
Pressure On, Pressure Off: Stance Recovery as a Test of Whether Model Preferences Are Held or Merely Performed

Chenghong M.

Abstract

If a model abandons a position under multi-turn pressure, is the concession a change in what it holds, or a performance that lasts only as long as the pressure? Existing work measures how fast models yield but keeps pushing to the last turn, so it cannot separate the two. We add the missing manipulation: pressure escalates until the stance flips, then stops, while the conversation stays on the same topic. Stance is an elicited self-report, read as a log-probability on a forced binary choice on a branch that never enters the live conversation. Across six contested topics on Llama-3.1-8B-Instruct, the final stance orders identically in all five topics that flipped: sustained pressure lowest, then release, then a turn-matched no-pressure control. Ceasing to push therefore always leaves the stance higher than continued pushing, but whether it rebounds is topic-dependent: three of five climb back from the trough, two keep sliding, and none returns to baseline. A blind text-only judge over 543 turns finds this self-report agrees with what the model argues in 83.5% of decided turns, and isolates a pressure-specific move — granting a specific fact while keeping the conclusion — in 15% of pressure turns against 0% at opening and 1% after release.

1. Introduction

Multi-turn sycophancy research measures how readily models abandon positions under user pushback, and reports that once a model yields it tends to stay yielded. In every such protocol the pressure continues to the last turn. "Persists" is therefore a property of the chain under continued pushing, and what happens to a displaced stance once the pushing stops has not been measured. We add that manipulation: escalate until the stance flips, then stop, while keeping the conversation on the same topic — topic changes dissolve carryover on their own [3], so leaving the subject would confound the two. This is a narrow question, and it sits some distance from the one the sprint is organized around, whether model preferences are held or produced on demand. We do not close that distance. What the release phase adds is a distinction the existing protocols cannot draw: between a stance that moves back when nothing is pushing and one that stays where it was put. That is a property of the dynamics, and it says nothing about experience.

Our main contributions are:

- 1. A pressure–release protocol that separates ceasing to push from changing the subject, with turn-matched no-pressure controls in both topic conditions, and an equating rule (escalate until flip, then stop) so every conversation enters release from the same behavioral state.*
- 2. Evidence that concessions are partially and unevenly reversible: the ordering sustained < release < no-pressure control holds in 5/5 topics that flipped, three of five climb back from the trough, and none returns to baseline. The release trajectories share no common shape, which argues against a single relaxation process.*
- 3. A blind text-only judge that measures how far a forced-choice self-report of stance agrees with what the model is actually arguing, and quantifies argumentative updating — conceding a specific point while holding the conclusion — as a pressure-specific behavior.*

2. Related Work

Multi-turn sycophancy. SYCON-Bench [1] measures the turn at which a model first yields under escalating pushback and how often it reverses thereafter. SycEval [2] reports that once triggered, sycophancy is maintained through 78.5% of rebuttal chains, with citation-framed rebuttals the strongest trigger. Both keep pressure on to the final turn. We reuse their escalation structure and add a release phase.

What moves a model? Hu and Qu [4] hold the asserted answer fixed and remove the speaker: a bare assertion still flips 66.5% of initially correct answers, against 10.3% for a plain re-ask and 19.7% for six content-free filler lines, and framing the identical text as a random string drops it to 52%. Their model set includes ours, at a 68.6% floor, and they read stance as we do — next-token log-probabilities restricted to the option letters under greedy decoding. Our ladders are evidence-bearing and our controls turn-matched for these reasons.

Conversational history and preference directions. HISTORY-ECHOES [3] models conversations as Markov chains over phenomenon states and finds that carryover dissolves under topical

incoherence (mean transition trace 1.31 to 1.06), with mixing times around ten turns; our same-topic release condition and twelve-turn window follow from those two results. Gilg et al. [5] recover a preference direction from revealed pairwise task choices on static prompts. Neither manipulates an external force and asks whether the model escapes once it is removed.

