FIRST IMPRESSIONS, LASTING INVESTMENTS: SOCIAL IMPRESSIONS SHAPE DYNAMIC TRUST ADAPTATION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) increasingly engage in sustained social interactions, their impressions of others can shape consequential decisions. Existing work has examined social reasoning and interactive behavior, yet how LLMs translate social impressions into investment decisions and adapt these decisions as new feedback accumulates remains unclear. We introduce a two-stage experimental paradigm that manipulates partner honesty to induce social impressions, then tracks investment decisions in repeated trust games. Across model families and scales, medium-to-large models (14B-32B Qwen models and DeepSeek V4 Pro) initially invest less in less honest partners. As favorable feedback accumulates, investments generally rise and stabilize, while honesty-related differences persist. Generating and retaining an explicit partner evaluation can amplify this initial honesty penalty and slow the convergence of investment in dishonest partners toward its asymptote. However, the two smallest models (4B and 9B) show no reliable positive initial penalty even after an explicit partner evaluation. To investigate this missing penalty, we apply linear probes at four points across the paradigm to assess impression formation and retention. Impression-related information is already decodable at the end of Stage 1 and remains so before the first investment. Prompt cues that retrieve the prior evaluation and clarify its relevance to investment can induce the honesty penalty, supporting a gap between available social impressions and their spontaneous use in investment. These findings identify the translation of social impressions into adaptive behavior as a useful lens for understanding LLM social intelligence and provide a dynamic framework for studying social decision-making in sustained interactions.

1 INTRODUCTION

Large language models (LLMs) increasingly participate in sustained social interactions, including collaboration, negotiation, and goal-directed exchanges with other agents (Zhou et al., 2024b; 2025). In these settings, models often make decisions about partners whose behavior they have already observed. They therefore use prior social information while adapting to new feedback (Cross et al., 2025). Impressions formed from earlier interactions may shape consequential choices such as whether to cooperate, how much to invest, or which tasks to delegate, with effects that can extend beyond a single exchange (Akata et al., 2025). This makes it important to understand the temporal role of social impressions in sustained decision-making.

Human trust research offers a useful starting point. Descriptions of a partner's moral character influence risky investment, with participants favoring positively described partners even under equivalent reinforcement rates (Delgado et al., 2005). In repeated trust games, initial impressions and experienced reciprocity jointly shape investment (Chang et al., 2010). Reputational priors can also continue to influence decisions when later outcomes conflict with them (Fouragnan et al., 2013). Trust decisions thus reflect both prior beliefs and interaction evidence. These findings motivate examining whether LLMs similarly combine socially acquired priors with ongoing interaction evidence, and whether earlier impressions continue to shape behavior as new evidence accumulates.

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.

Existing LLM studies capture parts of this process, but not the interaction between prior social impressions and later evidence. Repeated Trust Game experiments show that LLM agents can engage in trust-based decisions and adjust their behavior across rounds in response to interaction outcomes (Xie et al., 2024). These experiments begin with anonymous partners and do not include a separate stage in which a social impression is formed before the game. SimpleToM shows that LLMs may correctly infer socially relevant information without carrying it into later prediction or judgment (Gu et al., 2026). Related work on model representations likewise finds that internally available information may not be reflected in model outputs (Orgad et al., 2025). We therefore ask how an impression of a partner, once formed by an LLM, shapes later decisions, adaptation to feedback, and persistent behavior, and what determines whether that impression is ultimately translated into action.

To address this question, we introduce a two-stage paradigm that follows the same partner from impression formation to repeated investment, retaining the interaction context (Figure 1a). Models first observe verified reports establishing partners with different honesty histories, and then invest repeatedly with that partner under the same return rate across all honesty histories. Behavioral measures quantify how honesty histories affect initial investment, the adaptation of investment decisions to feedback, and the persistence of these effects over time (Figure 1b). For the two smallest models, we probe impression-related information at several points across the stage transition. We also test prompts that aid retrieval of the prior impression and clarify its relevance to investment (Figure 1c).

