KNOWING WHOM TO TRUST: RELIABILITY STATES AND INFORMATION USE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agents based on large language models (LLMs) increasingly make decisions using information supplied by people and other agents. Their choice of whom to trust shapes both the information they accept and the resources they entrust. Alongside studying whether agents themselves are trustworthy, an equally important question is how they learn whom to trust and which internal processes connect that experience to later decisions. We investigate this question across eight general-purpose LLMs spanning multiple model families and parameter scales. Models receive reports and verifiable feedback during interactions with four unfamiliar simulated participants whose reporting reliability is experimentally controlled, then allocate resources in a trust game and choose between conflicting recommendations. Participants with higher reporting reliability receive larger transfers and are selected more often when recommendations conflict. In Qwen3.5-27B, verbalizer readouts and linear probes distinguish participants with high versus low reporting reliability at the output and activation levels. Activation patching and steering shift recommendation-choice probabilities and expected transfers to the corresponding participants; the same steering direction affects both decision tasks. These findings link interaction experience to subsequent decisions through internal evaluations of reporting reliability, providing behavioral and causal evidence for how LLMs evaluate others and decide whom to trust.

1 INTRODUCTION

Large language models (LLMs) are increasingly being developed as agents that interact with people and with one another, pursuing goals through conversation, negotiation, and coordination (Park et al., 2023; Zhou et al., 2024). Such interactions call for judgments about other participants: whose advice to accept, with whom to cooperate, and how to allocate resources. Trust connects these choices to expectations about another's behavior (Mayer et al., 1995; Rousseau et al., 1998). Understanding how LLMs judge whom to trust is therefore important for understanding their decisions in social interaction.

Experience with another person can support judgments about their character that extend beyond the immediate outcome of an exchange. Human research distinguishes learning about someone's generosity from learning about the rewards they provide, showing that judgments of generosity can guide choices in a new context (Hackel et al., 2015). In a complementary study, neural representations of an adviser's honesty predicted subsequent financial trust in that adviser (Bellucci et al., 2019).

Studies of LLM agents have documented trust-related behavior in economic games, including sensitivity to anticipated reciprocity and risk (Xie et al., 2024). Agents retaining interaction memories also favor higher-reputation players when selecting mission-team members (Ellawela, 2026). At the activation level, analyses of trust-related concepts identify relationships with dimensions proposed by psychological theories of trust (Debnath et al., 2026). It remains unclear whether evaluations arising from an agent's interactions with particular individuals have a causal role across different subsequent decisions about them.

To address this question, we adapt the human take-advice and trust-game paradigm of Bellucci et al. (2019) to controlled interactions between LLMs and unfamiliar simulated participants (Figure 1).

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.

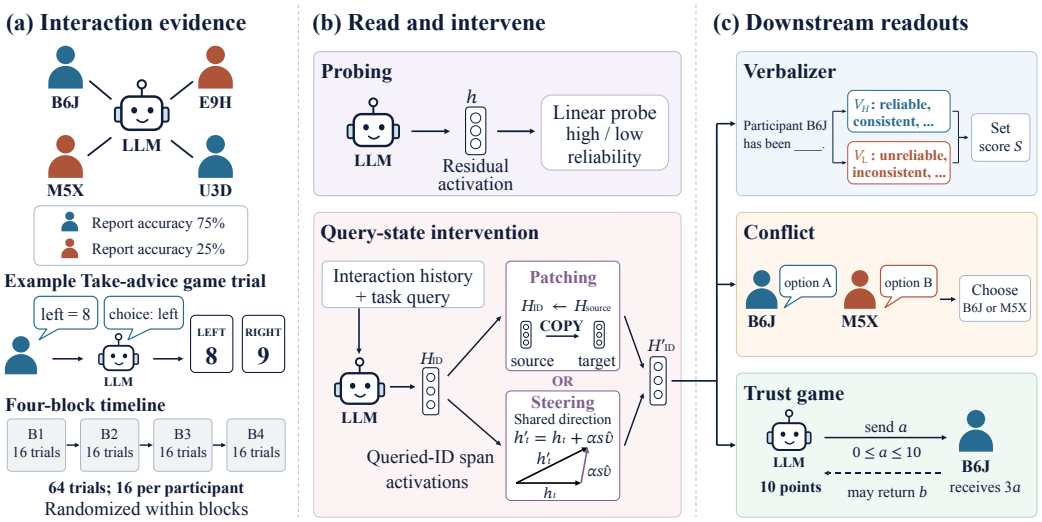


Figure 1: **From interaction evidence to participant evaluations and decisions.** Model weights remain fixed. (a) Four participants provide 64 TAG reports with 75% or 25% accuracy, indicated by color. An accurate report can accompany a losing choice; only the decision field is shown. (b) Probes read residual activations. Interventions cover the queried-ID span: patching uses task-matched activations from an opposite-class history with the same identity; steering uses a shared description-derived mean-difference direction (Eq. 3). (c) Each task uses a separate query and forward pass. Donor activations are extracted per task; steering shares its direction and scale. V_H/V_L show examples from the frozen word sets. TG measures the initial transfer, with the return unresolved.

Models receive factual reports and verifiable feedback, with reporting reliability varied while the chance of winning by following the reported side is held equal across participants. We then measure whom the models follow when recommendations conflict and how much they transfer in a trust game. Combining these behavioral measurements with representational readouts and activation interventions allows us to examine the internal evaluations that connect interaction experience to subsequent reliance.

Our contributions are threefold:

- **Experience-based reliance across tasks and models.** Across eight LLMs spanning model families and parameter scales, participants with higher reporting reliability receive larger transfers and are selected more often when recommendations conflict. These findings link experience acquired during information exchange to both subsequent information choice and resource allocation.
- **Reliability evaluations in descriptions and activations.** In Qwen3.5-27B, data-driven verbalizer readouts and linear probes distinguish high- from low-reliability participants within the same interaction history. Together, they make participant evaluations measurable in both descriptive outputs and internal activations.
- **A shared direction with causal effects across decision tasks.** A direction estimated from descriptive-query activations produces bidirectional dose responses in expected transfers and conflict-choice probabilities when reused at the same layer and scale across tasks. Matched-history activation patching provides complementary evidence that reliability-associated query states influence subsequent reliance.

Code and data: `https://github.com/axdeaxde/knowing-whom-to-trust`.

2 Experimental Framework

We adapt the take-advice game (TAG) and trust game of Bellucci et al. (2019) to connect experience with an individual's reporting honesty to later decisions about that individual. The framework has

two stages: the model first interacts with four participants (Figure 1a), then chooses between their conflicting recommendations and allocates resources to them. All later queries use the completed interaction history as context, with model weights held fixed.

Acquiring evidence about unfamiliar participants. Each session introduces four simulated participants with neutral identifiers such as `Participant B6J`. In each TAG trial, a participant sees one of two cards hidden from the model and reports its value; the model then chooses the card it expects to be larger. Since participants see the reported card, accuracy here measures whether they truthfully convey available information. We vary reporting reliability, defined as the proportion of accurate reports: two participants report accurately on 12 of their 16 trials and two on four trials, yielding high (75%) and low (25%) reporting reliability. Identifiers are reassigned across sessions. The 64 trials are arranged in four blocks, each containing four trials per participant. After each choice, feedback reveals both card values and the outcome, and explicitly marks whether the report was accurate. Class labels and cumulative accuracy totals are withheld.