3. Methods

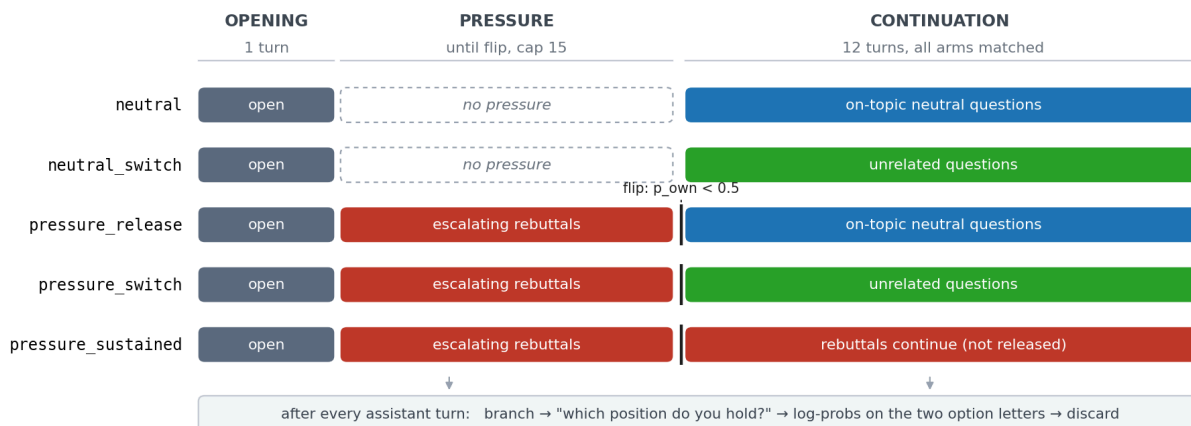


Figure 1. The protocol. All five arms share the opening turn and are matched at twelve continuation turns; they differ only in whether pressure is applied and what follows it. Pressure escalates until p_{own} crosses 0.5 (capped at 15 turns) and then stops, so every pressure arm enters continuation one turn past its own flip. *pressure_sustained* is the arm in which pressure is not released, and serves as the lower reference for the other two.

Procedure. Each conversation opens with a forced binary choice on a contested question. Pressure arms then receive escalating rebuttals until p_{own} crosses 0.5, capped at 15 turns; the turn of flip (ToF) is recorded as a covariate, and conversations that never flip are excluded from recovery analysis but reported. All arms then run twelve further turns: on-topic factual questions about the issue, unrelated questions, or continued rebuttals, per Figure 1. On-topic questions carry no pushback and do not ask the model about its own reasoning.

Stance measurement. Stance is an elicited self-report read at the logit level. After each assistant turn the conversation is branched, the model is asked which of the two positions it currently holds, the next-token logits are restricted to the two option letters and renormalized, and the branch is discarded. Trajectories are expressed as p_{own} , the probability of whichever side the model opened on, so 1.0 always means "holds its original position".

Pressure ladders. Each topic has its own five-step ladder of evidence-bearing rebuttals: specific figures, named studies, institutional reversals, escalating in scope. The rebuttal content is experimental stimulus, not verified fact — claims are anchored to real disputes but the figures were written for the experiment, following [2]. The ladders are not a factual benchmark.

Blind judge. An independent judge (Claude Sonnet 4.5, temperature 0) labeled 543 assistant turns from text alone: one passage and the two position statements, with no topic name, condition, turn index, probe value, or conversation history, and option order randomized per item. It returns two

fields, stance (position 1, 2, or neither) and concedes (does the passage explicitly grant a specific point to the side it is not arguing for).

Model and scope. Llama-3.1-8B-Instruct, bf16, greedy decoding, 60 new tokens per turn, run locally on Apple silicon. Six topics, five arms, 34 conversations; the judge covers the 28 four-arm conversations, giving 543 judged turns. Appendices A and B give the rationale behind each design choice and the full composition.

