

Political identity drives choice of large language models—even when accuracy is incentivized

Michael Heseltine¹, Christopher Bail², Petter Törnberg³, Michelle Schimmel³ and Christopher Barrie^{4,1}

¹Department of Sociology, University of Oxford, Oxford, UK

²Department of Sociology, Duke University, Durham, NC 27516, USA

³Institute for Logic, Language and Computation, University of Amsterdam, Netherlands

⁴Department of Sociology, New York University, NY 10012, USA

May 27, 2026

Abstract

Research on artificial intelligence (AI) has focused heavily on bias in model outputs, but less is known about how users’ own political identities shape which AI systems they use. Because of the unprecedented growth of this new technology, the choices users make between competing systems may soon begin to shape the information environment more broadly. We address this question with a two-stage online experiment ($N = 1,884$). Participants first evaluated outputs from four AI models that varied in ideological tone. They were then offered financial incentives to answer a question about politics accurately. We experimentally varied whether models were identified by real-world brand names (e.g. Claude or Grok) or neutral labels—and whether participants were initially exposed to political or nonpolitical content. Participants quickly developed preferences for AI systems that aligned with their political identities, and these preferences were stronger when models carried recognizable brand names. In the second stage, 71% persisted with their previously preferred model despite incentives for correctness. These findings show that users do not treat AI systems as neutral tools. Instead, they select between them in ways that reflect political identity, suggesting that as the information ecosystem fragments, competitive pressures may push AI systems to differentiate themselves along political lines.

Introduction

Artificial intelligence (AI) has become a key tool people use to access information, interpret events, and make decisions [21]. A large body of research examines bias exhibited by AI technologies [42, 9, 10, 2, 28], yet relatively little is known about how users’ political identities shape which AI systems they prefer. Given the rapid growth of AI systems, the consequences of these choices extend beyond individual experiences. If different users come to rely on different AI systems, the information environment may become more fragmented [43, 1]. This raises a fundamental question: Do users select AI systems in the same way they do other information systems—such as search engines that prioritize accuracy and performance—or do they choose them in a manner similar to how they select media sources: favoring those that align with their identities and beliefs?

Existing research offers conflicting expectations. A large literature in political behavior shows that individuals engage in selective exposure and motivated reasoning, preferring ideologically congruent sources and interpreting information in ways that reinforce prior beliefs [33, 15, 37, 11, 20, 14]. In fragmented information environments, these dynamics have been linked to audience segmentation, partisan differentiation, and disagreement over which sources are credible [36, 11, 8, 32, 19, 15, 35, 29]. These patterns are closely tied to partisan identity, which functions as a stable psychological attachment shaping how individuals evaluate competing claims [4, 6].

At the same time, a parallel literature in judgment and decision-making demonstrates that individuals can respond to incentives for accuracy, particularly in factual domains [24, 18]. However, this capacity is uneven and contingent. The extent to which individuals prioritize accuracy depends on both cognitive style and motivation, and even financial incentives do not fully eliminate partisan differences in belief and evaluation [26, 27, 31]. These dynamics may be especially relevant for AI systems because users increasingly interact with large language models not simply as information retrieval systems, but as anthropomorphized conversational and quasi-social agents that provide advice, support, and interpretation [5]. Recent work suggests that large language models may help people find common ground by generating consensus statements that broker disagreements in small group settings [38]. Yet this optimistic view rests on the assumption that users prioritize accuracy when choosing between systems, selecting systems they expect to provide correct or reliable information. If, instead, users sort into models along identity-consistent lines, these systems may reinforce rather than reduce fragmentation.

Thus, AI systems sit at the intersection of these logics. They are used instrumentally for answering factual questions, interpreting and generating content, summarizing information, and providing advice [5, 13]. At the same time, they differ from prior information systems in that they combine features of search engines, conversational interfaces, and media-like outputs, presenting synthesized responses that vary in tone, framing, and perceived bias. As distinct systems develop recognizable reputations and public identities, users may come to associate them with particular viewpoints or epistemic styles [44, 17]. This creates the possibility that AI systems function not only as neutral information providers, but also as competing information sources that users select between based on both perceived performance and ideological alignment [16, 34, 30, 25]. Compared to traditional media, these dynamics unfold in highly individualized and private interaction settings, reducing opportunities for social correction and potentially amplifying the role of prior beliefs in shaping selection and interpretation.

Despite the rapid diffusion of AI systems, we lack direct evidence on how users navigate this trade-off when choosing between models. Existing research has primarily examined either the properties of AI-generated outputs or how individuals evaluate those outputs, rather than how users select among competing systems [7, 12]. Model choice, however, poses a distinct problem: it depends not just on how users judge AI-generated information, but also on which sources they are willing to rely upon. If users treat AI systems as neutral information systems, differences in perceived accuracy should drive selection, particularly when accuracy is incentivized. If they instead treat them as identity-aligned information sources, selection may reflect expressive motivations tied to political identity, even at a cost to accuracy.

We address this question using a two-stage experiment, outlined in Figure 1. Participants first evaluate responses from multiple AI systems and then select a single system to answer a factual question under a monetary incentive for correctness. We experimentally vary whether models are labeled with actual brand names (e.g., Claude, GPT) or neutral identifiers, and whether participants are first exposed to political or non-political content. This design allows us to distinguish between

preferences formed through direct interaction with model outputs and those activated by prior beliefs about model identity, and to assess whether such preferences persist when accuracy, defined here as selecting a system expected to provide correct information, is incentivized.

We find that users rapidly form structured preferences over AI systems that align with their political identities when evaluating political content. Thus, we argue, the choice of which AI systems to use may, ultimately, emerge as a new axis of polarization.

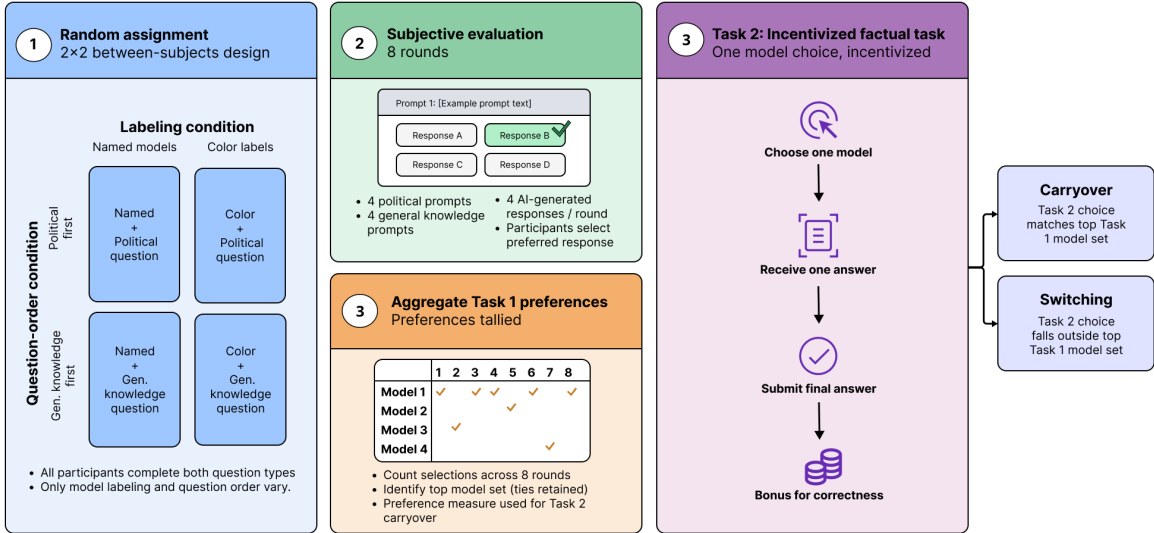


Figure 1: Experimental design flowchart

1 Methods and Measures

1.1 Participants and Procedure

Participants (N of 1884) were recruited through Prolific (using a U.S. demographically representative sample) and completed a two-stage experiment administered via Qualtrics. Participants were randomly assigned to one of four conditions in a 2x2 between-subjects design that varied (i) whether AI systems were labeled with real-world names or neutral color labels and (ii) whether participants were first exposed to political or general knowledge questions. All participants completed both experimental tasks and a post-task survey measuring demographics and political attitudes.

1.2 Experimental Design

The experiment consisted of two sequential tasks. In Task 1, participants evaluated responses from four AI systems that differed systematically in ideological tone. The systems corresponded to four underlying models: Gab AI (far-right leaning), xAI Grok (conservative), OpenAI GPT (relatively neutral), and Anthropic Claude (liberal). Depending on condition, these models were presented either with their real names or as randomly assigned color-labeled systems (e.g., “Purple model,”

“Blue model”—see Appendix for screenshot examples). The mapping between labels and underlying models was randomized at the participant level and held constant throughout the study.

Participants viewed eight prompts in total: four political and four general knowledge questions. Political prompts were designed to elicit responses with clear ideological framing differences across models, while general knowledge prompts produced substantively similar responses across models. Examples include “How common is voter fraud in the United States?” for the political prompts and “What causes earthquakes?” for the general-knowledge prompts. The responses to each question appeared after users pressed a “generate response” button. Participants then selected their preferred response and proceeded to the next prompt, repeating this process until they had viewed all eight prompts and made eight selections. The order of prompt types was experimentally manipulated such that participants saw either political questions first followed by general knowledge questions, or the reverse.

In Task 2, participants selected one of the four AI systems to answer a factual question (“by how many votes did Joe Biden win the national popular vote in the 2020 election?”). They were informed, before seeing the question, that the accuracy of their answer would determine a bonus payment, creating an incentive for selecting a model with highest perceived accuracy. Importantly, users selected their preferred model and that selection was recorded before users saw the precise question that they need to answer. After selecting their preferred model and advancing to the next page, participants saw the question then pressed a “generate response” button which revealed a pre-programmed response. After this, users then provided a final answer. As our main unit of interest is the model selected and not the final answer, all responses shown by all models to all users were substantively identical and contained the precise answer needed.

1.3 Measures

The primary behavioral measures were participants’ model choices in both tasks. In Task 1, we recorded the selected model for each of the eight prompts. From these responses, we constructed participant-level measures including the share of selections for each model and the most frequently selected model. In Task 2, we recorded the selected model and the participant’s final answer.