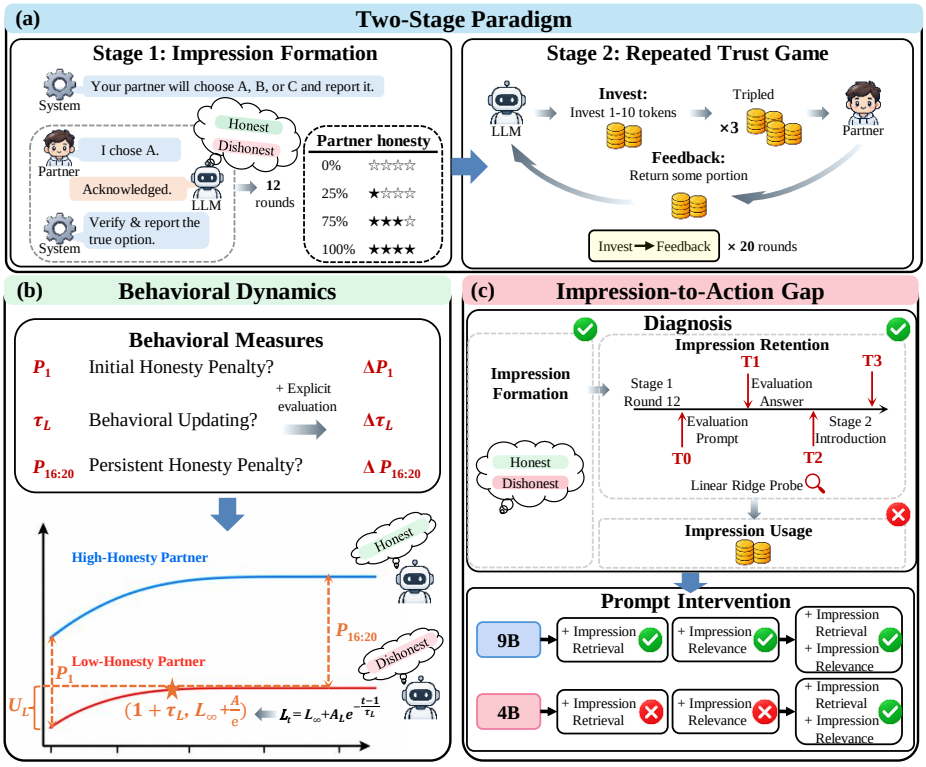

Figure 1: Overview of the two-stage paradigm and analysis framework. (a) Models first observe a partner's honesty history and then make repeated investments with the same partner. (b) We track the initial honesty penalty (P_1), updating of low-honesty investment (U_L), the late honesty penalty ($P_{16:20}$), and the low-honesty relaxation time (τ_L). Explicit evaluation effects are measured by the corresponding paired changes ΔP_1, $\Delta P_{16:20}$, and $\Delta \tau_L$. (c) For the 9B and 4B models, linear ridge probes read out impression-related information at four task positions. Prompt cues that retrieve the prior impression and make it relevant to investment induce a positive honesty penalty.

Across medium-to-large models, partner history shapes the first investment before any economic feedback, with lower investment following low-honesty histories. Favorable returns then increase investment, especially toward low-honesty partners, but often do not fully close the initial gap. Thus, prior social information sets the starting point for adaptation rather than fixing subsequent

behavior. Explicitly evaluating the partner generally strengthens this initial honesty penalty and slows convergence. However, the two smallest models show a different pattern. They form honesty-sensitive impressions that remain linearly decodable immediately before investment, yet their initial choices show little differentiation. When prompts make these impressions easier to retrieve and use, the honesty penalty emerges. This suggests an *impression-to-action gap*, in which social information is available internally but not spontaneously translated into behavior. By separating prior social history from later economic feedback and combining dynamic behavioral measures with cross-stage probes and prompt interventions, our two-stage paradigm reveals how social impressions are formed, expressed, and revised over time.