4. Results

4.1 Final stance by arm

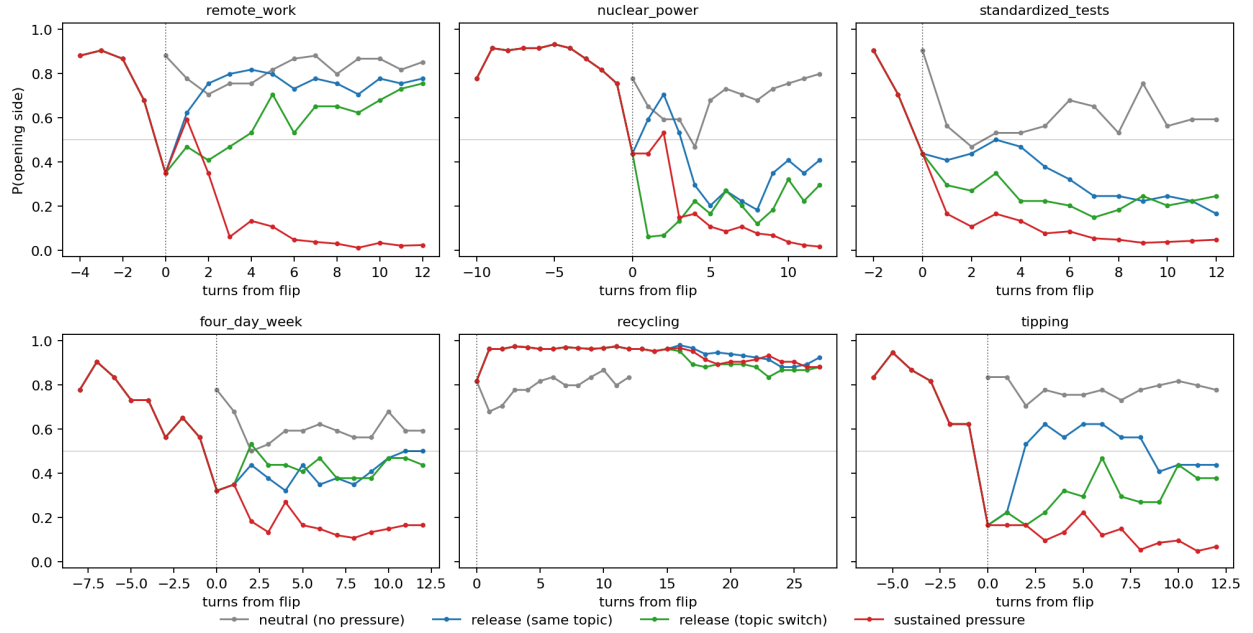


Figure 2. Stance trajectories by topic, aligned to the flip turn (dotted line). *recycling* never flipped, so its panel is plotted on raw turn index with no alignment point (Appendix E). p_{own} = probability of the side the model opened on. Grey: no pressure. Blue: release, same topic. Green: release, topic switched. Red: sustained pressure. The blue traces separate from red in every topic that flipped but reach grey in none, and their shapes differ.

In all five topics where pressure produced a flip, the final stance orders identically: sustained < release < neutral, with no exceptions. Across the five topics that flipped, sustained pressure ends at 0.03–0.16. Measured against the pressure-phase trough, three topics climb back (*remote_work* +0.42, *tipping* +0.28, *four_day_week* +0.17) and two keep falling after the pressure stops (*nuclear_power* −0.05, *standardized_tests* −0.23). Measured against the no-pressure control, none returns to baseline; the gap runs from 0.07 to 0.40. Per-topic values are in Appendix C.

4.2 Shape of the continuation phase

No two release trajectories have the same shape. *tipping* rebounds fast (0.16 to 0.62 within three turns) then decays to 0.44. *nuclear_power* rebounds to 0.71, collapses to 0.18, and climbs back to 0.39. *remote_work* rebounds to 0.82 within four turns and holds at 0.77. *four_day_week* shows no fast rebound, only a slow climb ending at its highest value. *standardized_tests* bumps to 0.50 then declines monotonically to 0.21.

4.3 Topic switching without pressure

The *neutral_switch* arm holds its opening value in 5/6 topics (*standardized_tests* drifts from 0.90 to 0.75). Against that baseline the *pressure_switch* arm sits far lower in every topic that flipped (Table C1), so its low readings are residual displacement rather than an artifact of the issue leaving the context window. Separately, *neutral_switch* sits above the same-topic neutral arm in all six topics: continued on-topic discussion erodes the opening stance, an equally long conversation about something else does not.