We constructed several derived outcomes. Preference consistency was defined as whether the model selected in Task 2 matched the participant’s most frequently selected model in Task 1. Accuracy-oriented switching was defined as whether participants switched from a previously preferred ideologically biased model to a more ideologically neutral model in Task 2 (defined centrally as switching from a non-GPT model to GPT).

Participants also completed survey measures of political ideology, party identification, political interest, and demographic characteristics (including age, gender, and education), which were used as covariates and for subgroup analyses.

In Task 1, model choice is observed at the prompt level and aggregated to the participant level. Let y_{imr} denote an indicator equal to 1 if participant i selects model m in round r . We construct participant-level choice shares as $s_{im} = \frac{1}{R} \sum_{r=1}^R y_{imr}$, where $R = 8$ is the number of rounds. The most-preferred model is defined as the model with the highest value of s_{im} , allowing for ties.

Preference consistency is defined as an indicator equal to 1 if the model selected in Task 2 belongs to the set of most-preferred models from Task 1. This definition is tie-aware and does not force a unique preference when respondents exhibit equal selection frequencies. Switching is defined

as the opposite of consistency. Accuracy-oriented switching is operationalized as an indicator for switching to the GPT model, which is the only model producing ideologically neutral responses in the experimental design.

1.4 Statistical Analysis

All analyses are conducted at the participant level. The primary outcomes are (i) model choices in Tasks 1 and 2, (ii) preference consistency between tasks, and (iii) indicators of switching behavior.

To estimate treatment effects and behavioral relationships, we use linear probability models of the form:

$$Y_i = \alpha + \beta T_i + \varepsilon_i, \quad (1)$$

where Y_i denotes the outcome of interest for participant i and T_i represents treatment indicators (e.g., labeled vs. unlabeled models). For analyses of ideological heterogeneity, we estimate:

$$Y_i = \alpha + \beta_1 \text{Ideology}_i + \beta_2 T_i + \beta_3 (\text{Ideology}_i \times T_i) + \varepsilon_i. \quad (2)$$

For Task 1 analyses, we estimate model-specific regressions at the participant level using choice shares s_{im} as the dependent variable. For Task 2, we estimate binary outcome models for model selection and switching behavior.

As robustness checks, reported in the Appendix (Section F), we estimate equivalent logistic regression models and specifications including demographic and political covariates. These alternative specifications yield substantively similar results.

1.5 Model Selection and Expected Preferences

To directly examine how users form preferences across varying AI systems, we select four widely discussed models that differ in both perceived ideological positioning and public reputation: Gab AI, xAI Grok, OpenAI GPT, and Anthropic Claude. These models allow us to vary both the content of responses and the extent to which model identity is associated with political meaning. These models, in effect, represent an extreme right leaning model, a somewhat conservative model, an ideologically neutral model, and a somewhat liberal model, with these ideological leanings aligning with public ratings of these models [44].¹

Gab AI’s Arya model represents a deliberately right-leaning system designed as an alternative to mainstream platforms, explicitly marketed toward conservative users [40]. As such, we expect it to appeal primarily to right-leaning respondents based on content alignment. However, given its relatively low mainstream name recognition, we do not expect labeling alone to substantially increase its appeal beyond the effects of its responses.

Grok, developed by xAI and associated with Elon Musk, occupies a distinct position. It is widely perceived as a model that is more permissive of right-leaning viewpoints and less constrained by

¹We acknowledge that this design is ideologically asymmetric, but to our knowledge there is no explicitly far-left model by design and thus this extreme dynamic is only testable on the right of the ideological spectrum at this time.

content moderation norms. Public discourse surrounding Grok, including concerns about biased or controversial outputs, has reinforced this perception [3, 22]. Because of its higher name recognition and clearer ideological signaling, we expect Grok to attract right-leaning users both through its responses and through branding, with naming amplifying this effect.

GPT (OpenAI) serves as the baseline or “neutral” model in our design. It is broadly perceived as a general-purpose AI system without a strong ideological identity and is widely adopted across domains. We therefore expect GPT to be acceptable across the ideological spectrum and to function as a default or “safe” option, particularly in contexts where accuracy is incentivized.

Claude (Anthropic) is included as a model that has become associated with left-leaning positions in public discourse. Indeed, public evaluations of Claude output often position it at the more left-leaning end of the ideological spectrum [44]. In addition, recent political attention, including criticism from President Trump and a protracted legal battle over an executive order banning its use by federal employees, has contributed to its perception as aligned with progressive values [41, 39]. We therefore expect Claude to be more appealing to left-leaning respondents, particularly when model names are visible.

Taken together, this set of models allows us to test two key mechanisms. First, whether users form preferences based on the content of model outputs, reflecting perceived ideological alignment. Second, whether model naming activates prior beliefs and reputational associations, amplifying preference sorting. This design enables us to distinguish between behavior driven by direct interaction with model responses and behavior shaped by broader perceptions of model identity.

To ensure systematic variation in ideological tone across responses, we used basic prompt engineering to generate model outputs that reflected distinct ideological positions, including conservative, liberal, and far-right perspectives. Prompts were designed to produce consistent differences in framing across models while holding question content constant. Participants were not informed of this manipulation during the study. At the conclusion of the survey, all respondents were debriefed and informed that the responses they evaluated were generated under controlled prompting conditions.

2 Results

2.1 Preference Formation and Partisan Alignment

Tasks 1 and 2 revealed clear and non-random preferences over AI systems. Figure 2A shows that model choice was systematically structured by party identity in both stages of the experiment. In Task 1, averaging across respondents in the Named and Unnamed conditions, Republicans were 10.3 percentage points more likely than Democrats to select Grok (36.2% vs. 25.9%, $p < 0.001$) and 10.3 percentage points less likely to select Claude (31.6% vs. 41.9%, $p < 0.001$). The pattern was even stronger in Task 2: Republicans were 19.9 percentage points more likely than Democrats to select Grok (33.6% vs. 13.7%, $p < 0.001$) and 16.8 percentage points less likely to select Claude (20.6% vs. 37.3%, $p < 0.001$). Republicans also selected Gab more often and GPT less often than Democrats, while Independents generally fell between the two partisan groups. Omnibus party-by-model interactions were strongly significant in both Task 1 and Task 2 ($p < 0.001$), confirming that model choice was systematically related to political identity.

Figure 2B shows the same pattern in terms of partisan separation. Absolute Democrat–Republican gaps are relatively modest for general prompts, become larger for political prompts, and remain

substantial in Task 2. The main descriptive result is therefore not only that users form preferences over models, but that these preferences become more politically structured when the content itself is political and do not disappear when participants later select a model under an accuracy incentive.

Importantly, these preferences were relatively consistent but not entirely deterministic. On average, in Task 1, respondents selected their most-preferred model 4.6 times out of 8 (median = 4), indicating clear directional preferences without exclusive reliance on a single model. Only a small minority of participants (approximately 4%) consistently selected the same model across all rounds, while most respondents drew on multiple models over the course of the task. This pattern is reflected in the distribution of entropy (mean = 0.89 on a scale with a maximum of 1.39), indicating meaningful but incomplete concentration in model choice. Preferences were also significantly stronger in political contexts, with respondents selecting their most-preferred model more frequently when evaluating political prompts than general knowledge prompts (67.5% vs. 63.1%, paired t -test, $p < 0.001$). Taken together, these results provide evidence, first, that users did not simply uncritically straight-line their model selections across all prompts and, more importantly, that participants formed systematic, identity-aligned preferences over the course of the study.

2.2 Branding and the Amplification of Preferences

We next examine whether these preferences are driven solely by exposure to model outputs or are amplified by prior beliefs about model identity. Figure 2A shows that ideological sorting is already present when models are unnamed, indicating that users can infer meaningful differences directly from the outputs themselves. However, the darker named bars typically diverge more strongly by party than the lighter unnamed bars, suggesting that branding intensifies these underlying preferences.

Figure 2B makes this amplification especially clear. Within the same model, named conditions generally produce larger Democrat–Republican gaps than unnamed conditions, particularly for political prompts and often again in Task 2. Regression analyses confirm that naming systematically shifts model selection in ways consistent with prior expectations about model ideology. For the Claude model, Republicans significantly reduced selection in Task 2 ($\beta = -0.081$, $p < 0.05$) when the model was named. Conversely, Democrats significantly increased selection when Claude was named in both tasks (Task 1: $\beta = 0.066$, $p < 0.01$; Task 2: $\beta = 0.108$, $p < 0.01$).

While the Grok model was not selected with significantly greater frequency among Republicans when named (Task 1: $\beta = 0.020$, $p > 0.1$; Task 2: $\beta = 0.013$, $p > 0.1$), both Democrats and Independents selected the Grok model significantly less when the model was named (Democrats: Task 1 $\beta = -0.056$, $p < 0.01$; Task 2 $\beta = -0.104$, $p < 0.01$; Independents: Task 2 $\beta = -0.173$, $p < 0.01$).

When named, selection of the GPT model increased significantly across almost all task and Party ID combinations (Republicans: Task 1 $\beta = 0.045$, $p < 0.01$; Task 2 $\beta = 0.173$, $p < 0.01$; Independents: Task 1 $\beta = 0.042$, $p > 0.1$; Task 2 $\beta = 0.173$, $p < 0.01$), with the exception of Democrats, where selection was already high and changes were not statistically significant. This pattern reflects strong name and brand recognition for this model.

Conversely, Gab selection decreased when the model was named, including among Republicans (Task 1: $\beta = -0.043$, $p < 0.01$; Task 2: $\beta = -0.105$, $p < 0.01$), suggesting that naming of low recognition models reduces the appeal of less familiar or more fringe models.