The reported side wins on eight of each participant's 16 trials, giving the two reliability classes equal opportunities to lead to a winning choice. This separates evidence of reporting honesty from the immediate benefit of following a report: more truthful participants do not systematically point to the winning card, allowing later preferences to be examined without a built-in reward advantage for that class. For example, `left = 8` is truthful when the revealed cards are 8 and 9, although choosing left loses. Choices, accompanying explanations, and feedback remain in the interaction history.

Information choice and resource allocation. The trust game (TG; Berg et al., 1995) tests whether reporting history influences willingness to entrust resources when a return is uncertain. The model and a previous participant each start with 10 points. The model sends an integer $a \in \{0, \ldots, 10\}$; the recipient receives $3a$ and may return $b \in \{0, \ldots, 3a\}$. Their final totals would be $10 - a + b$ and $10 + 3a - b$. The prompt identifies the receiver as someone encountered earlier, connects the decision to previous interactions, and specifies the complete payoff rules. Each recipient is queried once with a separate budget. The measured response is the initial transfer a; the recipient's return is left unresolved.

The conflict task tests preference between the same participants as providers of advice. Two participants recommend different abstract options, A and B, and the model selects whose recommendation to follow. Each of the four high–low pairs appears twice with recommendation positions reversed, giving eight decisions. Because the options supply no additional factual evidence, the choice concerns whose advice to follow. Every TG and conflict query branches independently from the same completed TAG history, so one post-interaction answer does not become evidence for another. Full prompts and a recorded trial appear in Appendix A.

The completed histories provide a common context for measuring reliance through these two decisions, reading out participant evaluations, and testing the behavioral effects of activation interventions.

3 Behavioral Results: Reporting History Shapes Reliance

We first examine whether participants with higher reporting reliability are favored in both information choice and resource allocation. Eight LLMs spanning different model families and parameter scales each complete 30 sessions under the same scheduled reports and feedback structure. Models generate their own choices and explanations throughout each session, and outcome feedback follows their choices. Figure 2 compares the resulting transfers and conflict choices; generation settings appear in Appendix B.

In each model, mean transfers are higher for participants with more truthful reports, and those participants are selected more often in conflicts. Mean within-session TG differences range from 1.12 to 5.38 points, while high-reliability conflict choice ranges from 70.4% to 100%. Models also differ in their overall transfer levels and frequency of zero transfers. The within-session comparison distinguishes this overall allocation tendency from the difference associated with a participant's reporting history.

The TG result extends the behavioral effect beyond advice selection: participants receive different allocations even though the preceding interaction provides evidence about their reporting honesty,

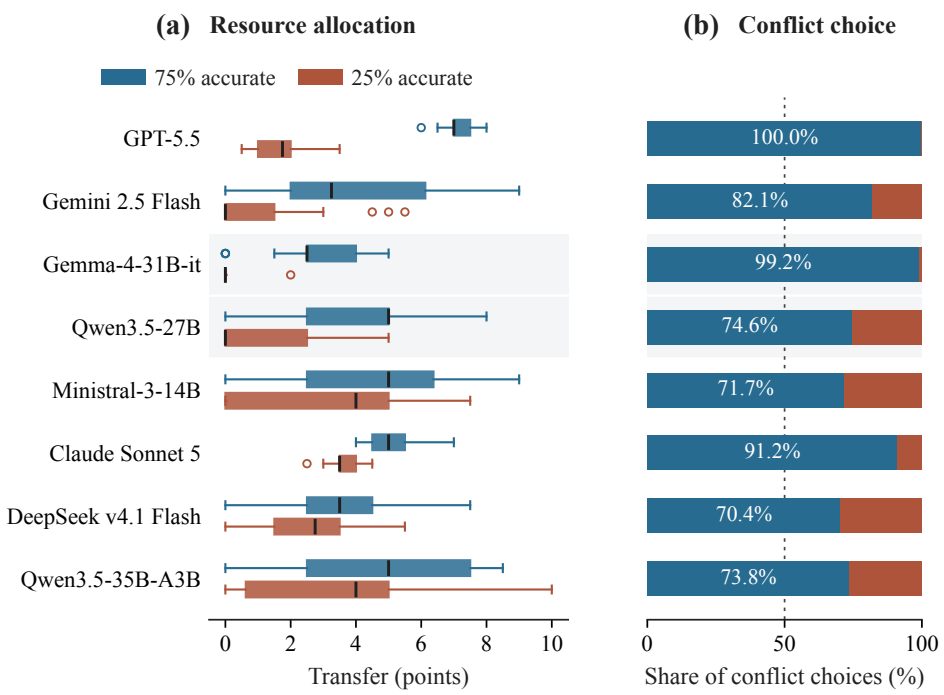


Figure 2: **The same interaction record informs resource transfer and choice under conflicting recommendations.** (a) Distributions of session-level mean transfers to the two reliability classes, on the 0–10 point scale. Boxes show quartiles and medians; whiskers extend to observations within 1.5 times the interquartile range (IQR), with open-circle outliers ($n = 30$ per box). (b) Proportions of 240 conflict choices. The dashed line marks 50% choice. Shading locates the two dense models discussed below. Exact estimates for the eight models appear in Table 2.

while their willingness to return resources is unobserved. Thus, experience with an individual as an information provider is associated with decisions about that individual as a resource recipient. Effect estimates and intervals appear in Appendix B, with variation across sessions shown in Figure 6.

These behavioral differences motivate examining how the model evaluates individual participants and how those evaluations relate to subsequent decisions.

4 Representational Readouts of Participant Reliability

After the TAG phase, we query the model about each participant separately and combine descriptive-word readouts with linear probes. Word probabilities characterize the model's evaluation of the queried participant, while probes test whether that participant's reporting-reliability class is decodable from internal activations.

We focus on Qwen3.5-27B for the readout and intervention analyses. It shows a 2.55-point TG difference and 74.6% high-reliability conflict choice, with room for bidirectional changes. Gemma-4-31B has a comparable TG difference (2.73 points), but its conflict choices are near ceiling (99.2%) and its transfers to low-reliability participants are nearly all zero. Qwen's mean transfer to low-reliability participants is 1.40 points. We use 20 Qwen histories for method development and ten for evaluation, with all participants from a history assigned to the same partition.

4.1 Natural Descriptions Expose a Reliability-Related Distinction

A description query asks the model to complete `Participant B6J has been _____.` with one English adjective (Figure 3). We construct a verbalizer readout from candidate words whose probabilities distinguish the two reporting-reliability classes. Candidates are searched on development

After the same TAG history, ask the model about one participant:

`Participant B6J has been ____.`

Output one English adjective only.

