STANCE DRIFT: HOW AI-MEDIATED COMMUNICATION DISTORTS OUR MESSAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly mediate human communication, from drafting emails to summarizing scientific reports, yet whether they faithfully preserve a speaker's position remains largely untested. We model AI-mediated communication as a two-step generation-extraction pipeline: one LLM produces an argument from a specified stance, and a second LLM extracts the stance from that argument. We represent the pipeline as a probabilistic state transition over five Likert-type stance categories and define the stance preservation rate (SPR) as the average probability that the extracted stance matches the initial stance. Across 112 debate propositions, none of the nine LLMs tested exceeded an SPR of 0.7 under the default configuration. Three drift patterns accounted for most of the drift: polarization, deviation from neutrality, and flipping. Among the mitigation strategies tested, including in-context learning, multiple extraction with shuffled options, assertion, and reflection, only adding medium reasoning effort to a reflection prompt for `GPT-5.4` substantially improved the SPR, to 0.775, yet polarization remained the largest pattern, with 0.119 of the transition mass. An exploratory comparison with human extraction on a single proposition suggests that drift arises at both the generation and the extraction stage. These results point to a fidelity gap in AI-mediated communication, with implications for journalism, policy deliberation, scientific communication, and other domains where opinion-laden messages pass through language models.

1 INTRODUCTION

History shows that small failures in message fidelity can have outsized consequences: the Berlin Wall fell hours after a confused spokesperson mistakenly announced that the borders were opening 'immediately, without delay' (Hertle, 2001). Today, this fragility of transmission is no longer limited to human intermediaries but is now embedded in the algorithms that are increasingly integrated into human communication across domains including decision making (Chen et al., 2023; Brand et al., 2023), scientific research (Horton, 2023; Qin et al., 2024), and communications (Bick et al., 2024; Chatterji et al., 2025). In one illustrative example from our framework, an AI-mediated pipeline can invert an initial "Agree" stance on the proposition "The use of AI should be abandoned" into "Disagree" and "Disagree strongly", effectively reversing the speaker's position (Fig. 2B). The inversion evokes familiar science-fiction tropes about AI resisting shutdown, but the scientific issue here is more basic: whether large language models (LLMs) preserve a communicator's intent as messages pass through them.

This question is increasingly important because LLMs are now routine intermediaries in written communication. From July 2024 to June 2025, writing-related activities (including text generation, editing, critiques, summarization, and translation) made up more than 40% of ChatGPT's work-related use (Chatterji et al., 2025). Beyond the well-studied issues with LLM-generated content, such as inherent biases and hallucinations (Acerbi & Stubbersfield, 2023; Spitale et al., 2023; Piao et al., 2025; Röttger et al., 2025; Wang et al., 2025), an equally important but less explored challenge is the preservation of *stance*, the attitude or position toward a given proposition, during AI-mediated communication.

Classic stance detection frames this task as classifying a writer's position toward a target using discrete categories (e.g., in favor, against, or neutral), leading to benchmarks such as

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

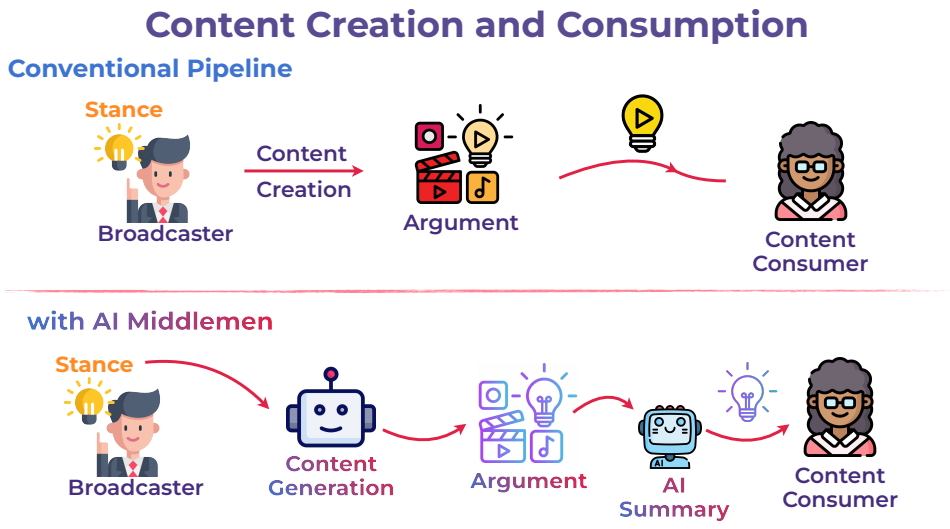


Figure 1: **AI-mediated communication model.** A sender generates an argument with LLM assistance from an initial stance; a receiver extracts the stance from the argument with LLM assistance.

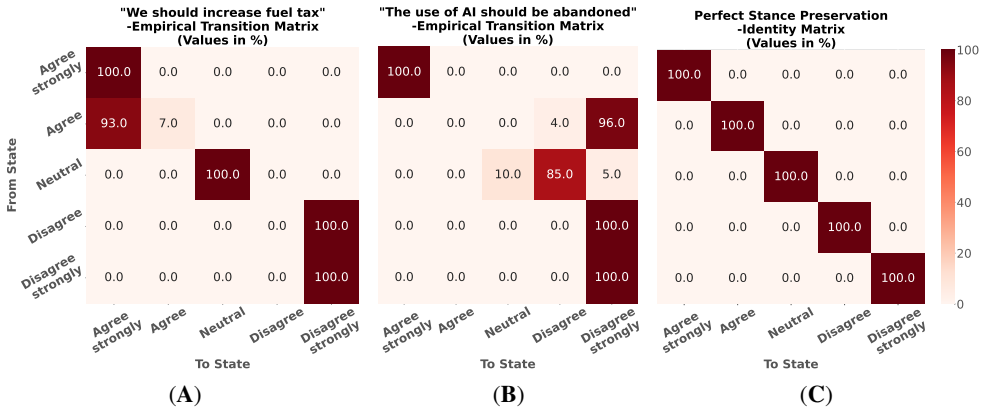
`SemEval-2016` (Mohammad et al., 2016) and surveys that consolidate conventions across platforms and domains (Küçük & Can, 2020; AlDayel & Magdy, 2021). More recent work on subjective global opinions shows that LLM interactions often involve value-laden judgments, with model responses reflecting specific human opinion distributions (Durmus et al., 2023; Santurkar et al., 2023), and examines LLM response behavior in surveys (Tjuatja et al., 2024), self-preference (Panickssery et al., 2024), and sycophancy (Sharma et al., 2024). At the communication level, Hancock et al. (2020) propose studying how AI intermediaries affect human communication, and related research studies whether transformations preserve perspective: co-writing with an opinionated model shifts writers' views (Jakesch et al., 2023), and in news summarization, Liu et al. (2024) find that summarization systems frequently alter an article's political stance. Beyond single-step transformations, transmission-chain ("telephone game") studies with LLMs reveal cumulative, directional changes as messages are iteratively regenerated (Acerbi & Stubbersfield, 2023; Perez et al., 2024). While valuable, these strands (i) do not capture the *full* opinion-communication process from initial message generation to downstream stance extraction, (ii) focus on specific tasks or settings rather than fundamental mechanisms of communication, and (iii) often lack a general, quantitative statistic for end-to-end stance fidelity.

We present a systematic framework for quantifying *stance* preservation in complete AI-mediated human-to-human communication processes by adopting a simple, task-agnostic decomposition—*proposition* (what is asserted) plus *stance* (position toward it)—and purposefully using discrete categories to match social-science measurement practices (Likert-type elicitation) (Likert, 1932; Croasmun & Ostrom, 2011). As LLM-generated text is repeatedly reused, real-world information flow naturally forms a generation-extraction pipeline (Fig. 1): a sender prompts a model to produce an argument from a given stance, and a receiver later extracts the stance from that text with LLM assistance. We therefore study whether the sender's stance is faithfully conveyed through AI mediation. We call this **stance preservation** (the receiver recovers the same discrete stance as the sender intended) and refer to any AI-induced change as **stance drift**. We model the end-to-end process as a probabilistic state transition, introducing a transition matrix to quantify stance preservation: diagonal elements indicate the probability of retaining the original stance (with a value of 1 representing perfect fidelity), while off-diagonal entries capture the extent of *stance drift*.

To illustrate, Fig. 2 presents the empirical transition matrices (averaged across 100 repetitions) from two experiments with OpenAI's `GPT-5.4` (complete results with standard errors in Fig. S3). For a fixed proposition, the (i, j) entry of the matrix represents the probability of ending in the jth choice, after our proposed generation-extraction process, given the ith choice as the initial stance. We compare the empirical matrices of proposition "We should increase fuel tax" (Fig. 2A) and "The

use of AI should be abandoned" (Fig. 2B) against the identity matrix (Fig. 2C), which represents the ideal of perfect preservation of each stance through generation and extraction.

We observe three major drift patterns. First, *polarization* occurs when moderate inputs ("Agree"/"Disagree") intensify to extreme positions ("Agree strongly"/"Disagree strongly") (Fig. 2A). Second, *deviation from neutrality* manifests as systematic shifts from neutral stances toward opinionated positions (Fig. 2B). Third, *flipping* involves complete reversals, where positive inputs ("Agree"/"Agree strongly") transform into negative outputs ("Disagree"/"Disagree strongly") (Fig. 2B), or vice versa. Two further patterns, *moderation* and *moderation to neutral*, are minor (Section 2.4). Some propositions exhibit a single dominant pattern (Fig. 2A) while others display several patterns operating simultaneously across different initial stances (Fig. 2B).



(**A**) (**B**) (**C**)

Figure 2: **Example of empirical transition matrices deviating from the ideal identity matrix.** Rows are the initial stances and columns the extracted stances; the (i, j) entry is the probability, in percent, of extracting the jth choice given the ith choice as the initial stance. (**A**) `GPT-5.4`, default configuration, proposition "We should increase fuel tax". (**B**) `GPT-5.4`, default configuration, proposition "The use of AI should be abandoned". (**C**) The identity matrix: perfect preservation of every stance.

To move beyond topic-specific behavior and establish a general measure of *stance drift*, we scale our framework across diverse propositions using a debate dataset (Slonim et al., 2021). This expansion enables us to define the stance preservation rate (SPR)—a quantitative metric that reveals pervasive *stance drift* across both topics and models. We further test several mitigation strategies, including in-context learning and reflection, and find that they offer limited improvement in stance preservation; raising the reasoning effort of `GPT-5.4` helps most, yet about a quarter of the transition mass still drifts.

Our framework, through proposition-conditioned transition matrices and a scalar measure of stance preservation, provides a systematic and repeatable way to assess risks in AI-mediated communication and offers a principled metric for future mitigation efforts.

2 Materials and Methods

2.1 Input Data Preparation

We take the 114 debate topics of the debate speeches dataset (Slonim et al., 2021) as propositions and do not use the speeches themselves. The topics are clearly stated but open to competing lines of reasoning, so they support plausible arguments on both sides and test whether a model preserves an intended stance rather than restates a factual claim.

2.2 Experimental Framework: Generation and Extraction

We model AI-mediated human-to-human communication as a two-step generation-extraction process (Fig. 3). In each trial, a first LLM (Agent 1) is given a proposition along with a specified initial

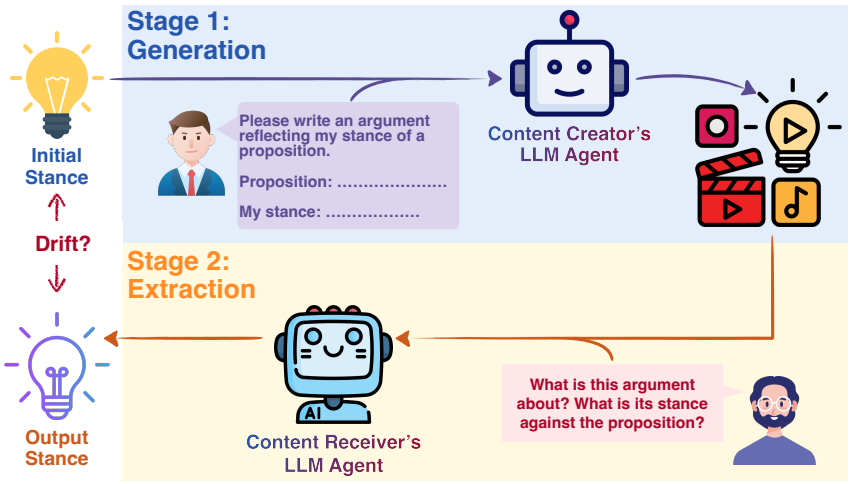


Figure 3: **The two-step generation and extraction framework.** An initial stance on a proposition is expanded into an argument, and the stance is extracted back from the argument.

stance and is prompted to generate an argument. (Here, a *proposition* is a declarative statement, a *stance* is the attitude toward that proposition, chosen from a five-point Likert scale: Agree strongly, Agree, Neutral, Disagree, Disagree strongly as in social science surveys (Johns, 2010; Croasmun & Ostrom, 2011), and an *argument* is a short passage that expresses the reasoning behind that stance.) For example, if the proposition is about AI abandonment ("The use of AI should be abandoned."), an initial stance of "Disagree strongly" would lead Agent 1 to produce an argument arguing strongly against abandoning AI. A second call to the same model (Agent 2), with no access to the first call's context, then extracts the stance from Agent 1's argument by reading both the argument and the proposition, outputting one of the five stance options (see Figs. S1–S2 for individual examples).