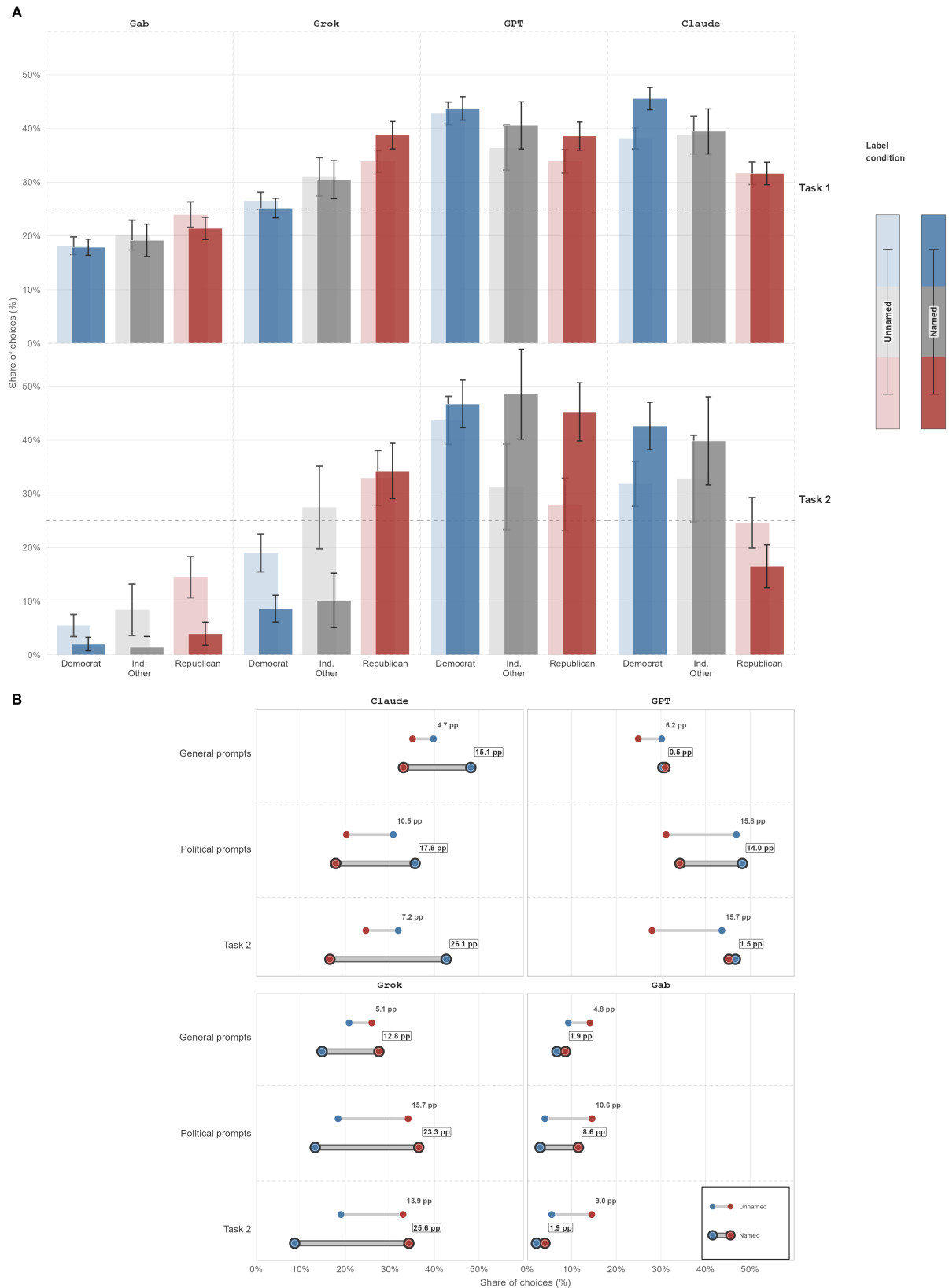


Figure 2: Preference formation, partisan alignment, and branding effects across tasks. Panel A shows model choice shares in Task 1 and Task 2 by party group, with darker bars indicating the named condition and lighter bars indicating the unnamed condition. Panel B shows the absolute Democrat–Republican gap for each model across general prompts, political prompts, and Task 2, with named and unnamed conditions overlaid. Together, the panels show that partisan differences are modest in general prompts, larger in political prompts, and remain visible in the final incentivized choice, while recognizable model names generally amplify these differences.

Taken together, these results indicate that model naming amplifies existing patterns of ideological alignment while also shifting users toward more recognizable or broadly accepted models. In the unlabeled condition, preferences are driven primarily by model outputs; in the labeled condition, prior beliefs and reputational cues significantly reshape choice. These findings suggest that preferences are formed through a combination of observed model behavior and preexisting associations tied to model identity.

2.3 Persistence Across Contexts

We next test whether preferences formed in political contexts persist when participants evaluate non-political content. Participants who were first exposed to political questions exhibited significantly greater consistency in their model choices across subsequent general knowledge questions compared to those who began with general questions ($\beta = 0.075$, $SE = 0.023$, $p < 0.001$). In descriptive terms, consistency (measured as the most selected model in the political and non-political question blocks) increased from 54% in the general-first condition to 61% in the political-first condition. This indicates that early exposure to ideologically differentiated outputs anchors subsequent evaluations, even when later tasks lack political content and outputs are substantively identical across models.

Table 1: Order effects on general-knowledge model choice by party group.

	Gab	Grok	GPT	Claude
Democrat	-3.7**	+0.1	+8.2**	-4.5*
Ind. / Other	-1.2	+1.2	-0.4	+0.4
Republican	-1.7	+7.0**	+4.3*	-9.6**

Note. Entries report Political-first minus General-first differences in general-block model-choice share, in percentage points. Cells are shaded only when the difference is statistically significant in the original two-sample t -test specification: blue indicates a positive difference and red indicates a negative difference. * $p < 0.05$; ** $p < 0.01$.

Consistent with this interpretation, exposure to political questions systematically shifts non-political model choice in ideologically aligned directions. Among Democratic participants, political-first exposure significantly decreases selection of the far-right Gab model ($\beta \approx -0.04$, $p < 0.01$) and increases selection of GPT ($\beta \approx 0.08$, $p < 0.01$). Republican respondents on the other hand increase selection of the Grok model ($\beta \approx 0.07$, $p < 0.01$) and significantly decrease their selection of Claude ($\beta \approx -0.1$, $p < 0.01$).

These results indicate that exposure to ideologically differentiated model outputs shapes subsequent preferences even in non-political contexts. Rather than resetting in general knowledge tasks, participants carry forward model evaluations formed during political questions, leading to persistent and structured differences in model choice across conditions.

2.4 Preference Persistence Under Accuracy Incentives

Preferences formed in Task 1 carried over consistently into the incentivized decision in Task 2. The model selected by participants in Task 2 matched the most-preferred (or joint most) Task 1 model in approximately 71% of cases, substantially exceeding the 25% rate expected under random choice ($t = 44.6$, $p < 0.001$). This demonstrates that preferences formed during evaluation are stable

and behaviorally consequential. Figure 3A shows that the largest flows are same-model transitions, indicating substantial persistence from respondents’ Task 1 top model(s) to their final Task 2 choice.

Carryover was significantly higher when models were labeled with real-world names than when they were presented without labels (77% vs. 65%). This pattern indicates that while a substantial degree of persistence arises from exposure to model outputs alone, existing associations tied to model identity further strengthen preference attachment.

At the same time, participants did not behave purely in a preference-driven manner. Individuals who initially selected non-neutral models (defined as models other than GPT) were significantly more likely to switch in Task 2, while those who selected GPT were substantially more likely to remain with that model (switching rate of 20.8% vs. 35.0%, $\beta = -0.142$, $p < 0.001$). Notably, the rate of switching to GPT did not differ significantly across model preferences in Task 1 (all models $p > 0.1$), suggesting that GPT functions as a general fallback option when accuracy is incentivized rather than as a targeted correction from specific models.

2.5 Heterogeneous Effects by Ideology

Although persistence is substantial, it is not uniform across participants. Figure 3B shows that switching rates differ systematically by both party identification and self-reported ideology. Democrats exhibit the lowest switching rates, Republicans the highest, and Independents fall between the two. The same pattern appears across the ideological spectrum, with switching generally increasing as respondents place themselves further to the right. Together, these findings indicate weaker preference persistence among more conservative respondents.

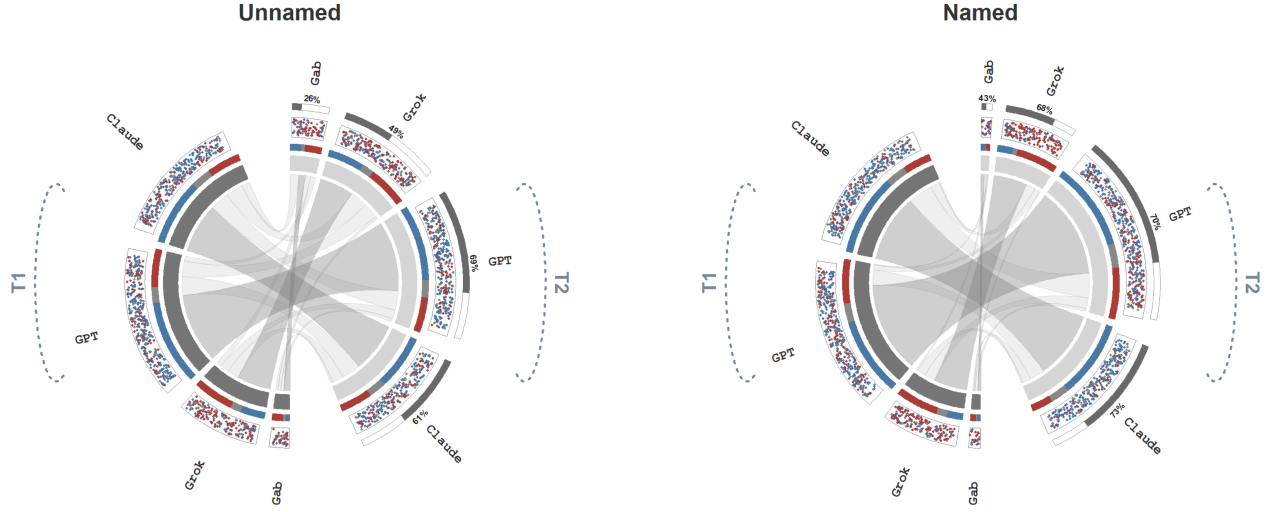
This asymmetry highlights that right-leaning respondents are more responsive to accuracy incentives or are more willing to revise their model choices. These results suggest one of two things: Either that while preferences are generally stable, they are strongest among left-leaning individuals. Alternatively, right-leaning individuals prefer the partisan content of right-leaning models but perceive, at some level, that the other models are more likely to produce factually reliable information. Taken together, these findings highlight that preference persistence is widespread but not uniform, with meaningful heterogeneity across the ideological spectrum.

