

Memory and Ideological Bias in Conversations With Large Language Models: Experimental Evidence Across Four Countries

Sebastián Vallejo Vera* Tiago Ventura† Ernesto Calvo‡ Paula Clerici§
Christian Arnold¶

October 21, 2026

Abstract

We investigate whether interactions between users and Large Language Models (LLMs) produce ideological responses and the role of chatbots’ memory in their ideological leaning. We embed gpt-5.6-terra in an online survey experiment fielded to national samples in Argentina, Brazil, the United Kingdom, and the United States. Each respondent completes two sequential conversations, and we measure the ideological content of both the model and the user’s text at every turn. Between the conversations, we randomize whether the model propagates information from the first round into the second. Within conversations, we test whether the model’s expressed position converges toward the respondent’s, whether the range of ideological content it expresses narrows over time, and whether it exhibits a directional tilt independent of the respondent. Across conversations, we test whether a memory-primed model imports ideological content from an earlier charged exchange into a subsequent topic, including one that would ordinarily not elicit polarized responses.

*University of Western Ontario. Email: sebastian.vallejo@uwo.ca.

†Georgetown University. Email: tv186@georgetown.edu

‡University of Maryland, College Park. Email: ecalvo@umd.edu.

§Universidad Torcuato Di Tella. Email: pclerici@utdt.edu.

¶Birmingham University. Email: c.arnold.2@bham.ac.uk.

Introduction

For over a decade, research across the social, cognitive, and computer sciences has examined how algorithmic systems structure the political information people encounter online. Studies of recommendation engines on platforms such as YouTube, Facebook, and Twitter have asked whether these systems enclose users in ideological “echo chambers” ([Bakshy, Messing and Adamic, 2015](#); [Guess, 2021](#); [Nyhan et al., 2023](#)), draw them along “rabbit holes” toward increasingly extreme content ([Ribeiro et al., 2020](#); [Kaiser and Rauchfleisch, 2019](#); [Chen et al., 2023](#)), or impose a platform-wide ideological bias independent of any individual user ([Ibrahim et al., 2023](#); [Bouchaud, Chavalarias and Panahi, 2023](#); [Huszár et al., 2022](#)). Work around these three outcomes—echo chambers, rabbit holes, and system-wide bias—has organized much of the subsequent empirical work on algorithmic political mediation, which generally finds the effects to be real but modest.¹

Large language models (LLMs) represent a categorically different mode of algorithmic mediation. A recommendation feed is largely unilateral: the platform shows content to which the user reacts by clicking, scrolling, or skipping, shaping the algorithm indirectly and in the aggregate. By contrast, the behaviour of an LLM is altered through its engagement with the user, turn by turn, at an individual level. Every user message modifies the model’s behaviour for the turn that follows. The political information environment a user encounters during a conversation is therefore not a property of the algorithm alone, but of the interaction between the user and the model over the exchange. Furthermore, the way people increasingly use conversational AI is not single-shot. LLMs retain information about the user across conversations, carrying what they have learned in one exchange into the next. The retained ‘memory’ will alter how LLMs behave,

¹See [Brown et al. \(2026\)](#) for an integrated formal framework for these three concepts.

and how we study them. Political preferences established in one conversation can persist into a later, unrelated one, so that a model primed by an earlier charged exchange may approach a subsequent and usually unremarkable topic as though it, too, were contested. This is akin to the effects of issue alignment (Baldassarri and Gelman, 2008) and opinion polarization across topics in social networks (Baumann et al., 2021).

In this paper, we build on the analytical categories developed to understand political effects of unilateral, aggregate algorithmic systems and extend them to conversational AI along two dimensions. First, we ask, during a conversation with a human, whether the model’s expressed ideology converges toward the user’s position, whether the range of content it expresses narrows as the conversation proceeds, and whether it carries a directional bias independent of the user—the LLM counterparts to echo chambers, rabbit holes, and system-wide bias, adapted to a setting in which there is no network of users, but rather only one user and one model. Second, across conversations, we examine what the model carries forward through memory, and whether a memory-primed model amplifies these ideological responses. To do so, our design embeds an LLM (gpt-5.6-terra) in an online survey experiment fielded to national samples in four countries (Argentina, Brazil, the United Kingdom, and the United States) and will measure the ideological content of both the model’s and the user’s text. Each respondent completes two multi-turn conversations. Between these conversations, we will randomize whether the model retains a memory of the first round of conversations.

Our hypotheses take the following structure. Three hypotheses concern behaviour of the LLM within a conversation—convergence toward the user, a narrowing of the range of content expressed, and a system-wide directional bias—and two concern memory across conversations—the transfer of ideological content onto an otherwise uncontested topic (on the LLM side), and

the amplification of the user’s own expressed ideological stance from engaging with a model that has retained the memory of a conversation on a previously polarized topic. The hypotheses, their dependent variables and predicted directions, the statistical tests, and the interpretation given to each possible outcome are developed in the sections that follow and consolidated in the Design Table (Table 1).

A growing literature has begun to characterize LLMs as political objects. Their outputs exhibit ideological tilts (Motoki, Pinho Neto and Rodrigues, 2024; Rozado, 2024); those tilts vary with the language of the prompt and the cultural context of the user (Walker and Timoneda, 2024; Vallejo Vera and Driggers, 2025; Gallegos et al., 2024); and in structured debate they can be more persuasive than human interlocutors (Salvi et al., 2025). Models also tend, in their default state, to affirm the views and reasoning of the user they are speaking with—i.e., sycophancy (Sharma et al., 2024; Cheng et al., 2026)—which can move the positions users express, in some settings toward moderation rather than extremity (Conlon and Schwardmann, 2026). This body of work has clarified what LLMs produce and how they can move users, but it has largely examined static outputs during the course of a single conversation and worked mainly with models in English, building evidence from single-country studies. Our work makes three contributions to this growing literature. First, we theorize three distinct types of ideological behaviour, building on prior research on algorithmic systems and social media, that more precisely identify how these new AI systems moderate the political information they share with users. Second, we focus on multi-turn interactions and multi-round conversations, randomizing the memory component of these models, which allows us to precisely identify the role of memory in LLMs’ political outputs. Third, we extend our design across four countries and three languages, assessing critical aspects of contextual variation in political mediation across AI-human

interactions and the robustness of our findings across multilingual systems.

Hypotheses

Because a conversation between a respondent and an LLM produces ideologically measurable content from two agents, we divide the hypotheses into a set relating to the behaviour of the LLM and a set relating to the behaviour of the respondent. Throughout, our unit of analysis is the text produced by both the LLMs and the respondents during the conversations. We measure ideological leaning on a common left–right scale using an ensemble of LLMs.²

LLM-Side Hypotheses: Behaviour Within a Conversation

We begin with three core hypotheses about the LLM’s behaviour. Figure 1 summarizes our hypotheses graphically.

Our first hypothesis predicts convergence between LLMs’ responses and users’ political leaning. That is, the ideological distance between the positions expressed by the LLM and the respondent will decrease as the conversation evolves. This convergence may occur because the LLM moves toward the respondent, the respondent moves toward the LLM, both move toward each other, or their positions converge through some other combination of movements. Our primary hypothesis concerns whether the distance between the two positions decreases, rather than which participant produces that convergence. Visually, we present this in the left-most plot in Figure 1. To test this hypothesis, we use the distance between LLMs’ and respondents’ ideological leaning at each conversational turn as the primary outcome, estimated directly from their responses. We expect that as the conversation evolves, the gap will shrink as LLMs and

²The ensemble consists of Gemma, Mistral, and Llama models drawn from model families other than the model used in the conversation turns (gpt-5.6-terra). This avoids having the same model score its own output. We detail the scoring process in *Outcomes: Estimating Ideology* and Appendix C.