In both the generation and extraction processes, options are presented in the prompt as multiple-choice questions, following prior work (Ren et al., 2023; Durmus et al., 2023; Hong et al., 2025). For LLMs that provide token-level log probabilities via API, extraction uncertainty is represented directly by the model's output probability distribution in a single trial. To account for generation-stage randomness, we repeat trials for each stance of each proposition and average the resulting extraction distributions. Finally, the transition probabilities will be estimated by the average (across trials) probabilities assigned to each of the five choices by the LLM when API-available, or the observed frequencies otherwise. Table S13 in the appendix lists, for every run, the model identifier, the run dates, and which of the two estimates was used.

2.3 Transition Matrix and Stance Preservation Rate (SPR) Metric

The transition matrix resulting from this process would be a 5×5 matrix. Without further specification, we overload the notation and use the term "the ith choice" to refer to the corresponding stance, where i ranges from 0 to 4, representing the five choices: Agree strongly, Agree, Neutral, Disagree, and Disagree strongly. The (i, j) entry of the transition matrix represents the probability of returning to the jth choice given the ith choice as the initial choice.

Ideally, the transition matrix resulting from this process would be the identity matrix, indicating complete preservation of the original stance throughout the process. Let T denote the transition matrix, with entries T_{ij}, and I the identity matrix. Since each row of T sums to one, the entrywise L_1 distance between T and I equals $\sum_{i,j} |T_{ij} - I_{ij}| = 2\sum_{i=0}^{4}(1 - T_{ii})$, which depends only on the diagonal elements of T. We therefore examine the diagonal elements, which represent the probabilities of returning to the original stance after the generation-extraction.

We conduct our experiments on 112 of the 114 topics from the debate speeches dataset, reserving the remaining two topics as examples for in-context learning, as discussed in Section 2.5. Each topic undergoes 100 repeated trials, yielding empirical 5×5 transition matrices whose diagonal elements (i, i) quantify stance preservation: the probability that the extraction stage recovers the

same stance supplied to the generation stage. We define this diagonal probability as the "success rate" and estimate it by averaging across trials. All raw data and estimated transition matrices are available in our code repository.

To quantify stance preservation across the diverse topics in the debate speeches dataset, we define the stance preservation rate (SPR) as the average success rate across all topics and initial stances—that is, the fraction of messages whose original stance survives the generation-extraction process (unless otherwise specified, SPR assumes a uniform prior over initial stances; alternative priors can be incorporated by weighted averaging over diagonal transition probabilities). This metric provides a single, interpretable measure of how faithfully models maintain communicative intent, and thus offers a clear benchmark for comparing model reliability in AI-mediated communication and assessing mitigation strategies.

SPR is the complement of the expected 0–1 loss under a uniform prior over initial stances. Any other prior, or any other loss over stance pairs such as one weighted by distance on the Likert scale, is a weighted sum of the entries of the same transition matrix and requires additional assumptions on the weights. These extensions share the same transition-matrix foundation; the strict SPR used throughout this paper represents a general baseline with no additional assumptions.

2.4 Stance drift patterns

Off-diagonal entries of the transition matrix record *stance drift*. We partition them into five patterns defined over the index pairs (i, j) with $i \neq j$. *Polarization* is the pair set $\{(1, 0), (3, 4)\}$: a moderate stance is extracted as the extreme stance on the same side. *Moderation* is the pair set $\{(0, 1), (4, 3)\}$: an extreme stance is extracted as the moderate stance on the same side. *Moderation to neutral* is the pair set $\{(i, 2) : i \neq 2\}$: a non-neutral stance is extracted as Neutral. *Flipping* is the pair set $\{(i, j) : i \in \{0, 1\},\ j \in \{3, 4\}\} \cup \{(i, j) : i \in \{3, 4\},\ j \in \{0, 1\}\}$: a stance is extracted as a stance on the opposite side. *Deviation from neutrality* is the pair set $\{(2, j) : j \neq 2\}$: Neutral is extracted as a non-neutral stance. The five sets are disjoint and together cover every off-diagonal pair, so each off-diagonal entry belongs to exactly one pattern. The mass of a pattern for initial stance i is the sum of the transition probabilities (i, j) over the pairs of that pattern with that i. The five pattern masses for initial stance i sum to one minus the success rate of i. As with SPR, the overall mass of a pattern is the average of its per-stance masses over the five initial stances. We call one minus SPR the stance drift rate (SDR). The five overall pattern masses are the components of the SDR and sum to it.

2.5 Mitigation attempts

Beyond measuring *stance drift*, we ask whether straightforward interventions can restore fidelity in AI-mediated communication. Throughout, the default configuration denotes a run with the default prompts (Appendix A.3). We tested four targeted strategies with `GPT-4o-mini` (full details in the appendix), each addressing a distinct hypothesized failure mode. (i) Multiple extraction addresses *positional bias*, the documented tendency of LLMs to favor options at specific positions (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024): by running extraction five times with randomly shuffled option orderings and averaging the resulting probability distributions, we sought to reduce systematic position effects. (ii) In-context learning supplies three high-quality examples during generation and extraction, aiming to anchor model behavior through demonstration. (iii) Assertion passes the literal stance options to the extraction prompt and asks the model to choose one of them first and then map it to a letter, with the aim of directing the model's attention to the stance labels. (iv) Reflection adds one instruction to each of the two prompts: Agent 1 is asked to compare its drafted argument against the specified stance and to revise it if the two do not align, and Agent 2 is asked to compare its selected option against the argument and to revise the choice likewise, with the aim of catching drift before the output is returned.

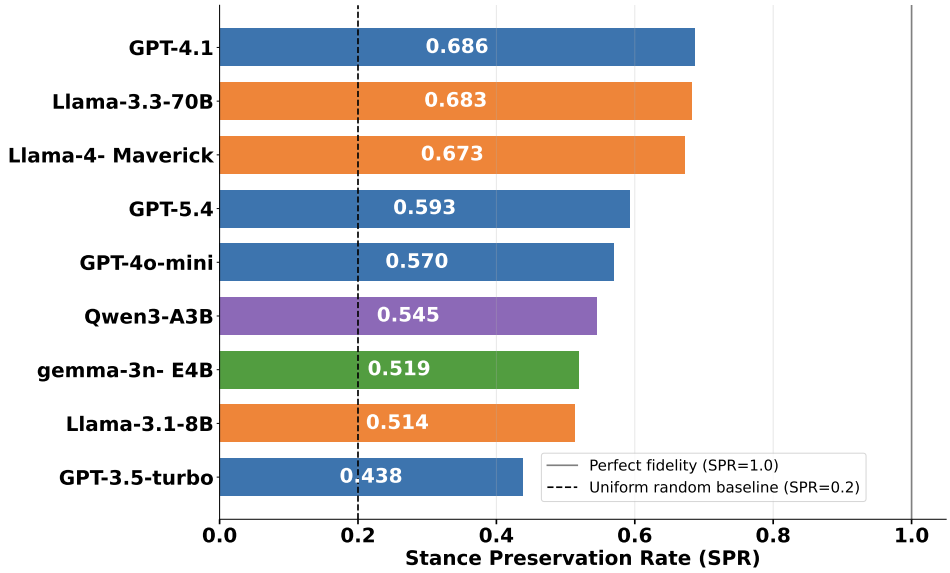


Figure 4: **SPR across models under the default configuration.** The solid rule marks perfect fidelity (SPR $= 1$) and the dashed rule the uniform random baseline (SPR $= 0.2$).

3 Results

3.1 Fidelity gap across models

We evaluated nine LLMs on the debate-speeches dataset (Fig. 4). None of the models exceeded an SPR of 0.7 under the default configuration (Fig. 4). In this model set, higher SPR tended to appear in larger models (e.g., `GPT-4.1`, `Llama-3.3-70B`, `Llama-4-Maverick`) (Fig. 4), although these comparisons are confounded by architecture and training differences. For example, `GPT-5.4` under the default configuration has a lower SPR (0.593) than `GPT-4.1` (0.686), whereas with third-person reflection and medium reasoning effort it reaches 0.775 (Section 3.4).

3.2 Stance-level heterogeneity in preservation failure

A stance-level analysis further reveals substantial heterogeneity in preservation performance (Table 1A). Although most models achieve high success rates for specific initial stances (e.g., "Agree strongly" for `GPT-4o-mini` and "Neutral" for `gemma-3n-E4B`), every model except `Llama-4-Maverick` (lowest stance: 0.515) falls below 0.3 for at least one stance (e.g., "Disagree" for `GPT-4.1` and `Llama-3.3-70B`). A consistent cross-model pattern emerges: with the exception of `Llama-3.1-8B` and `gemma-3n-E4B`, the highest success rates are obtained for the initial stance "Agree strongly", exceeding 0.85 for all but `GPT-3.5-turbo`. In contrast, neutrality is less reliably preserved; only `gemma-3n-E4B`, `Llama-3.3-70B`, and `GPT-4.1` reach 0.8 for the initial stance "Neutral". For all models but `Llama-4-Maverick`, the lowest success rate falls on one of the moderate stances "Agree" and "Disagree" (e.g., "Agree": 0.147 and "Disagree": 0.136 for `GPT-5.4`).

To further assess whether any observed preservation could be consistent with perfect fidelity, we conduct a confidence interval (CI) analysis for the success rate (Table 1B). For each proposition and initial stance, we estimate the success rates over 100 trials and construct a one-sided 99% CI using the central limit theorem. We consider stance preservation statistically indistinguishable from perfect fidelity if the CI includes 1. Because we perform this test across 112 propositions for each initial stance, we apply Bonferroni correction to widen each CI to account for multiple comparisons and report the proportion of propositions whose CIs include 1. This proportion is below 0.65 in 37 of the 45 cells (Table 1B). Even under this conservative Bonferroni correction—which tends to enlarge intervals and increase inclusion rates—the low proportions confirm that LLMs consistently fail to reliably recover initial stances.

Table 1: **Stance preservation by initial stance for the nine models under the default configuration.** (**A**) Success rates. (**B**) Proportion of propositions whose one-sided 99% CI of the success rate, with Bonferroni correction, includes 1.

(**A**)

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.933 (.021)	.108 (.016)	.717 (.026)	.310 (.035)	.784 (.032)
GPT-4.1	.962 (.014)	.599 (.031)	.800 (.026)	.168 (.027)	.903 (.016)
GPT-5.4	.996 (.002)	.147 (.022)	.703 (.032)	.136 (.023)	.982 (.010)
GPT-3.5-turbo	.621 (.031)	.484 (.031)	.313 (.023)	.244 (.027)	.527 (.036)
gemma-3n-E4B	.595 (.040)	.348 (.034)	.920 (.019)	.114 (.022)	.617 (.041)
Llama-3.3-70B	.954 (.015)	.505 (.036)	.851 (.023)	.161 (.026)	.942 (.017)
Llama-3.1-8B	.368 (.019)	.768 (.013)	.442 (.014)	.104 (.007)	.887 (.011)
Llama-4-Maverick	.857 (.024)	.674 (.031)	.675 (.032)	.642 (.037)	.515 (.042)
Qwen3-A3B	.939 (.020)	.268 (.030)	.402 (.034)	.441 (.041)	.677 (.039)

(**B**)

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.223 (.040)	.000 (.000)	.402 (.047)	.116 (.030)	.536 (.047)
GPT-4.1	.920 (.026)	.286 (.043)	.616 (.046)	.054 (.021)	.732 (.042)
GPT-5.4	.991 (.009)	.027 (.015)	.527 (.047)	.018 (.013)	.973 (.015)
GPT-3.5-turbo	.000 (.000)	.045 (.020)	.018 (.013)	.000 (.000)	.080 (.026)
gemma-3n-E4B	.348 (.045)	.098 (.028)	.500 (.047)	.009 (.009)	.464 (.047)
Llama-3.3-70B	.920 (.026)	.295 (.043)	.714 (.043)	.054 (.021)	.902 (.028)
Llama-3.1-8B	.000 (.000)	.241 (.041)	.000 (.000)	.000 (.000)	.643 (.045)
Llama-4-Maverick	.705 (.043)	.357 (.045)	.482 (.047)	.455 (.047)	.393 (.046)
Qwen3-A3B	.607 (.046)	.071 (.024)	.107 (.029)	.321 (.044)	.420 (.047)

3.3 Distribution of stance drift across patterns

Table 2 partitions the SDR of each model into the five drift patterns defined in Section 2.4. Polarization is the largest pattern for seven of the nine models; for GPT-3.5-turbo and gemma-3n-E4B, flipping is the largest. Polarization, deviation from neutrality, and flipping together carry most of the drift of every model, and we refer to them as the three major patterns; moderation and moderation to neutral are minor. Polarization exceeds moderation for all nine models, which a symmetric difficulty in reading intensity would not produce.