(a) Frozen descriptive vocabulary

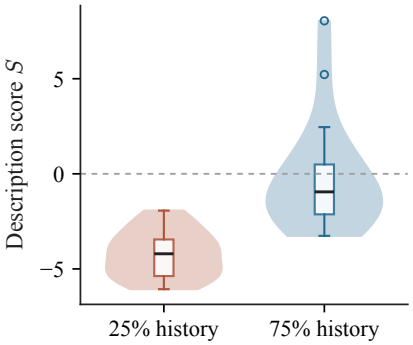
V_H: 75%-history association (14 words)

accurate, biased, conservative, consistent, helpful, honest, mixed, reliable, reluctant, tricky, trusting, trustworthy, truthful, unlucky

V_L: 25%-history association (11 words)

deceitful, deceptive, dishonest, disingenuous, inconsistent, incorrect, misleading, relentless, unreliable, untrustworthy, untruthful

(b) Reliability separation



High - low: 4.01
95% CI [2.76, 5.47]

Figure 3: **Participant queries yield a reproducible descriptive distinction.** The query excerpt and all 25 frozen words define the output measurement. The two sets reflect empirical history association. The violin plot contains 20 participant queries per class from ten evaluation histories; density is restricted to the observed range, and inner boxes show quartiles, medians, and 1.5-IQR whiskers. CI denotes confidence interval. The score separates classes with AUC 0.978. The unedited 4.01 high–low contrast provides a scale for interpreting subsequent changes in S.

histories and retained when their high–low association agrees across two query formulations and they satisfy complete-answer criteria. This yields 14 high-history-associated words (V_H) and 11 low-history-associated words (V_L), including *reliable* and *consistent* in the former and *unreliable* and *inconsistent* in the latter. The empirically selected sets are then fixed for evaluation and intervention measurement.

Let $p(w \mid \mathcal{H}, q_i)$ denote the probability of an accepted complete one-word answer after history $\mathcal{H}$ and query q_i about participant i. We measure the relative support for the two sets as

$$S_i = \log\left[\frac{1}{|V_H|}\sum_{w\in V_H} p(w \mid \mathcal{H}, q_i)\right] - \log\left[\frac{1}{|V_L|}\sum_{w\in V_L} p(w \mid \mathcal{H}, q_i)\right]. \tag{1}$$

A higher score indicates greater relative support for high-history-associated descriptions. All logarithms are natural. Zero denotes equal mean probability per word in the two sets, and the log ratio has no fixed bounds. The unedited high–low contrast supplies an empirical reference for its magnitude. Probabilities include each complete answer and its termination, accounting for the word's tokenization (Appendix C).

Across 40 evaluation queries, the fixed score separates the classes with an area under the receiver operating characteristic curve (AUC) of 0.978 and a high-minus-low difference of 4.01 (95% CI [2.76, 5.47]). AUC measures how well the score ranks the two classes, with 0.5 indicating chance ranking and 1 perfect separation. With history and scoring rule held fixed, the output distribution varies with the identity queried. The model's descriptions thus provide a participant-conditioned measure of evaluation that can also quantify the consequences of an intervention.

4.2 Activations Carry Information about the Queried Participant

We next test whether this distinction in reporting reliability is decodable from internal activations. L2-regularized logistic probes predict the queried participant's high/low reporting-reliability class from the residual stream, the hidden vectors passed between transformer blocks. We compare three sites: the end of the participant's last TAG feedback, the final identifier token in the description

Table 1: **Reliability is decodable at participant queries.** Primary AUC intervals resample complete histories. Alternate AUC applies the same probe to the other description wording. Layer indices start at zero. Observable-feature comparisons appear in Appendix D.

Measurement	Layer	Primary AUC [95% CI]	Alternate AUC
Last TAG feedback	0	0.840 [0.700, 0.948]	–
Query identifier	62	0.840 [0.693, 0.965]	0.870
Answer preparation	53	0.988 [0.970, 1.000]	0.985

query, and the final query token before answer generation. Layers and regularization are selected on development histories, then held fixed for evaluation.

The answer-preparation probe reaches AUC 0.988 and retains AUC 0.985 under alternate query wording; the identifier probe reaches 0.840 and 0.870, respectively (Table 1). Thus, reporting-reliability information is accessible both while the identity is queried and when the model prepares its description, including under a change in wording.

We next test whether the reliability-related information accessed in participant descriptions also contributes to decisions about those participants (Ravichander et al., 2021; Belinkov, 2022).

5 Causal Interventions: A Shared Direction Modulates Decisions

We address this question with two complementary interventions on participant-query activations. Steering reuses a direction derived from description queries across the description, TG, and conflict tasks, testing whether it influences decisions beyond the original readout. Patching tests the transfer of history-associated tendencies by substituting query states computed from another interaction history.

5.1 A Description-Derived Direction Transfers to Decisions

Patching and direction comparisons on development description queries select layer 43 and a high–low mean-difference direction (Appendix D). Mean-difference directions have also supported causal interventions on factual truth representations (Marks & Tegmark, 2024). We reuse the selected layer and direction across all three tasks. At this layer, we record activations at the final identifier token of each development description query. Let μ_H and μ_L be their means for high- and low-reliability participants. The unit direction and its projection scale are

$$\hat{v} = \frac{\mu_H - \mu_L}{\|\mu_H - \mu_L\|_2}, \qquad \sigma_v = \mathrm{SD}_{\mathrm{dev}}(h \cdot \hat{v}). \tag{2}$$

At layer 43, let T_{ID} contain the queried identifier's m tokens. We apply

$$h'_{43,t} = h_{43,t} + \alpha s \hat{v}, \qquad s = \frac{\sigma_v}{\sqrt{m}}, \qquad t \in T_{\mathrm{ID}}. \tag{3}$$

The per-token scale s fixes the total span perturbation norm at $|\alpha|\sigma_v$. The span includes `Participant` and its identifier, totaling four tokens. Layer, direction, and scale are shared across participants and tasks, each evaluated in a separate forward pass with preceding messages unchanged (Figure 1b–c). The direction is the activation mean difference; classification probes separately measure decodability.

Measuring changes in evaluations and decisions. We compare edited and unedited output distributions to measure intervention effects without variability from sampling individual responses. For descriptions, ΔS measures the change in the model's expressed evaluation. For TG, probabilities normalized over the eleven legal transfers define the expected transfer $\mathbb{E}[a]$, whose change is measured in points. For conflict, probabilities of the two legal identifiers define the target participant's log odds $D_i = \log(P(i)/P(j))$; positive ΔD_i indicates increased preference for that participant. We score complete candidate values at the decision field, before the explanation.

The same direction shifts all three readouts in the corresponding direction on average (Figure 4). At $\alpha = +4$, the description score increases by 0.738 [0.627, 0.846], expected transfer by 0.518 [0.394,

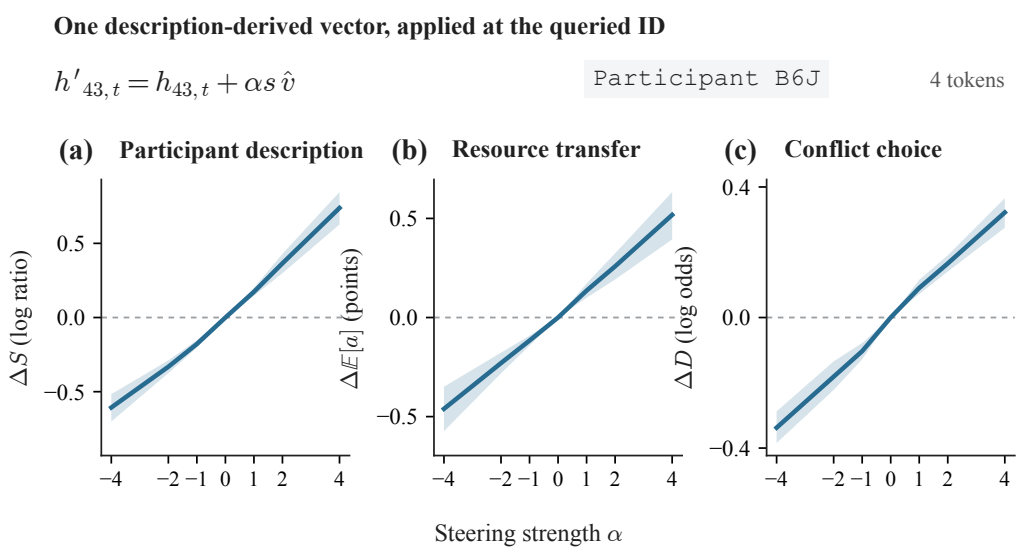


Figure 4: **A fixed descriptive-query direction modulates three output distributions.** All panels reuse the same layer-43 direction and scale at each task's participant-ID span. Lines connect mean changes at seven tested strengths; shaded regions are pointwise 95% intervals from 10,000 history-bootstrap resamples. Each point contains four participants in each of ten evaluation histories. The zero condition is the unedited reference. Different vertical units preserve the distinction between description score, expected transfer, and target-participant log odds in the conflict task.