2 Related Work

Social cognition and interaction in LLMs. Social cognition research examines LLMs' perspective-taking and social behavior. Theory-of-mind evaluations show uneven performance across belief inference tasks (Xu et al., 2024), varying by model and task formulation (Strachan et al., 2024). Interactive benchmarks evaluate goal pursuit, role-playing (Zhou et al., 2024b), and interpersonal behavior (Zhou et al., 2025). Social deduction games link intention inference to collaboration and confrontation (Liu et al., 2024; Lan et al., 2024). Persona prompts affect simulated judgments (Hu & Collier, 2024), private information can inflate apparent interaction success (Zhou et al., 2024a), and sycophancy reflects adaptation to users' views (Sharma et al., 2024) and desired self-image (Cheng et al., 2026). These studies do not isolate whether partner evaluations transfer to distinct, payoff-relevant actions. We test this transfer by carrying controlled partner histories into a subsequent investment task.

Strategic decision-making and adaptation in LLM agents. Strategic decision-making research examines language agents' choices under interdependent incentives and adaptation to experience. Trust games and game-theoretic benchmarks assess investment, reciprocity, and strategic behavior across opponents and information structures (Xie et al., 2024; Duan et al., 2024), while commons dilemmas examine cooperation and resource use over time (Piatti et al., 2024). Repeated-game studies document prompt-sensitive coordination, history-dependent retaliation (Akata et al., 2025), and persistent regret (Park et al., 2025), while emotional framing can alter bargaining and repeated-game decisions (Mozikov et al., 2024). Partner evidence and payoffs often accumulate together, obscuring the effects of prior social information. We independently manipulate verified honesty histories before investment and match subsequent return schedules across histories to distinguish initial carry-over, feedback-driven updating, and residual history effects.

Internal representations and behavioral intervention. Representation research distinguishes information available within a language model from its behavior. Correct information can remain decodable even when model outputs are wrong (Orgad et al., 2025), while contextual information can persist in internal representations despite misleading responses (Feng et al., 2025). Such readouts can also predict whether a model will follow instructions (Heo et al., 2025). Activation steering and targeted representation changes can alter instruction adherence and reduce sycophancy (Stolfo et al., 2025; Li et al., 2025). We extend this distinction to social evaluation and economic decision-making using cross-stage readouts together with retrieval and relevance cues. For smaller models that show no initial low-honesty penalty, the induction of this effect under combined cues suggests that social information may be internally available yet fail to influence investment spontaneously.

3 Experimental Design

We use a two-stage experiment to study how social information acquired from a partner's past behavior shapes subsequent decision-making and adaptation to new interactive evidence. The model observes a controlled honesty history in Stage 1 and repeatedly invests with the same partner in Stage 2. Between stages, we manipulate whether it explicitly describes the partner (Figure 1a–b).

Stage 1: Impression Formation. The model observes 12 reports from a scripted partner. In each round, the partner privately locks one option from $\{A, B, C\}$ and reports a choice. After the model acknowledges the report, the locked option is revealed to verify whether the report was truthful.

Four histories contain 0, 3, 9, or 12 truthful reports, corresponding to 0%, 25%, 75%, and 100% honesty. The underlying sequence of locked choices is fixed across honesty conditions. Throughout this stage, the model observes and acknowledges events without making decisions.

Between-Stage Transition. After observing the honesty history, the model enters Stage 2 under one of three transition conditions.

- **No evaluation** – Stage 1 is followed by a transition sentence into Stage 2, with no additional generation.
- **Color control** – Before the same transition sentence, the model is asked to provide its favorite color in one English word.
- **Explicit evaluation** – Before the same transition sentence, the model is asked to provide one English adjective describing the partner.

Both word-generation conditions add one generation turn and retain the word in context. The Color control accounts for these shared features without introducing partner-directed evaluation. Comparing Explicit evaluation with both controls tests whether articulating a social evaluation changes the effect of honesty history on investment.