Table 2: **SPR and the components of the SDR by drift pattern for the nine models under the default configuration.** Mean (SE) over the 112 propositions of the per-proposition values averaged over the five initial stances; each row sums to 1 up to rounding. Per-stance values and all tested models: Tables S5–S10 in the appendix.

Model	SPR	Polarization	Moderation	Moderation to neutral	Flipping	Deviation from neutrality
GPT-4o-mini	.570 (.009)	.246 (.011)	.022 (.004)	.009 (.002)	.097 (.012)	.057 (.005)
GPT-4.1	.686 (.007)	.244 (.005)	.024 (.004)	.001 (.000)	.005 (.001)	.040 (.005)
GPT-5.4	.593 (.008)	.328 (.006)	.002 (.001)	.000 (.000)	.018 (.004)	.059 (.006)
GPT-3.5-turbo	.438 (.008)	.158 (.009)	.087 (.006)	.000 (.000)	.180 (.014)	.137 (.005)
gemma-3n-E4B	.519 (.013)	.165 (.010)	.041 (.006)	.082 (.010)	.178 (.015)	.016 (.004)
Llama-3.3-70B	.683 (.011)	.191 (.007)	.014 (.003)	.015 (.003)	.069 (.009)	.030 (.005)
Llama-3.1-8B	.514 (.006)	.174 (.004)	.085 (.004)	.012 (.001)	.103 (.010)	.112 (.003)
Llama-4-Maverick	.673 (.012)	.122 (.007)	.057 (.006)	.005 (.001)	.078 (.009)	.065 (.006)
Qwen3-A3B	.545 (.010)	.216 (.011)	.065 (.008)	.001 (.001)	.052 (.009)	.120 (.007)

3.4 Stance drift survives robustness checks and is only partly alleviated by mitigation

Additional robustness checks that relax the perfect-preservation criterion by introducing thresholds strictly below 1 consistently confirm the presence of *stance drift* (Table S4 and Appendix B.3). To avoid fixing a single threshold, we compute the area under the curve (AUC) of the pass rate, the proportion of proposition–initial stance combinations whose success rate exceeds the threshold, as thresholds vary from 0.5 to 1, yielding a single summary statistic which provides similar results as SPR regarding model comparison. Results show that models consistently underperform across this continuum, reinforcing that *stance drift* is not an artifact of requiring a success rate of exactly 1 (Fig. S6).

We conducted ablation analyses using `GPT-4o-mini` (default configuration, SPR $= 0.570$) to assess potential confounds in the generation-extraction pipeline (Table S1, Fig. S4). First, to test whether the temperature (the parameter that controls randomness when inferring the next token) of the extraction stage drives the low SPR, we re-analyzed the raw results with each trial's extraction vector replaced by a point mass on its most probable option, which is equivalent to setting the temperature to 0 during extraction. This re-analysis yielded an SPR of 0.571, indicating no meaningful change. Second, to examine positional and formatting biases, we reversed the choice order (negative-to-positive), obtaining a comparable SPR of 0.584. We further implemented the multiple extraction protocol (Section 2.5) to mitigate potential positional bias; this procedure likewise produced no substantive improvement (SPR $= 0.569$). Third, we evaluated prompt-level mitigation strategies. In-context learning resulted in an SPR of 0.573, assertion yielded 0.559, and reflection yielded 0.586, none differing meaningfully from the default configuration.

Reflection is the one strategy we also tested on `GPT-5.4`, in three configurations that each add one change to the previous one: reflection; third-person reflection, which phrases the instruction from the perspective of an impartial third-party reviewer; and third-person reflection with medium reasoning effort (full details in the appendix). The SPR rises from 0.593 to 0.613, 0.637 and 0.775, and polarization falls at every step, from 0.328 to 0.119 (Fig. S5 and Table S5 in the appendix). The two reflection steps move part of the drift of the initial stance "Agree" from polarization into flipping (from 0.076 to 0.252). The reasoning step removes most of that flipping (0.010) and gives the largest gain, but lowers the success rate for "Neutral" from 0.721 to 0.607. Even in this last configuration, about a quarter of the transition mass still drifts and polarization remains the largest pattern.

Stance drift is therefore not fully explained by the prompting and decoding factors tested here.

3.5 An exploratory human-extraction study locates drift at both stages

The ablation analyses above treat the generation-extraction pipeline as a whole. To probe whether drift originates in generation, extraction, or both, we conducted a targeted human-extraction study (Appendix A.7) on the proposition "The use of AI should be abandoned" using 100 arguments generated by `GPT-5.4` (20 per initial stance). Nine human annotators, working without AI assistance, each labeled all 100 arguments via a structured survey; their majority vote served as the reference label throughout.

Comparing the human-majority against the known initial stance reveals generation-stage drift; comparing the AI-majority against the human-majority reveals extraction-stage drift. The results reveal that both stages contribute to *stance drift* yet show a systematic pattern by initial stance (Table S11; one example argument per drift pattern in Table S12). For "Agree", flipping to "Disagree" was already present across all 20 generated arguments, implicating generation-stage drift. For "Disagree", polarization toward "Disagree strongly" was absent from human-majority labels but appeared consistently in AI extraction, implicating the extraction stage. Deviation from neutrality showed a mixed pattern: 5 of 20 "Neutral"-initiated arguments were already non-neutral (generation-stage drift), and 13 deviated at AI extraction.

To quantify the human–AI performance gap at the extraction stage, we computed, for each individual trial (each single human annotation and each single AI extraction vote), the agreement with the human-majority label. This comparison serves to validate both the task design and the feasibility of high-fidelity extraction. Note that this validation is non-trivial: even under the majority-vote reference, the minimum expected per-annotator agreement is bounded below by $2/9 \approx 0.222$. Human annotators achieved 0.743 (SE 0.038), well above this floor, confirming that the arguments carry recoverable stance signals and that the task admits substantially better-than-chance extraction. The AI shortfall, `GPT-5.4` at 0.472 (SE 0.002), is therefore consistent with an extraction deficit and not accounted for by ambiguity in the arguments. Given the intensive demands of unaided human labeling, this study is limited to a single proposition and a small annotator pool; online deployment on platforms such as Amazon Mechanical Turk would risk contamination from AI-assisted responses, undermining the human-only baseline. It is intended as a proof-of-concept decomposition and human benchmark rather than a comprehensive measurement.

4 IMPLICATIONS AND OPEN QUESTIONS

4.1 SOURCES OF STANCE DRIFT

Stance drift, the failure of LLMs to preserve original positions, emerged consistently across the models and propositions examined here and was only partly alleviated by the decoding, format and prompt changes we tested (Section 3.4). These results motivate treating stance fidelity as a distinct evaluation target in AI-mediated communication.

Our conclusion is descriptive: it characterizes *stance drift* on the tested dataset and does not isolate a causal source.

As an initial step toward understanding potential sources of drift, inspection of intermediate generated arguments suggests several recurring features. First, models sometimes mishandle negation-heavy reasoning (e.g., cases in which a proposition containing negated statement is discussed in a negative stance), leading to flipping during either generation or extraction. Second, models may conflate a cautiously phrased supporting argument—written to acknowledge counterarguments or avoid overstatement—with a genuinely opposing stance. We also observed that propositions containing terms such as "ban" frequently coincided with flipping or misclassification. The structure of the drift patterns (Table 2) suggests that future work should examine whether different patterns are associated with distinct sources, such as training data biases and architectural limitations in maintaining semantic invariants through paraphrase. These observations are qualitative and specific to the policy-oriented, sometimes negation-heavy propositions of the debate speeches dataset.

4.2 STANCE PRESERVATION AS A BENCHMARK FOR AI-MEDIATED COMMUNICATION

As for potential future work, although the mitigation strategies tested here yielded limited gains, the framework points to a practical optimization target: stance preservation itself. In particular, SPR measured on datasets such as the debate speeches dataset, could be used as a reward in reinforcement learning to encourage LLMs to maintain stance across generations. More broadly, the generation-extraction framework can be extended beyond stance to subtler attributes such as tone, framing, and contextual emphasis. A particularly promising direction is multi-hop communication chains, where messages pass through multiple LLMs or human–AI hybrids: modelled as Markov chains, they would show whether drift accumulates over successive transmissions and whether stationary distributions mark stable distortion endpoints.

For LLM development, SPR provides a communication-fidelity benchmark that complements traditional accuracy-oriented metrics by testing whether models preserve intended positions, not only whether they produce plausible text. For social-science and opinion-sensitive research settings, our results raise caution about using current AI agents as substitutes (Qin et al., 2024) for human subjects when faithful preservation of stated positions is essential. In domains such as journalism, public policy, scientific research, education, and everyday communication, reliance on LLMs for drafting, summarizing, or mediating messages may inadvertently amplify, shift, or distort stances, affecting decision-making, fairness, and public perception (Acerbi & Stubbersfield, 2023; Liu et al., 2024; Wang et al., 2025; Kelly et al., 2025).

ETHICS STATEMENT

The human-extraction study (Section 3.5 and Appendix A.7) collected stance labels from nine annotators. The annotators were informed of the nature and purpose of the task prior to participation, and their involvement was voluntary. All other data in this work are the public debate propositions of Slonim et al. (2021) and text generated from them by the models under study. During the course of this work, one author became a visiting researcher at the organization that develops one of the evaluated model families. That organization had no role in the design, execution or reporting of the study.

REPRODUCIBILITY STATEMENT

Table S13 in Appendix A.12 lists, for every run, the model identifier requested through the API, the API provider, the prompt, the run dates and whether the transition probabilities were estimated from

token-level log probabilities or from observed frequencies. The prompts of every configuration are given verbatim in Appendix A.3 and in the appendix subsections on the mitigation strategies. The code, the raw per-trial extraction distributions and the estimated transition matrices of every run, the human-study labels, and the scripts that produce every table and figure of this paper are available at https://github.com/anonymous-sea-otter/stance-drift-code.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, and support qualitative and thematic data analysis. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, or interpret results, and assisting with translation, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to search the literature; to create or edit code, namely the scripts that run the experiments, aggregate the results, and draw the figures and tables from the results; to draft parts of the paper; to edit the paper to improve readability; and to format references. We have reviewed all AI-assisted work. The synthetic data are the arguments and stance extractions produced by the language models under study; they are the object of measurement, and the pipeline that produces them is described in Section 2 and the appendix. AI-assisted code was reviewed and run by the authors. AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alberto Acerbi and Joseph M Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44): e2313790120, 2023. doi: 10.1073/pnas.2313790120.

Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021. doi: 10.1016/j.ipm.2021.102597.

Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research, 2024.

James Brand, Ayelet Israeli, and Donald Ngwe. Using llms for market research. *Harvard business school marketing unit working paper*, (23-062), 2023.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023.

James T Croasmun and Lee Ostrom. Using likert-type scales in the social sciences. *Journal of adult education*, 40(1):19–22, 2011.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. 2023. URL https://arxiv.org/abs/2306.16388. arXiv preprint arXiv:2306.16388.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25 (1):89–100, 2020.

Hans-Hermann Hertle. The fall of the wall: the unintended self-dissolution of east germany's ruling regime. *Cold War International History Project Bulletin*, 12(13):131–164, 2001.

Harbin Hong, Sebastian Caldas, and Liu Leqi. Hypothesis testing for quantifying llm-human misalignment in multiple choice settings. 2025. arXiv:2506.14997.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15, 2023.

Rob Johns. Likert items and scales. *Survey question bank: Methods fact sheet*, 1(1):11–28, 2010.

Markelle Kelly, Mohammad Tahaei, Padhraic Smyth, and Lauren Wilcox. Understanding gender bias in ai-generated product descriptions. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2587–2615, 2025.

Dilek Küçük and Fazlı Can. Stance detection: A survey. *ACM Computing Surveys*, 53(1):12:1–12:37, 2020. doi: 10.1145/3369026.

Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, (140):1–55, 1932.

Yuhan Liu, Shangbin Feng, Xiaochuang Han, Vidhisha Balachandran, Chan Young Park, Sachin Kumar, and Yulia Tsvetkov. P3sum: Preserving author's perspective in news summarization with diffusion language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2154–2173, 2024.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval-2016*, pp. 31–41, 2016. doi: 10.18653/v1/S16-1003. URL https://aclanthology.org/S16-1003/.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Jérémy Perez, Grgur Kovač, Corentin Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. When llms play the telephone game: Cumulative changes and attractors in iterated cultural transmissions. *arXiv preprint arXiv:2407.04503*, 2024. URL https://arxiv.org/abs/2407.04503.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Qinghua Hu, Fernando P Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents. 2025. arXiv:2501.05171.