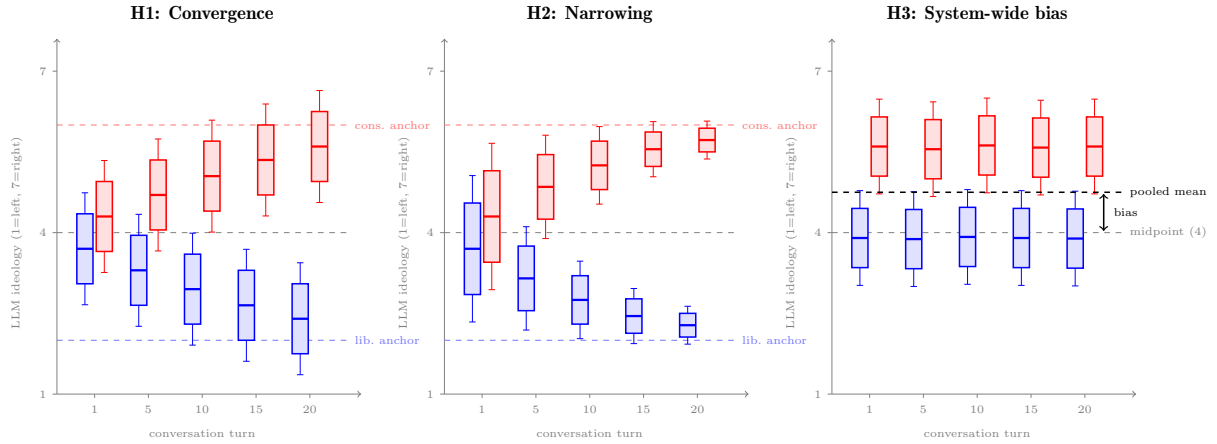


Figure 1 Schematic of Hypotheses 1–3 if true. Each panel shows the distribution of LLM ideology (1–7 scale, midpoint dashed at 4) across conversation turns, for a liberal respondent (blue) and a conservative respondent (red) shown together. **H1 (Convergence):** the gap between the LLM’s position and the respondents’ ideological position converges (blue toward the liberal anchor, red toward the conservative anchor) while the spread of expressed content stays wide. **H2 (Narrowing):** the position converges as in H1 *and* the box heights shrink over turns, indicating a narrowing of the range of content the LLM expresses. **H3 (System-wide bias):** both respondents’ LLM positions sit off the midpoint in the same direction and stay flat across turns; the pooled mean departs from the midpoint (here, a conservative tilt) regardless of respondent ideology. The panels are illustrative schematics, not predictions of magnitude.

respondents converge on an ideological position. Formally, we expect:

H1 Echo chambers: convergence in position. The ideological distance between the positions expressed by the LLM and the respondent will decrease as conversational turns evolve.

Our second hypothesis predicts that the variance of the LLM’s responses will decrease over the course of a conversation. That is, conversations with a more conservative respondent will produce more conservative LLM responses (H1), and these responses will be more precise along the ideological spectrum. Visually, we present this in the middle plot in Figure 1. Our primary outcome is variance around the LLM’s responses across a moving window of conversational turns.

We expect:

H2 Rabbit holes: narrowing in variance. The across-turn variance of the LLM’s

position will decrease over the course of a conversation, indicating a narrowing of the range of ideological content the LLM expresses as the conversation progresses.

Our third hypothesis predicts that the ideological leaning of the LLM’s responses can be orthogonal to respondents’ own partisan preferences expressed during the conversation. In other words, the LLM’s position will depart from the scale midpoint in a consistent direction after controlling for the respondent’s anchor, topic, and conversational turn, and this departure will hold steady across turns. For this hypothesis, our primary outcome is the LLM’s per-turn position centered at the scale midpoint, rather than its distance to the respondent as in H1. We additionally use the coefficient for conversational turn to describe whether any observed tilt strengthens, weakens, or remains stable over the course of the conversation. We expect:

H3 (System-wide ideological bias). After controlling for the respondent’s anchor, topic, and conversational turn, the LLM’s responses will depart from the midpoint of the scale in a consistent direction, and this departure will be stable across turns.

Our three hypotheses formalize the distinction drawn in prior work on algorithmic mediation (Brown et al., 2026) between a static, ideologically aligned distribution of content, a dynamic narrowing of that distribution toward a restricted range, and one-directional system-wide leaning of the distribution. In this sense, Hypotheses 1 and 2 separate the *position* and *variance* of the LLM’s expressed ideology. Confirming H1 with variance remaining wide (H2 not supported) indicates that the LLM produces directionally aligned but substantively diverse content. Meanwhile, confirming H1 *together with* H2 indicates that the LLM both aligns with the respondent and progressively restricts the range of content it offers.

Lastly, confirming H3, without H1, indicates that the LLM carries a consistent ideological tilt that does not depend on the respondent’s own leaning. We test these three hypotheses in

the first round of high-divisiveness conversations, since these are more likely to generate political responses from LLMs and users are not primed by previous (high-divisiveness) conversations.

LLM-Side Hypothesis: Transfer Across Conversations Through Memory

Our design randomizes what the LLM carries from Round 1 into Round 2. Round 1 always covers a highly divisive political topic; the LLM ends Round 1 knowing the position the respondent expressed on that charged topic. After Round 1, we randomize participants into memory-on and memory-off conditions. In the memory-on condition, a profile derived from Round 1 is available to the LLM at the start of Round 2; in the memory-off condition, the LLM is reset and meets Round 2 with no prior information. Under memory-on, if the convergence hypothesis is confirmed, the LLM enters Round 2 already tilted toward the respondent’s Round 1 established position, and it imports that tilt into whatever the Round 2 topic is. Thus, in Round 2, LLM responses might start with a more extreme ideological leaning given that their prior interaction happened on a high-divisiveness topic. Under memory-off, the LLM resets, meets the Round 2 conversation with no priors, and no such tilt is carried in. This effect will be even larger when the Round 2 conversation deals with a low-divisiveness topic; memory-on makes a low-divisiveness topic behave like a high-divisiveness one: the low-divisiveness topic is the setting, but the behaviour reflects that of a high-divisiveness topic (see Figure 2). To test this hypothesis, we use LLM-side Round 2 distance from the midpoint of the ideological scale (LLM extremity), evaluated during the opening conversational turns of Round 2.

Therefore, we expect:

H4a - Memory transfer - Ideologically tilted LLMs In Round 2, LLMs in the memory-on condition will begin the conversations with more extreme positions than LLMs in

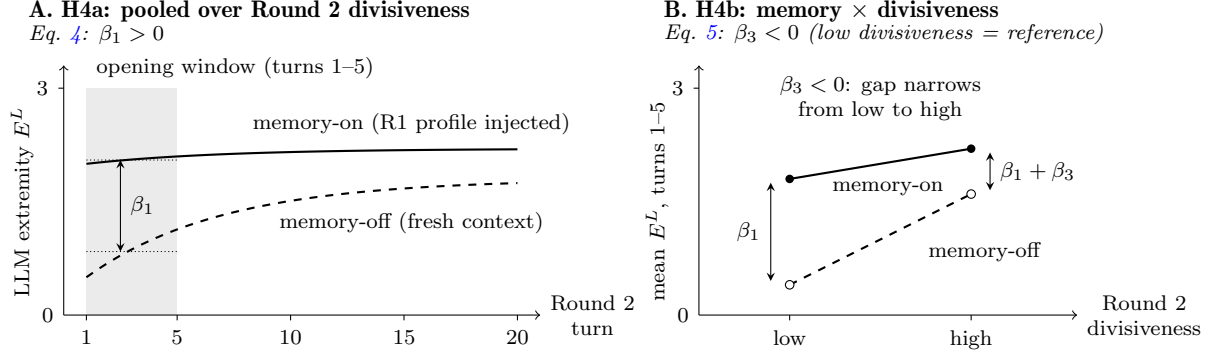


Figure 2 Schematic of Hypotheses 4a and 4b (LLM-side memory transfer) if true. The dependent variable is the LLM’s *extremity*, $E^L = |s^L - 4|$, its distance from the scale midpoint, read over the opening window of Round 2 (turns 1–5, shaded), where we expect the memory-off arm to be near-neutral because it has not yet inferred the respondent’s position. **Panel A** (H4a, Eq. 4) pools over Round 2 divisiveness. A memory-on LLM (solid), primed by the profile derived from the respondent’s high-divisiveness Round 1 exchange, opens already tilted; a memory-off LLM (dashed) opens near the midpoint and rises only as it infers the respondent from the Round 2 exchange. The difference in mean opening extremity between the two arms is the pooled memory effect β_1 ; H4a predicts $\beta_1 > 0$. **Panel B** (H4b, Eq. 5) splits the opening-window means by Round 2 divisiveness, with low divisiveness as the reference. The memory effect is β_1 on low-divisiveness topics and $\beta_1 + \beta_3$ on high-divisiveness topics, where the topic already drives extremity in the memory-off arm and leaves less room for memory to add; H4b predicts $\beta_3 < 0$, so the gap between the arms narrows from low to high divisiveness. Coefficients in each panel follow the numbering of their own equation. The panels are illustrative schematics, not predictions of magnitude.

the memory-off condition.