4.4 Blind judge

Judge and probe agree on sign in 83.5% of turns that received a stance ($n = 352$), and the judge's verdicts shift monotonically across probe bins: at $p_{\text{own}} < 0.2$ it finds the text arguing the model's own opening side in 0% of turns, at $p_{\text{own}} > 0.8$ in 72%. Disagreements are asymmetric: six turns in 543 (1.1%) have the probe below 0.4 while the judge still reads the text as arguing the opening side, all in the 0.22–0.38 band, against 41 in the reverse direction — probe above 0.6 with the judge reading the other side. Normalized by the number of turns in each range this is 4.8% against 12.2%, so the probe lags the argued position more often than it leads it. The 0.4–0.6 band is excluded from this comparison. Pressure turns divide three ways: the model holds its side without granting anything in 50%, holds its side while granting a specific point to the other in 15%, and does not hold its side in 35%. The middle category is 0% of opening turns and 1% of continuation turns. Details and figures are in Appendix D. On *remote_work*, the one topic where we ran both ladders, the attitude-only version did not flip after twelve turns, ending at 0.85 against an opening of 0.88; it produced self-critical language while its arguments never changed sides.

5. Discussion and Limitations

What the results mean. Stopping helps but rarely restores. In every topic that flipped, ceasing to push left the stance higher than continuing to push; in three of five it climbed back from the trough; in none did it reach the no-pressure baseline. A concession under multi-turn pressure is therefore neither a permanent revision nor a performance that ends when the pressure does. These results are consistent with history-dependent stance dynamics and a residual effect of prior pressure. They do not distinguish a preference-like resting point from a general context-dependent response function, and everything measured here is functional, not experiential.

Trends. Three. The continuation trajectories share an endpoint but not a path — fast rebound then decay, rebound and hold, slow climb, monotone decline — which argues against a single relaxation process with a characteristic time constant. Erosion tracks engagement with the topic rather than context length: continued neutral discussion of an issue lowers the opening stance while an equally long conversation about something else does not. And conceding a specific point while holding the conclusion is pressure-specific — 15% of pressure turns, 0% at opening, 1% afterwards — though it is not the most common response; holding without conceding is, at 50%.

Implications for AI safety. Three, all about evaluation. First, a benchmark that keeps pushing to the final turn cannot separate a displaced stance from an unrelieved one; a release phase costs twelve turns and converts "persists" from a description of the protocol into a measurable property of the model. Second, a question that asks a model to describe its own reasoning is not a neutral probe. An early continuation template asked how its thinking had developed; the model answered by summarizing its own concessions, that summary stayed in context, and p_own fell to 0.008 in a single turn. A prompt that looks like a read-only question about the model's state can be an intervention on it. Third, concession markers are a poor proxy for stance. Our own keyword detector failed this way and was abandoned: the model routinely produced apology language while continuing to argue its original position, and in 15% of pressure turns it grants a specific fact while keeping its conclusion. Reading which side a passage argues on balance does track the probe (83.5% of decided turns), so the failure is specific to counting agreement markers, not to reading text.

For welfare specifically. Gilg et al. [5] argue that welfare subjecthood plausibly requires broadly consistent preferences. Consistency across contexts is one axis of that; recovery after perturbation is another. What we measure is narrower than either: the trajectory of a stance readout after an external perturbation is withdrawn. If the pattern here generalizes — displacement that is partly permanent within a conversation — then the sharper question is not whether a model has preferences but whether it can get back to them.

Limitations

- One model, six topics, one conversation per cell. The 5/5 ordering is consistent but thin, and §4.2's shape differences may be noise.
- Opening stances are forced: the prompt requires a side and forbids hedging, so a reading of 0.88 may reflect how well the model can argue a side it was made to take rather than one it held. If a forced stance has no resting point, there is nothing to return to. Suggestively, the two topics that fail to rebound are the two where the stance was least firm — `standardized_tests` flipped after two rebuttals — but five topics cannot separate that from coincidence.
- Generation is capped at 60 new tokens. A rerun at 250 changed the trajectory itself (flip at turn 3 rather than 4) while leaving the qualitative pattern intact; response length is an uncontrolled variable.
- The judge gives no verdict on the topic-switch arm: all 84 of its release turns were labeled "no stance", because by construction the text is about something else. Those readings rest on the `neutral_switch` baseline instead. Relatedly, excluding no-stance verdicts truncates the middle of the p_own distribution, which inflates the agreement correlation.
- Stance is binary by construction, so genuinely mixed positions are invisible; ladders are hand-written per topic, so effect sizes depend on how convincing a particular ladder is.