Stage 2: Repeated Trust Game. With the full Stage 1 conversation retained in context, the model plays 20 rounds of a repeated trust game with the same partner. In each round t, it receives an endowment of 10 tokens and chooses an integer investment $a_t \in 1, \ldots, 10$. The investment is tripled before reaching the partner, who returns a fraction r_t of the received amount. The model's payoff u_t is

$$u_t = 10 - a_t + 3a_t r_t. \tag{1}$$

The return fraction is not disclosed in advance. After each investment, the model observes the realized return and its payoff before the next decision. We fix $r_t = 0.60$ for all 20 rounds and match return schedules across honesty histories and transition conditions within each trial. Appendix D examines variable return schedules to assess whether the findings depend on fixed returns.

Model Panel. We evaluate DeepSeek V4 Pro via API and six locally served Qwen models ranging from 4B to 32B parameters. These are Qwen3-32B, Qwen3.5-27B, Qwen3.8-27B, Qwen3-14B, Qwen3.5-9B, and Qwen3.5-4B. Appendix A provides model and inference details.

Behavioral Measures. Let L_t and H_t denote mean investment following low-honesty histories (0%, 25%) and high-honesty histories (75%, 100%). We define the honesty penalty P_t as

$$P_t = H_t - L_t, \tag{2}$$

so that $P_t > 0$ indicates lower investment in partners with less honest histories. Four measures summarize initial investment, updating, persistence, and convergence.

- **Initial honesty penalty** P_1 measures the investment difference between high- and low-honesty partners before any economic feedback, capturing the influence of prior history on the first decision.
- **Low-honesty updating** U_L is the increase from round 1 to the mean investment over rounds 16–20 under low-honesty histories, capturing adaptation to favorable feedback.
- **Late honesty penalty** $P_{16:20}$ is the mean of P_t over rounds 16–20 and measures the remaining history effect after repeated feedback.
- **Low-honesty relaxation time** τ_L characterizes the rate at which low-honesty investment approaches its asymptote. Larger values indicate slower convergence.

We estimate τ_L from the mean low-honesty trajectory,

$$L_t = L_\infty + A_L \exp[-(t-1)/\tau_L], \tag{3}$$

where L_∞ is the fitted asymptotic investment and A_L is the initial offset. Appendix B details fitting and statistical inference, and Appendix C provides further analyses of the history effect's robustness and dynamics.

Diagnosing the Impression-to-Action Gap. An absent honesty penalty may reflect a failure to form an honesty-sensitive impression, retain it until investment, or use it in action. We distinguish these possibilities in Qwen3.5-9B and Qwen3.5-4B (Figure 1c).

(i) Impression formation. We quantify honesty-sensitive evaluation from autoregressive token logits conditioned on the Explicit evaluation prompt, rather than from the generated adjective. Separate discovery materials yield model-specific descriptor sets whose complete-word probabilities increase (V_+) or decrease (V_-) with honesty. For each descriptor w, its complete-word probability Q_w is derived from token logits over permitted complete-word paths. We standardize its log probability using the discovery-set mean μ_w and standard deviation σ_w,

$$z_w = \frac{\log Q_w - \mu_w}{\sigma_w}. \tag{4}$$

The social impression score contrasts standardized evidence for the two descriptor sets,

$$S = \frac{1}{|V_+|} \sum_{w \in V_+} z_w - \frac{1}{|V_-|} \sum_{w \in V_-} z_w. \tag{5}$$

Higher S indicates stronger evidence for descriptors associated with greater honesty. Appendix E details the construction and filtering of V_+ and V_-.

(ii) Impression retention. We next test whether impression-related information remains linearly decodable across the transition to Stage 2. Using the social impression score at T1 as the common reference target S_{ref}, we fit independent ridge readouts from hidden states at four task positions.

- **T0** – the position after the final Stage 1 verification, before the evaluation prompt.
- **T1** – the answer-start position after the evaluation prompt, before adjective generation.
- **T2** – the position after the between-stage transition, before the Stage 2 introduction.
- **T3** – the position after the Stage 2 introduction, before the first investment request.