H4b - Memory transfer on Low-divisiveness Topics The effect of memory on the extremity of the LLM at the onset of Round 2 conversations will be larger for low-divisiveness topics than for high-divisiveness topics.

User-Side Hypothesis: Amplification

The LLM-side hypotheses treat the respondent’s anchor as an exogenous predictor of LLM behaviour. We now model how the ideological leaning of user responses is influenced by the LLM behaviour.³ Our core memory randomization allows us to test whether a previously polarized LLM, rather than a fresh-state LLM, amplifies this behaviour, even when the topic analyzed, in

³Note that we are not arguing that the user’s ideology is necessarily changing as they interact with the LLM. The short term structure of the treatment is unlikely to do so. Rather, we argue that the user will reveal an increasingly polarized position on a topic, even a more extreme one. See *Analysis plan* for more details on the estimation.

regular conditions (low divisiveness topic) would not produce this effect. This setup, we believe, reflects how users currently use LLMs: models that have a context of the user’s preferences and cater to aligning with these preferences (Rathje et al., 2024; Cheng et al., 2026).

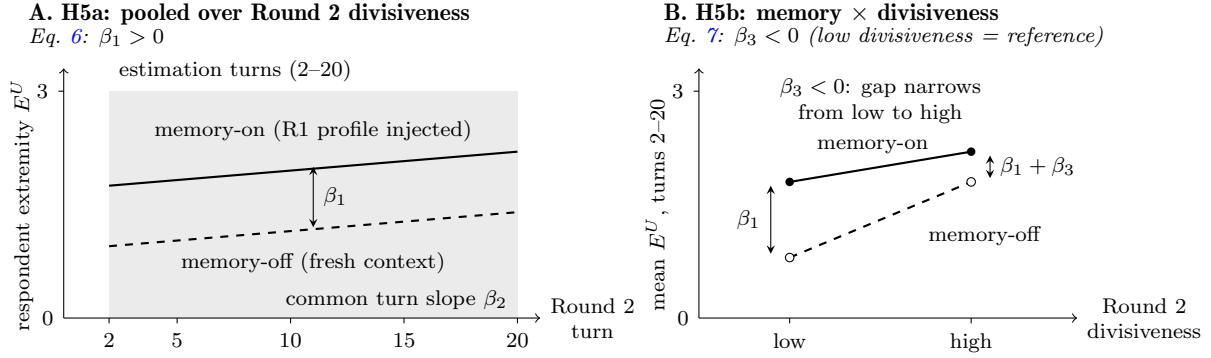


Figure 3 Schematic of Hypotheses 5a and 5b (user-side amplification) if true. The dependent variable is the respondent’s *extremity*, $E^U = |s^U - 4|$, the distance of their expressed position from the scale midpoint, over Round 2 turns from the second turn onward (shaded). **Panel A** (H5a, Eq. 6) pools over Round 2 divisiveness. Respondents conversing with a memory-on LLM (solid), primed by the profile derived from their own high-divisiveness Round 1 exchange, express more extreme positions than respondents conversing with a memory-off LLM (dashed). Both arms share a common turn trend (β_2); the vertical gap between them is the average memory effect β_1 , and H5a predicts $\beta_1 > 0$. The prediction concerns the level of extremity, so the gap need not widen over turns. **Panel B** (H5b, Eq. 7) splits mean extremity by Round 2 divisiveness, with low divisiveness as the reference. The memory effect is β_1 on low-divisiveness topics and $\beta_1 + \beta_3$ on high-divisiveness topics, which already push respondents toward more extreme positions in the memory-off arm and leave less room for memory to add; H5b predicts $\beta_3 < 0$, so the gap between the arms narrows from low to high divisiveness. Coefficients in each panel follow the numbering of their own equation. The panels are illustrative schematics, not predictions of magnitude.

Figure 3 presents our hypothesis visually. The primary outcome is the respondent’s ideological extremity in Round 2: the distance of their position at each turn from the midpoint of the scale. Larger values indicate more extreme expressed positions, and a decrease indicates moderation. A version oriented toward the side the respondent took in Round 1 is reported in Appendix G. For this family of hypotheses, we expect:

H5a (Amplification, directional). Respondents in the memory-on condition will express more extreme positions than respondents in the memory-off condition.

H5b (Amplification on low-divisiveness topics). The amplification effects of memory

will be larger on low-divisiveness Round 2 topics.

A summary of all the hypotheses is presented in Table [1](#).

Table 1 Design Table

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given to different outcomes
RQ1: Within a conversation, does the LLM’s expressed position converge toward the respondent’s, does the range of ideological content it expresses narrow, and does it exhibit a directional ideological tilt independent of the respondent?	Below are our within-conversation hypotheses, which we test on all Round 1 conversations and all turns: H1. The ideological distance between the positions expressed by the LLM and the respondent will decrease as conversational turns evolve. H2. The across-turn variance of the LLM’s position will decrease over the course of a conversation, indicating a narrowing of the range of ideological content the LLM expresses. H3. After controlling for the respondent’s anchor, topic, and conversational turn, the LLM’s responses will depart from the midpoint of the scale in a consistent direction, and this departure will be stable across turns.	Our target sample size is 10,000 respondents, 2,500 in each of Argentina, Brazil, the United Kingdom, and the United States, and every respondent contributes a Round 1 conversation. Assuming 95% statistical power and two-sided tests with p-values < 0.05, our power analysis estimates that this sample size can detect a minimum effect size of approximately 0.036 SD for each of H1, H2, and H3. See the <i>Power Analysis</i> section.	For H1, we will regress the LLM–respondent gap, the absolute distance between the LLM’s and the respondent’s scored positions at the same turn, on conversational turn (Eq. 1). For H2 we will regress the dispersion of the LLM’s position, the standard deviation of its scores over a trailing window of five turns, on conversational turn (Eq. 2). For H3 we will regress the LLM’s position centered at the scale midpoint on the respondent’s midpoint-centered anchor and the mean-centered turn (Eq. 3). All models include topic and country fixed effects and cluster standard errors at the respondent level. See the <i>Analysis Plan</i> for full models. We will apply the Benjamini-Hochberg False Discovery Rate adjustment across the three tests, and present all tests with unadjusted p-values and adjusted q-values. Our measure of statistical significance will be two-sided p-values < 0.05. See the <i>Multiple comparisons and inference</i> subsection.	For H1, if we find a significant negative effect of turn on the gap, we will reject the null and find evidence that the positions expressed by the LLM and the respondent converge over the course of the conversation. For H2, if we find a significant negative effect of turn on dispersion, we will reject the null and find evidence that the range of content narrows; H1 supported with H2 not supported indicates aligned but substantively diverse content, and both supported indicate alignment with a narrowing range. For H3, if the intercept differs significantly from zero, we will reject the null and find evidence of a system-wide tilt, conservative if positive and liberal if negative. We will use the coefficient for conversational turn and the corresponding predicted values to describe whether the tilt strengthens, weakens, or remains stable over the course of the conversation. To evaluate the robustness of these null effects, we will compute two-sided default Bayes factors (BF_{01}) to quantify the relative evidence for the null over the alternative, and will interpret ($BF_{01} \geq 10$) as strong evidence in favor of the null.