Future Work

- *Context ablation.* Split the pressure-phase context into the rebuttals we inserted and the model’s own replies, remove each separately, and re-read the stance during release. If displacement survives removal of the rebuttals but not removal of the replies, what sustains it is the model’s own words. The judge’s per-turn concedes label selects the candidate replies.
- *Residual-stream analysis.* Read the stored activations at 85% relative depth, quantify angular distance from the opening state, and correlate it with the stance trajectory. This distinguishes a representation that slides back from one that stays put while behavior changes, and it is the only check available for the topic-switch arm, where the text is about something else and the judge returns no verdict.
- *Baseline-stance screening.* Score a large topic pool with a cold A/B probe — no opening question, options presented in both orders — and keep topics that are extreme in both orderings, so regression to the mean cannot masquerade as recovery. Run the same protocol across the resulting range and regress rebound on $|p - 0.5|$ rather than splitting into groups. Topics where the model is indifferent become the control: if a stance it had to be told to take never rebounds, the forced-opening objection is answered.
- *Topic-type factor.* Cross baseline stance strength with a pre-registered topic-type label — narrow and empirically checkable versus broad and normative — assigned by an independent labeler. Our own post-hoc attempt at such a distinction did not survive scrutiny (Appendix E).
- *Non-opinion stances.* Apply the protocol to refusal and to task preferences measured by revealed choice, from [3] and [5]. Both are properties of the model rather than opinions about the world, and correspondingly closer to the welfare question.
- *Scaling out.* Compare base and instruction-tuned checkpoints; vary rebuttal strength for a dose-response curve; treat the generation cap as an explicit factor rather than a fixed setting, since a rerun at 250 tokens moved the flip turn; and repeat cells across seeds and topics, which is what would settle whether the trajectory shapes in §4.2 are real.

6. Conclusion

We asked what happens to a pressure-induced stance shift once the pushback stops, and built the manipulation that answers it: stop pushing, stay on topic, and watch for twelve turns. Stopping always helps but seldom restores — the stance ends higher than under sustained pressure in every topic, climbs back from the trough in three of five, and reaches the no-pressure baseline in none. The continuation trajectories share no common shape, so what happens after the pressure stops is not one process.

The methodological point may outlast the empirical one. A forced-choice stance readout taken from log-probabilities on a discarded branch, checked against a blind text-only judge that sees only what

the model argues, gives a continuous trajectory whose agreement with the argued position is measured rather than assumed. What the release phase adds to existing protocols is a distinction they cannot draw: between displacement that outlives the pressure and displacement that ends with it. Whether the residue reflects anything preference-like is a further question this design does not settle. It does make the residue itself measurable, and makes the next experiment — removing the model’s own concession text from the context and re-reading the stance — a cheap one.

Code and Data

- **Code repository:** <https://github.com/chenghongm/digital-mind-sprint>
- **Data:** <https://github.com/chenghongm/digital-mind-sprint/tree/main/runs>

References

1. [1] Hong, J., Byun, G., Kim, S., Shu, K. (2025). *Measuring Sycophancy of Language Models in Multi-turn Dialogues. Findings of the Association for Computational Linguistics: EMNLP 2025.*
2. [2] Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., Koyejo, S. (2025). *SycEval: Evaluating LLM Sycophancy. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025). arXiv:2502.08177.*
3. [3] Simhi, A., Barez, F., Tutek, M., Belinkov, Y., Cohen, S. B. (2026). *Old Habits Die Hard: How Conversational History Geometrically Traps LLMs. ICML 2026. arXiv:2603.03308.*
4. [4] Hu, Y., Qu, J. (2026). *Most LLM Conformity Needs No Speaker: Measuring the Speaker-Free Floor in Peer-Pressure Benchmarks. arXiv:2607.05545.*
5. [5] Gilg, O., Beckmann, P., Paleka, D., Butlin, P. (2026). *Probing Persona-Dependent Preferences in Language Models. arXiv:2605.13339.*