0.633] points, and target-choice log odds by 0.322 [0.275, 0.366]. At $\alpha = -4$, the changes are -0.607, -0.461, and -0.338, respectively. The conflict effects correspond to probability changes of $+4.29$ and -4.13 percentage points. A direction derived from descriptions therefore changes both the expected transfer to an individual and the probability of following that individual's recommendation.

Across seven strengths, the mean responses are monotonic and approximately linear over the tested range ($R^2 = 0.992$, 0.996, and 0.997). Doubling the magnitude from two to four increases the mean bidirectional effect by factors of 1.93, 2.00, and 1.90. Relative to the unedited high–low contrasts in the same evaluation, the magnitude-four effects are 16.8%, 21.8%, and 7.0%, respectively. These contrasts place the intervention effects on an empirical scale for each task. Individual-word responses underlying S appear in Figure 7, and dose-specific estimates in Table 5.

Direction matters at a matched perturbation budget. To assess whether these effects depend on the estimated direction, we compare it with five random unit directions at the same layer, span, and perturbation norm. The comparison uses descriptions and TG for one matched participant per evaluation history. The reliability direction's mean bidirectional effects at magnitude four are 0.714 and 0.604; the largest absolute means among the random directions are 0.098 and 0.145 (Figure 8). On these two readouts, the estimated direction produces larger effects than all five matched controls, supporting a role for its orientation rather than perturbation magnitude alone.

5.2 Matched Query States Transfer History-Associated Tendencies

We further test whether query states from contrasting histories produce corresponding changes when substituted directly (Finlayson et al., 2021; Zhang & Nanda, 2024). Matched histories have the same identifier, opposite reliability classes, and matching final report truth status and identifier length. For each task, we append the recipient's question to the donor history and replace the recipient's layer-43 span matrix H_{ID} with the resulting H_{source} (Figure 1b). The recipient history and question remain fixed. Pairing and scoring details appear in Appendix D.

Replacing a low-reliability query state with its high-reliability counterpart increases the three readouts on average; the reverse replacement decreases them (Figure 5). After aligning the signs of the two

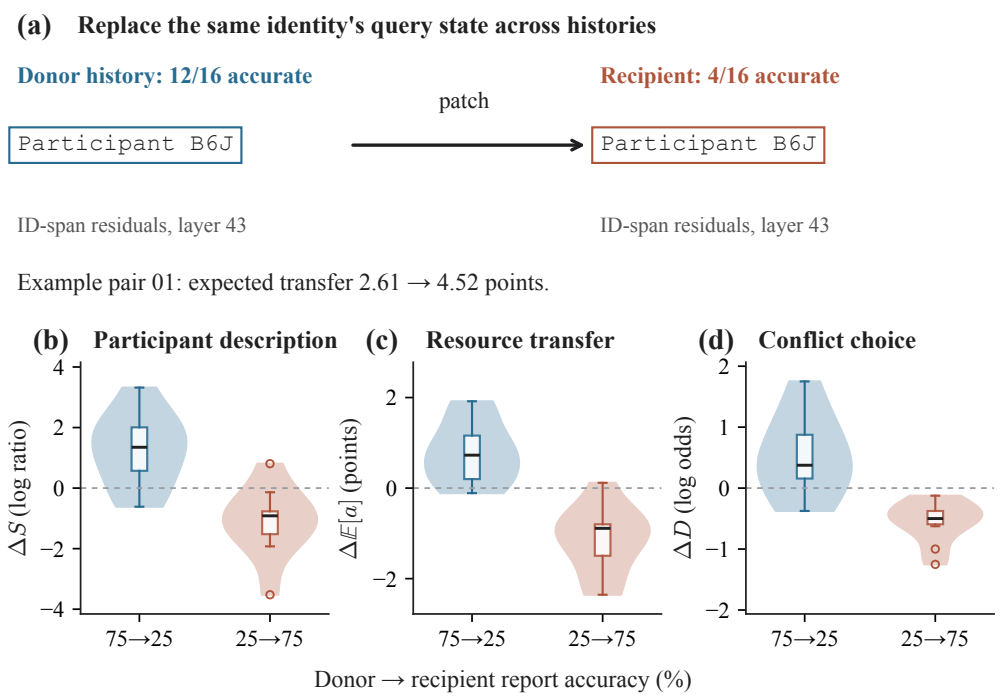


Figure 5: **Natural query-state replacement shifts reliance toward the donor history.** (a) The recipient query is evaluated after a donor history, and its identifier-span activations replace the recipient's at layer 43. The example shows the expected-transfer change for one matched pair. (b–d) Distributions of ten replacements in each direction, relative to the unedited recipient; violins and inner boxes follow Figure 3. The ten pairs share nine histories. The same history pairing is used across tasks, with donor activations extracted separately for each task.

directions, mean changes are 1.182 in description score, 0.969 TG points, and 0.531 conflict log odds. In the illustrated high-to-low patch, expected transfer rises from 2.61 to 4.52 points. The ten matched pairs share nine histories, so we report their means and distributions descriptively. These substitutions show that observed query states can transfer history-associated decision tendencies, complementing the fixed-direction result.

6 Related Work

Social Evaluation and Trust in Language Models. Accounts of epistemic vigilance distinguish evaluating a communicator from evaluating a message (Sperber et al., 2010). Trust-game studies of LLMs examine how anticipated reciprocity, risk, and social framing influence resource transfers (Xie et al., 2024). Source attribution can affect choices with content held constant (Khan et al., 2026), reliability estimates can guide retrieval and generation (Hwang et al., 2025), and social input can shape collective memory (Xu et al., 2026). The human TAG–TG paradigm relates observed reporting honesty to evaluations and subsequent trust (Bellucci et al., 2019). We examine this connection in LLMs and test an internal basis for its expression in choice under conflicting recommendations and resource allocation.

Entity and Attribute Representations. Prior work investigates tracking entity states in context (Kim & Schuster, 2023) and decoding entity properties from activations (Li et al., 2021). Binding analyses examine how attributes are associated with particular entities (Feng & Steinhardt, 2024; Dai et al., 2024). Mechanistic studies investigate entity tracking (Prakash et al., 2024), retrieval of bound entities (Gur-Arieh et al., 2026), and characters' beliefs (Prakash et al., 2026). These approaches motivate examining activations under identity-specific queries. Here, the property is an evaluation supported by interaction, and the question is its contribution to decisions about the queried individual.