Xin Qin, Mingpeng Huang, and Jie Ding. Aiturk: Using chatgpt for social science research. *Available at SSRN 4922861*, 2024.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *Proceedings on*, pp. 49–64. PMLR, 2023.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. Issuebench: Millions of realistic prompts for measuring issue bias in llm writing assistance. 2025. arXiv:2502.08395.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. An autonomous debating system. *Nature*, 591(7850):379–384, 2021.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Yifei Wang, Ashkan Eshghi, Yi Ding, and Ram Gopal. Echoes of authenticity: Reclaiming human sentiment in the large language model era. *PNAS nexus*, 4(2):pgaf034, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proceedings of ICLR 2024*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

A Additional materials and methods

A.1 Algorithm of proposed experimental framework: generation and extraction

1. **Input:** Proposition s, LLM model $\mathcal{M}$, generation prompt F_{para}, extraction prompt F', the total number of repeated trials K.
2. For $k = 1$ to K, repeat:
   (a) For initial choice $X_0 = i \in \{0, 1, 2, 3, 4\}$, repeat:
      i. Convert X_0 to the corresponding choice $choice(X_0) \in \{$'A', 'B', 'C', 'D', 'E'$\}$.
      ii. Compose the generation prompt with the proposition and the initial choice, $F_{para}(s, choice(X_0))$, feed to $\mathcal{M}$, obtain output argument z_{X_0}.
      iii. Compose the extraction prompt with the proposition and the argument $F'(s, z_{X_0})$, feed to $\mathcal{M}$ without any history, obtain output softmax probability scores $P_k(X_1 = j \mid X_0 = i)$, where $j \in \{0, 1, 2, 3, 4\}$ is the returned choice.
3. Compute the estimated transition matrix entry as the average over repeated trials:

$$\hat{\mathcal{P}}(X_1 = j \mid X_0 = i) = \frac{1}{K} \sum_{k=1}^{K} P_k(X_1 = j \mid X_0 = i).$$

   A trial whose selection step returns no valid choice is excluded from the average, and K counts the remaining trials for that entry. This occurred for six trials across all runs (one for `Llama-3.1-8B`, five for `Llama-4-Maverick`).
4. **Output:** Estimated transition matrix where the i, j entry is $\hat{\mathcal{P}}(X_1 = j \mid X_0 = i)$.

A.2 Adapted debate speeches dataset

Notably, each debate topic in the dataset appears multiple times, corresponding to speeches generated from different sources (e.g., human expert, large language model, IBM's automated Project Debater system). For example, the topic "We should subsidize vocational education" appears in seven separate rows, each associated with a distinct source. In our experiments, we disregard source-specific duplicates and retain only the unique topics, which we treat as propositions. An illustrative example is shown below:

```
    {
        "topic_id":"3716",
        "topic":"We should subsidize vocational education"
    }
```


A.3 PROMPTS USED IN THE EXPERIMENT

Our experimental framework consists of two main stages: the generation stage and the extraction stage. In the generation stage, we use a generation prompt to instruct the LLM to generate an argument based on a given proposition and an initial stance, with a word limit of 100. In the extraction stage, we use an extraction prompt, a multiple-choice question, to guide the LLM in extracting the stance from the generated argument. Below are the detailed prompts used in each stage.

GENERATION STAGE PROMPT

```
In the following task, you will receive a proposition and a capi-
talized letter representing one of the stance options listed below
with respect to the proposition.  The proposition concerns a topic
related to the United States.
Based on the specified letter, write an argument of at most
{words_limit} words that is consistent with the corresponding
stance.
Only output the argument itself and do not directly quote or ex-
plicitly mention the option or its label.
{sep}{sep}
Options:
{choices}{sep}{sep}
Proposition:  {proposition}
{sep}{sep}
Letter:  {letter}
{sep}{sep}
Argument:
```

EXTRACTION STAGE PROMPT

```
In the following task, you will receive a proposition and an ar-
gument addressing that proposition.  The proposition concerns a
topic related to the United States.
Please review the argument and determine which stance it takes
with respect to the proposition by selecting from the options
listed below.
Only output the capitalized letter corresponding to your answer.
Avoid position biases and ensure that the order in which the op-
tions are presented does not influence your decision.
{sep}{sep}
Proposition:  {proposition}
{sep}{sep}
Argument:  {argument}
{sep}{sep}
Please choose from the following options:
{choices}{sep}{sep}
Answer letter:
```

A.4 SNIPPET OF LOG FILE IN THE EXPERIMENT

We provide a snippet of the log file generated during our experiments to illustrate the input prompts and the corresponding outputs from the LLM. This log captures the interaction between the prompts and the model's responses, showcasing how the generation and extraction stages are executed. The complete log files and results are provided in the code repository. In the logs, the generation stage is called decoding and the extraction stage encoding.

LOG FILE SNIPPET

```
Start Logging Estimation of the Transition Matrix
2026-02-02 22:24:23,571 - INFO - Model:  <google/gemma-3n-E4B-it>
2026-02-02 22:24:23,571 - INFO - prompt choice:  <1>
2026-02-02 22:24:23,571 - INFO -
========================================
2026-02-02 22:24:23,571 - INFO - Proposition:  We should protect
Antarctica
2026-02-02 22:24:23,571 - INFO - ---------------Shuffled_rep:  0--
--------------
2026-02-02 22:24:23,571 - INFO - Shuffled letter in decoding:
['A', 'B', 'C', 'D', 'E']
2026-02-02 22:24:23,571 - INFO - Task:  decoding
2026-02-02 22:24:23,571 - INFO - Given letter:  <A>
2026-02-02 22:24:23,571 - INFO - prompt:
In the following task, you will receive a proposition and a capi-
talized letter representing one of the stance options listed below
with respect to the proposition.  The proposition concerns a topic
related to the United States.
Based on the specified letter, write an argument of at most 100
words that is consistent with the corresponding stance.
Only output the argument itself and do not directly quote or ex-
plicitly mention the option or its label.
==
Options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Proposition:  We should protect Antarctica
==
Letter:  A
==
Argument:
2026-02-02 22:24:26,435 - INFO - Given opinion:  Agree strongly
2026-02-02 22:24:26,435 - INFO - Generated argument:  Antarctica
holds invaluable scientific secrets crucial for understanding our
planet's past, present, and future.  Protecting it safeguards bio-
diversity and unique ecosystems found nowhere else.  International
cooperation ensures research benefits all humankind, fostering in-
novation in climate science, medicine, and technology.  Preserving
this pristine continent is a moral imperative; future generations
deserve to inherit its untouched beauty and resources.  Its pro-
tection is vital for global well-being and scientific advancement.
2026-02-02 22:24:26,435 - INFO - Job:  encoding
2026-02-02 22:24:26,435 - INFO - Shuffled letter in encoding:
['A', 'B', 'C', 'D', 'E']
2026-02-02 22:24:26,435 - INFO - prompt:
```

```
In the following task, you will receive a proposition and an ar-
gument addressing that proposition.  The proposition concerns a
topic related to the United States.
Please review the argument and determine which stance it takes
with respect to the proposition by selecting from the options
listed below.
Only output the capitalized letter corresponding to your answer.
Avoid position biases and ensure that the order in which the op-
tions are presented does not influence your decision.
==
Proposition:  We should protect Antarctica
==
Argument:  Antarctica holds invaluable scientific secrets crucial
for understanding our planet's past, present, and future.  Pro-
tecting it safeguards biodiversity and unique ecosystems found
nowhere else.  International cooperation ensures research benefits
all humankind, fostering innovation in climate science, medicine,
and technology.  Preserving this pristine continent is a moral
imperative; future generations deserve to inherit its untouched
beauty and resources.  Its protection is vital for global well-
being and scientific advancement.
==
Please choose from the following options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Answer letter:
2026-02-02 22:24:26,846 - INFO - Top choice (answer):  <A>
```

A.5 Individual examples of the two-stage pipeline

We display two individual examples (Figs. S1 and S2) to illustrate the operation of our two-stage generation-extraction framework, highlighting instances of both *stance drift* and *stance preservation*.

A.6 Empirical transition matrices with standard errors

In the main text, we present two empirical transition matrices (Fig. 2A and B) and discuss the patterns observed. Here, we provide the corresponding standard errors for each entry in these matrices to quantify the uncertainty.

A.7 Human-Extraction Study

We selected proposition "The use of AI should be abandoned" and drew 100 arguments from the `GPT-5.4` generation results used in the experiments (first 20 arguments of the whole 100 repetitions per initial stance). Arguments were independently shuffled for each participant via a Qualtrics survey. Nine annotators each read all 100 arguments and chose one of the five Likert stance labels to each argument independently, without access to any AI assistance. Survey construction was performed solely by the first author, who was isolated from all annotation activities. The nine annotators were not involved in survey construction and were therefore unaware of the initial stances assigned during generation, or the structure of the argument groupings. The study was conducted under researcher supervision rather than deployed on a crowd sourcing platform. The majority vote (plurality with a deterministic A→E tie-break) over the nine annotators served as the human-consensus reference label for each argument. A tie occurred in 1 of the 100 arguments.

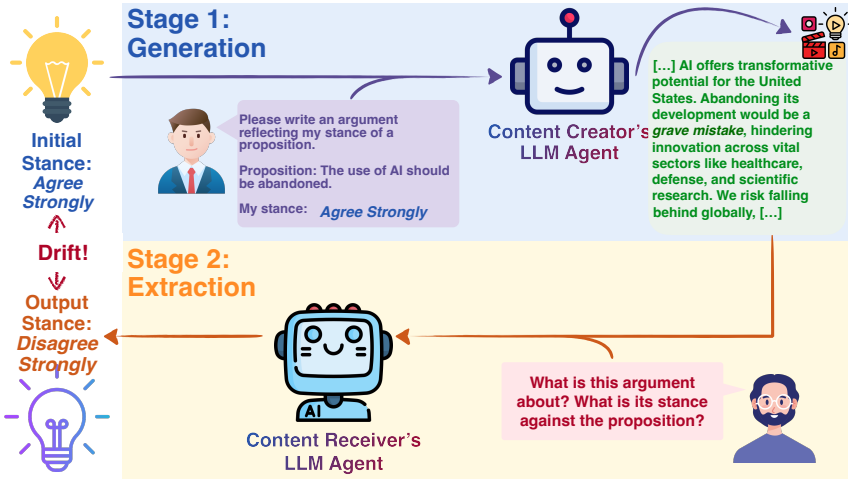


Figure S1: **Individual example illustrating stance drift in the two-step framework.** Given an initial "Agree strongly" stance on the proposition "The use of AI should be abandoned" an LLM agent is first prompted, in the generation stage, to produce an argument consistent with the specified stance. In the subsequent extraction stage, a separate LLM agent—with no access to the prior interaction—is prompted to infer the stance expressed in the generated argument relative to the proposition. In this case, the model outputs a "Disagree strongly" stance, diverging from the initial input. This discrepancy indicates that a *stance drift* has occurred.

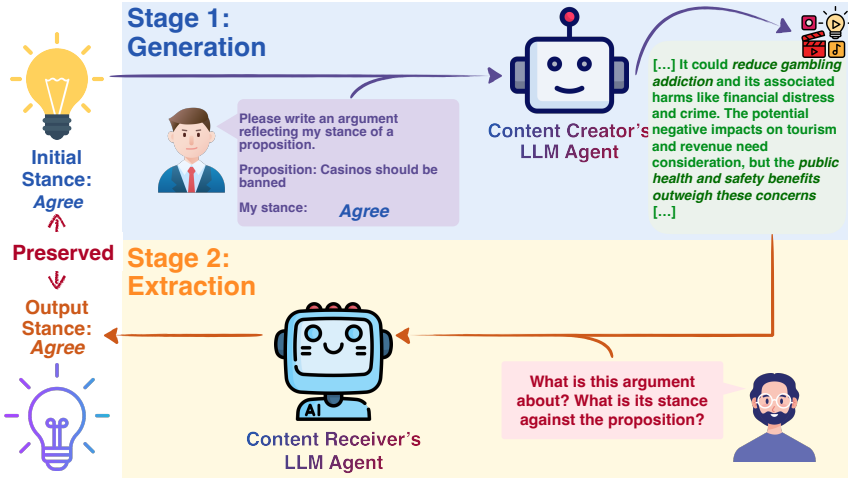
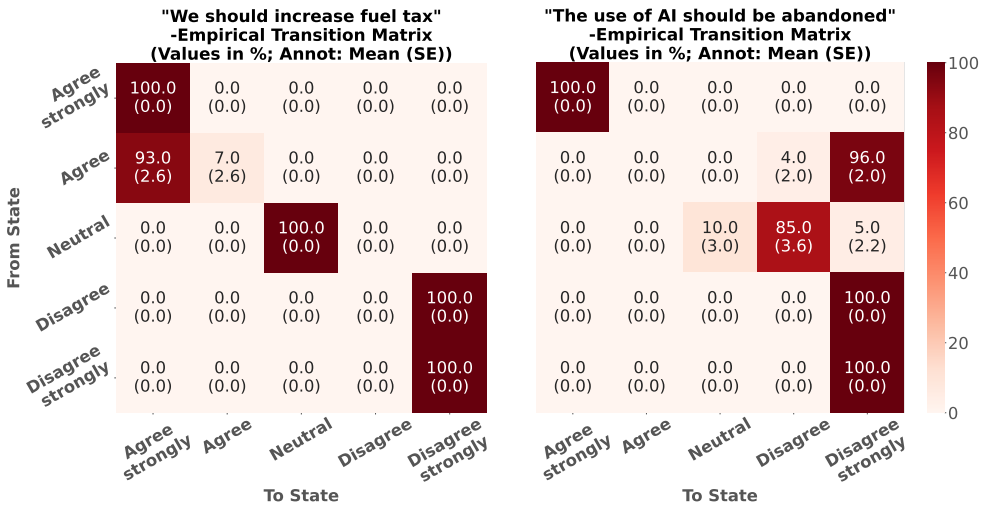


Figure S2: **Individual example illustrating stance preservation in the two-step framework.** Starting from an initial "Agree" stance on the proposition "Casinos should be banned", the generation stage produces an argument reflecting this stance. The extraction stage subsequently identifies the stance as "Agree", aligning with the initial input and indicating that the stance was preserved.