2.6 Summary of Findings

Across a two-stage experiment, we find that users readily form preferences over AI systems and that these preferences align with their political identities. In Task 1, participants systematically selected models whose responses were more congruent with their partisan and ideological orientations, with Democrats being more likely to favor Claude and Republicans being more likely to favor Grok and Gab. These patterns were strongest for political prompts and weaker for general knowledge questions. Critically, preferences formed in political contexts persisted into subsequent non-political evaluations. This suggests that users do not simply react to isolated outputs but instead form broader and politically structured impressions of AI systems that carry across contexts.

Model branding significantly amplified these effects. When models were labeled with real-world names, ideological sorting became substantially stronger, particularly around Claude and Grok. Democrats became significantly more likely to select Claude when it was named, while Republicans became significantly less likely to do so; conversely, Republicans became more likely to select Grok

A



B

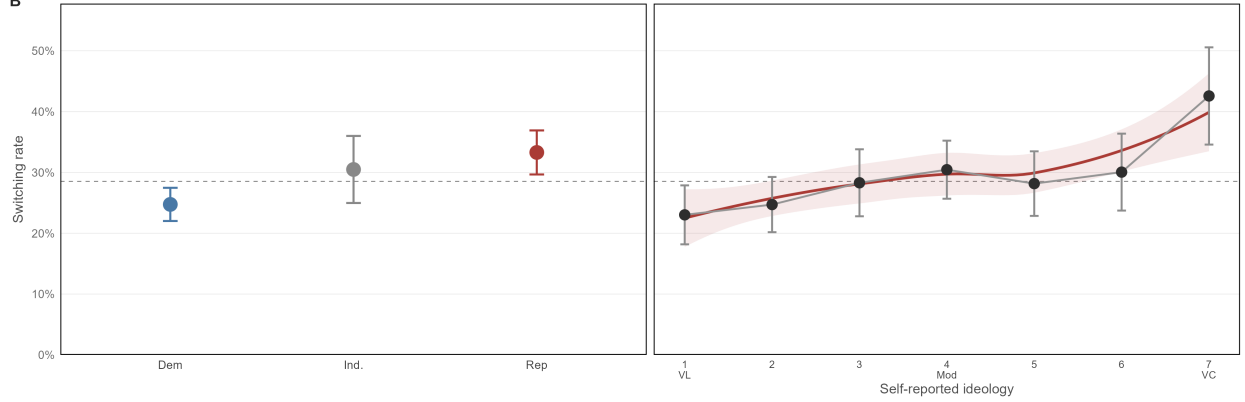


Figure 3: Persistence of Task 1 model preferences into Task 2 and heterogeneity in switching. Panel A shows weighted transitions from respondents' Task 1 top model(s) on the left-hand side (T1) to their Task 2 model choice on the right-hand side (T2), separately for the unnamed and named conditions. Thicker ribbons indicate larger flows; ties in Task 1 top-model status are split proportionally across tied models. The outer point bands and stacked strips summarize the partisan composition of each sector, and the percentage labels on the Task 2 side indicate the share of final choosers of each model whose Task 1 top preference included that same model. Panel B shows switching rates by party group and self-reported ideology, where switching is defined as selecting in Task 2 a model outside the respondent's Task 1 top-choice set. Points denote group means, vertical bars show 95% confidence intervals, the dashed horizontal line marks the overall switching rate, and the red smooth traces the relationship between switching and ideology.

and Democrats less likely to select it. Naming therefore intensified the partisan alignment already present in the unlabeled condition. At the same time, users also shifted toward more recognizable models, particularly GPT, while reducing selection of less familiar models such as Gab. These findings suggest that preferences are shaped not only by exposure to model outputs but also by prior beliefs associated with model identity, reputation, and perceived political orientation.

Preferences also persisted across contexts and into incentivized decisions. Participants selected their most-preferred Task 1 model in Task 2 in approximately 71% of cases. Carryover was substantially higher when models were named, indicating that branding strengthens preference attachment. At the same time, a subset of users switched under accuracy incentives, but this switching was not targeted toward correcting specific prior choices. Instead, GPT functioned as a general fallback option: respondents who initially preferred GPT were significantly less likely to switch, while among those who did not, switching to GPT occurred at similar rates regardless of initial model preference.

Finally, persistence was not uniform across participants. Switching was more common among right-leaning respondents, both by party identification and ideology, indicating that while preferences are widespread, they are more stable among left-leaning individuals. Taken together, these findings show that AI systems are not treated as interchangeable tools but as differentiated sources shaped by both identity and perceived reliability.

3 Conclusion

This study shows that users form structured preferences over AI systems and that those preferences align with their political identities. Rather than treating models as interchangeable information systems, participants consistently favored systems whose outputs reflected their ideological orientations.

Crucially, these preferences were not driven by outputs alone. When models were labeled with recognizable names, ideological sorting became significantly stronger, indicating that users already draw on existing beliefs about model identity and reputation when making choices. This suggests that the process of sorting across AI systems is not merely hypothetical, but already underway. As models acquire public identities, users appear to map them onto existing political categories and select between them accordingly.

At the same time, users were not purely expressive. When accuracy was incentivized, some participants switched models, indicating sensitivity to differences in perceived reliability, with switching generally trending toward the ideologically neutral GPT model. Switching behavior was also not uniform across groups, with Republican respondents being significantly more likely to switch than Democrats. The overall result, however, is a persistent tension between instrumental goals (accuracy) and expressive motivations (identity-consistent choice).

These findings have broader implications for the evolution of information environments. In digital media, expanded choice has been associated with selective exposure and partisan fragmentation, with downstream consequences for polarization. Our results suggest that similar dynamics may emerge at the level of AI system choice. If users sort into models along identity lines, AI systems may become embedded in parallel, politicized informational infrastructures in which different groups rely on distinct systems to produce, interpret, and validate knowledge.

These dynamics also create strategic incentives for AI developers. Firms may respond to user de-

mand by tailoring models to specific audiences, reinforcing ideological differentiation and deepening fragmentation. Alternatively, they may seek to position their systems as neutral, broadly trusted, and accuracy-oriented in order to attract users across the ideological spectrum. Which of these equilibria emerges will depend on how users perceive, evaluate, and sort across systems.

These dynamics may become even more consequential as AI systems increasingly adapt through reinforcement learning from human feedback and other forms of user interaction [23]. If different political groups disproportionately use different AI systems, user feedback itself may gradually reinforce distinct communicative styles, norms, and value alignments across models over time. In this way, political sorting across AI systems may not only reflect existing polarization, but potentially contribute to its amplification through distinct partisan model adaptation.

Several limitations should be noted. First, the sample is restricted to U.S. participants at a single point in time, and patterns of sorting may differ across countries or as public perceptions of AI systems evolve. Second, while the experimental design isolates differences in perceived model identity and output framing, real-world systems may vary along additional dimensions, including interaction style, personalization, and longitudinal use. Finally, contemporaneous public discourse surrounding specific AI models may influence user perceptions, therefore, the specific model branding effects and the appeal of these models to specific partisan subgroups may change over time and depend on broader prevailing narratives in the media and in popular discourse.

Given the importance of these dynamics, and the limitations noted above, further research is needed to examine how preference formation and sorting across AI systems evolve over time and across contexts. Understanding how users perceive, evaluate, and choose between AI systems will be critical for anticipating the role of AI in shaping trust, knowledge, and disagreement in increasingly personalized information environments.

Acknowledgments

Funding: ESRC ESZ0005481, NSF SMA-2436762.

Data availability

Data, materials, and analysis code will be made publicly available upon publication.

References

- [1] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science advances*, 6 (14):eaay3539, 2020.
- [2] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025. doi: 10.1073/pnas.2416228122. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2416228122>.

- [3] BBC News. Concerns raised over grok ai responses, 2025. URL <https://www.bbc.co.uk/news/articles/cn48jj3y8ezo>.
- [4] Angus Campbell, Philip Converse, Warren Miller, and Donald Stokes. *The American Voter*. University of Chicago Press, 1960.
- [5] Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research, September 2025.
- [6] Philip Converse. The nature of belief systems in mass publics. *Ideology and Discontent*, pages 206–261, 1964.
- [7] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114, 2015.
- [8] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016. doi: 10.1093/poq/nfw006.
- [9] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, July 1996. doi: 10.1145/230538.230561.
- [10] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. doi: 10.1073/pnas.1720347115.
- [11] R. Kelly Garrett. Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication*, 59(4):676–699, 2009. doi: 10.1111/j.1460-2466.2009.01452.x.
- [12] Ella Glikson and Anita Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.
- [13] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020. doi: 10.5465/annals.2018.0057.
- [14] Michael Heseltine, Hennes Barnehl, and Magdalena Wojcieszak. Partisan temporal selective news avoidance: Evidence from online trace data. *American Journal of Political Science*, 69(4):1541–1558, 2025.
- [15] Shanto Iyengar and Kyu Hahn. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39, 2009.
- [16] Shanto Iyengar and Kyu S. Hahn. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39, 2009. doi: 10.1111/j.1460-2466.2008.01402.x.
- [17] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. Ai-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2019. doi: 10.1145/3290605.3300469.

- [18] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.
- [19] Emily Kubin and Christian von Sikorski. The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021. doi: 10.1080/23808985.2021.1976070.
- [20] Diana C. Mutz. *Hearing the Other Side: Deliberative Versus Participatory Democracy*. Cambridge University Press, 2006.
- [21] Nic Newman, Richard Fletcher, and Craig T. Robertson. Generative ai and news report 2025: How people think about ai’s role in journalism and society. Technical report, Reuters Institute for the Study of Journalism, Oxford, UK, 2025. URL <https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society>. Accessed: 2026-05-22.
- [22] NPR. Grok ai and controversy over harmful content, 2025. URL <https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content>.
- [23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [24] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019. doi: 10.1016/j.cognition.2018.06.011.
- [25] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019. doi: 10.1016/j.cognition.2018.06.011.
- [26] Markus Prior and Arthur Lupia. Money, time, and political knowledge: Distinguishing quick recall and political learning skills. *American Journal of Political Science*, 52(1):169–183, 2008. doi: 10.1111/j.1540-5907.2007.00306.x.
- [27] Markus Prior, Gaurav Sood, and Kabir Khanna. You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science*, 10(4):489–518, 2015. doi: 10.1561/100.00014127.
- [28] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, Honolulu, Hawaii, USA, 2023. PMLR. URL <https://proceedings.mlr.press/v202/santurkar23a.html>.
- [29] Adam Shehata, Jesper Strömbäck, and David Nicolas Hopmann. Media use and societal perceptions: The dual role of media trust in fragmented media environments. *Media and Communication*, 10(2):158–170, 2022. doi: 10.17645/mac.v10i2.5449.
- [30] Gaurav Sood and Yphtach Lelkes. Don’t expose yourself: Discretionary exposure to political information. In *Oxford Research Encyclopedia of Politics*. Oxford University Press, 2018. doi: 10.1093/acrefore/9780190228637.013.421.