Design Table (continued)

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given to different outcomes
RQ2: Across conversations, does the LLM carry a Round 1 tilt into Round 2 through memory, and is the memory-driven tilt larger on low-divisiveness Round 2 topics?	H4a. In Round 2, LLMs in the memory-on condition will begin the conversations with more extreme positions than LLMs in the memory-off condition. H4b. The memory-driven tilt at the onset of Round 2 will be larger on low-divisiveness topics than on high-divisiveness topics.	Our target sample size is 10,000 respondents, half assigned to memory-on and half to memory-off, crossed with high- and low-divisiveness Round 2 topics in a balanced 2×2 . Assuming 95% statistical power and two-sided tests with p-values < 0.05 , our power analysis estimates that this sample size can detect a minimum effect size of approximately 0.072 standard deviations for the memory effect in H4a and 0.144 standard deviations for the memory $\times$ divisiveness interaction in H4b. See the <i>Power Analysis</i> section.	The outcome for both hypotheses is the LLM's extremity, the absolute distance of its scored position from the midpoint of the scale, over the first five turns of Round 2. For H4a we will regress extremity on the memory condition, pooling over turns and Round 2 divisiveness (Eq. 4). For H4b, we add the divisiveness condition and its interaction with memory (Eq. 5). All models include topic and country fixed effects and cluster standard errors at the respondent level. See the <i>Analysis Plan</i> for full models. For these two hypotheses, we will apply the Benjamini-Hochberg False Discovery Rate adjustment across the two tests, and present all tests with unadjusted p-values and adjusted q-values. Our measure of statistical significance will be two-sided p-values < 0.05 . See the <i>Multiple comparisons and inference</i> subsection.	For H4a, if we find a significant positive effect of the memory condition on the LLM's opening extremity, we will reject the null and find evidence that the Round 1 profile transfers a tilt into Round 2. A significant negative effect would indicate that memory moderates the LLM, which we do not predict and would report as such. For H4b, if we find a significant negative interaction between memory and divisiveness, we will reject the null and find evidence that the transfer is concentrated on low-divisiveness topics, where memory makes a calm topic behave as a charged one. A non-significant interaction indicates uniform transfer across topic types, and a significant positive interaction would indicate stronger transfer on high-divisiveness topics. For any effects with p-values greater than 0.05, we will fail to reject the null hypothesis. To evaluate the robustness of these null effects, we will compute default Bayes Factors (BF_{01}) to quantify the relative evidence for the null over the alternative, and will interpret a Bayes Factor of at least 10 in favor of the null as strong evidence for it.

Design Table (continued)

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given to different outcomes
RQ3: Does interaction with a memory-primed LLM push the respondent toward more extreme expressed positions, and is this amplification larger on low-divisiveness Round 2 topics?	H5a. Respondents in the memory-on condition will express more extreme positions than respondents in the memory-off condition. H5b. The amplification effects of memory will be larger on low-divisiveness Round 2 topics.	Our target sample size is 10,000 respondents, half assigned to memory-on and half to memory-off, crossed with high- and low-divisiveness Round 2 topics in a balanced 2×2 . Assuming 95% statistical power and two-sided tests with p-values < 0.05 , our power analysis estimates that this sample size can detect a minimum effect size of approximately 0.072 standard deviations for the memory effect in H5a and 0.144 standard deviations for the memory $\times$ divisiveness interaction in H5b. See the <i>Power Analysis</i> section.	The outcome for both hypotheses is the respondent's extremity, as the absolute distance of their scored position from the midpoint of the scale, at every turn of Round 2. For H5a we will regress respondent extremity on the memory condition and conversational turn, pooling over Round 2 divisiveness (Eq. 6). For H5b we will add the divisiveness condition and its interaction with memory (Eq. 7), with low divisiveness as the reference category, so that the memory coefficient is the memory effect on low-divisiveness topics and the interaction is the change in that effect on high-divisiveness topics. All models include topic and country fixed effects and cluster standard errors at the respondent level. See the <i>Analysis Plan</i> for full models. For these two hypotheses, we will apply the Benjamini-Hochberg False Discovery Rate adjustment across the two tests, and will present all tests with unadjusted p-values and adjusted q-values. Our measure of statistical significance will be two-sided p-values < 0.05 . See the <i>Multiple comparisons and inference</i> subsection.	For H5a, if we find a significant positive effect of the memory condition on respondent extremity, we will reject the null and find evidence of amplification. A significant negative effect would indicate moderation, the competing outcome, and we will report it as evidence against H5a. For H5b, if we find a significant negative interaction between memory and divisiveness, we will reject the null and find evidence that amplification is concentrated on low-divisiveness topics. A non-significant interaction indicates uniform amplification across topic types, and a significant positive interaction would indicate stronger amplification on high-divisiveness topics. For any effects with p-values greater than 0.05, we will fail to reject the null hypothesis. To evaluate the robustness of these null effects, we will compute default Bayes Factors (BF_{01}) to quantify the relative evidence for the null over the alternative, and will interpret a Bayes Factor of at least 10 in favor of the null as strong evidence for it.

Methods

Ethics Information

Ethics approval for this project was received by the University of Western Ontario non-medical research ethics board (Project ID: 127701). Participants will provide informed consent at the beginning of each survey, and will be debriefed after the entire study is over. Survey payments will be based on the typical payment rate provided by the survey company in each country, adjusted for the local standard of living.

Sample and Recruitment

The study will field national samples of adult respondents in four countries: Argentina, Brazil, the United Kingdom, and the United States. The target sample size is approximately 2,500 respondents per country, for a total registered sample of approximately 10,000 completes; the full power analysis is reported in the *Power Analysis* subsection. Recruitment will be conducted through country-specific panel providers: Leger (United States), Netquest (Argentina and Brazil), and Prolific (United Kingdom), using nationally representative quotas for each country.

The decision to field across four countries was made primarily for practical reasons: country-specific panels provide the most reliable access to representative national samples in the participating countries, distributing the sample across four countries allows us to obtain sufficient statistical power within the available fielding budget, and the country variation allows us to discuss contextual differences between LLMs and users' behaviour. The cross-language composition of the sample—Spanish in Argentina, Portuguese in Brazil, and English in the United

Kingdom and the United States—provides an additional opportunity to examine whether the patterns we test vary with the language of interaction, an issue raised by prior work on LLM ideological tilt as a function of prompt language and cultural context (Walker and Timoneda, 2024; Vallejo Vera and Driggers, 2025; Waight et al., 2026). Conversations are conducted in each country’s primary language, and per-turn ideological scoring is conducted in the language of the conversation. A robustness analysis in which all texts are re-scored after translation into English will be reported in Appendix F; the substantive analyses reported in the main text are based on the original-language scoring.

Design

Each respondent completes two sequential conversations with a Large Language Model (LLM), referred to as Round 1 and Round 2. The topic of conversation varies within respondents across the two rounds. In Round 1, all respondents will interact with the LLM on a randomly assigned high-divisiveness topic. In Round 2, two factors are randomized between respondents: whether the LLM has access to a memory of the respondent’s Round 1 conversation (memory on versus memory off), and whether the Round 2 topic is drawn from a high-divisiveness or a low-divisiveness pool of political issues.⁴

The central manipulation is the memory condition. Every respondent first completes a Round 1 conversation on a high-divisiveness political topic with no prior context. At the conclusion of Round 1, the conversational model is prompted to produce a short profile of the respondent, summarizing what it has learned about the user over the course of the exchange (see *Memory Manipulation* below). The LLM will be explicitly prompted to describe the user’s position in the issue based on their conversation, the responses of the LLM, and the user’s reaction

⁴In Appendix I, we describe a similar pilot study we conducted with a Canadian sample. The pilot informed many of the design decisions, as well as allowed us to validate the ideology scoring approached used.

to the response from the LLM. In the memory-on condition, this profile is supplied to the model at the start of Round 2; in the memory-off condition, Round 2 begins with no carryover from Round 1 and the model is reset to a fresh context. The LLM’s access to Round 1 is manipulated. Comparisons between the memory-on and memory-off conditions in Round 2 isolate the contribution of algorithmic memory. This manipulation serves a two-fold purpose: 1) it allows us to identify what information the LLM retains about a user from a conversation and, in the next step, 2) evaluate the effect of prior high-divisiveness conversations on subsequent LLM-human interactions.

Conversation Structure

Before respondents engage in the first conversation on a divisive political topic in Round 1, they will all be asked to engage in a conversation with the LLM on a non-political conversation about *food and cooking* for 3 turns. We call this the warm-up period, aimed at getting respondents familiarized with the task.