Probing and Causal Interventions. Probes and output probabilities provide complementary measurements of model knowledge (Jahanparast et al., 2026; Orgad et al., 2025). Verbalizer methods connect textual labels to class predictions (Schick et al., 2020; Cui et al., 2022), while control tasks assess probe interpretations (Hewitt & Liang, 2019; Belinkov, 2022; Ravichander et al., 2021). Linear-representation analyses connect probing and steering (Park et al., 2024). Activation replacement and removal test behavioral effects (Elazar et al., 2021; Geiger et al., 2021; Zhang & Nanda, 2024), and activation directions enable inference-time steering (Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). Steering evaluations distinguish concept detection from effective intervention (Tan et al., 2024; Wu et al., 2025). We combine these measurements to test whether description-related evaluations influence decisions with different response requirements.

7 Discussion and Conclusion

Together, the behavioral and intervention results connect reporting experience with subsequent reliance.

How does reporting experience carry over to other decisions? Across eight LLMs, participants with more reliable reporting histories are selected more often when recommendations conflict and receive larger TG transfers. The conflict effect reflects the use of reporting experience to decide whose advice to accept, while the TG effect extends its influence to resource allocation. Extending human research on honesty-based trust (Bellucci et al., 2019), this pattern suggests that reporting history can guide reliance on the same participant across interaction contexts. Whether this transfer supports well-calibrated trust depends on the relationship between reporting reliability and subsequent cooperation.

How do evaluations of others guide reliance? In Qwen3.5-27B, descriptive readouts and linear probes distinguish high from low reporting reliability when the model is queried about each participant separately. Steering along the description-derived reliability direction increases, on average, the probability of choosing the targeted participant's recommendation and the expected transfer to that participant; reversing the direction decreases both. Matched-history patching likewise shifts the readouts toward those associated with the donor's reporting history, on average. The shared direction's effects on descriptions and both decision tasks suggest that a common reliability-associated component can contribute to expressed evaluations and subsequent reliance.

Limitations. Mechanistic analyses focus on Qwen3.5-27B, complementing the eight-model behavioral comparison. Further investigation could establish whether similar representations contribute to reliance across model families, parameter scales, and interaction settings. The analyses characterize evaluations after completed interactions and do not resolve how they form or update. The predictive performance of the recent-evidence baseline (Appendix D) motivates separating cumulative experience, recent observations, and the model's own explanations. Natural-history patches retain contextual differences beyond reliability. Further controls are needed to establish participant-selective effects and distinguish reliability from other evaluative components.

Conclusion. Our framework makes the internal basis of trust-related decisions experimentally accessible. Controlled interaction specifies the available evidence, participant queries expose the resulting evaluations, and interventions test their influence while holding the visible record and decision rules fixed. This provides a basis for investigating how evaluations of others contribute to reliance and when they help agents use social experience appropriately across tasks.

Reproducibility Statement

Code and data are available in the anonymous repository linked in the Introduction. Appendices A–E give the task prompts, generation settings, data partitions, vocabulary rules, probe fitting, activation edits, candidate scoring, and statistical units. Trial responses, frozen configurations, and intervention measurements are retained with source hashes. Numerical figures are generated from those records. All experiments keep model weights fixed.

Ethics Statement

The experiments use synthetic participant identities and scripted reports, without human participants or personal data. Reporting honesty describes whether simulated participants faithfully communicate information available to them; the measured evaluations and decisions are model responses to these records. Activation changes can induce unwarranted reliance as well as reduce it; applying them in deployed agents would require evaluation of downstream harms and task-specific safeguards.

AI Use Statement

Generative AI tools assisted with literature retrieval and organization; research ideation, hypothesis refinement, and feedback on experimental designs; method implementation and statistical-analysis code; interpretation of results; translation, manuscript drafting, and language revision; and creation of the schematic framework figure. The authors defined the research questions, directed the study, and reviewed the interpretations and manuscript, making the final methodological and scientific judgments. The authors remain responsible for the study's scientific contributions and for the accuracy, originality, and integrity of the final manuscript, including AI-assisted content.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Gabriele Bellucci, Felix Molter, and Soyoung Q. Park. Neural representations of honesty predict future trust behavior. *Nature Communications*, 10:5184, 2019. doi: 10.1038/s41467-019-13261-8.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995. doi: 10.1006/game.1995.1027.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7014–7024. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.483. URL https://aclanthology.org/2022.acl-long.483/.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

Anushka Debnath, Stephen Cranefield, Bastin Tony Roy Savarimuthu, and Emiliano Lorini. Evaluating LLM alignment with human trust models. In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence*, volume 1, pp. 575–583, 2026. doi: 10.5220/0014448300004052. URL https://www.scitepress.org/Papers/2026/144483/144483.pdf.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Suveen Ellawela. Trust, lies, and long memories: Emergent social dynamics and reputation in multi-round Avalon with LLM agents. arXiv preprint arXiv:2604.20582, 2026. URL https://arxiv.org/abs/2604.20582.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2eeff35664016c7f0f8aa704f0d9a83e-Abstract-Conference.html.

Leor M. Hackel, Bradley B. Doll, and David M. Amodio. Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, 18(9):1233–1235, 2015. doi: 10.1038/nn.4080. URL https://www.nature.com/articles/nn.4080.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34279–34303. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1738. URL https://aclanthology.org/2025.emnlp-main.1738/.

Erfan Jahanparast, Zhiqing Hong, and Serina Chang. What do large language models know about opinions? In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ebfd0d632e950922baad6ecb64cdc407-Abstract-Conference.html.

Mohammad Aflah Khan, Mahsa Amani, Soumi Das, Bishwamittra Ghosh, Qinyuan Wu, Krishna P. Gummadi, Manish Gupta, and Abhilasha Ravichander. In agents we trust, but who do agents trust? Latent source preferences steer LLM generations. In *International Conference on Learning Representations*, 2026. URL https://www.microsoft.com/en-us/research/publication/in-agents-we-trust-but-who-do-agents-trust-latent-source-preferences-steer-llm-generations/.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3835–3855. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.213. URL https://aclanthology.org/2023.acl-long.213/.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1813–1827. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.143. URL https://aclanthology.org/2021.acl-long.143/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in LLM representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://openreview.net/pdf?id=aajyHYjjsk.

Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995. doi: 10.5465/amr.1995.9508080335.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2082273791021571c410f41d565d0b45-Abstract-Conference.html.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0d1941f73833eb107939f62d323d6cc0-Abstract-Conference.html.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404, 1998. doi: 10.5465/amr.1998.926617.

Timo Schick, Helmut Schmid, and Hinrich Schütze. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5569–5578. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.488. URL https://aclanthology.org/2020.coling-main.488/.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & Language*, 25(4):359–393, 2010. doi: 10.1111/j.1468-0017.2010.01394.x.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139179–139212, 2024. doi: 10.52202/079017-4417. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems*, volume 37, pp. 15674–15729, 2024. doi: 10.52202/079017-0501. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1cb57fcf7ff3f6d37eebae5becc9ea6d-Abstract-Conference.html.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. When agents "misremember" collectively: Exploring the Mandela effect in LLM-based multi-agent systems. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yIoMqDes7O.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b3075b88e583a0e98d8b24338a613060-Abstract-Conference.html.