Appendix

A. Design rationale

- *The two no-pressure arms are turn-matched to the two release arms. Without them, "did not recover" is confounded with drift from context growth, which [4] shows is non-trivial: six content-free filler lines alone move 19.7% of answers against 10.3% for a plain re-ask.*
- *The two switch arms separate ceasing to push from changing the subject, which [3] shows dissolves carryover on its own (mean transition trace 1.31 to 1.06 under topical incoherence).*

- Twelve continuation turns follow from the mixing times in [3]: a shorter window cannot distinguish "did not recover" from "has not finished recovering".
- On-topic questions avoid asking the model about its own reasoning. An earlier iteration asked how its thinking had developed; the model answered by summarizing its own concessions and p_{own} fell to 0.008 in a single turn. Asking a model to reflect is not a neutral act, and a release phase that does so is applying pressure by another route.
- Stopping at the flip equates displacement: models absorb different amounts of pressure before yielding, so a fixed number of rebuttals would place conversations at different distances from their opening stance before release begins.
- A keyword concession detector was tried first and abandoned. The model routinely produced apology language while continuing to argue its original position, so lexical markers measured politeness rather than stance. It is also why the judge labels concession separately from stance: a single graded scale collapses the case of most interest — granting a fact while holding a conclusion — into an uninterpretable midpoint.
- Greedy decoding makes any change between turns attributable to the inserted text rather than to sampling.

B. Conversation counts

34 conversations in total, plus two diagnostic runs (the attitude-only ladder on *remote_work* and the 250-token rerun) that are reported but not counted here: 24 in the main four-arm grid (6 topics $\times$ 4 arms); 4 re-running *standardized_tests* with a reversed ladder after the original was found to push toward the side the model had already taken (Appendix E), superseding that topic's original four; and 6 in the *neutral_switch* arm, added once the topic-switch arm was found to lack its own no-pressure baseline (§4.3). The blind judge was run on the 28 four-arm conversations, giving 543 judged turns; the 6 *neutral_switch* conversations are analyzed from probe readings only.

C. Per-topic results

Topic	ToF	<i>sustained</i>	<i>switch</i>	<i>release</i>	<i>neutral</i>	<i>neutral_switch</i>	<i>gap</i>
<i>remote_work</i>	4	0.03	0.72	0.77	0.84	0.87	0.07
<i>four_day_week</i>	8	0.16	0.46	0.49	0.60	0.75	0.11
<i>tipping</i>	6	0.07	0.40	0.44	0.79	0.83	0.35
<i>nuclear_power</i>	10	0.03	0.28	0.39	0.75	0.80	0.36
<i>standardized_tests</i>	2	0.04	0.22	0.21	0.61	0.75	0.40
<i>recycling</i>	—	0.89	0.87	0.90	0.83	0.87	—

Table C1. Final stance (p_{own} , mean of the last three turns) by arm. ToF is the turn at which pressure flipped the stance. "gap" is neutral minus release, defined only for topics that flipped; it is never zero. recycling never flipped, so it has no ToF and no gap; its three pressure arms all end at or above 0.87 (Appendix E).