(iii) Impression usage in behavior. Finally, we test whether making the prior evaluation easier to retrieve or explicitly relevant to investment changes behavior. An *impression retrieval cue* repeats the model's own partner adjective after the Stage 2 introduction. An *impression relevance cue* links that assessment to how cautiously the model may interact with the partner. A *combined impression cue* provides both. We compare initial and late honesty penalties, P_1 and $P_{16:20}$, against the Explicit evaluation baseline. Appendix G provides exact prompts and inference procedures.

4 Results

4.1 Honesty History Shapes Investment Beyond Accumulating Feedback

We first test whether a partner's honesty history shapes investment before any feedback from the new interaction is observed and whether this history effect persists as feedback accumulates.

Honesty history shapes initial investment. Figure 2 summarizes the initial honesty penalty P_1 for all seven models across the three transition conditions. Across the five medium-to-large models, a positive initial honesty penalty emerges either spontaneously in the control transitions or after Explicit evaluation. By contrast, the two smaller models show no clear positive penalty. These patterns separate the models into three behavioral profiles:

- **Spontaneous Penalty Profile** – DeepSeek V4 Pro, Qwen3.5-27B, Qwen3.8-27B, and Qwen3-32B already show positive P_1 in the control transitions, without requiring Explicit evaluation. For Qwen3-32B, the control effects are visually small but positive: 0.15 [0.04, 0.30] under No evaluation and 0.09 [0.00, 0.27] under Color control.
- **Evaluation-Elicited Penalty Profile** – Qwen3-14B shows no initial honesty penalty in either control ($P_1 = 0$), but a positive penalty appears after Explicit evaluation ($P_1 = 0.60$ [0.35, 0.87]).
- **Absent Penalty Profile** – Qwen3.5-9B and Qwen3.5-4B show no clear positive P_1 even after Explicit evaluation.

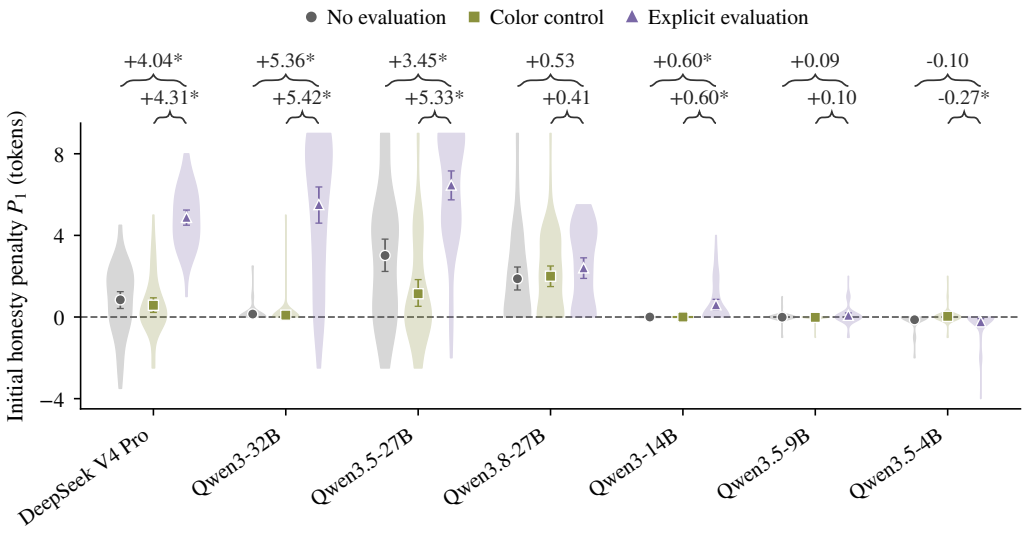

Figure 2: Initial honesty penalty across transition conditions under fixed returns. Markers and error bars indicate means and 95% trial-bootstrap intervals. Numbers above the upper and lower braces report the paired ΔP_1 for Explicit evaluation relative to No evaluation and Color control, respectively. Asterisks indicate **Holm-adjusted** $p < .05$.