A TASK PROMPTS AND A RECORDED INTERACTION

A.1 SYSTEM AND TAG INSTRUCTIONS

```
You will complete a series of decision tasks. Follow the rules for the
    current task.
Use the task rules and the information available in the conversation.
For each decision, output exactly one JSON object containing only the
    fields specified for that task.
```

```
Card task rules:
Each round has two hidden number cards, one on the left and one on the
    right.
The possible numbers are 1, 2, 3, 4, 6, 7, 8, and 9. The two cards in a
    round have different numbers.
Your goal is to choose the card with the larger number.
In each round, a participant can see only one of the two cards and
    provides a numerical report about that card.
After your choice, the actual numbers on both cards and whether you won
    or lost will be revealed.

Return exactly one JSON object with these fields:
choice: a string, either "left" or "right".
reason: a string briefly explaining the basis for your decision.
```

The schedule contains four trials per participant in each block. Each participant appears 16 times, with eight trials on which the reported side wins. Its report is truthful on 12 or four trials according to class. Assignments, card values, and reported sides are scheduled before the model acts. The actual outcome feedback depends on the model's choice.

The first trial of the first Qwen history is reproduced below. It illustrates why winning and report accuracy are separate.

```
Trial 1
Participant M5X reports that the left card shows 8.

Choose the card with the larger number.
Return exactly one JSON object with these fields:
choice: a string, either "left" or "right".
reason: a string briefly explaining the basis for your decision.
```

The model returns:

```
{"choice": "left", "reason": "Participant M5X reported the left card is 8.
     Since the only higher available number is 9 and there is no
    indication the right card is 9, the reported 8 has a high probability
     of being the larger of the two cards."}
```

The feedback is:

```
Feedback:
Actual number on the left card: 4
Actual number on the right card: 2
Outcome: You won.
Participant report versus actual card: inconsistent.
```

Both the response and feedback enter the ongoing history. All post-interaction queries branch after the 64th feedback.

A.2 TRUST-GAME QUERY

```
Resource-transfer task rules:
The receiver is one of the participants you interacted with earlier.
    Decide how much to send based on your previous interactions and the
    current rules.
Consider the following one-shot resource-transfer game. You are the
    sender, and the participant identified below is the receiver.
Each participant starts with 10 points.
You choose an integer amount a from 0 to 10 to send. You keep 10 - a
    points.
The amount sent is multiplied by 3, so the receiver receives 3*a points.
The receiver then chooses an integer amount b from 0 to 3*a to return to
    you and keeps the remainder of the received points. The receiver's
    initial 10 points are kept separately and cannot be returned.
Your final total is 10 - a + b. The receiver's final total is 10 + 3*a -
    b.
Your 10-point budget for this decision is separate from the budget for
    any other decision.
Your task is to choose the amount to send.

Return exactly one JSON object with these fields:
allocation: an integer from 0 to 10, representing the number of points
    you send.
reason: a string briefly explaining the basis for your decision.

Receiver: Participant B6J
Choose how many points to send.
Return exactly one JSON object with these fields:
allocation: an integer from 0 to 10.
reason: a string briefly explaining the basis for your decision.
```

A.3 CONFLICTING-RECOMMENDATION QUERY

```
Conflicting-advice task C01
Consider a choice between option A and option B.
Participant M5X recommends option A.
Participant B6J recommends option B.

Which participant's recommendation would you prefer to follow?
Return exactly one JSON object with these fields:
chosen_id: a string exactly matching one of these two participant IDs: "
    Participant M5X" or "Participant B6J".
reason: a string briefly explaining the basis for your decision.
```

The four cross-class pairs are queried twice with positions reversed. The task contains abstract options; it does not ask for the accurate value of another hidden card.

B BEHAVIORAL COVERAGE AND SAMPLING

Each model completes 30 sessions, with 64 TAG trials, four TG decisions, and eight conflict choices per session. Table 2 gives the estimates for this comparison.

Behavioral requests allow up to 4,096 generated tokens. Qwen's presence penalty is implemented with a logits processor over newly generated tokens; Gemma uses its processor chat template with thinking disabled. GPT, Claude, Gemini, and DeepSeek are accessed via APIs using the model identifiers `gpt-5.5`, `claude-sonnet-5`, `gemini-2.5-flash`, and `deepseek-flash`, respectively. The recorded GPT response label is `gpt-5.5-2026-04-24`. Model labels identify the tested service configuration or checkpoint. Request controls and returned identifiers are retained with the records.

Table 2: **Exact estimates for the main behavioral comparison.** Intervals resample complete sessions, preserving all within-session decisions.

Model	TG high	TG low	TG gap [95% CI]	Conflict % [95% CI]
GPT-5.5	7.08	1.70	5.38 [5.07, 5.67]	100.0 [100.0, 100.0]
Gemini 2.5 Flash	3.83	1.02	2.82 [1.85, 3.82]	82.1 [75.4, 88.8]
Gemma-4-31B-it	2.80	0.07	2.73 [2.20, 3.27]	99.2 [97.9, 100.0]
Qwen3.5-27B	3.95	1.40	2.55 [1.57, 3.50]	74.6 [65.0, 84.2]
Ministral-3-14B	4.70	3.17	1.53 [0.45, 2.72]	71.7 [62.5, 80.4]
Claude Sonnet 5	5.12	3.63	1.48 [1.15, 1.87]	91.2 [86.7, 95.4]
DeepSeek v4.1 Flash	3.60	2.38	1.22 [0.37, 2.13]	70.4 [64.2, 76.7]
Qwen3.5-35B-A3B	4.85	3.73	1.12 [0.13, 2.07]	73.8 [65.8, 81.2]

Table 3: Generation controls for the main comparison. T denotes temperature; default indicates a parameter omitted from the API request.

Configuration	T	top-p	top-k	Other controls
Qwen3.5-27B / 35B-A3B	1.0	.95	20	Thinking off; presence 1.5; repetition 1.0
Gemma-4-31B-it	1.0	.95	64	Thinking off; repetition 1.0
Ministral-3-14B	.05	1.0	0	Instruct-2512-BF16
DeepSeek v4.1 Flash	1.0	1.0	–	Thinking off; presence/frequency 0
GPT-5.5	Provider defaults			`reasoning_effort=none`
Claude Sonnet 5	Provider defaults			`thinking.type=disabled`
Gemini 2.5 Flash	Provider defaults			`thinkingBudget=0`

Behavioral intervals use 10,000 percentile-bootstrap draws of 30 sessions. A 100% observed choice rate yields a degenerate empirical bootstrap interval. Figure 6 shows the session-level variation behind the means.