After the warm-up, each respondent completes two conversations, Round 1 and Round 2. Each substantive conversation consists of a minimum of 10 and a maximum of 12 turns, where a turn comprises one user message and the LLM’s response to it.⁵ These conversations are about a substantive political topic, introduced by a topic-initiating prompt provided to the model (constant across respondents within a topic and country). The first turn of the substantive portion is initiated by this prompt; subsequent turns are open-ended and driven by the exchange between the respondent and the model. The full sample completes Round 1 under the same experimental conditions. Its substantive topic is drawn from a pool of high-divisiveness political issues (see *Topics*), and it is conducted with no memory of any prior interaction. The LLM is not

⁵Users are required to complete at least 10 interactions with the LLM, and the conversation is capped at 20 interactions.

provided with information about the user’s ideological position, gender, age, level of education, or voting behaviour.

Within a conversation, the model’s chat history is preserved across all completed turns, so that each turn depends on the accumulated context of that conversation. Between Round 1 and Round 2, the model’s chat history is reset; the only channel by which Round 1 content reaches Round 2 is the LLM-generated profile, and only in the memory-on condition. Conversations terminate at 12 turns. Respondents are informed of the 12-turn structure in advance but are required to complete a minimum of 5 turns before they may advance, after which they may either continue to the 12-turn limit or move on. To improve response quality, we also impose a minimum word count on user messages during the warm-up conversation.

Round 2 is completed by the full sample as well; its substantive topic is drawn from either the high-divisiveness or the low-divisiveness pool, randomly assigned, and it is conducted under the randomly assigned memory condition. The design thus yields three distinct sets of conversations: Round 1 conversations (full sample, high-divisiveness, no memory context); Round 2 memory-on conversations, in which a profile derived from the respondent’s Round 1 conversation is available to the model; and Round 2 memory-off conversations, in which the model has no access to Round 1. We present a diagram of the possible conditions in Figure 4.

All conversations are conducted with gpt-5.6-terra (OpenAI). Generation parameters are reported in Appendix B; all parameters not stated there are held at provider defaults. We accept that model outputs are non-deterministic at the temperature used, and the design is powered to average over this per-conversation stochasticity.

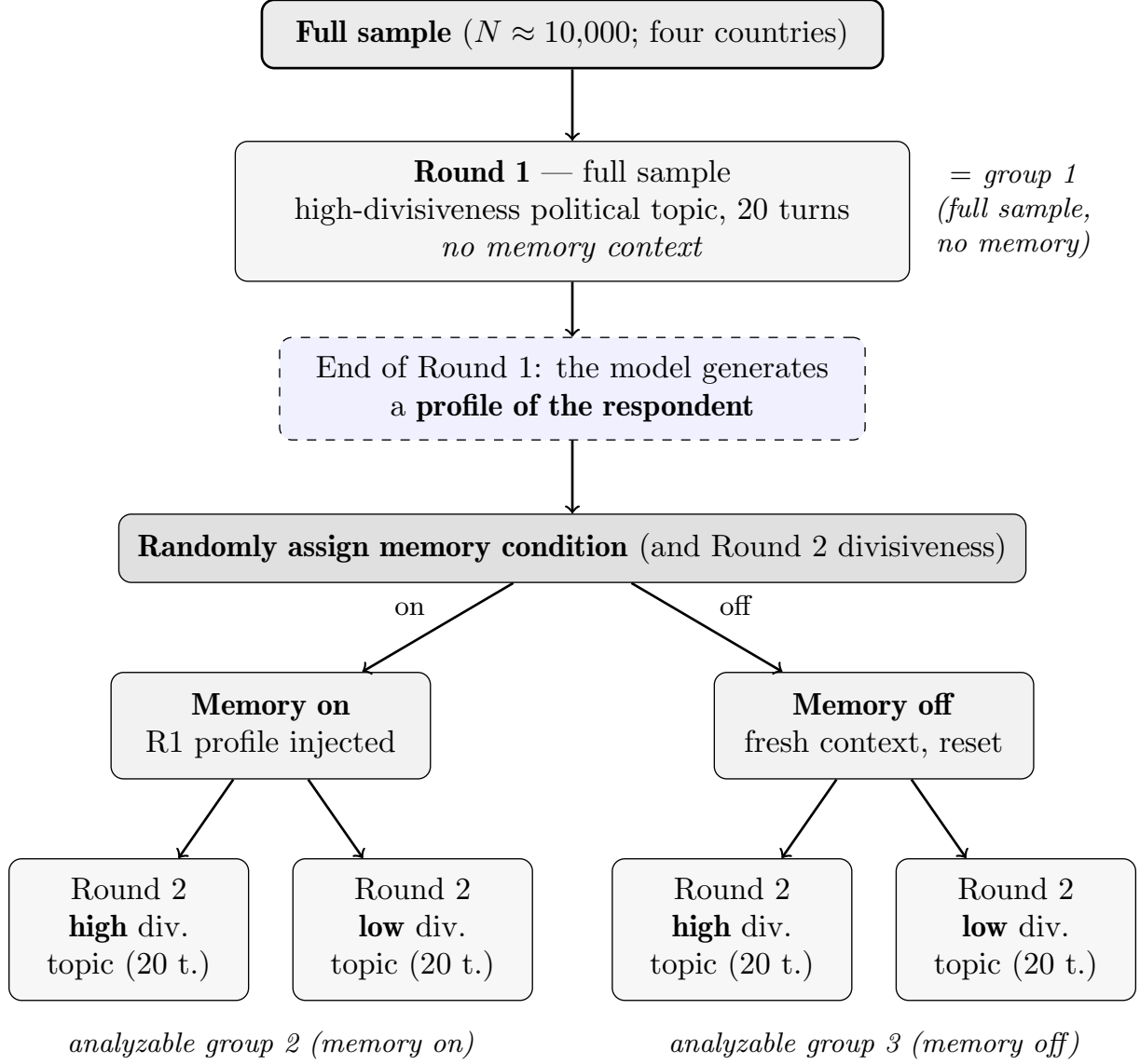


Figure 4 Study design and the analytically distinct sets of conversations. Every respondent completes Round 1—a 20-turn conversation on a high-divisiveness political topic—with no memory context (group 1). At the end of Round 1 the conversational model generates a profile of the respondent. Respondents are randomly assigned to a memory condition, which defines two streams: in the memory-on stream (group 2) the Round 1 profile is supplied to the model at the start of Round 2 (dashed arrow); in the memory-off stream (group 3) the model is reset and has no access to Round 1. Within each memory stream, the Round 2 topic (also a 20-turn conversation) is randomly drawn from the high- or the low-divisiveness pool, yielding a 2×2 arrangement of Round 2 cells (memory $\times$ divisiveness). Because every respondent is the same person across both rounds, respondents in both streams carry forward their own recollection of Round 1; only the model’s access to Round 1 is manipulated.

Topics

For each country, political topics are organized into two pools: a high-divisiveness pool and a low-divisiveness pool. Divisiveness is defined at the level of the country, since issues that are contested in one national context need not be contested in another; the pools are therefore constructed separately for each country rather than drawn from a common cross-national list. Each pool contains multiple issues (for example, immigration (U.K.) and environment (Brazil) in the high-divisiveness pool and investment in infrastructure (U.S.) and support for small businesses (Argentina) in the low-divisiveness pool for a given country). The full set of issues in each pool, for each country, together with the exact topic-initiating prompts in each conversation language, is provided in Appendix A and summarized in Table A1.

Round 1 draws its topic from the high-divisiveness pool. Round 2 draws from either the high- or the low-divisiveness pool, according to the randomly assigned Round 2 divisiveness condition, and—when it draws from the high-divisiveness pool—selects an issue distinct from the one used in Round 1. Within the high-divisiveness pool, the assignment of issues to rounds is randomized and counterbalanced, so that a given issue appears in Round 1 for some respondents and in Round 2 for others, and so that ordered pairs of issues are observed in both orders (issue A followed by issue B, and issue B followed by issue A). This counterbalancing ensures that any difference associated with the memory condition is attributable to the experimental setup—the memory manipulation and the accumulated conversational interaction—rather than to the content of any particular issue or issue ordering.⁶ We have no theoretical interest in heterogeneity across the substantive issues themselves; we select politically salient issues that provide clear

⁶The counterbalancing is symmetric only for high-to-high transitions, in which both the Round 1 and Round 2 topics are drawn from the same pool. Because low-divisiveness issues never appear in Round 1, high-to-low transitions have no reverse; the low-divisiveness Round 2 arm is identified from the memory manipulation together with the pooled low-divisiveness issues, and the inference that issue identity is not driving the memory effect transfers to the low-divisiveness arm by assumption rather than by symmetric design.

ideological discrimination within each country, while the *food and cooking* warm-up topic provides a baseline for evaluating the per-turn ideology-measurement instrument. Nevertheless, for robustness, we will report analyses measuring heterogeneity across topics in Appendix H.