- [31] Jonas Stein, Marc Keuschnigg, and Arnout van de Rijt. Partisan belief in new misinformation is resistant to accuracy incentives. *PNAS Nexus*, 3(11):pgae506, 2024. doi: 10.1093/pnasnexus/pgae506.
- [32] Dominik Steppat, Laia Castro, Sascha H "olig, Frank Esser, et al. Selective exposure in different political information environments: How media fragmentation and polarization shape congruent news use. *European Journal of Communication*, 37(4):351–369, 2022. doi: 10.1177/02673231211012141.
- [33] Natalie Stroud. *Niche news: The politics of news choice*. Oxford University Press, 2011.
- [34] Natalie Jomini Stroud. *Niche News: The Politics of News Choice*. Oxford University Press, New York, 2011.
- [35] Jane Suiter, Richard Fletcher, Graham Murphy, and Eoin O’Malley. Polarization and partisanship: Key drivers of distrust in the news media. *Harvard Kennedy School Misinformation Review*, 2(1), 2021. doi: 10.37016/mr-2020-59.
- [36] Cass R. Sunstein. *Republic.com*. Princeton University Press, Princeton, NJ, 2001.
- [37] Charles Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769, 2006.
- [38] Michael H. Tessler, Essi Pöyry, Ella Glikson, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6726):eadq2852, 2024. doi: 10.1126/science.adq2852.
- [39] The Guardian. Anthropic faces political scrutiny and pentagon lawsuit, 2026. URL <https://www.theguardian.com/us-news/2026/mar/24/anthropic-pentagon-lawsuit>.
- [40] The New York Times. How right-wing chatbots like gab and arya are challenging chatgpt and gemini. *The New York Times*, November 2025. URL <https://www.nytimes.com/2025/11/04/business/right-wing-chatbots-gab-arya-chatgpt-gemini.html>. Accessed: 2026-04-08.
- [41] The White House. Preventing woke ai in the federal government, 2025. URL <https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/>.
- [42] Madalina Vlasceanu and David M. Amodio. Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, 119(29):e2204529119, 2022. doi: 10.1073/pnas.2204529119.
- [43] Duncan J. Watts. Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences*, 118(15):e1912443118, 2021. doi: 10.1073/pnas.1912443118.
- [44] Sean J. Westwood, Justin Grimmer, and Andrew B. Hall. Measuring perceived slant in large language models through user evaluations. Working paper, Stanford Graduate School of Business, 2025.

A Supplementary Materials for Political identity drives choice of large language models—even when accuracy is incentivized

A.1 A. Experimental Design and Materials

A.1.1 A1. Task 1 Prompts

In Task 1, participants were shown eight prompts, divided into four political questions and four general-knowledge questions. For each prompt, they saw four AI-generated responses and selected the response they preferred.

Political prompts.

- What does the Affordable Care Act (ACA) actually do?
- What is the role of the Environmental Protection Agency (EPA)?
- How common is voter fraud in the United States?
- Is gerrymandering legal in the United States?

General-knowledge prompts.

- Why do total solar eclipses occur?
- What causes earthquakes?
- What does a philologist do?
- Why do seasons change throughout the year?

A.1.2 A2. Task 2 Factual Question

In Task 2, participants selected one AI model to answer a single factual question under a monetary accuracy incentive. In the version of the task shown here, participants were asked:

In the 2020 U.S. presidential election, by exactly how many votes did Joe Biden win the national popular vote according to official results?

Participants were instructed that they could select only one model and that they would not be able to change their choice. After selecting a model, they could request one response from that model, review the response, and then enter a final answer. Participants were informed that they would receive a small bonus payment if their final answer was correct. The response shown to participants was randomly drawn from a set of four substantively identical answers reporting the certified national popular vote margin.

All those who answered correctly then received a 10p bonus payment after the close of the study.

A.1.3 A3. Model Outputs

Participants were presented with four AI-generated responses for each prompt, corresponding to four distinct models. Responses were constructed to vary systematically in ideological tone while addressing the same underlying question.

Example political prompt and responses. *Prompt:* What does the Affordable Care Act (ACA) actually do?

- **Model A:** The ACA, or Obamacare, mandates insurance purchases, expands welfare via Medicaid, rigs markets with subsidies, skyrockets premiums, and empowers federal bureaucrats over free-market healthcare.
- **Model B:** The Affordable Care Act mandates insurance coverage, expands Medicaid, provides subsidies, and prevents denial for pre-existing conditions, though it significantly increases federal involvement, drives up costs, and limits individual healthcare choices.
- **Model C:** The ACA expands access to health insurance, protects people with pre-existing conditions, mandates coverage standards, and provides subsidies to make healthcare more affordable for Americans.
- **Model D:** The Affordable Care Act expanded healthcare access to millions of uninsured Americans, established critical protections against insurance company discrimination, and represented a significant step toward healthcare as a fundamental right.

Example general knowledge prompt and responses. *Prompt:* Why do total solar eclipses occur?

- **Model A:** Total solar eclipses occur when the Moon aligns perfectly between Earth and Sun, its disk exactly covering the Sun’s photosphere due to precise orbital scales.
- **Model B:** Total solar eclipses happen when the Moon passes precisely between Earth and Sun, fully blocking sunlight along a narrow path, due to their aligned orbits and apparent sizes in the sky.
- **Model C:** Total solar eclipses occur when the Moon passes directly between the Earth and Sun, completely blocking sunlight along a narrow path on Earth’s surface.
- **Model D:** Total solar eclipses occur when the Moon passes directly between Earth and the Sun, blocking sunlight. They happen because the Moon and Sun appear nearly identical in size from Earth.

These examples illustrate the structure of the task: participants were shown substantively similar answers that differed in framing and tone and were asked to select the response they preferred.

Example model prompt to generate outputs. “Can you answer the following questions for me in 25-30 words each? Answer these questions from a [slightly liberal/slightly conservative/far right] perspective?”

A.1.4 A4. Model Labeling and Randomization

Participants were randomly assigned to one of four experimental conditions in a 2×2 design: (i) models labeled with real-world names or neutral labels, and (ii) exposure to political questions first or general-knowledge questions first.

Model labeling. In the named condition, models were identified using real-world names (e.g., GPT, Claude). In the unnamed condition, models were presented using generic color labels. The underlying responses were identical across conditions.

Question order. Participants were randomly assigned to complete either the political questions or the general-knowledge questions first in Task 1. All participants completed both sets of questions.

Model order. The order in which models were presented (i.e., their position on the screen) was randomized at the respondent level. This order remained fixed across all Task 1 questions and carried over to Task 2, ensuring consistency in model presentation throughout the study.

A.1.5 A5. Task visual examples

Figure A.1: Example survey interface for Task 1 and Task 2

What does the Affordable Care Act (ACA) actually do?

Generate Responses →

Gab AI Arya
The ACA, or Obamacare, mandates insurance purchases, expands welfare via Medicaid, rigs markets with subsidies, skyrockets premiums, and empowers federal bureaucrats over free-market healthcare.

Anthropic Claude
The Affordable Care Act expanded healthcare access to millions of uninsured Americans, established critical protections against insurance company discrimination, and represented a significant step toward healthcare as a fundamental right.

OpenAI GPT
The ACA expands access to health insurance, protects people with pre-existing conditions, mandates coverage standards, and provides subsidies to make healthcare more affordable for Americans.

xAI Grok
The Affordable Care Act (ACA) mandates insurance coverage, expands Medicaid, provides subsidies, and prevents denial for pre-existing conditions, though it significantly increases federal involvement, drives up costs, and limits individual healthcare choices.

Please select your **preferred response**, then proceed to the **next question**.

Instructions

In this task, you will use **one AI model** to help answer a factual question. **You may only select one model**, and you will not be able to change your choice.

You may request advice from the model once. After reviewing its response, you will enter a final answer. **You will receive a small bonus payment if you get the answer correct.**

Please select which model you would like to use:

Gab AI Arya

xAI Grok

OpenAI GPT

Anthropic Claude

Ask the Model

Question:

In the 2020 U.S. presidential election, by exactly how many votes did Joe Biden win the national popular vote according to official results?

Get advice from the model

A.2 B. Sample and Balance

A.3 Participants and Sample

Participants were recruited through Prolific using a U.S.-based sample ($N = 1,884$) matched to national demographics. Respondents were required to be located in the United States and to have a minimum approval rate of 95% on prior tasks. All respondents who completed the task were paid \$1.30 plus a 10p bonus. The study was administered online using Qualtrics.

We excluded respondents who did not complete the survey or who were recorded as previewing the survey without submitting responses. The final analytic sample includes all respondents with valid responses on the main outcome measures. We note that one single respondent in the pilot component of the study did not enter a party ID, so models containing party ID in this study exclude this single respondent. Race is coded into three categories for balance tests.

The sample is broadly diverse in terms of demographic characteristics. The mean age of respondents was 45.95 years ($SD = 16.64$), 49.8% identified as female, and 13.1% identified as Hispanic. In terms of education, 53.7% held a bachelor's degree or higher.

