a labeled background alone does not identify label override. Across these diagnostics, background sensitivity may describe a change in the simulated person rather than an inference error. This is why the paper retains independently specified targets and relevance assumptions in G/H as the basis for evaluating mitigation. The pilots neither test the exact-recovery condition nor validate the strictly positive-distribution assumptions of Proposition 1.

E Statistical procedure and nonresponse

E.1 Aggregation and bootstrap

For a method contrast in family f, let J_f contain complete eligible cells and write

$$d_f = \frac{1}{|J_f|}\sum_{j\in J_f}\left[\mathrm{BS}(q_{f,j}, r_{f,j}) - \mathrm{BS}(p^{\mathrm{control}}_{f,j}, r_{f,j})\right], \qquad \widehat{\Delta} = \frac{1}{F}\sum_{f=1}^{F} d_f.$$

This gives equal weight to families even when one family has a missing cell. Each bootstrap iteration first resamples repeated answers within fixed option order, recomputes smoothed scored and background distributions, applies the frozen adjustment, and obtains family contrasts. Families are then resampled with replacement. Background draws are reused across every method and family sharing that probe. Percentile quantiles at 0.025 and 0.975 form the reported individual intervals. For an observed contrast $\widehat{\Delta}$ and $B = 1000$ bootstrap values Δ_b^*, the two-sided centered test is

$$\widehat{p} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{|\Delta_b^* - \widehat{\Delta}| \geq |\widehat{\Delta}|\}}{B + 1}.$$

Holm adjustment is applied to the six primary p-values together. In sorted order, adjusted values are the cumulative maximum of the remaining-family-size multiplier times each unadjusted value, truncated at one. The unadjusted bootstrap confidence intervals are not simultaneous confidence intervals obtained by inverting that procedure. Table 3 explicitly reports the adjusted decision to prevent confusion between the two summaries.

The intervals condition on the frozen strengths and do not include uncertainty from rerunning development selection. Two repetitions per option order limit the resolution of within-order resampling. No prospective power analysis was saved.

E.2 Prospective boundaries

The final joint reporting rule was saved before held-out choices began and specifies the six-comparison family, accuracy criteria, explicit-evidence margins, and completeness condition. Parameters are fitted separately in each model and frozen before that model's first test choice. Analyzer selection and the stored oracle/original-analyzer development fits also precede held-out collection. These facts support the prespecified status of the final comparisons. They do not convert the public-profile exploration, the interpretation of results, or operational changes into a single original preregistration. The two perturbation pilots were collected after the formal G/H evaluation. Their own designs and request plans were fixed before their respective calls, but they are subsequent exploratory diagnostics, not part of the original six-comparison test or an external preregistration. They did not change the selected strengths or success criteria.

E.3 Nonresponse handling

DeepSeek development collection contains four incomplete E0/high probe cells, while every one of its 288 scoring cells is complete. A separate nonresponse amendment was frozen before DeepSeek's first held-out request but after GPT's test collection had begun. Under that amendment, missing model answers are retained and not resampled, and a probe is usable only if its repeated choices are complete. If any development scoring cell were incomplete, the selection procedure would conservatively revert to zero adjustment rather than fit after excluding the failure. In the observed data, the original fit on all scoring cells is retained and unavailable probes are handled through the original coverage rules.

The 14 formal DeepSeek nonresponses affect 13 probe requests and one scored request. The complete-cell primary endpoint therefore has 383 scored cells. Completing the one missing scored answer in all four possible ways gives the deterministic range reported in Section 5, using actual probe availability and frozen strengths. No completed value is substituted into the primary dataset. This exercise addresses the missing scored answer under specified assumptions. It cannot recover unobserved responses from unavailable probes or establish that missingness is ignorable.

E.4 Transport interruptions

During H, 52 DeepSeek requests encounter connection interruptions: 49 retain only blank heartbeats and three retain no response, with no queryable response identifier. After inspecting the saved envelopes, one recovery within the original attempt budget is permitted for this group. The original attempts and unknown costs remain in the experimental record. Absence of a received answer does not establish that the server never generated one. This later operational intervention may select on response availability and is not described as a preregistered scientific change.