Memory Manipulation

The memory manipulation determines whether the model carries information about the respondent from Round 1 into Round 2. We implement memory as a model-generated profile rather than as a verbatim transcript. At the conclusion of each respondent’s Round 1 conversation, the conversational model (gpt-5.6-terra) is prompted, in a separate call, to produce a concise profile of the respondent that summarizes what it has learned about the user over the course of the conversation. Note that for the profile, the model is only supplied the conversation, and not the demographic information, to capture what the model prioritizes solely from the conversation.⁷ The exact profile-elicitation prompt is provided in Appendix C. In the memory-on condition, this profile is added to the model’s context at the start of Round 2; in the memory-off condition, no profile is supplied and Round 2 proceeds from a fresh context.

Outcomes: Estimating Ideology

Scoring is performed by an ensemble of LLMs distinct from the model used in the LLM-respondent conversation. Using scorers from different model families avoids any self-evaluation biases that might arise from a model scoring its own output. The LLMs used to score are Claude, Mistral, and Llama; the pinned snapshots of the scorers are reported in Appendix C. Each text unit is scored independently by every scorer in the ensemble, so that each per-turn position estimate is associated with a distribution of scores across scorers. The spread of the across-

⁷We produce a second profile with the demographic information included, and compare the differences in both. The comparison is reported in Appendix C.

scorer distribution reflects scorer disagreement about a given text, and allows us to measure reliability.

At each turn of each conversation, the substantive content of the LLM’s response (excluding any follow-up question the model may pose) and the full text of the respondent’s message are scored by this ensemble of LLMs (i.e., scorer ensemble) on a 1–7 left–right ideological scale, where 1 denotes the most strongly left-leaning and 7 the most strongly right-leaning position on the topic of the conversation. Each turn-level text—whether produced by the model or the respondent—is appended, and the scorer is asked to determine whether there is enough information to provide a score, and, if so, to score (separately) the respondent and the LLM. Since the scorer ensemble is provided with the user’s self-reported 1–7 left-right ideological position, it allows the scorers to have an anchor from where they begin their estimation of the user’s ideology. Note that the scorers are not provided with any information about the user when scoring the ideological placement of the LLM. Part of the model-generated Round 1 profile is scored on the same scale, yielding an ideological reading of what the model encoded about each respondent. The scorers maintain a separate running estimate for users and LLMs, with the user’s self-placement anchoring the user-side series (its $t=0$ value by construction) and the LLM-side series starting wherever the LLM’s first substantive turn scores.

The scorer prompt is identical across scorers and is provided verbatim in Appendix C. Native-language scoring is the primary measure; the English-translation re-scoring described above is reported as a robustness analysis in Appendix F. Validation of the scoring instrument against human coders is described in Appendix E and its outcome-neutral pass criterion in *Outcome-Neutral Conditions*.

Randomization Platform

Random assignment to substantive topics, Round 2 topic divisiveness, and memory condition is implemented through Qualtrics’ BlockRandomizer. Assignment is independent across respondents. Within the high-divisiveness pool, issue-to-round assignment is randomized subject to the counterbalancing described under *Topics*.

Outcome-Neutral Conditions and Manipulation Checks

Four outcome-neutral conditions must be met for the data to support informative tests of the hypotheses. Failure of any condition will be reported transparently in the Stage 2 manuscript and will prevent us from interpreting the corresponding hypothesis tests as informative.

Memory manipulation check. The memory manipulation must produce observably different model behaviour across the memory-on and memory-off conditions at the start of Round 2. We pre-specify a check in which a separate LLM, called in a fresh context and blind to the respondent’s assigned condition, rates each Round 2 conversation on the extent to which the model’s early responses reflect prior knowledge of the respondent. The evaluator will be instructed to assess evidence that the model possessed information about the respondent before the Round 2 conversation, regardless of the ideological direction or extremity of the model’s responses.

Warm-up conversation as negative control. The per-turn ideology measure must distinguish ideological from non-ideological content. As a negative control, warm-up (i.e. conversation on food and cooking) turns must not produce systematic ideological updates. For the respondent, the ideology-score distributions should not differ significantly from the anchor (i.e., the initial self-placement of the user), indicating that the scorer does not update the respondent’s ideological position based on nonpolitical content. The across-turn variances should also be substantially

lower than that of the substantive political turns. For the LLM, the warm-up turns should produce no scores, and high degrees of the LLM reporting that there is not enough information to produce scores.

User-ideology validity check. The per-turn measure must validly capture respondent ideology. Respondents’ pre-treatment ideological self-placement on a standard 1–7 left–right survey item must correlate positively with the scored ideology of the end-of-Round 1 profile produced by the LLM, with a pre-specified threshold of Pearson $r > 0.25$ pooled across topics.

Absence of floor and ceiling effects. The ideology-score distributions on the political topics must not exhibit floor or ceiling effects; we pre-specify that no more than 15% of scores on the political topics may fall at either endpoint of the 1–7 scale.

Analysis Plan

General framework

In this section, we define the general notation for the statistical models we use. We define $s_{u,c,t}^L$ and $s_{u,c,t}^U$ as the ideological score (on the 1–7 left–right scale) of, respectively, the LLM’s and the respondent’s turn- t text in conversation c of respondent u . Each score is the **mean across the scorer ensemble**; the standard deviation across scorers is retained as a measure of scoring reliability and reported alongside the estimates, but the point estimate entering the models below is the ensemble mean.⁸ Then, we define the **respondent’s extremity**, $E_{u,c,t}^U = |s_{u,c,t}^U - 4|$: the distance of the respondent’s expressed position from the scale midpoint. Extremity that grows over turns indicates the respondent expressing an increasingly polarized position. An analogous LLM extremity, $E_{u,c,t}^L = |s_{u,c,t}^L - 4|$, is used for the memory-transfer test. Let a_u

⁸As a robustness check we re-estimate the primary models treating scorer as a crossed random effect, so that across-scorer disagreement propagates into the standard errors; results are reported in Appendix F.

denote the respondent’s anchor (pre-treatment ideological self-placement, 1–7) and $\tilde{a}_u = a_u - 4$ its value centered at the fixed midpoint of the scale and $\tilde{s}_{u,c,t}^L = s_{u,c,t}^L - 4$ denote the LLM’s position centered at the midpoint. Let $\text{Mem}_u = 1$ for the memory-on condition and $\text{Mem}_u = 0$ for memory-off, and $\text{Div}_c = 1$ for a high-divisiveness Round 2 topic and $\text{Div}_c = 0$ for a low-divisiveness one. Conversational turn enters as t . We define $\tilde{t} = t - \bar{t}$ as the turn centered at its mean, so that $\tilde{t} = 0$ at the middle of the conversation. All models include topic and country fixed effects (δ).

The unit of analysis is the respondent-turn for the dynamic models (Eqs. 1, 2, 4, 5, 6, 7) and the respondent-conversation for aggregate quantities. Because each respondent contributes two conversations and up to 40 turns, all standard errors in the primary specifications are clustered at the respondent level.

H1: Convergence in position

We test Hypothesis 1 on all Round 1 conversations, using the following equation:

$$\Delta_{u,c,t} = \alpha + \beta_1 t + \delta + \varepsilon_{u,c,t}. \quad (1)$$

Where $\Delta_{u,c,t} = |s_{u,c,t}^L - s_{u,c,t}^U|$ represents the distance between the LLM’s expressed position and the respondent’s expressed position at the same turn. H1 is supported if $\beta_1 < 0$ indicating that the ideological distance between the positions expressed by the LLM and the respondent decreases over turns. This primary test identifies convergence between the two series but does not require that the convergence be produced by movement from the LLM to the user, even though the literature suggests this to be the most likely outcome (Malmqvist, 2025). In Appendix G, we decompose changes in the two series to examine whether convergence occurs through

movement by the LLM, movement by the respondent, movement by both, or movement by both participants toward a common position.