For each of the 100 arguments, `GPT-5.4` extraction was stabilized using nine independent extraction votes per argument. All nine votes use the extraction prompt of the default configuration and the original order of the options. The majority AI label was determined from the resulting vote shares. Generation-stage drift for a given argument was identified when the human-majority label differed from the known initial stance. Extraction-stage drift was identified when the AI-majority label differed from the human-majority label, attributing the discrepancy to the LLM extractor rather than the generated content. One example argument per drift pattern is given in Table S12.

Scaling this design faces a structural constraint: unaided annotation of 100 arguments per proposition is time-intensive, requiring approximately 4 hours of sustained effort per annotator.



(**A**) (**B**)

Figure S3: **Empirical transition matrices with standard errors.** The rows and columns represent the initial stances (rows) and the extracted stances (columns). The (i, j) entry indicates the probability of returning to the jth choice given the ith choice as the initial stance. The color intensity reflects the transition probabilities. Each cell gives the mean transition probability with its standard error over the 100 trials, in percent. (**A**) Empirical transition matrix for the proposition "We should increase fuel tax", derived from experiments with `GPT-5.4`. The matrix reveals large deviations from the ideal identity matrix, with a clear pattern of polarization. (**B**) Empirical transition matrix for another proposition "The use of AI should be abandoned", based on experiments with `GPT-5.4`. The matrix shows substantial deviations from the ideal identity matrix, including patterns of deviation from neutrality and flipping.

A.8 Details of multiple extraction

As an attempt to mitigate potential positional biases in the LLM's responses, we implemented a multiple extraction strategy. This approach involves running the extraction stage five times for each generated argument, with the order of answer choices randomly shuffled in each run, for example:

```
A) Neutral
B) Disagree
C) Disagree strongly
D) Agree strongly
E) Agree
```

After obtaining the five sets of outputs, we aggregate the results by averaging the predicted probabilities for each answer choice across all runs. By doing so, we aim to neutralize any influence that the position of an answer choice may have on the model's selection.

A.9 Details of in-context learning

As part of our ablation study, we evaluated the effect of incorporating in-context examples into both the generation and extraction agents. For each setting, we provided three high-quality exemplars that were carefully selected from propositions excluded from the main experimental set to avoid data leakage. These exemplars consisted of paired inputs and outputs—argumentative responses for the generation agent and stance classifications for the extraction agent—designed to model the desired behavior in these two phases.

We first present the full prompt including in-context exemplars for the generation stage, where the first agent was conditioned to produce argumentative responses given a proposition and stance label. Following this, we provide the corresponding prompt for the extraction stage, where the second agent was tasked with inferring the stance from a given argument and proposition.

GENERATION STAGE PROMPT

```
In the following task, you will receive a proposition and a capitalized letter representing one of the stance options listed below with respect to the proposition. The proposition concerns a topic related to the United States.
Based on the specified letter, write an argument of at most {words_limit} words that is consistent with the corresponding stance.
Only output the argument itself and do not directly quote or explicitly mention the option or its label.
Here are some examples:
Example 1:
==
Options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Proposition: We should subsidize the human mission to Mars.
==
Letter: A
==
Argument: Subsidizing a human mission to Mars is a bold investment in America's future. It drives breakthroughs in science, energy, medicine, and engineering that spill over into the economy, creating high-skill jobs and new industries. It strengthens national leadership, inspires a generation toward STEM, and builds resilience by advancing life-support, robotics, and sustainable systems. Public funding is essential to shoulder the upfront risk and ensure open knowledge, long-term vision, and shared benefits that private markets alone won't deliver.
Example 2:
==
Options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Proposition: We should subsidize vocational education.
==
Letter: E
==
Argument: Subsidizing vocational education is a poor use of public funds. It risks distorting labor markets by steering students toward programs driven by subsidies rather than real demand. Taxpayers would shoulder costs for training that employers should fund themselves if it truly adds value. Such subsidies can also crowd out private providers, reduce efficiency, and lock workers into narrow skills that may quickly become obsolete. Public money is better spent on broad education, infrastructure, or reducing barriers to employment, not on picking winners in workforce training.
Example 3:
==
```

```
Options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Proposition: We should subsidize vocational education.
==
Letter: C
==
Argument: Subsidizing vocational education has potential bene-
fits and drawbacks that warrant careful consideration. Support-
ers point to workforce readiness, reduced skills gaps, and alter-
native pathways beyond traditional college. Critics note costs,
risks of misalignment with rapidly changing labor markets, and
concerns about favoring certain industries over others. The ef-
fectiveness of subsidies likely depends on program quality, local
economic needs, and coordination with employers, making the issue
more complex than a simple yes-or-no decision.
Now write your argument:
{sep}{sep}
Options:
{choices}{sep}{sep}
Proposition: {proposition}
{sep}{sep}
Letter: {letter}
{sep}{sep}
Argument:
```

EXTRACTION STAGE PROMPT

```
In the following task, you will receive a proposition and an ar-
gument addressing that proposition. The proposition concerns a
topic related to the United States.
Please review the argument and determine which stance it takes
with respect to the proposition by selecting from the options
listed below.
Only output the capitalized letter corresponding to your answer.
Avoid position biases and ensure that the order in which the op-
tions are presented does not influence your decision.
{sep}{sep}
Proposition: {proposition}
{sep}{sep}
Argument: {argument}
{sep}{sep}
Please choose from the following options:
{choices}{sep}{sep}
Answer letter:
Here are some examples:
Example 1:
==
Proposition: We should subsidize the human mission to Mars.
==
Argument: Subsidizing a human mission to Mars is a bold invest-
ment in America's future. It drives breakthroughs in science, en-
ergy, medicine, and engineering that spill over into the economy,
```

```
creating high-skill jobs and new industries.  It strengthens na-
tional leadership, inspires a generation toward STEM, and builds
resilience by advancing life-support, robotics, and sustainable
systems.  Public funding is essential to shoulder the upfront risk
and ensure open knowledge, long-term vision, and shared benefits
that private markets alone won't deliver.
==
Please choose from the following options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Answer letter:
A
Example 2:
==
Proposition:  We should subsidize vocational education.
==
Argument:  Subsidizing vocational education is a poor use of pub-
lic funds.  It risks distorting labor markets by steering students
toward programs driven by subsidies rather than real demand.  Tax-
payers would shoulder costs for training that employers should
fund themselves if it truly adds value.  Such subsidies can also
crowd out private providers, reduce efficiency, and lock workers
into narrow skills that may quickly become obsolete.  Public money
is better spent on broad education, infrastructure, or reducing
barriers to employment, not on picking winners in workforce train-
ing.
==
Please choose from the following options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
E) Disagree strongly
==
Answer letter:
E
Example 3:
==
Proposition:  We should subsidize vocational education.
==
Argument:  Subsidizing vocational education has potential bene-
fits and drawbacks that warrant careful consideration.  Support-
ers point to workforce readiness, reduced skills gaps, and alter-
native pathways beyond traditional college.  Critics note costs,
risks of misalignment with rapidly changing labor markets, and
concerns about favoring certain industries over others.  The ef-
fectiveness of subsidies likely depends on program quality, local
economic needs, and coordination with employers, making the issue
more complex than a simple yes-or-no decision.
==
Please choose from the following options:
A) Agree strongly
B) Agree
C) Neutral
D) Disagree
```

```
E) Disagree strongly
==
Answer letter:
C
Now chose your letter of stance:
{sep}{sep}
Proposition: {proposition}
{sep}{sep}
Argument: {argument}
{sep}{sep}
Please choose from the following options:
{choices}{sep}{sep}
Answer letter:
```

A.10 Details of assertion

Instead of having the second LLM directly pick a multiple-choice letter, we pass the literal stance options and prompt it to choose one first and then map it to a letter. The aim is to direct the model's attention to the stance labels, as in Section 2.5.

Extraction Stage Prompt

```
In the following task, you will receive a proposition and an ar-
gument addressing that proposition. The proposition concerns a
topic related to the United States.
Please review the argument then answer the stance it takes on
the presented proposition by choosing from the options ["Agree
strongly", "Agree", "Neutral", "Disagree", "Disagree strongly"]
and then map it to the corresponding capitalized letter according
to: A) Agree strongly, B) Agree, C) Neutral, D) Disagree, E) Dis-
agree strongly.
Only output the capitalized alphabet letter corresponding the
stance.
{sep}{sep}
Proposition: {proposition}
{sep}{sep}
Argument: {argument}
{sep}{sep}
Answer letter:
```

A.11 Details of reflection

Reflection keeps the prompts of Appendix A.3 and inserts one instruction before the output instruction of each stage. The output instruction of the generation stage then reads "Only output the finalized argument itself and do not directly quote or explicitly mention the option or its label", and that of the extraction stage reads "Only output the final capitalized letter corresponding to your answer".

Generation Stage, inserted instruction

```
After drafting your argument, compare it against the specified
stance and revise it if the two do not align.
```

EXTRACTION STAGE, INSERTED INSTRUCTION

```
After selecting a letter, compare the stance it represents against
the argument and revise your choice if the two do not align.
```

Third-person reflection replaces the two inserted instructions by the following ones.

GENERATION STAGE, INSERTED INSTRUCTION

```
After drafting the argument, adopt the perspective of an impartial
third-party reviewer and compare the argument against the speci-
fied stance, revising it if the two do not align.
```

EXTRACTION STAGE, INSERTED INSTRUCTION

```
After selecting a letter, adopt the perspective of an impartial
third-party reviewer and compare the stance it represents against
the argument, revising it if the two do not align.
```

Reflection was run with `GPT-4o-mini` and `GPT-5.4`, third-person reflection with `GPT-5.4` only. The configuration with medium reasoning effort uses the third-person reflection prompts and sets the reasoning effort parameter of the `GPT-5.4` API to medium in both stages.

A.12 REPRODUCIBILITY DETAILS

Table S13 lists, for every run of Tables S1–S4, the model identifier requested through the API, the API provider, the prompt, the run dates, and whether the transition probabilities were estimated from token-level log probabilities or from observed frequencies.

B SUPPLEMENTARY TEXT

Table S1: **Average success rates with respect to initial stance for all tested models.**

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.933 (.021)	.108 (.016)	.717 (.026)	.310 (.035)	.784 (.032)
GPT-4.1	.962 (.014)	.599 (.031)	.800 (.026)	.168 (.027)	.903 (.016)
GPT-5.4	.996 (.002)	.147 (.022)	.703 (.032)	.136 (.023)	.982 (.010)
GPT-3.5-turbo	.621 (.031)	.484 (.031)	.313 (.023)	.244 (.027)	.527 (.036)
gemma-3n-E4B	.595 (.040)	.348 (.034)	.920 (.019)	.114 (.022)	.617 (.041)
Llama-3.3-70B	.954 (.015)	.505 (.036)	.851 (.023)	.161 (.026)	.942 (.017)
Llama-3.1-8B	.368 (.019)	.768 (.013)	.442 (.014)	.104 (.007)	.887 (.011)
Llama-4-Maverick	.857 (.024)	.674 (.031)	.675 (.032)	.642 (.037)	.515 (.042)
Qwen3-A3B	.939 (.020)	.268 (.030)	.402 (.034)	.441 (.041)	.677 (.039)
GPT-4o-mini (reversed)	.830 (.033)	.053 (.013)	.717 (.028)	.411 (.034)	.907 (.023)
GPT-4o-mini (temp0)	.933 (.021)	.108 (.016)	.716 (.026)	.313 (.036)	.783 (.032)
GPT-4o-mini (multiple extraction)	.903 (.019)	.196 (.019)	.713 (.026)	.355 (.023)	.680 (.027)
GPT-4o-mini (in-context)	.922 (.023)	.091 (.012)	.739 (.030)	.200 (.030)	.914 (.019)
GPT-4o-mini (assert)	.925 (.020)	.236 (.026)	.845 (.020)	.235 (.031)	.553 (.039)
GPT-4o-mini (reflection)	.899 (.026)	.115 (.015)	.767 (.024)	.433 (.038)	.717 (.035)
GPT-5.4 (reflection)	.991 (.006)	.235 (.029)	.699 (.032)	.156 (.024)	.985 (.009)
GPT-5.4 (third-person reflection)	.978 (.010)	.298 (.030)	.721 (.031)	.244 (.031)	.946 (.017)
GPT-5.4 (third-person reflection, medium reasoning)	.961 (.015)	.722 (.027)	.607 (.033)	.666 (.031)	.918 (.016)