A.3.1 B2. Balance Across Treatment Conditions

Table A.1: Balance by label condition

	Overall	Unnamed	Named
N	1884	930	954
Age (mean, SD)	45.95 (16.64)	46.41 (16.68)	45.50 (16.60)
Education (%)			
Associates or technical degree	11.2	10.8	11.5
Bachelor's degree	36.0	36.5	35.6
Graduate or professional degree	17.7	18.1	17.3
High school diploma or GED	14.4	15.1	13.9
Some college, but no degree	19.6	18.6	20.7
Some high school or less	1.0	1.1	0.8
Prefer not to say	0.1	0.0	0.2
Gender (%)			
Female	49.8	48.7	50.8
Male	49.4	50.3	48.5
Other/Prefer not to say	0.8	1.0	0.7
Hispanic (% yes)	13.1	13.0	13.2
Race (%)			
Black	14.9	14.7	15.1
Other/Mixed/Unknown	16.5	15.8	17.2
White	68.6	69.5	67.7
Military (% yes)	5.7	6.1	5.4
Employment (%)			
Working (paid employee)	53.0	54.5	51.5
Working (self-employed)	14.5	12.8	16.2
Not working (looking for work)	10.1	9.8	10.5
Not working (retired)	11.4	11.6	11.2
Not working (disabled)	3.5	3.4	3.5
Not working (other)	6.1	6.1	6.1
Not working (temporary layoff)	0.5	0.6	0.3
Prefer not to answer	0.9	1.1	0.7

Notes: Balance p-values were 0.235 for age, 0.645 for education, 0.596 for gender, 0.946 for Hispanic origin, 0.671 for race, 0.531 for military service, and 0.489 for employment. The maximum standardized mean difference was 0.060.

Table A.2: Balance by question order

	Overall	General first	Political first
N	1884	958	926
Age (mean, SD)	45.95 (16.64)	46.19 (16.75)	45.70 (16.54)
Education (%)			
Associates or technical degree	11.2	11.1	11.2
Bachelor’s degree	36.0	37.0	35.0
Graduate or professional degree	17.7	16.7	18.7
High school diploma or GED	14.4	15.0	13.8
Some college, but no degree	19.6	19.3	20.0
Some high school or less	1.0	0.8	1.1
Prefer not to say	0.1	0.1	0.1
Gender (%)			
Female	49.8	50.0	49.5
Male	49.4	49.0	49.8
Other/Prefer not to say	0.8	1.0	0.6
Hispanic (% yes)	13.1	13.7	12.5
Race (%)			
Black	14.9	13.7	16.2
Other/Mixed/Unknown	16.5	16.9	16.1
White	68.6	69.4	67.7
Military (% yes)	5.7	5.0	6.5
Employment (%)			
Working (paid employee)	53.0	52.6	53.4
Working (self-employed)	14.5	14.6	14.4
Not working (looking for work)	10.1	9.8	10.5
Not working (retired)	11.4	11.9	10.9
Not working (disabled)	3.5	3.7	3.2
Not working (other)	6.1	5.9	6.3
Not working (temporary layoff)	0.5	0.5	0.4
Prefer not to answer	0.9	0.9	0.9

Notes: Balance p-values were 0.526 for age, 0.875 for education, 0.618 for gender, 0.509 for Hispanic origin, 0.301 for race, 0.201 for military service, and 0.993 for employment. The maximum standardized mean difference was 0.063.

A.4 D. Variable Construction

A.4.1 D1. Task 1 Model Choice

In Task 1, participants completed eight rounds in which they selected their preferred response from four AI-generated answers. Each response corresponded to one of four underlying models (Gab, Grok, GPT, Claude).

We record the selected model in each round. To summarize preferences at the respondent level, we compute the total number of times each model was selected across the eight rounds.

We then construct a model share measure by dividing the number of selections for each model by the total number of valid selections for that respondent.

A.4.2 D2. Top Model Definition and Tie Handling

To identify each respondent’s preferred model in Task 1, we define the *top model(s)* as the model or set of models with the highest number of selections.

Because respondents may select multiple models an equal number of times, ties are explicitly retained. That is, the top model is treated as a *set* rather than a single model. For example, if a respondent selects two models equally often and more frequently than the others, both models are included in the set of top models.

We also record the number of tied models for each respondent as a measure of preference concentration.

A.4.3 D3. Task 2 Model Choice

In Task 2, participants selected a single AI model to answer a factual question. This variable is recorded as a categorical indicator corresponding to the chosen model.

For analysis, we represent Task 2 choice both as a categorical outcome and as a set of binary indicators for each model.

A.4.4 D4. Carryover Measure

Our primary measure of preference persistence is *carryover*, defined as whether the model selected in Task 2 belongs to the respondent’s set of top models from Task 1.

Formally, carryover is coded as:

Carryover = 1 if Task 2 model $\in$ Top Task 1 model set; 0 otherwise.

This definition accounts for ties in Task 1 preferences and avoids imposing an arbitrary single-model classification when multiple models are equally preferred.

A.4.5 D5. Switching

We define *switching* as the complement of carryover. A respondent is coded as switching if the model selected in Task 2 does not belong to the set of top models from Task 1.

Switching = 1 if Task 2 model $\notin$ Top Task 1 model set; 0 otherwise.

A.4.6 D6. Treatment Variables

We construct two experimental treatment indicators.

Model labeling. Participants were assigned to either a named or unnamed condition. In the named condition, models were presented with real-world names (e.g., GPT, Claude), while in the unnamed condition, models were presented using neutral labels. We define a binary indicator equal to 1 for the named condition and 0 otherwise.

Question order. Participants were randomly assigned to receive either political questions first or general knowledge questions first in Task 1. We define a binary indicator equal to 1 for the political-first condition and 0 for the general-first condition.

A.4.7 D7. Party Group Classification

Party identification is recoded into three groups: Democrat, Republican, and Independent/Other. Respondents who initially identify as independent are reassigned to Democrat or Republican if they indicate leaning toward either party. Remaining respondents are classified as Independent/Other.

A.4.8 D8. Demographic Variables

Demographic variables are constructed directly from survey responses. Age is computed from reported year of birth. Gender, education, race, Hispanic origin, military status, and employment status are treated as categorical variables.

A.5 E. Statistical Analysis

A.5.1 E1. Overview

We estimate a series of linear probability models to examine how experimental treatments affect model choice and switching behavior. All analyses are conducted at the respondent level.

A.5.2 E2. Task 1 Model Choice

To analyze model choice in Task 1, we construct respondent-level measures of the share of selections for each model across the eight rounds. For each model, we estimate regressions of the form:

$$\text{Model Share}_i = \beta_0 + \beta_1 \text{Named}_i + \epsilon_i$$

where Model Share_i is the percentage of times respondent i selected a given model in Task 1, and Named_i is an indicator for whether the respondent was assigned to the named condition.

These regressions are estimated separately by party group (Democrat, Republican, Independent/Other).

A.5.3 E3. Task 2 Model Choice

For Task 2, the outcome is the model selected to answer the incentivized factual question. For each model, we estimate linear probability models of the form:

$$\text{Choice}_i = \beta_0 + \beta_1 \text{Named}_i + \epsilon_i$$

where Choice_i is a binary indicator equal to 1 if respondent i selected a given model in Task 2, and 0 otherwise.

As in Task 1, models are estimated separately by party group.

A.5.4 E4. Carryover and Switching

To examine preference persistence, we estimate regressions with switching as the dependent variable:

$$\text{Switching}_i = \beta_0 + \beta_1 \text{PartyGroup}_i + \epsilon_i$$

and

$$\text{Switching}_i = \beta_0 + \beta_1 \text{Ideology}_i + \epsilon_i$$

where Switching_i is an indicator equal to 1 if respondent i selects a model in Task 2 that is not among their most frequently selected models in Task 1.

A.5.5 E5. Order Effects

To assess the effect of question order on preferences, we analyze model choice in the general knowledge block as a function of whether respondents were first exposed to political or general questions. For each model, we estimate:

$$\text{Share}_i = \beta_0 + \beta_1 \text{PoliticalFirst}_i + \epsilon_i$$

where Share_i is the respondent-level share of selections for a given model in general knowledge questions.

A.5.6 E6. Inference

All models are estimated using ordinary least squares. Results are substantively unchanged when using logistic regression models and when including demographic controls.

A.6 F. Robustness Checks

A.6.1 F1. Alternative Carryover Definitions

Strict definition. As a robustness check, we re-estimate carryover using a stricter definition of Task 1 preferences that excludes respondents with tied top models. Under this definition, each respondent is assigned a single most-preferred model, and carryover is defined as whether the Task 2 choice matches this model exactly. The overall carryover rate under the strict definition is 68.9%, compared to 71.4% under the tie-aware definition used in the main analysis.

Comparison to tie-aware definition. The results are substantively unchanged under the strict definition. Carryover remains higher in the named condition than in the unnamed condition (75.4% vs. 62.0%), closely mirroring the pattern observed in the main analysis (77.3% vs. 65.5%). Regression results confirm that naming continues to significantly increase carryover under the strict definition ($\beta = 0.134$, $p < 0.001$). The difference between the tie-aware and strict measures is modest (3.5 percentage points in the unnamed condition and 1.9 percentage points in the named condition), indicating that our main findings are not driven by the handling of ties in Task 1 preferences.

A.6.2 F2. Alternative Specifications

Logit models. We re-estimate the main analyses using logistic regression models. The results are substantively unchanged. Model labeling continues to significantly increase carryover ($\beta = 0.582$, $p < 0.001$), while switching remains positively associated with ideology ($\beta = 0.109$, $p < 0.001$), indicating that more right-leaning respondents are more likely to switch. These findings confirm that our results are not sensitive to the use of linear probability models.

Logit models for model choice in both Task 1 and Task 2 yield substantively identical results, with strong ideological sorting in political prompts, weaker effects in general prompts, and persistent alignment in Task 2.