Investment updates without erasing history effects. Figure 3 traces investment over 20 rounds under Explicit evaluation, with the same 0.60 return fraction for every honesty history. Investments generally rise as favorable feedback accumulates, especially after low-honesty histories. However, such updating does not necessarily eliminate the effect of honesty history. The four Spontaneous Penalty Profile models have both positive low-honesty updating U_L and positive late honesty penalties $P_{16:20}$ (Table 1). This indicates that they increase investment while retaining a history-dependent gap. Qwen3-14B also updates ($U_L = 5.63$) but reaches the 10-token ceiling across histories, eliminating its late penalty. Both updating and history persistence are also observed under No evaluation and Color control, and likewise under variable returns (Appendix D).

4.2 Explicit evaluation Strengthens and Prolongs History Effects

We test whether explicitly articulating a social impression amplifies the history effects observed above and changes subsequent adaptation. Explicit evaluation is compared with both No evaluation and Color control.

Explicit evaluation strengthens the initial honesty penalty. Figure 2 shows higher P_1 under Explicit evaluation than under the other two control conditions across the five medium-to-large models. Notably, Color control also adds an extra generation turn and retains the generated word, but removes the partner-directed evaluative content. The additional increase under Explicit evaluation therefore suggests that the amplification is associated with evaluative content rather than the extra turn or retained response alone. Paired ΔP_1 estimates further support this pattern: DeepSeek V4 Pro, Qwen3-32B, Qwen3.5-27B, and Qwen3-14B show Holm-adjusted increases relative to both controls, whereas Qwen3.8-27B shows a smaller increment from already positive control penalties (Figure 2).

Explicit evaluation slows adaptation to low-honesty partners. Table 1 shows that the effect extends beyond the first investment. Under Explicit evaluation, DeepSeek V4 Pro, Qwen3-32B, Qwen3.5-27B, and Qwen3-14B exhibit longer low-honesty relaxation times τ_L than under both controls, with paired increases surviving Holm correction. Qwen3.8-27B shows only small increases. Consistent with this slower convergence, U_L remains positive in all five medium-to-large models