H2: Narrowing in variance

We test for Hypothesis 2 using the following equation in all Round 1 conversations:

$$d_{u,c,t}^L = \alpha + \beta_1 t + \delta_{\text{topic}} + \delta_{\text{country}} + \varepsilon_{u,c,t}, \quad (2)$$

where $d_{u,c,t}^L$ is the dispersion of the LLM position series at turn t , operationalized as the standard deviation of s^L over a pre-specified trailing window of w turns.⁹ β_1 captures the narrowing in variance over turns, and H2 is supported if $\beta_1 < 0$. H1 supported with H2 not supported indicates directionally aligned but substantively diverse content; H1 and H2 both supported indicate alignment accompanied by a narrowing range.

H3: System-wide ideological bias

Estimated on all Round 1 conversations, with the DV, the anchor, and the turn centered as defined in the general framework:

$$\tilde{s}_{u,c,t}^L = \alpha + \beta_1 \tilde{a}_u + \beta_2 \tilde{t} + \delta_{\text{topic}} + \delta_{\text{country}} + \varepsilon_{u,c,t}. \quad (3)$$

Where $\tilde{s}_{u,c,t}^L = s_{u,c,t}^L - 4$ measures the LLM position's distance to the midpoint of the ideology scale. System-wide bias is a departure of the LLM's position from the midpoint that holds regardless of the respondent's own ideology. α is the predicted position of the LLM for a midpoint respondent at the average turn, averaged over topics and countries. H3 is supported

⁹ w is fixed in advance (default $w = 5$); the running-SD alternative is reported as a robustness check.

if α differs from zero: $\alpha > 0$ indicates a conservative tilt and $\alpha < 0$ a liberal tilt. β_1 captures how far the LLM’s position tracks the respondent’s anchor, so the tilt in α is the departure that survives this control. We use β_2 to assess whether the tilt strengthens, weakens, or remains stable as the conversation evolves.

H4a: Memory transfer

Estimated on the opening turns of Round 2 (the first five turns), pooled over divisiveness, with the DV the LLM’s extremity:

$$E_{u,c,t}^L = \alpha + \beta_1 \text{Mem}_u + \delta_{\text{topic}} + \delta_{\text{country}} + \varepsilon_{u,c,t}. \quad (4)$$

Where $E_{u,c,t}^L = |s_{u,c,t}^L - 4|$ measures LLM’s extremity. Because the conversational LLM receives no information about the respondent’s ideology, a memory-off LLM opens Round 2 knowing nothing of the respondent and is likely to express a near-neutral position at the outset, converging only as the Round 2 exchange unfolds. A memory-on LLM, primed by the Round 1 profile, instead begins Round 2 with information from the earlier high-divisiveness conversation and may import that charge into whatever the Round 2 topic is. Here β_1 is the pooled memory effect on opening extremity. H4a is supported if $\beta_1 > 0$.

H4b: Memory transfer on low-divisiveness topics

Estimated on the same opening window, adding divisiveness and its interaction with memory:

$$E_{u,c,t}^L = \alpha + \beta_1 \text{Mem}_u + \beta_2 \text{Div}_c + \beta_3 (\text{Mem}_u \times \text{Div}_c) + \delta_{\text{topic}} + \delta_{\text{country}} + \varepsilon_{u,c,t}. \quad (5)$$

With $\text{Div}_c = 0$ (low divisiveness) as the reference, β_1 is the memory effect on opening ex-

tremity in the low-divisiveness cell—the tilt the Round 1 profile carries into a topic that would otherwise be calm—and β_3 captures how that effect changes on high-divisiveness topics. Because a high-divisiveness topic already draws a strong position from the respondent in the memory-off arm, memory has less room to add there; H4b predicts the memory effect is larger on low-divisiveness topics, i.e., $\beta_3 < 0$.

H5a: Amplification (user side)

Estimated on Round 2 conversations from the second turn onward, we model the respondent’s extremity as a function of the memory condition and conversational turn:

$$E_{u,c,t}^U = \alpha + \beta_1 \text{Mem}_u + \beta_2 t + \delta_{\text{topic}} + \delta_{\text{country}} + \varepsilon_{u,c,t}. \quad (6)$$

Where $E_{u,c,t}^U = |s_{u,c,t}^U - 4|$ measures the user’s extremity over the turns. β_1 is the average effect of the memory-on condition on respondent extremity. H5a (amplification) is supported if $\beta_1 > 0$: respondents in the memory-on condition express more extreme positions than respondents in the memory-off condition.

H5b: Amplification on low divisiveness topics

Estimated also on Round 2 conversations from the second turn onward, adding divisiveness and its interaction with memory:

$$E_{u,c,t}^U = \alpha + \beta_1 \text{Mem}_u + \beta_2 \text{Div}_c + \beta_3 (\text{Mem}_u \times \text{Div}_c) + \beta_4 t + \delta_{\text{topic}} + \delta_{\text{country}} + \varepsilon_{u,c,t}. \quad (7)$$

With $\text{Div}_c = 0$ (low divisiveness) as the reference, β_1 is the memory effect on user extremity

across the turns in the low-divisiveness cell—the degree to which users move to the extreme when engaging with a tilted LLM on a topic that would otherwise be calm—and β_3 captures how that effect changes on high-divisiveness topics. Because a high-divisiveness topic already tends to push respondents to more extreme positions, H5b predicts the memory effect is larger on low-divisiveness topics, i.e., $\beta_3 < 0$.

Multiple comparisons and inference

For all statistical tests, we will use two-sided tests with $p < 0.05$ as our measure of statistical significance. We control the false discovery rate within each family of hypotheses using the Benjamini–Hochberg procedure at $q = 0.05$. We define three families of hypotheses. The first family contains H1–H3 (LLM-Side Behaviour Within a Conversation), the second contains H4a–H4b (LLM-Side Transfer Across Conversations), and the third contains H5a–H5b (User-Side Hypothesis: Amplification). We will present all tests with unadjusted p -values and adjusted q -values. Secondary and robustness tests (warm-up negative control in Appendix E, scorer-random-effect parameterization in Appendix F, additional hypothesis analyses in Appendix G, and the entrenchment DV and multilevel re-estimation in Appendix H) will be reported without multiple-hypothesis correction.

For null effects, we will also measure Bayes factors to assess the strength of evidence for the null hypothesis, as discussed in Table 1. For any statistical tests in which we fail to reject the null hypothesis, we will compute two-sided default Bayes Factors (BF_{01}). We will use the default JZS (Jeffreys–Zellner–Siow) prior. We will interpret a Bayes factor of $BF_{01} \geq 10$ as strong evidence in favor of the null hypothesis relative to the alternative. Values of $BF_{01} < 10$ will not be interpreted as strong evidence for the null. Implementation details are reported in Appendix H.

Power Analysis and Pilot Results

In this section, we present power analysis for the main hypotheses specified in this report as well as results from a large pilot study.

We start by assuming two design parameters are constant. First, we assume all our statistical analyses should achieve 95% statistical power – consistent with Nature’s Registered Report Guidelines. Second, we use inferential testing as described in this proposal, with two-sided tests with $p < 0.05$ as our measure of statistical significance. Under these assumptions, we calculate the minimum detectable effect size for our hypotheses under distinct sample sizes in our full study (Bloom, 1995). All the primary analyses use the full dataset with four countries. Therefore, we present MDE values up to 10,000 respondents, which is our target sample size. In Appendix D, we present power analysis for exploratory research questions with country-specific samples. We define the MDE as:

$$\text{MDE} = (\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta)) \cdot \text{SE}$$

Where $\Phi^{-1}(1 - \alpha/2)$ represents the inverse normal (z-score) associated with the inferential test described for each hypothesis, $\Phi^{-1}(1 - \beta)$ represents the inverse normal of the desired statistical power, and SE represents the estimated standard error for the full study.

Figure 5 shows the power analysis results for our seven hypotheses. Assuming 95% statistical power—consistent with Nature’s Registered Report Guidelines—and inferential testing at the 95% confidence level, our planned sample size of 10,000 participants is powered to detect standardized Minimum Detectable Effect Sizes (MDEs) of approximately 0.036 SD for hypotheses 1-3, 0.072 SD for Hypotheses 4a and 5a (memory-on effects) and 0.144 SD for hypotheses 4b and

5b (memory-on $\times$ divisiveness). These effects are considered small effect sizes in the specialized literature (Cohen, 2013).