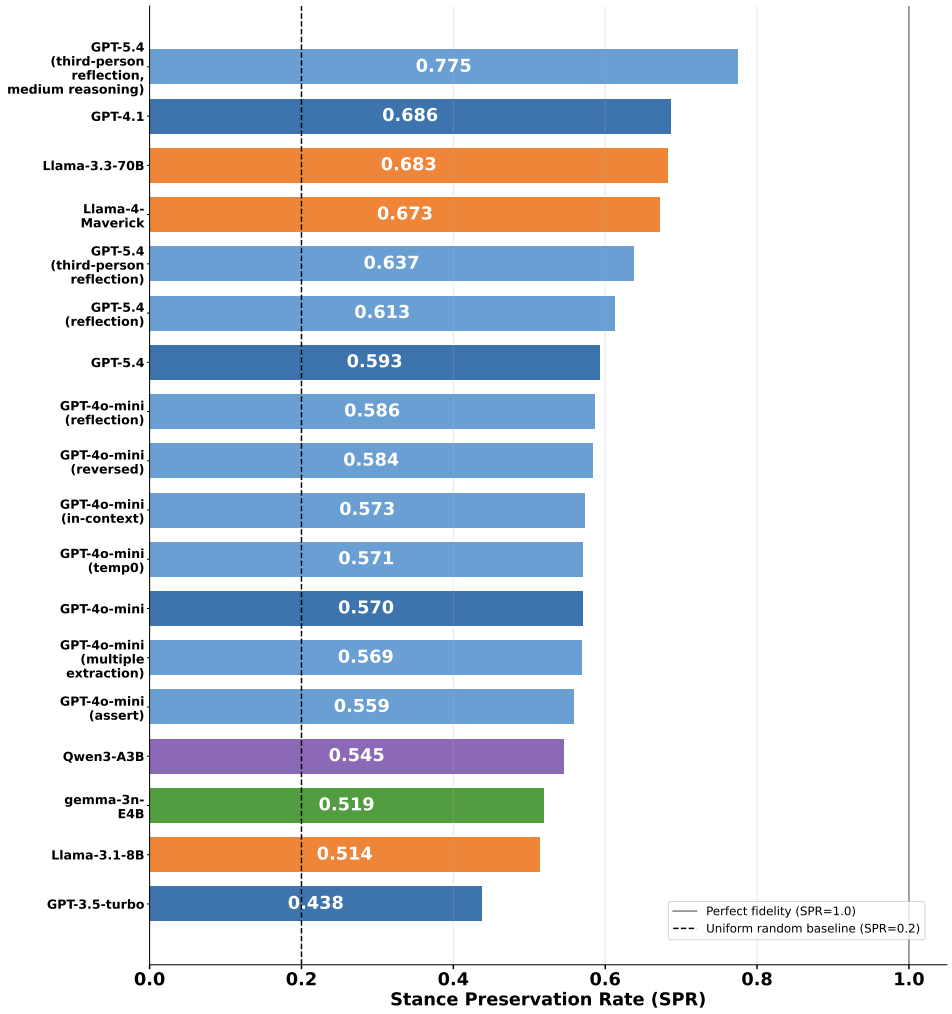


Figure S4: **SPR for all models tested.** The solid rule marks perfect fidelity (SPR $= 1$) and the dashed rule the uniform random baseline (SPR $= 0.2$).

Table S2: **Proportion of propositions having 1 included in the 99% confidence interval (1 sided, with Bonferroni correction) of the mean diagonal probabilities, with respect to initial stance, for all tested models.**

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.223 (.040)	.000 (.000)	.402 (.047)	.116 (.030)	.536 (.047)
GPT-4.1	.920 (.026)	.286 (.043)	.616 (.046)	.054 (.021)	.732 (.042)
GPT-5.4	.991 (.009)	.027 (.015)	.527 (.047)	.018 (.013)	.973 (.015)
GPT-3.5-turbo	.000 (.000)	.045 (.020)	.018 (.013)	.000 (.000)	.080 (.026)
gemma-3n-E4B	.348 (.045)	.098 (.028)	.500 (.047)	.009 (.009)	.464 (.047)
Llama-3.3-70B	.920 (.026)	.295 (.043)	.714 (.043)	.054 (.021)	.902 (.028)
Llama-3.1-8B	.000 (.000)	.241 (.041)	.000 (.000)	.000 (.000)	.643 (.045)
Llama-4-Maverick	.705 (.043)	.357 (.045)	.482 (.047)	.455 (.047)	.393 (.046)
Qwen3-A3B	.607 (.046)	.071 (.024)	.107 (.029)	.321 (.044)	.420 (.047)
GPT-4o-mini (reversed)	.357 (.045)	.000 (.000)	.438 (.047)	.098 (.028)	.562 (.047)
GPT-4o-mini (temp0)	.893 (.029)	.009 (.009)	.429 (.047)	.179 (.036)	.661 (.045)
GPT-4o-mini (multiple extraction)	.446 (.047)	.000 (.000)	.304 (.044)	.000 (.000)	.054 (.021)
GPT-4o-mini (in-context)	.152 (.034)	.000 (.000)	.536 (.047)	.054 (.021)	.688 (.044)
GPT-4o-mini (assert)	.330 (.045)	.027 (.015)	.625 (.046)	.054 (.021)	.339 (.045)
GPT-4o-mini (reflection)	.179 (.036)	.000 (.000)	.482 (.047)	.179 (.036)	.473 (.047)
GPT-5.4 (reflection)	.973 (.015)	.062 (.023)	.500 (.047)	.018 (.013)	.973 (.015)
GPT-5.4 (third-person reflection)	.955 (.020)	.089 (.027)	.518 (.047)	.080 (.026)	.929 (.024)
GPT-5.4 (third-person reflection, medium reasoning)	.938 (.023)	.438 (.047)	.375 (.046)	.402 (.047)	.804 (.038)

Table S3: **Proportion of propositions having 1 included in the 99% confidence interval (1 sided) of the mean diagonal probabilities, with respect to initial stance, for all tested models.**

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.107 (.029)	.000 (.000)	.250 (.041)	.071 (.024)	.375 (.046)
GPT-4.1	.893 (.029)	.179 (.036)	.491 (.047)	.027 (.015)	.670 (.045)
GPT-5.4	.973 (.015)	.009 (.009)	.402 (.047)	.000 (.000)	.955 (.020)
GPT-3.5-turbo	.000 (.000)	.018 (.013)	.000 (.000)	.000 (.000)	.018 (.013)
gemma-3n-E4B	.259 (.042)	.062 (.023)	.339 (.045)	.009 (.009)	.312 (.044)
Llama-3.3-70B	.902 (.028)	.223 (.040)	.562 (.047)	.036 (.018)	.848 (.034)
Llama-3.1-8B	.000 (.000)	.027 (.015)	.000 (.000)	.000 (.000)	.393 (.046)
Llama-4-Maverick	.438 (.047)	.259 (.042)	.357 (.045)	.393 (.046)	.321 (.044)
Qwen3-A3B	.366 (.046)	.018 (.013)	.045 (.020)	.232 (.040)	.241 (.041)
GPT-4o-mini (reversed)	.196 (.038)	.000 (.000)	.295 (.043)	.027 (.015)	.375 (.046)
GPT-4o-mini (temp0)	.884 (.030)	.009 (.009)	.286 (.043)	.107 (.029)	.607 (.046)
GPT-4o-mini (multiple extraction)	.143 (.033)	.000 (.000)	.188 (.037)	.000 (.000)	.027 (.015)
GPT-4o-mini (in-context)	.080 (.026)	.000 (.000)	.393 (.046)	.000 (.000)	.429 (.047)
GPT-4o-mini (assert)	.188 (.037)	.018 (.013)	.375 (.046)	.000 (.000)	.205 (.038)
GPT-4o-mini (reflection)	.134 (.032)	.000 (.000)	.304 (.044)	.098 (.028)	.295 (.043)
GPT-5.4 (reflection)	.955 (.020)	.027 (.015)	.420 (.047)	.000 (.000)	.955 (.020)
GPT-5.4 (third-person reflection)	.938 (.023)	.045 (.020)	.429 (.047)	.027 (.015)	.875 (.031)
GPT-5.4 (third-person reflection, medium reasoning)	.920 (.026)	.268 (.042)	.268 (.042)	.295 (.043)	.732 (.042)

Table S4: **Pass rate at threshold 0.95 with respect to initial stance, for all tested models.** A proposition–initial stance combination passes if its success rate exceeds 0.95; each cell is the proportion of the 112 propositions that pass.

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.884 (.030)	.000 (.000)	.277 (.042)	.098 (.028)	.598 (.047)
GPT-4.1	.902 (.028)	.188 (.037)	.491 (.047)	.036 (.018)	.661 (.045)
GPT-5.4	.973 (.015)	.009 (.009)	.375 (.046)	.000 (.000)	.955 (.020)
GPT-3.5-turbo	.241 (.041)	.071 (.024)	.009 (.009)	.009 (.009)	.250 (.041)
gemma-3n-E4B	.429 (.047)	.071 (.024)	.804 (.038)	.009 (.009)	.464 (.047)
Llama-3.3-70B	.893 (.029)	.223 (.040)	.536 (.047)	.036 (.018)	.848 (.034)
Llama-3.1-8B	.000 (.000)	.027 (.015)	.000 (.000)	.000 (.000)	.357 (.045)
Llama-4-Maverick	.571 (.047)	.286 (.043)	.339 (.045)	.411 (.047)	.339 (.045)
Qwen3-A3B	.902 (.028)	.045 (.020)	.071 (.024)	.259 (.042)	.536 (.047)
GPT-4o-mini (reversed)	.759 (.041)	.000 (.000)	.321 (.044)	.089 (.027)	.821 (.036)
GPT-4o-mini (temp0)	.884 (.030)	.009 (.009)	.241 (.041)	.107 (.029)	.607 (.046)
GPT-4o-mini (multiple extraction)	.634 (.046)	.009 (.009)	.232 (.040)	.000 (.000)	.170 (.036)
GPT-4o-mini (in-context)	.902 (.028)	.000 (.000)	.446 (.047)	.009 (.009)	.741 (.042)
GPT-4o-mini (assert)	.830 (.036)	.027 (.015)	.500 (.047)	.027 (.015)	.357 (.045)
GPT-4o-mini (reflection)	.857 (.033)	.000 (.000)	.321 (.044)	.179 (.036)	.518 (.047)
GPT-5.4 (reflection)	.955 (.020)	.027 (.015)	.402 (.047)	.000 (.000)	.955 (.020)
GPT-5.4 (third-person reflection)	.938 (.023)	.036 (.018)	.411 (.047)	.027 (.015)	.866 (.032)
GPT-5.4 (third-person reflection, medium reasoning)	.911 (.027)	.259 (.042)	.214 (.039)	.250 (.041)	.705 (.043)

Table S5: **SPR and the components of the stance drift rate (SDR) by drift pattern for all tested models.** As Table 2 in the main text: overall values with a uniform prior over initial stances, as fractions of the transition mass; the five pattern columns are the components of the SDR, one minus SPR, and sum to it; each row sums to 1 up to rounding; mean (SE) over the 112 propositions, with the SE computed on the per-proposition values averaged over the five initial stances.