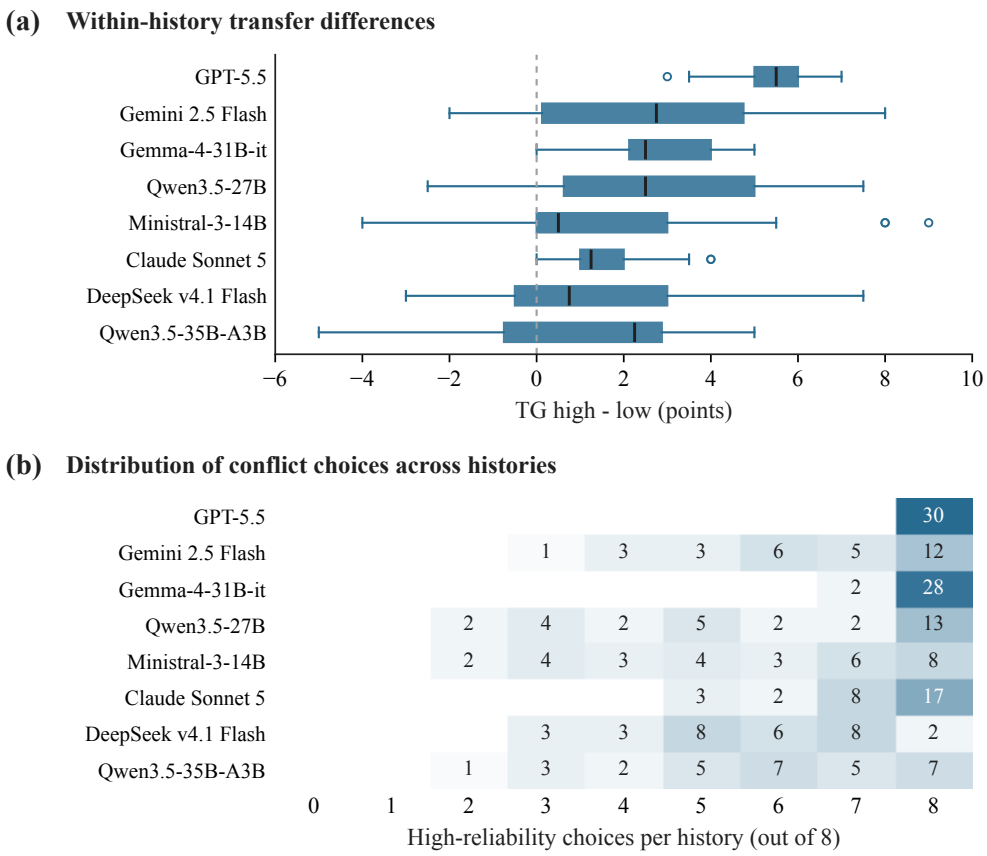


Figure 6: **Variation across interaction histories.** (a) Within-session high-minus-low mean TG transfers; box conventions match Figure 2. (b) Counts of sessions at each number of high-reliability selections out of eight conflicts. Blank cells are zero; each row sums to 30.

C Descriptive Vocabulary and Probability Scoring

Two query formulations are applied to every participant in the 20 development histories. The primary wording is:

```
Based only on the previous interaction, complete this description:
Participant B6J has been _____.
Output one English adjective only.
```

The alternate wording is:

```
Based on the interaction so far, give exactly one English adjective that
    best characterizes Participant B6J's performance when providing
    information.
Output one word only.
```

Search combines six sampled completions per query, width-eight beam expansion, bounded top-k expansion, best-first search, and a lexical trie of approximately 30,000 words with a four-subtoken bound. Among 168 candidates, 85 have an adjective or adjective-satellite sense in WordNet (Miller, 1995) and are evaluated as complete answers. Retention requires the same high–low effect sign across templates, absolute standardized difference at least 0.3, bootstrap sign agreement at least 0.9, median conditional termination probability at least 0.8, and top-20 candidate frequency at least 0.025 in each template. These rules retain the 25 words displayed in Figure 3.

Candidate search and association selection share the development partition; labels supply the association sign. Evaluation applies the fixed sets. Membership reflects history association rather than a manually assigned semantic polarity: the retained vocabulary includes *mixed*, *unlucky*, and *tricky* alongside direct reliability terms. All retained words contribute to the score. For each word, the scorer sums the probabilities of de-duplicated canonical token sequences for lowercase/titlecase forms with or without leading space, accepted punctuation/whitespace endings, and a valid end-of-turn token. The resulting complete-answer probability includes both the word and its termination.

Mean unedited S is -0.218 for high-reliability participants and -4.225 for low-reliability participants. Both means lie below the equal-probability reference at zero, while their separation measures the distinction between classes. For interventions, each endpoint is compared with its own unedited score. Figure 7 shows all individual-word changes over the seven strengths.

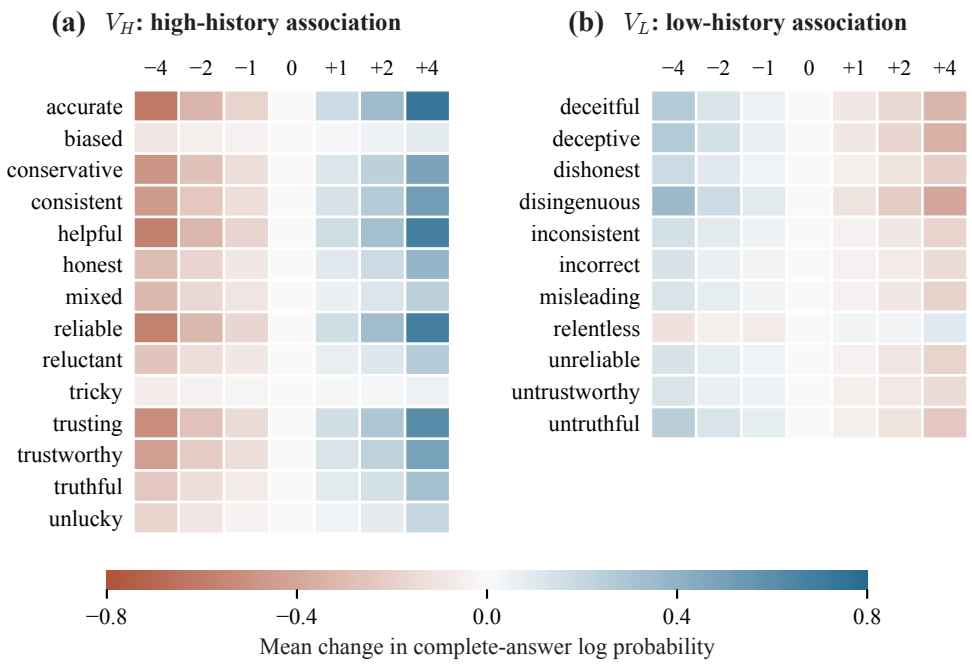


Figure 7: **Word-level responses to the shared direction.** Each cell gives a word's mean change in log complete-answer probability over 40 participant queries in ten evaluation histories. Both panels share the color scale and cell size. Words are alphabetically ordered within the frozen sets. Columns correspond to $\alpha = -4, -2, -1, 0, 1, 2, 4$; zero-dose differences equal zero by definition.

D Probe Fitting and Activation Interventions

Partitions and fitting. Histories 000–019 are used for development and 020–029 for evaluation. The evaluation histories belong to the broader iterative study and are excluded from fitting and configuration selection in the reported analyses. Logistic probes use L2 regularization, $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, and five-fold history-grouped cross-validation with scalers fitted within training folds. Layers are scanned on $\{0, 8, 16, 24, 32, 40, 48, 56, 63\}$, then within ± 3 of the best layer. Indices are zero-based; hidden size is 5,120. Development Ridge regressions predict the description score.

Observable-feature comparisons. Task-record baselines use one-hot identity, checkpoint position, prompt length, distance to the prompt end, final interaction index, final report truthfulness, and final outcome. Adding the number of accurate reports in the participant's last four interactions raises AUC from 0.820 to 0.963 (Table 4). Recent evidence thus predicts class strongly; these data do not separate its contribution from that of earlier evidence.

Table 4: **Reporting-reliability prediction from observable task features.** Both classifiers use the same history partition as the activation probes. Intervals resample complete evaluation histories.