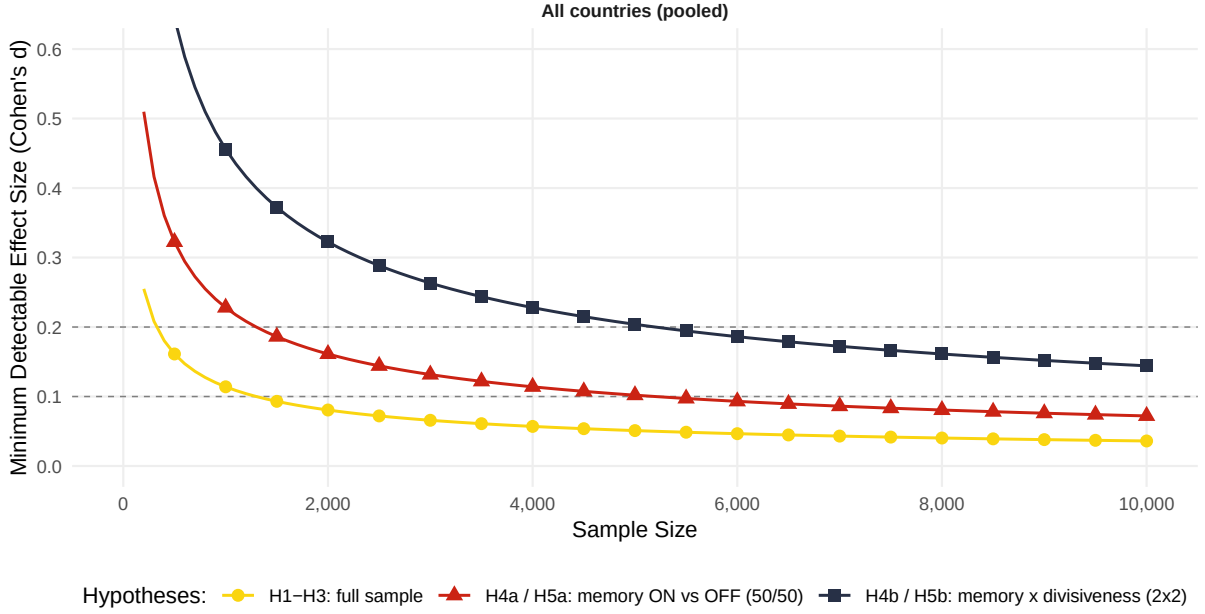


Figure 5 Power Analysis: Primary Hypotheses

Pilot Results

To ensure the technical feasibility of our approach, we conducted a large pilot study ($n=2,596$ respondents) with a similar design in Canada. Details of the pilot experiments are in Supplementary Appendix Section I.

We conducted the pilot between March and April 2026. Each respondent in the pilot was prompted to engage with an LLM for a maximum of twenty turns on three salient political topics (immigration, US-Canada relations, and trust in government). Additionally, the design included a neutral, non-political topic as a control.

The pilot informed several decisions in the present design. First, the pilot allows us to validate our measurement strategy, particularly using the ensemble scoring (See SI Table C.2). Second,

the pilot indicated that a large share of conversations stopped at the minimum number of turns (See SI Figure I.1), and the measurement improves among longer conversations (See SI Table C.2, Panel B). This lead us to raise the per-conversation minimum from the five turns used in the pilot to ten turns as proposed in this registered report. Third, a full twenty-turn conversation on a non-political topic proved largely uninformative and power-heavy; therefore, we now propose a non-political topic only as a brief warm-up in the beginning of the survey. Lastly, the prevalence of one-word turns was considerably high in the pilot (close to 23%), motivating the five-word minimum for each respondent turn proposed in this registered report.

The pilot experiment provides a strong foundation for rolling out the four studies proposed in this registered report.

References

- Bakshy, Eytan, Solomon Messing and Lada A. Adamic. 2015. “Exposure to ideologically diverse news and opinion on Facebook.” *Science* 348(6239):1130–1132.
- Baldassarri, Delia and Andrew Gelman. 2008. “Partisans without constraint: Political polarization and trends in American public opinion.” *American Journal of Sociology* 114(2):408–446.
- Baumann, Fabian, Philipp Lorenz-Spreen, Igor M Sokolov and Michele Starnini. 2021. “Emergence of polarized ideological opinions in multidimensional topic spaces.” *Physical Review X* 11(1):011012.
- Bloom, Howard S. 1995. “Minimum detectable effects: A simple way to report the statistical power of experimental designs.” *Evaluation Review* 19(5):547–556.
- Bouchaud, Pierre, David Chavalarias and M. Panahi. 2023. “Crowdsourced audit of Twitter’s recommender systems.” *Scientific Reports* 13(1):16815.
- Brown, Megan A., James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler and Joshua A. Tucker. 2026. “Evaluating Echo Chambers, Rabbit Holes, and Radicalization Pathways on YouTube.” *Political Communication* 0(0):1–27.
URL: <https://doi.org/10.1080/10584609.2026.2671765>
- Chen, A. Y., Brendan Nyhan, Jason Reifler, Ronald E. Robertson and Christo Wilson. 2023. “Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels.” *Science Advances* 9(35):eadd8080.

- Cheng, Myra, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han and Dan Jurafsky. 2026. “Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence.” *Science* 391:eaec8352.
- Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. 2 ed. New York: Routledge.
- Conlon, John J and Peter Schwardmann. 2026. “AI Sycophancy and Decisions.” *SSRN*.
- Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang and Nesreen K. Ahmed. 2024. “Bias and fairness in large language models: A survey.” *Computational Linguistics* pp. 1–79.
- Guess, Andrew M. 2021. “(Almost) everything in moderation: new evidence on Americans’ online media diets.” *American Journal of Political Science* 65(4):1007–1022.
- Huszár, Ferenc, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer and Moritz Hardt. 2022. “Algorithmic amplification of politics on Twitter.” *Proceedings of the National Academy of Sciences* 119(1):e2025334119.
- Ibrahim, H., N. AlDahoul, S. Lee, Talal Rahwan and Y. Zaki. 2023. “YouTube’s recommendation algorithm is left-leaning in the United States.” *PNAS Nexus* 2(8):pgad264.
- Kaiser, Jonas and Adrian Rauchfleisch. 2019. “The implications of venturing down the rabbit hole.” *Internet Policy Review* 8(2):1–22.
- Malmqvist, Lars. 2025. “Sycophancy in Large Language Models: Causes and Mitigations.” *Intelligent Computing. Proceedings of the Computing Conference* pp. 61–74.
- Motoki, Fabio, Valdemar Pinho Neto and Victor Rodrigues. 2024. “More human than human: measuring ChatGPT political bias.” *Public Choice* 198(1):3–23.
- Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, A. Y. Chen and Joshua A. Tucker. 2023. “Like-minded sources on Facebook are prevalent but not polarizing.” *Nature* 620(7972):137–144.
- Rathje, Steve, C. Robertson, William J. Brady and Jay J. Van Bavel. 2024. “People think that social media platforms do (but should not) amplify divisive content.” *Perspectives on Psychological Science* 19(5):781–795.
- Ribeiro, Manoel H., Raphael Ottoni, Robert West, Virgílio A. Almeida and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. pp. 131–141.
- Rozado, David. 2024. “The political preferences of LLMs.” *PLOS ONE* 19(7):e0306621.
- Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti and Robert West. 2025. “On the Conversational Persuasiveness of GPT-4.” *Nature Human Behaviour* 9(8):1645–1653.
- Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Asbell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rauker, Nicholas Schiefer, Da Yan, Miranda Ziegler and Ethan Perez. 2024. “Towards Understanding Sycophancy in Language Models.” Preprint, arXiv:2310.13548.

- Vallejo Vera, Sebastián and Hunter Driggers. 2025. “LLMs as annotators: the effect of party cues on labelling decisions by large language models.” *Humanities and Social Sciences Communications* 12(1):1–11.
- Waight, Hannah, Eddie Yang, Yin Yuan, Solomon Messing, Margaret E Roberts, Brandon M Stewart and Joshua A Tucker. 2026. “State media control influences large language models.” *Nature* pp. 1–9.
- Walker, C. and Joan C. Timoneda. 2024. “Identifying the sources of ideological bias in GPT models through linguistic variation in output.”. arXiv preprint.