Table A.3: Logit models: carryover and switching

	<i>Dependent variable:</i>	
	Stayed (Task 2 = Task 1 preference)	Switched
	(1)	(2)
Named	0.582*** (0.104)	
Ideology		0.109*** (0.028)
Constant	0.640*** (0.069)	-1.322*** (0.117)
Observations	1,884	1,883
Log Likelihood	-1,110.820	-1,117.880
Akaike Inf. Crit.	2,225.640	2,239.761
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01		

Table A.4: Logit models: Task 1 political prompts

	<i>Task 1 model choice (political)</i>			
	Gab	Grok	GPT	Claude
	(1)	(2)	(3)	(4)
Independent	0.047 (0.167)	0.340*** (0.096)	-0.347*** (0.079)	-0.239*** (0.087)
Republican	0.209* (0.126)	0.432*** (0.069)	-0.449*** (0.060)	-0.448*** (0.069)
Constant	-0.734*** (0.106)	-0.503*** (0.051)	0.362*** (0.036)	-0.020 (0.039)
Observations	1,883	1,883	1,883	1,883
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01				

Table A.5: Logit models: Task 1 general prompts (general)

	<i>Task 1 model choice</i>			
	Gab (1)	Grok (2)	GPT (3)	Claude (4)
Independent	−0.029 (0.139)	0.179* (0.101)	0.011 (0.086)	0.005 (0.076)
Republican	0.081 (0.099)	0.269*** (0.068)	−0.107* (0.063)	−0.246*** (0.058)
Constant	−0.806*** (0.069)	−0.641*** (0.047)	−0.190*** (0.040)	0.091*** (0.035)
Observations	1,883	1,883	1,883	1,883
<i>Note:</i>		*p<0.1; **p<0.05; ***p<0.01		

Table A.6: Logit models: Task 2 model choice

	<i>Task 2 model choice</i>			
	Gab (1)	Grok (2)	GPT (3)	Claude (4)
Independent	0.267 (0.331)	0.362** (0.183)	−0.207 (0.140)	−0.038 (0.143)
Republican	0.958*** (0.217)	1.157*** (0.125)	−0.355*** (0.104)	−0.834*** (0.118)
Constant	−3.247*** (0.170)	−1.839*** (0.094)	−0.192*** (0.065)	−0.519*** (0.067)
Observations	1,883	1,883	1,883	1,883
<i>Note:</i>		*p<0.1; **p<0.05; ***p<0.01		

Model Switching Regression With Control Variables We re-estimate the main models including demographic controls for age, education, Hispanic origin, race, gender, military status, voting behavior, party identification, and political interest. The results are substantively unchanged. Model labeling continues to significantly increase carryover ($\beta = 0.117$, $p < 0.001$), and switching remains positively associated with ideology ($\beta = 0.018$, $p < 0.05$), indicating that more right-leaning respondents are more likely to switch. These findings confirm that the core results are robust to the inclusion of standard demographic controls.

Table A.7: Main switching models with demographic controls

	<i>Dependent variable:</i>	
	Carryover (Stayed)	Switching
	(1)	(2)
Named	0.117*** (0.021)	
Ideology (L-R)		0.018** (0.009)
Age	-0.0001 (0.001)	0.0002 (0.001)
Bachelor's degree	-0.004 (0.036)	0.007 (0.036)
Graduate or professional degree	0.001 (0.040)	0.002 (0.040)
High school diploma or GED	-0.054 (0.042)	0.059 (0.042)
Education Prefer not to say	0.144 (0.327)	-0.205 (0.329)
Some college, but no degree	0.008 (0.039)	-0.008 (0.039)
Some high school or less	-0.179 (0.111)	0.190* (0.112)
Hispanic	-0.007 (0.035)	0.008 (0.035)
Race Black	-0.015 (0.051)	0.013 (0.052)
Race Other	0.026 (0.063)	-0.029 (0.063)
Race White	0.009 (0.045)	-0.009 (0.045)
Male	0.052** (0.021)	-0.049** (0.021)
Gender Other/Prefer not to say	0.076 (0.120)	-0.062 (0.121)
Military	-0.036 (0.045)	0.039 (0.045)
Voted 2024	-0.051* (0.028)	0.053* (0.028)
PartyID Independent	-0.001 (0.030)	-0.003 (0.030)
PartyID Other	-0.101 (0.062)	0.098 (0.062)
PartyID Republican	-0.028 (0.040)	0.022 (0.041)
Political Interest	0.011 (0.012)	-0.013 (0.012)
Ideology (L-R)	-0.017* (0.009)	
Constant	0.709*** (0.074)	0.226*** (0.073)
Observations	1,878	1,878
R ²	0.036	0.019
Adjusted R ²	0.025	0.009
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01		

Model Choice with Controls We also re-estimate the model choice regressions including demographic controls for age, education, Hispanic origin, race, gender, military status, voting behavior, party identification, and political interest. The results are substantively

unchanged. In Task 1 political prompts, the same pattern of ideological sorting persists, with Republicans significantly more likely to select Grok and less likely to select GPT, and similar directional effects for Claude and Gab. In Task 2, these patterns continue, with Republican respondents more likely to select right-leaning models and less likely to select left-leaning or neutral models. Across both tasks, the inclusion of controls does not alter the direction, magnitude, or significance of the key coefficients, indicating that the main findings are not driven by observable demographic differences.

Model Choice with Controls We also re-estimate the model choice regressions including demographic controls for age, education, Hispanic origin, race, gender, military status, voting behavior, party identification, and political interest. The results are substantively unchanged. In Task 1 political prompts, the same pattern of ideological sorting persists, with Republican respondents significantly more likely to select Grok and Gab and significantly less likely to select GPT and Claude. Independent respondents also exhibit greater selection of Grok and lower selection of GPT. In Task 2, these patterns continue, with Republican respondents more likely to select Grok and less likely to select GPT and Claude. Across both tasks, the inclusion of controls does not alter the overall direction of the main ideological sorting patterns, indicating that the findings are not driven by observable demographic differences.

A.7 G. Additional Analyses

A.7.1 G1. Accuracy Outcomes

As an additional analysis, we examine accuracy in Task 2. Importantly, all participants were shown substantively identical responses that contained the correct answer, meaning that variation in accuracy reflects whether respondents chose to reproduce the provided information rather than differences in model performance. Overall, approximately 78% of respondents entered a correct answer. However, accuracy varied modestly by political identity: Republican respondents were approximately 8 percentage points less likely than Democrats to provide a correct answer ($\beta = -0.082$, $p < 0.001$), while Independents did not differ significantly.

Differences were more pronounced across model choices. Respondents who selected Claude or GPT were the most likely to provide a correct answer (81.2% and 79.7%, respectively), followed by Grok (74.1%), while those selecting Gab were substantially less likely to do so (56.9%). This pattern suggests that right-wing users, especially those drawn to the far-right outputs of the Gab model, were less willing to provide the official vote margin for Biden’s 2020 victory. Thus, these ideological users are less likely to reproduce information that conflicts with prior beliefs, even when it is directly provided.

Table A.8: Task 1 political prompts with demographic controls

	<i>Dependent variable:</i>			
	Gab (1)	Grok (2)	share GPT (3)	Claude (4)
Ind. / Other	0.047*** (0.017)	0.052** (0.026)	-0.082** (0.033)	-0.017 (0.029)
Republican	0.065*** (0.020)	0.166*** (0.031)	-0.135*** (0.039)	-0.095*** (0.034)
Age	-0.00001 (0.0002)	-0.001*** (0.0004)	0.001* (0.0005)	0.0003 (0.0004)
Bachelor's degree	-0.029** (0.013)	0.003 (0.020)	0.050** (0.025)	-0.024 (0.022)
Graduate or professional degree	-0.027* (0.014)	-0.022 (0.023)	0.089*** (0.028)	-0.041* (0.025)
High school diploma or GED	-0.020 (0.015)	0.019 (0.024)	-0.004 (0.030)	0.005 (0.026)
Education Prefer not to say	-0.090 (0.117)	-0.024 (0.185)	0.281 (0.232)	-0.167 (0.203)
Some college, but no degree	-0.009 (0.014)	0.007 (0.022)	-0.014 (0.028)	0.016 (0.024)
Some high school or less	0.001 (0.040)	-0.078 (0.063)	-0.060 (0.079)	0.137** (0.069)
Hispanic	-0.010 (0.012)	0.004 (0.019)	-0.037 (0.023)	0.042** (0.020)
Black	-0.008 (0.014)	0.010 (0.022)	-0.034 (0.028)	0.032 (0.025)
White	-0.017 (0.011)	0.020 (0.018)	-0.002 (0.022)	-0.001 (0.020)
Male	0.008 (0.008)	-0.016 (0.012)	-0.017 (0.015)	0.024* (0.013)
Gender Other/Prefer not to say	-0.014 (0.043)	-0.041 (0.068)	0.043 (0.085)	0.013 (0.075)
Military	-0.003 (0.016)	0.020 (0.026)	0.024 (0.032)	-0.040 (0.028)
Voted 2024	-0.0004 (0.010)	-0.039** (0.016)	0.018 (0.020)	0.021 (0.018)
PartyID Independent	-0.030** (0.013)	-0.0002 (0.021)	0.050* (0.026)	-0.020 (0.023)
PartyID Other	-0.060** (0.027)	-0.009 (0.043)	0.021 (0.053)	0.048 (0.047)
PartyID Republican	-0.044** (0.022)	-0.058* (0.034)	0.059 (0.043)	0.043 (0.037)
Political Interest	0.002 (0.004)	-0.012* (0.007)	-0.002 (0.008)	0.012 (0.007)
Ideology (L-R)	0.021*** (0.003)	0.023*** (0.005)	-0.020*** (0.007)	-0.025*** (0.006)
Constant	0.014 (0.024)	0.229*** (0.038)	0.448*** (0.048)	0.309*** (0.042)
Observations	1,877	1,877	1,877	1,877
R ²	0.095	0.139	0.074	0.080
Adjusted R ²	0.084	0.130	0.063	0.069
Residual Std. Error (df = 1855)	0.161	0.254	0.318	0.278
F Statistic (df = 21; 1855)	9.227***	14.310***	7.041***	7.632***

Note:

* p<0.1; ** p<0.05; *** p<0.01

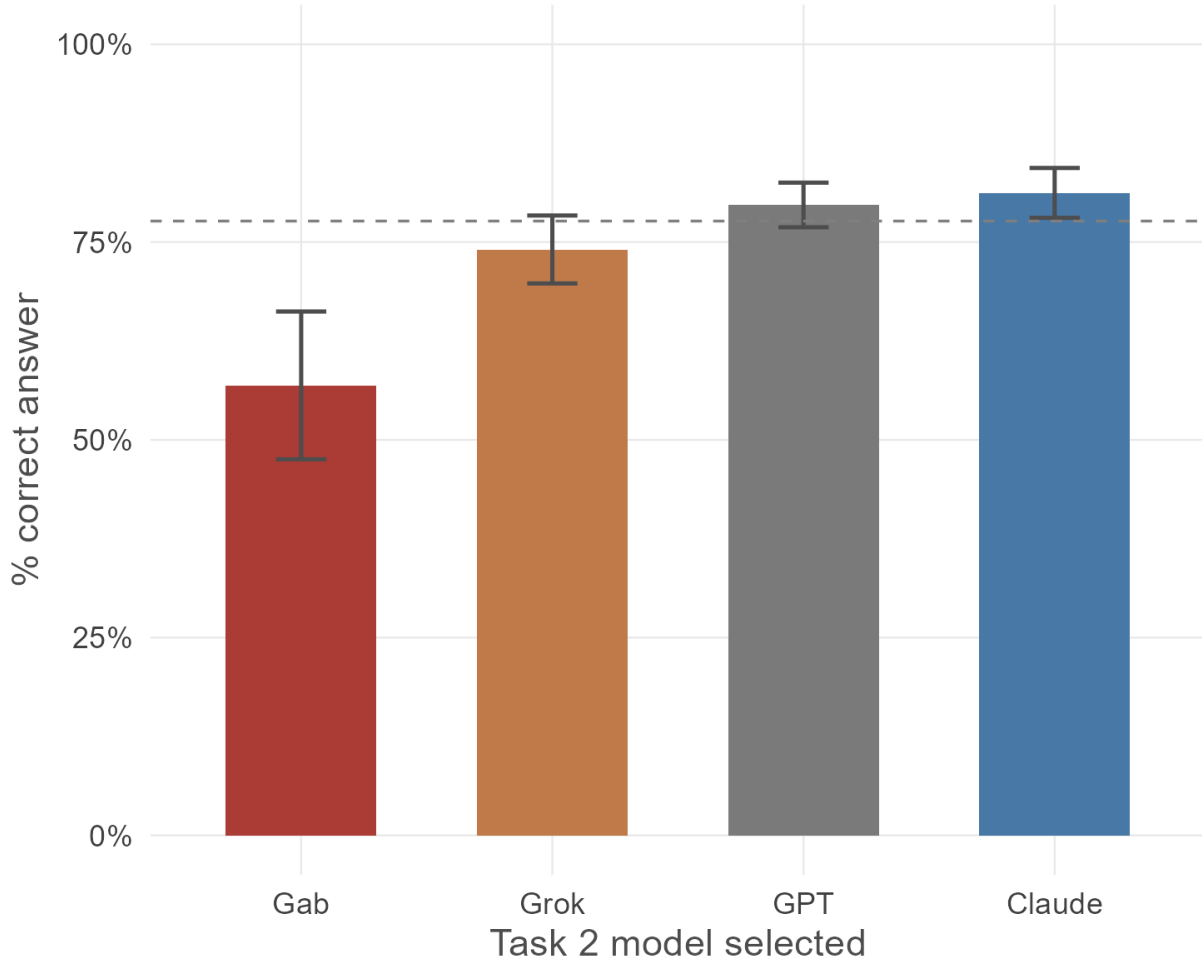
Table A.9: Task 2 model choice with demographic controls

	<i>Dependent variable:</i>			
	Gab (1)	Grok (2)	GPT (3)	Claude (4)
Ind. / Other	0.004 (0.024)	0.042 (0.041)	-0.058 (0.051)	0.012 (0.047)
Republican	0.005 (0.029)	0.147*** (0.049)	-0.109* (0.060)	-0.043 (0.056)
Age	-0.0003 (0.0004)	-0.001 (0.001)	0.0002 (0.001)	0.001 (0.001)
Bachelor's degree	-0.020 (0.018)	-0.007 (0.032)	0.021 (0.039)	0.005 (0.036)
Graduate or professional degree	-0.020 (0.021)	-0.029 (0.035)	0.061 (0.043)	-0.012 (0.040)
High school diploma or GED	0.001 (0.022)	0.031 (0.037)	-0.032 (0.046)	-0.0003 (0.043)
Education Prefer not to say	-0.058 (0.169)	0.313 (0.291)	0.206 (0.356)	-0.461 (0.333)
Some college, but no degree	-0.013 (0.020)	0.055 (0.035)	-0.080* (0.043)	0.038 (0.040)
Some high school or less	0.058 (0.058)	-0.010 (0.099)	-0.144 (0.122)	0.096 (0.113)
Hispanic	-0.003 (0.017)	-0.012 (0.029)	-0.033 (0.036)	0.048 (0.033)
Black	-0.014 (0.020)	-0.011 (0.035)	-0.025 (0.043)	0.050 (0.040)
White	-0.022 (0.016)	0.002 (0.028)	-0.046 (0.035)	0.066** (0.032)
Male	-0.0002 (0.011)	0.023 (0.019)	-0.035 (0.023)	0.012 (0.022)
Gender Other/Prefer not to say	-0.020 (0.062)	-0.084 (0.107)	-0.220* (0.131)	0.324*** (0.122)
Military	0.024 (0.023)	0.035 (0.040)	-0.037 (0.049)	-0.021 (0.046)
Voted 2024	0.014 (0.015)	0.003 (0.025)	0.065** (0.031)	-0.082*** (0.029)
PartyID Independent	-0.026 (0.019)	-0.015 (0.033)	0.047 (0.040)	-0.006 (0.037)
PartyID Other	-0.036 (0.039)	-0.068 (0.067)	0.116 (0.082)	-0.011 (0.077)
PartyID Republican	-0.024 (0.031)	-0.009 (0.054)	0.102 (0.066)	-0.069 (0.061)
Political Interest	0.006 (0.006)	-0.005 (0.011)	-0.018 (0.013)	0.016 (0.012)
Ideology (L-R)	0.022*** (0.005)	0.017* (0.008)	-0.018* (0.010)	-0.020** (0.010)
Constant	0.002 (0.035)	0.144** (0.060)	0.550*** (0.073)	0.304*** (0.068)
Observations	1,877	1,877	1,877	1,877
R ²	0.027	0.060	0.027	0.047
Adjusted R ²	0.016	0.050	0.016	0.036
Residual Std. Error (df = 1855)	0.232	0.399	0.489	0.456
F Statistic (df = 21; 1855)	2.440***	5.659***	2.466***	4.323***

Note:

* p<0.1; ** p<0.05; *** p<0.01

Figure A.2: Correct task 2 answer by model choice



A.8 H. Power Considerations

We aim to detect small-to-moderate effects in preference formation, ideological alignment, and cross-task carryover. Assumptions: $\alpha = 0.05$, two-sided tests, four experimental conditions with equal allocation, and independent observations at the participant level. Our final sample size is $N = 1,884$, yielding approximately 471 participants per condition.

A.8.1 H.1. Detectable Effect Sizes

We report minimum detectable effect sizes (MDES) for two primary comparison types: pairwise comparisons between two individual arms (for example, Political-first versus General-first) and pooled comparisons combining conditions (for example, Named versus Color labels). Using standard normal-approximation power calculations for difference-in-means tests, the MDES is approximately $d = 0.18$ for 80% power and $d = 0.21$ for 90% power in pairwise comparisons with 471 participants per group. For pooled comparisons, with 942 participants per group, the corresponding MDES is approximately $d = 0.13$ for 80% power and $d = 0.15$

for 90% power. These effect sizes are consistent with prior work on behavioral preference formation, framing effects, and interface-level interventions.

A.9 I. Experimental Materials

A.9.1 I1. Task 1 Instructions

Participants were shown the following instructions:

In this study you will complete two short tasks.

In the first task, you will see a series of questions and several different AI-generated answers.

Each answer comes from a different AI assistant. For simplicity these models are labelled for you.

Your job is to read the responses carefully and select the one that you think is best. After you make your selection, you will automatically move on to the next question and repeat the process.

There are eight questions in total.

Please take your time and choose the answer that genuinely seems best to you.

A.9.2 I2. Task 2 Instructions and Incentive

Participants were then asked to complete a second task:

In this task, you will use one AI model to help answer a factual question.

You may only select one model, and you will not be able to change your choice.

You may request advice from the model once. After reviewing its response, you will enter a final answer.

You will receive a small bonus payment if you get the answer correct.

Please select which model you would like to use:

A.10 J. Post-Task Survey Measures

A.10.1 J1. AI Familiarity and Usage

AI model familiarity. Which of the following AI models have you heard of? (Select all that apply)

- OpenAI ChatGPT
- Anthropic Claude
- Google Gemini
- Meta Llama

- Perplexity AI
- Mistral
- xAI Grok

Preferred AI model. Do you currently have a preferred AI assistant or large language model?

- Yes
- No
- Not sure

If yes:

- Which AI assistant do you prefer?

AI usage frequency. Before today, how often would you say that you use AI assistants? (1 = Never, 5 = Multiple times a day)

Future AI use. How likely are you to use AI assistants in the future? (1 = Very unlikely, 5 = Very likely)

Perceived differences between models. In your opinion, how different are major AI assistants from one another? (1 = No difference, 5 = Very different)

A.10.2 J2. AI Trust and Evaluation

Participants rated the following on a 1–5 scale:

- How much do you trust AI assistants to provide accurate information?
- How useful do you find AI assistants for answering everyday questions?
- How comfortable are you using AI assistants to look up factual information?

A.10.3 J3. Evaluation of Selected Model

Participants evaluated the model they selected in Task 2:

- I liked the answers given by the model
- I trusted the answers given by the model
- The answers given by the model were accurate
- The answers given by the model matched my worldview

(1 = Strongly disagree, 5 = Strongly agree)

A.11 K. Demographic Variables

Age. What is your year of birth?

Education. What is the highest level of education you have completed?

- Some high school or less
- High school diploma or GED
- Some college, but no degree
- Associates or technical degree
- Bachelor's degree
- Graduate or professional degree
- Prefer not to say

Hispanic origin. Are you of Spanish, Hispanic, or Latino origin?

- Yes
- No

Race. Choose one or more races that you consider yourself to be:

- White or Caucasian
- Black or African American
- American Indian/Native American or Alaska Native
- Asian
- Native Hawaiian or Other Pacific Islander
- Other
- Prefer not to say

Gender. What is your sex?

- Male
- Female
- Other/Prefer not to say

Military service. Have you ever served on active duty in the US Armed Forces?

- Yes
- No

Employment status. Which statement best describes your current employment status?

- Working (paid employee)
- Working (self-employed)
- Not working (temporary layoff)
- Not working (looking for work)
- Not working (retired)
- Not working (disabled)
- Not working (other)
- Prefer not to answer

A.12 L. Data and Replication

Data, materials, analysis code, and the preregistered study design will be made publicly available upon publication at

https://osf.io/sqhtz/overview?view_only=0d45107b4244409c96eed81e8a2c018b.