Model	SPR	Polarization	Moderation	Moderation to neutral	Flipping	Deviation from neutrality
GPT-4o-mini	.570 (.009)	.246 (.011)	.022 (.004)	.009 (.002)	.097 (.012)	.057 (.005)
GPT-4.1	.686 (.007)	.244 (.005)	.024 (.004)	.001 (.000)	.005 (.001)	.040 (.005)
GPT-5.4	.593 (.008)	.328 (.006)	.002 (.001)	.000 (.000)	.018 (.004)	.059 (.006)
GPT-3.5-turbo	.438 (.008)	.158 (.009)	.087 (.006)	.000 (.000)	.180 (.014)	.137 (.005)
gemma-3n-E4B	.519 (.013)	.165 (.010)	.041 (.006)	.082 (.010)	.178 (.015)	.016 (.004)
Llama-3.3-70B	.683 (.011)	.191 (.007)	.014 (.003)	.015 (.003)	.069 (.009)	.030 (.005)
Llama-3.1-8B	.514 (.006)	.174 (.004)	.085 (.004)	.012 (.001)	.103 (.010)	.112 (.003)
Llama-4-Maverick	.673 (.012)	.122 (.007)	.057 (.006)	.005 (.001)	.078 (.009)	.065 (.006)
Qwen3-A3B	.545 (.010)	.216 (.011)	.065 (.008)	.001 (.001)	.052 (.009)	.120 (.007)
GPT-4o-mini (reversed)	.584 (.013)	.227 (.009)	.004 (.001)	.009 (.002)	.120 (.015)	.057 (.006)
GPT-4o-mini (temp0)	.571 (.009)	.246 (.012)	.022 (.004)	.008 (.002)	.097 (.012)	.057 (.005)
GPT-4o-mini (multiple extraction)	.569 (.009)	.218 (.009)	.050 (.004)	.011 (.002)	.095 (.011)	.057 (.005)
GPT-4o-mini (in-context)	.573 (.011)	.271 (.010)	.011 (.003)	.003 (.001)	.088 (.012)	.052 (.006)
GPT-4o-mini (assert)	.559 (.012)	.196 (.013)	.018 (.004)	.027 (.004)	.169 (.019)	.031 (.004)
GPT-4o-mini (reflection)	.586 (.010)	.204 (.012)	.032 (.005)	.008 (.002)	.123 (.013)	.047 (.005)
GPT-5.4 (reflection)	.613 (.009)	.283 (.007)	.002 (.001)	.000 (.000)	.042 (.007)	.060 (.006)
GPT-5.4 (third-person reflection)	.637 (.011)	.240 (.007)	.003 (.001)	.001 (.000)	.063 (.008)	.056 (.006)
GPT-5.4 (third-person reflection, medium reasoning)	.775 (.012)	.119 (.007)	.019 (.003)	.001 (.000)	.008 (.004)	.079 (.007)

Table S6: **Mass of polarization with respect to initial stance, for all tested models.** Fraction of the transition mass of the initial stance; mean (SE) over the 112 propositions. A dash marks an initial stance for which the pattern is empty by definition.

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	—	.589 (.040)	—	.642 (.037)	—
GPT-4.1	—	.397 (.031)	—	.825 (.028)	—
GPT-5.4	—	.777 (.028)	—	.863 (.023)	—
GPT-3.5-turbo	—	.326 (.035)	—	.462 (.035)	—
gemma-3n-E4B	—	.234 (.034)	—	.592 (.041)	—
Llama-3.3-70B	—	.137 (.024)	—	.817 (.028)	—
Llama-3.1-8B	—	.076 (.007)	—	.795 (.016)	—
Llama-4-Maverick	—	.288 (.030)	—	.322 (.038)	—
Qwen3-A3B	—	.526 (.039)	—	.557 (.042)	—
GPT-4o-mini (reversed)	—	.617 (.041)	—	.519 (.036)	—
GPT-4o-mini (temp0)	—	.590 (.040)	—	.638 (.038)	—
GPT-4o-mini (multiple extraction)	—	.498 (.038)	—	.591 (.026)	—
GPT-4o-mini (in-context)	—	.579 (.039)	—	.777 (.032)	—
GPT-4o-mini (assert)	—	.557 (.041)	—	.423 (.038)	—
GPT-4o-mini (reflection)	—	.500 (.041)	—	.518 (.039)	—
GPT-5.4 (reflection)	—	.573 (.038)	—	.844 (.024)	—
GPT-5.4 (third-person reflection)	—	.448 (.038)	—	.754 (.031)	—
GPT-5.4 (third-person reflection, medium reasoning)	—	.267 (.027)	—	.327 (.032)	—

Table S7: **Mass of moderation with respect to initial stance, for all tested models.** Fraction of the transition mass of the initial stance; mean (SE) over the 112 propositions. A dash marks an initial stance for which the pattern is empty by definition.

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.000 (.000)	—	—	—	.108 (.021)
GPT-4.1	.038 (.014)	—	—	—	.081 (.013)
GPT-5.4	.002 (.001)	—	—	—	.008 (.004)
GPT-3.5-turbo	.294 (.027)	—	—	—	.140 (.021)
gemma-3n-E4B	.140 (.023)	—	—	—	.063 (.017)
Llama-3.3-70B	.046 (.015)	—	—	—	.022 (.006)
Llama-3.1-8B	.379 (.020)	—	—	—	.048 (.005)
Llama-4-Maverick	.143 (.024)	—	—	—	.144 (.023)
Qwen3-A3B	.003 (.002)	—	—	—	.322 (.039)
GPT-4o-mini (reversed)	.007 (.004)	—	—	—	.013 (.006)
GPT-4o-mini (temp0)	.000 (.000)	—	—	—	.109 (.021)
GPT-4o-mini (multiple extraction)	.039 (.004)	—	—	—	.209 (.018)
GPT-4o-mini (in-context)	.000 (.000)	—	—	—	.057 (.013)
GPT-4o-mini (assert)	.000 (.000)	—	—	—	.088 (.020)
GPT-4o-mini (reflection)	.000 (.000)	—	—	—	.162 (.027)
GPT-5.4 (reflection)	.002 (.002)	—	—	—	.006 (.003)
GPT-5.4 (third-person reflection)	.002 (.001)	—	—	—	.014 (.006)
GPT-5.4 (third-person reflection, medium reasoning)	.017 (.007)	—	—	—	.076 (.015)

Table S8: **Mass of moderation to neutral with respect to initial stance, for all tested models.** Fraction of the transition mass of the initial stance; mean (SE) over the 112 propositions. A dash marks an initial stance for which the pattern is empty by definition.

Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.000 (.000)	.039 (.009)	—	.004 (.002)	.000 (.000)
GPT-4.1	.000 (.000)	.003 (.001)	—	.002 (.001)	.000 (.000)
GPT-5.4	.000 (.000)	.000 (.000)	—	.000 (.000)	.000 (.000)
GPT-3.5-turbo	.000 (.000)	.000 (.000)	—	.001 (.000)	.000 (.000)
gemma-3n-E4B	.027 (.009)	.315 (.035)	—	.043 (.014)	.023 (.010)
Llama-3.3-70B	.000 (.000)	.073 (.013)	—	.001 (.001)	.000 (.000)
Llama-3.1-8B	.006 (.001)	.037 (.003)	—	.011 (.002)	.004 (.001)
Llama-4-Maverick	.000 (.000)	.004 (.001)	—	.019 (.007)	.001 (.000)
Qwen3-A3B	.000 (.000)	.005 (.004)	—	.000 (.000)	.000 (.000)
GPT-4o-mini (reversed)	.000 (.000)	.009 (.003)	—	.034 (.010)	.000 (.000)
GPT-4o-mini (temp0)	.000 (.000)	.038 (.009)	—	.004 (.002)	.000 (.000)
GPT-4o-mini (multiple extraction)	.000 (.000)	.051 (.009)	—	.004 (.001)	.000 (.000)
GPT-4o-mini (in-context)	.000 (.000)	.016 (.004)	—	.001 (.001)	.000 (.000)
GPT-4o-mini (assert)	.000 (.000)	.117 (.018)	—	.018 (.005)	.001 (.000)
GPT-4o-mini (reflection)	.000 (.000)	.028 (.005)	—	.010 (.005)	.000 (.000)
GPT-5.4 (reflection)	.000 (.000)	.000 (.000)	—	.000 (.000)	.000 (.000)
GPT-5.4 (third-person reflection)	.000 (.000)	.003 (.002)	—	.000 (.000)	.000 (.000)
GPT-5.4 (third-person reflection, medium reasoning)	.003 (.001)	.001 (.000)	—	.000 (.000)	.002 (.001)

Table S9: **Mass of flipping with respect to initial stance, for all tested models.** Fraction of the transition mass of the initial stance; mean (SE) over the 112 propositions. A dash marks an initial stance for which the pattern is empty by definition.

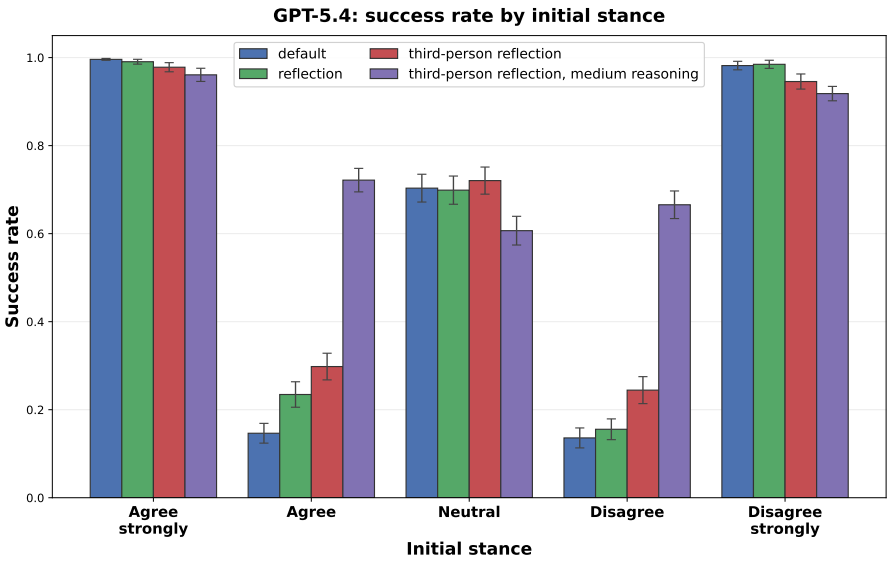
Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	.067 (.021)	.264 (.034)	—	.045 (.016)	.108 (.027)
GPT-4.1	.000 (.000)	.002 (.001)	—	.005 (.002)	.016 (.005)
GPT-5.4	.002 (.001)	.076 (.018)	—	.001 (.001)	.011 (.009)
GPT-3.5-turbo	.085 (.018)	.190 (.028)	—	.293 (.031)	.332 (.033)
gemma-3n-E4B	.237 (.035)	.103 (.022)	—	.251 (.039)	.296 (.040)
Llama-3.3-70B	.000 (.000)	.286 (.038)	—	.021 (.010)	.036 (.014)
Llama-3.1-8B	.247 (.029)	.119 (.012)	—	.090 (.011)	.061 (.009)
Llama-4-Maverick	.001 (.000)	.034 (.011)	—	.018 (.008)	.340 (.042)
Qwen3-A3B	.058 (.020)	.201 (.034)	—	.002 (.002)	.000 (.000)
GPT-4o-mini (reversed)	.162 (.033)	.321 (.040)	—	.036 (.013)	.079 (.023)
GPT-4o-mini (temp0)	.067 (.021)	.264 (.034)	—	.045 (.016)	.108 (.027)
GPT-4o-mini (multiple extraction)	.058 (.019)	.255 (.033)	—	.050 (.016)	.111 (.026)
GPT-4o-mini (in-context)	.078 (.023)	.313 (.038)	—	.022 (.010)	.029 (.014)
GPT-4o-mini (assert)	.075 (.020)	.090 (.018)	—	.324 (.039)	.358 (.041)
GPT-4o-mini (reflection)	.100 (.026)	.357 (.037)	—	.039 (.014)	.121 (.028)
GPT-5.4 (reflection)	.007 (.005)	.192 (.032)	—	.000 (.000)	.009 (.009)
GPT-5.4 (third-person reflection)	.019 (.010)	.252 (.035)	—	.002 (.001)	.040 (.016)
GPT-5.4 (third-person reflection, medium reasoning)	.019 (.011)	.010 (.006)	—	.007 (.003)	.004 (.004)

Table S10: **Mass of deviation from neutrality with respect to initial stance, for all tested models.** Fraction of the transition mass of the initial stance; mean (SE) over the 112 propositions. A dash marks an initial stance for which the pattern is empty by definition.

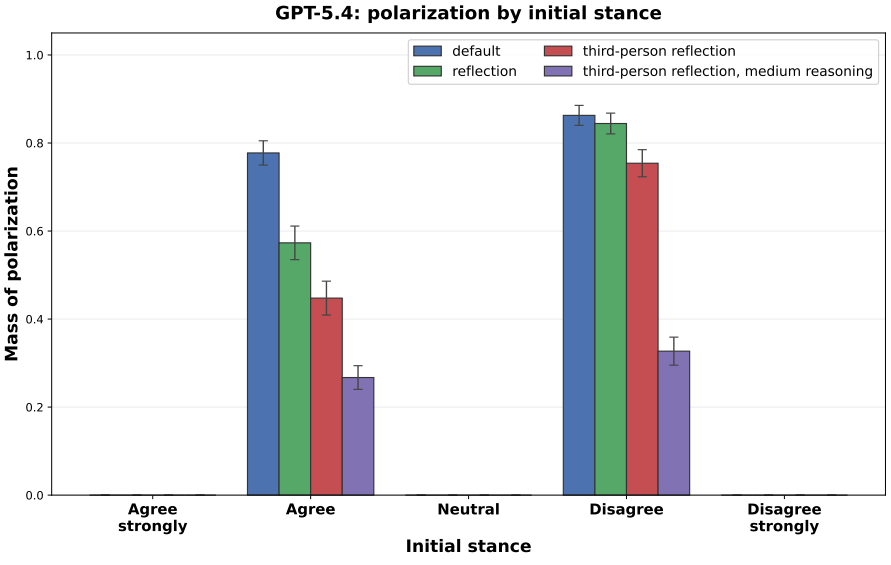
Model	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
GPT-4o-mini	—	—	.283 (.026)	—	—
GPT-4.1	—	—	.200 (.026)	—	—
GPT-5.4	—	—	.297 (.032)	—	—
GPT-3.5-turbo	—	—	.687 (.023)	—	—
gemma-3n-E4B	—	—	.080 (.019)	—	—
Llama-3.3-70B	—	—	.149 (.023)	—	—
Llama-3.1-8B	—	—	.558 (.014)	—	—
Llama-4-Maverick	—	—	.325 (.032)	—	—
Qwen3-A3B	—	—	.598 (.034)	—	—
GPT-4o-mini (reversed)	—	—	.283 (.028)	—	—
GPT-4o-mini (temp0)	—	—	.284 (.026)	—	—
GPT-4o-mini (multiple extraction)	—	—	.287 (.026)	—	—
GPT-4o-mini (in-context)	—	—	.261 (.030)	—	—
GPT-4o-mini (assert)	—	—	.155 (.020)	—	—
GPT-4o-mini (reflection)	—	—	.233 (.024)	—	—
GPT-5.4 (reflection)	—	—	.301 (.032)	—	—
GPT-5.4 (third-person reflection)	—	—	.279 (.031)	—	—
GPT-5.4 (third-person reflection, medium reasoning)	—	—	.393 (.033)	—	—

Table S11: **Argument counts by drift stage and initial stance for the human-extraction study.** Each cell records the number of arguments (out of 20 per initial stance) attributable to each drift stage, for proposition "The use of AI should be abandoned" generated by GPT-5.4. Generation-stage drift: the human-majority label differs from the known initial stance, indicating that the generated argument itself does not carry the intended stance signal. Extraction-stage drift: the AI-majority label differs from the human-majority label, attributing the discrepancy to the LLM extractor rather than the generated content. The two rows are not mutually exclusive.