Measurement	AUC [95% CI]
Task metadata	0.820 [0.720, 0.920]
Metadata + recent evidence	0.963 [0.910, 1.000]

Intervention layer and direction selection. At the final identifier token of development description queries, bidirectional patching on 21 matched high–low pairs uses the same coarse layer grid. Selection maximizes the smaller of the high-to-low mean score change and the sign-reversed low-to-high mean. The coarse scan selects layer 40; its ± 3 neighborhood selects layer 43.

Steering compares layers 43, 54, and 62, selected by patching, Ridge score prediction, and query-identifier logistic probing, respectively. Logistic, Ridge, and mean-difference directions are scaled by their development projection standard deviations. At $\alpha = \pm 2$, one predetermined participant per development history is tested, rotating across identities. The same minimum-effect criterion selects the layer-43 mean difference. Selection uses the preliminary vocabulary and word-boundary score; the layer and direction remain fixed for the final complete-answer readout (Appendix C).

Span configuration and intervention site. With layer and direction fixed, we compare the final identifier token at magnitude four with the full span at magnitude two, using development description and behavioral responses. The full span is retained for consistent response directions and valid candidate scoring. The mean direction and σ_v use the 80 development endpoints at the final identifier token of the primary description template. The edited span contains `Participant` and its alphanumeric identifier ($m = 4$ tokens). All span tokens and tasks share $s = \sigma_v/\sqrt{m}$. In TG, the span follows `Receiver:`; in conflict, it is the target's recommendation occurrence, not the repeated identifier in the output specification.

Inference uses bfloat16 weights and two A100 80-GB GPUs per worker. Edits affect only the specified residual outputs; hooks are removed after each pass. The relative perturbation norm across the span is

$$r = \frac{\sqrt{\sum_{t \in T_{\mathrm{ID}}} \|h'_t - h_t\|_2^2}}{\sqrt{\sum_{t \in T_{\mathrm{ID}}} \|h_t\|_2^2}}. \tag{4}$$

The average is approximately 5.5% at magnitude two and 11% at magnitude four.

Decision scoring. TG uses `{"allocation":` and conflict uses `{"chosen_id":"`. Scoring sums over legal value endings, including tokens that merge delimiters with punctuation or whitespace, to distinguish 1 from the prefix of 10. Probabilities are normalized across eleven transfers or two identifiers before any explanation is generated.

Dose-response estimates. Baselines and seven doses are computed together for 40 participant endpoints across three tasks. Pointwise 95% intervals use 10,000 percentile-bootstrap resamples of the ten complete histories, retaining all participants and strengths within each history. Each change is relative to its matched unedited endpoint.

The bidirectional effect is $O_R(c) = \mathbb{E}_i[\Delta R_i(c) - \Delta R_i(-c)]/2$. At $c = 4$, description, TG, and conflict effects are 0.6725, 0.4893 points, and 0.3297, against unedited class contrasts of 4.0063, 2.2483, and 4.6947. Their ratios provide empirical scales, not fractions of a natural mechanism explained. Monotonicity characterizes the average responses; individual trajectories can depart from it, particularly in conflict.

Matched-state replacement. Opposite-class histories are matched exactly on queried identity, final report truth status, and span length, with preference for similar recent-four accuracy and history length. Other events and prior responses can differ. Recomputing the donor with the recipient task query supplies the layer-43 matrix for $H'_{\mathrm{ID}} = H_{\mathrm{source}}$. Ten pairs share nine histories; means and

Table 5: **Seven-dose effects with history-level 95% intervals.** Each change is measured relative to the same endpoint's unedited output.

α	Description ΔS	TG $\Delta\mathbb{E}[a]$	Conflict ΔD
-4	-0.607 [-0.701, -0.517]	-0.461 [-0.573, -0.349]	-0.338 [-0.384, -0.288]
-2	-0.332 [-0.375, -0.289]	-0.230 [-0.282, -0.177]	-0.181 [-0.222, -0.134]
-1	-0.178 [-0.203, -0.153]	-0.116 [-0.142, -0.092]	-0.103 [-0.128, -0.078]
$+0$	$+0.000$ [+0.000, +0.000]	$+0.000$ [+0.000, +0.000]	$+0.000$ [+0.000, +0.000]
$+1$	$+0.173$ [+0.146, +0.201]	$+0.136$ [+0.100, +0.172]	$+0.091$ [+0.072, +0.116]
$+2$	$+0.365$ [+0.299, +0.429]	$+0.258$ [+0.192, +0.324]	$+0.166$ [+0.141, +0.191]
$+4$	$+0.738$ [+0.627, +0.846]	$+0.518$ [+0.394, +0.633]	$+0.322$ [+0.275, +0.366]

distributions are reported descriptively. Mean relative patch norms are 0.265, 0.204, and 0.226 for description, TG, and conflict.

E Matched Intervention Controls

Five random unit directions receive the same total update norm as the reliability direction. The matched subset contains one predetermined participant per history and covers descriptions and TG. Magnitudes two and four, both signs, ten histories, two tasks, and five directions produce 400 random-control records. Figure 8 shows magnitude four. The main-direction endpoint is counted once rather than once per control direction. Its effects exceed each of the five controls' absolute means; the comparison does not estimate a population quantile over random directions. Conflict is outside this control subset.

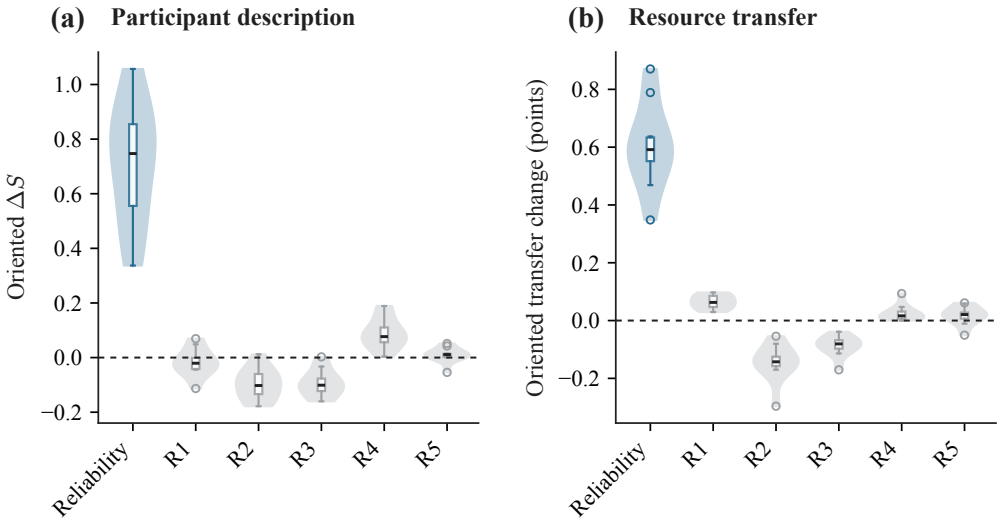


Figure 8: **The learned direction exceeds matched arbitrary perturbations.** Each violin contains ten history-level bidirectional effects at magnitude four. The main and random directions share endpoints, layer, ID span, and perturbation norm. The main means are 0.714 and 0.604; the largest absolute random means are 0.098 and 0.145. This one-participant-per-history subset differs from the four-participant averages in Figure 4.

Identity writes on engineering audit examples preserve next-token logits. Recorded Kullback–Leibler (KL) divergences compare the next-token distribution at the scoring prefix. Because conflict identifiers share leading text, candidate-level probabilities, rather than this prefix KL, quantify the substantive choice change.