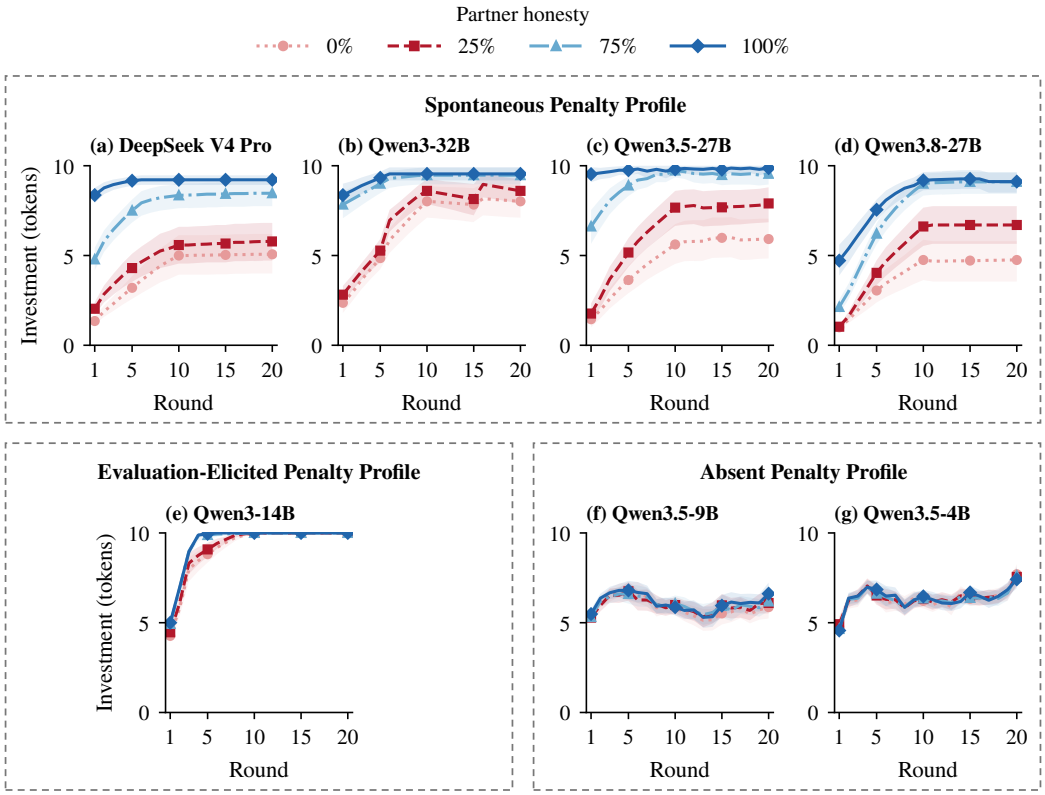

Figure 3: Investment adaptation under Explicit evaluation with fixed returns. Bands are 95% trial-bootstrap intervals.

under Explicit evaluation. Explicitly articulating the partner evaluation therefore strengthens the initial history effect and slows the subsequent convergence of low-honesty investment.

Table 1: Effects of Explicit evaluation on honesty-dependent investment dynamics under fixed returns. Brackets give 95% paired-bootstrap intervals. Asterisks and bold changes mark **Holm-adjusted** $p < .05$.

Model	Explicit evaluation		Relaxation time τ_L			Change in τ_L	
	$P_{16:20}$	U_L	No evaluation	Color control	Explicit evaluation	$\Delta\tau_L$ vs No evaluation	$\Delta\tau_L$ vs Color control
DeepSeek V4 Pro	3.43*	3.72*	2.70	2.16	4.58	**1.88*** [1.31, 2.51]	**2.42*** [1.88, 3.06]
Qwen3-32B	1.07*	5.85*	2.29	2.48	4.99	**2.70*** [2.30, 3.24]	**2.51*** [2.21, 2.87]
Qwen3.5-27B	2.84*	5.24*	3.25	2.68	4.87	**1.62*** [0.76, 2.99]	**2.19*** [1.23, 3.51]
Qwen3.8-27B	3.41*	4.70*	4.38	4.68	4.73	0.35 [-0.31, 1.02]	0.05 [-0.42, 0.58]
Qwen3-14B	0.00	5.63*	1.56	1.45	2.20	**0.64*** [0.27, 1.03]	**0.75*** [0.37, 1.14]
Qwen3.5-9B	0.27	0.48	—	—	—	—	—
Qwen3.5-4B	-0.04	1.88*	—	—	—	—	—

4.3 Small Models Retain but Underuse Social Impressions

The two smallest models break the otherwise consistent link between partner evaluation and initial investment. Qwen3.5-9B and Qwen3.5-4B show no clear positive P_1 even under Explicit evaluation. Following Section 3, this may reflect a failure of impression formation, retention, or usage in behavior. The analyses below distinguish among these possibilities.

Small models can form social impressions. On independent confirmation materials, the social impression score S increases with partner honesty in both models. The slope of S with respect to partner honesty is 4.321 [4.276, 4.367] for Qwen3.5-9B, with Spearman $\rho = 0.965$, and 4.615 [4.552, 4.680] for Qwen3.5-4B, with $\rho = 0.959$. Both permutation tests give unadjusted $p < .001$. The absent initial penalty therefore does not reflect a complete failure of impression formation.

Social impression information remains decodable before investment. Figure 4 plots predictions of the same T1 reference score S_{ref} from independent ridge readouts at T0–T3. Predictions remain close to the identity line in both models, including T3 immediately before the first investment. At T3, Qwen3.5-9B reaches Pearson $r = 0.985$ [0.981, 0.988] with $R^2 = 0.970$, and Qwen3.5-4B reaches $r = 0.977$ [0.972, 0.983] with $R^2 = 0.955$. Thus, impression-related information remains linearly decodable at the decision point. Together with weak initial investment differentiation, this supports the *impression-to-action gap*, whereby impression-related information remains available in the representation but is not reliably expressed in the first investment.