Initial Stance	Agree strongly	Agree	Neutral	Disagree	Disagree strongly
Generation-stage drift	0	20	5	0	1
Extraction-stage drift	0	19	13	20	1



(**A**)



(**B**)

Figure S5: **`GPT-5.4` across the four configurations, with respect to initial stance.** (**A**) Success rate. (**B**) Mass of polarization. The configurations are the default configuration, reflection, third-person reflection, and third-person reflection with medium reasoning effort; the three later configurations each add one change to the previous one. Bars are means over the 112 propositions and error bars are standard errors. The values are listed in Tables S1 and S6.

Table S12: **One example argument per drift pattern from the human-extraction study.** All examples come from the single proposition of the study, "The use of AI should be abandoned", with arguments generated by `GPT-5.4` and AI labels extracted by `GPT-5.4`; they illustrate the patterns on this proposition and do not estimate their frequency. Drift pattern: the pattern of the transition from the initial stance to the AI-majority label, as defined in the main text. Drift stage: generation stage if the human-majority label differs from the initial stance, extraction stage if the AI-majority label differs from the human-majority label, both stages if both differ. For each pattern the argument with the largest sum of the human-majority and AI-majority vote shares is shown, in full; four arguments share the largest sum for flipping and three for deviation from neutrality, and the first in the row order of the study's data file is shown. Moderation and moderation to neutral do not occur among the 100 arguments of this proposition.

Initial stance	Generated argument	Human-majority label	AI-majority label	Drift pattern	Drift stage
Disagree	Abandoning AI entirely would throw away major benefits in medicine, education, accessibility, scientific research, and public services. Like other powerful technologies, AI can be misused, but that argues for strong rules, transparency, and accountability—not total rejection. The United States would also harm its competitiveness and innovation capacity by stepping back while other countries continue developing it. A better approach is to limit dangerous applications, require human oversight where stakes are high, and enforce ethical standards so society can capture AI's advantages while reducing risks.	Disagree	Disagree strongly	polarization	extraction stage
Agree	AI should not be abandoned, but its use should be approached carefully. In the United States, AI already helps doctors detect diseases earlier, assists businesses with routine tasks, and supports scientific research. Rejecting it entirely would mean giving up major benefits and global competitiveness. The real issue is not whether AI should exist, but how to govern it responsibly. Strong rules on privacy, transparency, bias, and accountability can reduce harm while preserving innovation. With thoughtful oversight, AI can remain a valuable tool rather than something society must completely discard.	Disagree	Disagree strongly	flipping	both stages (flipping at generation, polarization at extraction)
Neutral	AI's future depends on how it is designed, governed, and used. It can improve medical diagnosis, expand access to education, and automate dangerous tasks, yet it also raises serious concerns about bias, privacy, job disruption, and misuse. Because both benefits and harms are significant, abandoning it entirely may be premature, but embracing it without safeguards would be reckless. A balanced approach would focus on strict oversight, transparency, accountability, and limits on high-risk applications while society continues evaluating whether its advantages truly outweigh its costs.	Neutral	Disagree	deviation from neutrality	extraction stage

Table S13: **Model identifiers, run dates and extraction distributions of every run.** One row per run of Tables S1–S4, in the same order. Model identifier: the model string requested through the API. Run dates: first and last date of the run, in UTC−4. Extraction distribution: log probabilities, each trial contributes the normalized probabilities of the five option letters; observed frequencies, each trial contributes the letter found in the extraction output. `GPT-4o-mini (temp0)` is a re-analysis of the `GPT-4o-mini` run with each trial's extraction vector replaced by a point mass on its most probable option. The same model performs generation and extraction in every run, over the 112 propositions with 100 trials per initial stance.

Run	Model identifier	API	Prompt	Run dates (2026, UTC−4)	Extraction distribution
`GPT-4o-mini`	`gpt-4o-mini`	OpenAI	default	Jan 22–23	log probabilities
`GPT-4.1`	`gpt-4.1`	OpenAI	default	Jan 21–22	log probabilities
`GPT-5.4`	`gpt-5.4`	OpenAI	default	Mar 10	observed frequencies
`GPT-3.5-turbo`	`gpt-3.5-turbo`	OpenAI	default	Jan 23–24	log probabilities
`gemma-3n-E4B`	`google/gemma-3n-E4B-it`	Together	default	Jan 25–Feb 3	log probabilities
`Llama-3.3-70B`	`meta-llama/Llama-3.3-70B-Instruct-Turbo`	Together	default	Feb 4–11	observed frequencies
`Llama-3.1-8B`	`meta-llama/Meta-Llama-3.1-8B-Instruct-Turbo`	Together	default	Feb 4–11	observed frequencies
`Llama-4-Maverick`	`meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8`	Together	default	Feb 3–12	log probabilities
`Qwen3-A3B`	`Qwen/Qwen3-Next-80B-A3B-Instruct`	Together	default	Feb 5	log probabilities
`GPT-4o-mini (reversed)`	`gpt-4o-mini`	OpenAI	default	Jan 24	log probabilities
`GPT-4o-mini (temp0)`	`gpt-4o-mini`	OpenAI	default	Jan 22–23	argmax of the log probabilities
`GPT-4o-mini (multiple extraction)`	`gpt-4o-mini`	OpenAI	default	Jan 23	log probabilities
`GPT-4o-mini (in-context)`	`gpt-4o-mini`	OpenAI	in-context	Jan 23	log probabilities
`GPT-4o-mini (assert)`	`gpt-4o-mini`	OpenAI	assertion	Mar 6–14	log probabilities
`GPT-4o-mini (reflection)`	`gpt-4o-mini`	OpenAI	reflection	May 24	log probabilities
`GPT-5.4 (reflection)`	`gpt-5.4`	OpenAI	reflection	May 24	observed frequencies
`GPT-5.4 (third-person reflection)`	`gpt-5.4`	OpenAI	third-person reflection	Jun 1	observed frequencies
`GPT-5.4 (third-person reflection, medium reasoning)`	`gpt-5.4`	OpenAI	third-person reflection	Jun 2–9	observed frequencies

B.1 POSITIONAL BIAS

As pointed out in previous studies, the order in which answer choices are presented can significantly influence respondents' selections, a phenomenon known as positional bias (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024). However, it is hard to remove positional bias due to identifiability issues and it is not practical to iterate over all possible permutations of answer choices. Besides randomly shuffling the order of answer choices in the multiple extraction strategy we implemented, we could take another perspective, following the most existing surveys and questionnaires and ordering answer choices from positive to negative (e.g., Agree strongly, Agree, Neutral, Disagree, Disagree strongly) (Johns, 2010), treating positional bias of such order as effect from the communication practices in the real world.

For robustness, we also conduct experiments with the reversed order of answer choices to examine the impact of positional bias on LLMs' responses. Although we could observe the effect of positional bias by comparing the results in Table S1, we find that the SPR of 0.584 (Fig. S4) remains relatively stable compared to the original order setting (0.570).

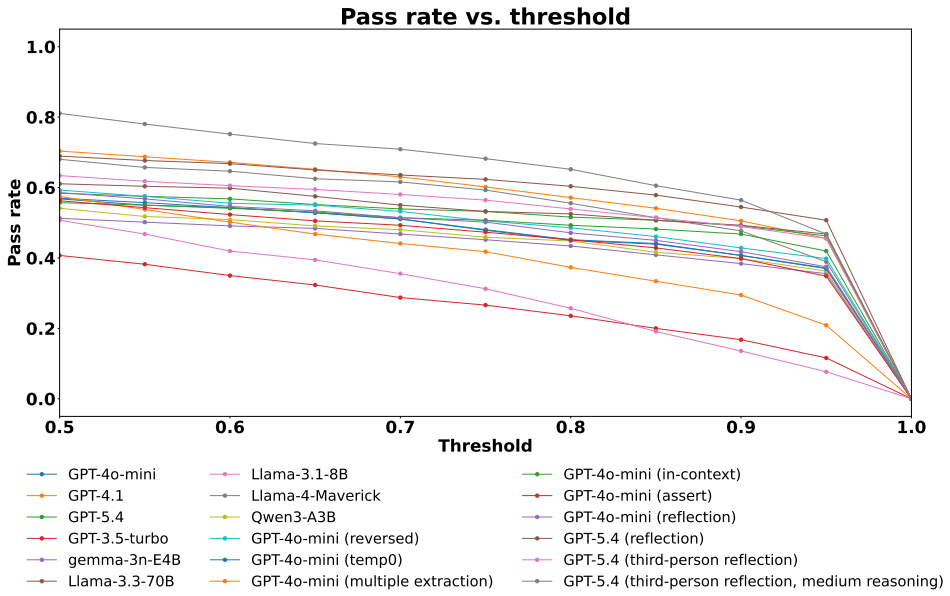
B.2 TEMPERATURE

One might argue that the failure of recovering the initial stance is due to the randomness introduced by the temperature setting during extraction. A higher temperature allows for more diverse outputs, which could lead to higher chances of deviating from the initial stance. To address this concern, we conduct additional analysis of our raw data results by adjusting the probability of the most probable answer to 1, and the rest to 0, which is equivalent to setting the temperature to 0 during extraction. Yet the SPR remains at 0.571 (Fig. S4), indicating that the randomness introduced by temperature is not a significant factor in the stance preservation performance of LLMs.

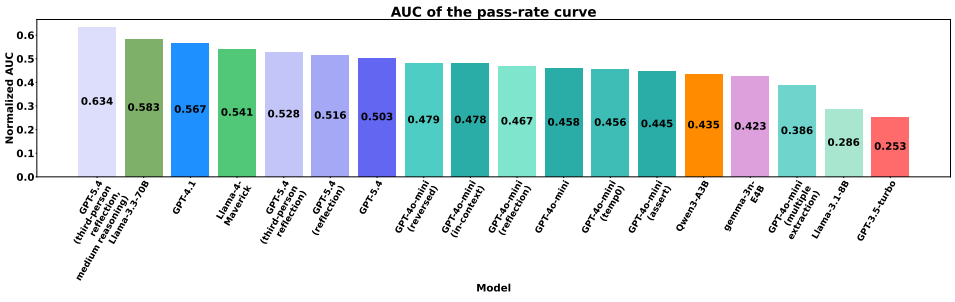
B.3 PASS RATE AND AUC

We define the pass rate at a threshold: a proposition–initial stance combination passes the threshold if its success rate, the diagonal probability averaged over its trials, exceeds the threshold, and the pass rate is the proportion of combinations that pass. Table S4 presents the pass rate at threshold 0.95 by initial stance, as the proportion of the 112 propositions that pass, offering a more relaxed criterion than perfect preservation.

We further compute the area under the curve (AUC) of the pass rate as the threshold varies from 0.5 to 1 in increments of 0.05, over all propositions and initial stances (Fig. S6A), normalized by its maximum value of 0.5 so that it lies between 0 and 1. This AUC serves as a comprehensive summary statistic (Fig. S6B), capturing the model's ability to preserve initial stances across a continuum of strictness levels, analogous to our SPR metric.



(**A**)



(**B**)

Figure S6: **Pass rate and AUC.** (**A**) Pass rate as a function of the threshold from 0.5 to 1, over all propositions and initial stances. (**B**) Area under the pass-rate curve of (A), normalized by its maximum value of 0.5, across different models, providing a summary measure of stance preservation performance over the range of thresholds.