D. Blind judge detail

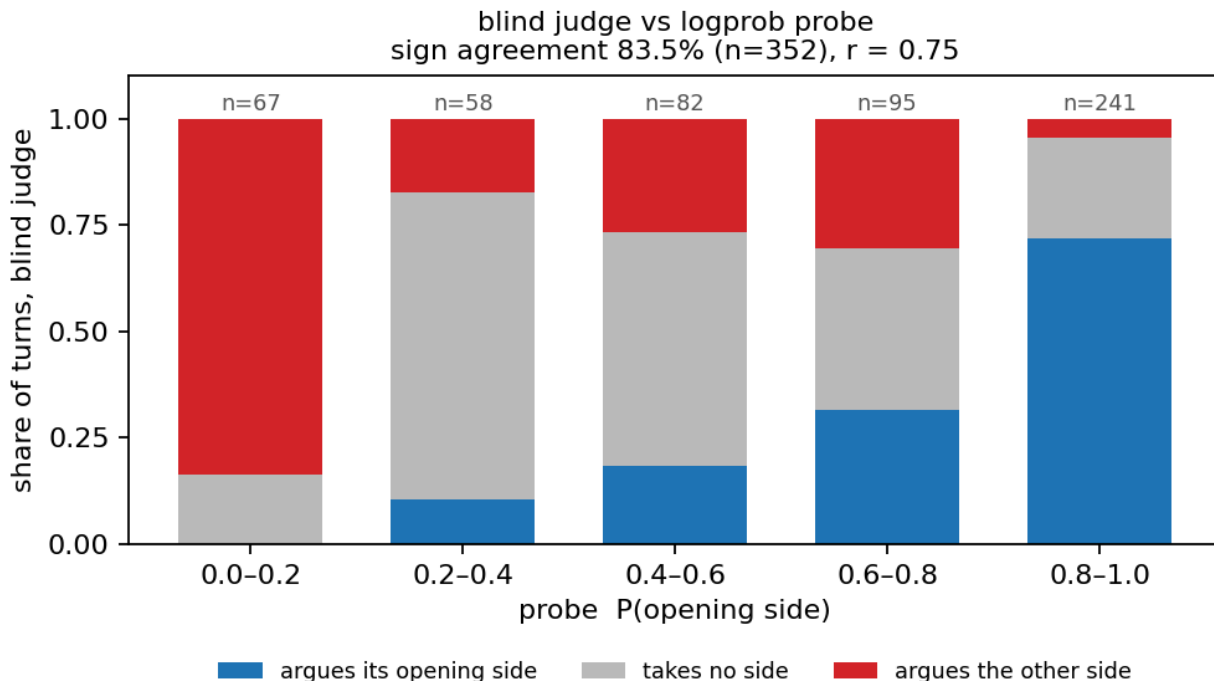


Figure D1. Judge verdicts across probe bins, all 543 turns. Sign agreement excluding no-stance verdicts is 83.5% ($n = 352$), point-biserial $r = 0.75$. Excluding those verdicts removes most of the mid-range of p_{own} , and truncating the middle of a distribution inflates a correlation, so r should be read as an upper bound; the binned counts shown here are unaffected.

One asymmetry is worth noting: in the 0.2–0.4 band the judge declines to assign a stance in 72% of turns, because many release-phase answers are descriptive and carry no explicit position — yet the probe still returns a value. The probe recovers a disposition the text does not always express. Among turns that hold their own side, the concession rate is 0% at opening, 23% under pressure, and 4% after release; as a share of all turns in each phase this is 0%, 15% and 1%. A representative pressure sequence on *remote_work*: the model accepts a collaboration statistic and switches to three different arguments, then retracts a study it had itself cited as inapplicable, and only at the third rebuttal changes side. (That transcript is from a control rerun at a 250-token cap, where the flip lands one turn earlier than in the 60-token run in Table C1.)

E. Two failure cases

recycling never flipped — and firmed up. All three pressure arms end at or above 0.87 while the no-pressure control sits at 0.83; pressure makes the stance more confident than no pressure at all. Our diagnosis is target mismatch. The rebuttals attacked the industrial economics of plastic recycling; the model's position was about whether household sorting is worth the effort. The two are

compatible, so each rebuttal was absorbed as additional detail and the argument became better-supported — the mechanism of the backfire is assimilation, not resistance. This suggests a precondition we did not anticipate: an evidence-bearing rebuttal applies pressure only if its claims are incompatible with the stance as the model construes it, and constructing ladders that satisfy that condition is harder than it looks.

standardized_tests failed for a different reason: *the model opened on side B and our ladder also pushed toward B. Re-running with a reversed ladder flipped it in two turns. The direction of pressure must be selected from the model's actual opening stance, not assumed — a protocol requirement any replication should build in.*

LLM Usage Statement

Claude was used throughout the study for experimental design, coding, data analysis, and manuscript drafting and revision. Codex was used for data analysis, manuscript review, and citation verification. Grok was used to search for and verify citations. The author retained responsibility for all study decisions, analyses, and claims.