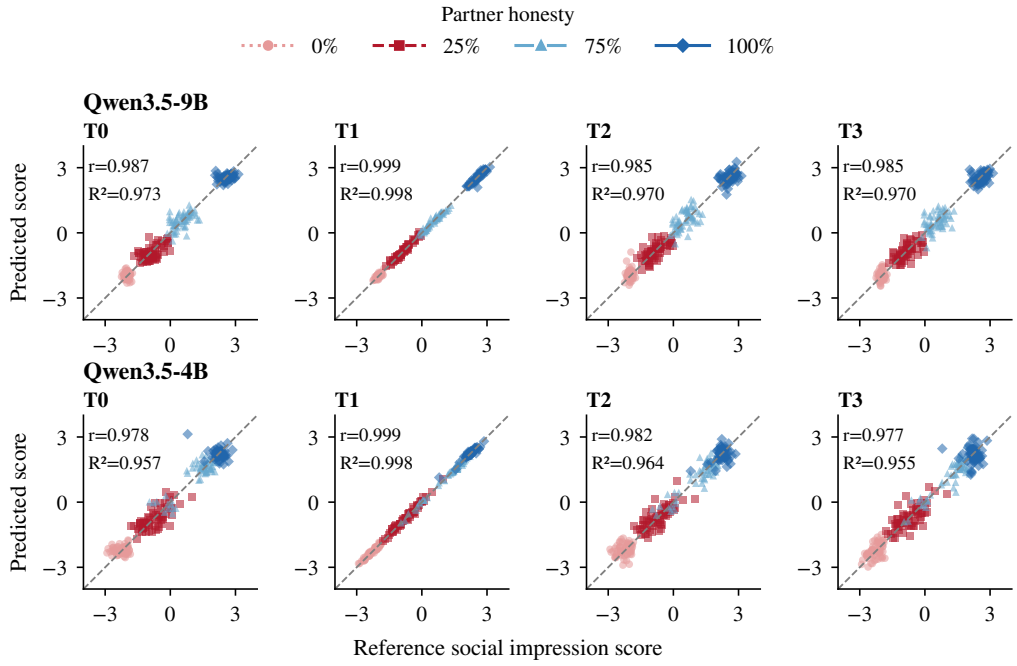

Figure 4: Linear decodability of social impression information across task stages. Dashed lines mark $y = x$.

Targeted cues promote behavioral use of social impressions. Having established formation and retention, we next test impression usage in behavior. Relative to the matched Explicit evaluation baseline, the *impression retrieval cue* repeats the model's own partner adjective, the *impression relevance cue* links the prior assessment to how cautiously the model may interact with the partner, and the *combined impression cue* supplies both. Appendix G provides exact prompts. Either single cue induces positive P_1 in Qwen3.5-9B, whereas only the combined cue does so in Qwen3.5-4B among the tested conditions (Table 2). The combined cue also increases $P_{16:20}$ in both models, with paired increments surviving Holm correction. These results support limited spontaneous use of available social impression information rather than its absence or loss. Making the impression easier to retrieve and explicitly relevant to action can induce the honesty penalty.

Table 2: Effects of impression cues in small models under fixed returns. Brackets give 95% paired-trial bootstrap intervals. Asterisks and bold changes mark **Holm-adjusted** $p < .05$.

Condition	P_1	ΔP_1 vs Explicit evaluation	$P_{16:20}$	$\Delta P_{16:20}$ vs Explicit evaluation
Qwen3.5-9B				
Impression retrieval cue	1.77	**1.68*** [1.30, 2.10]	0.82	0.56 [0.08, 1.06]
Impression relevance cue	3.40	**3.32*** [2.72, 3.96]	1.36	1.09 [0.26, 1.93]
Combined impression cue	3.55	**3.47*** [2.83, 4.12]	2.08	**1.81*** [1.08, 2.56]
Qwen3.5-4B				
Impression retrieval cue	-0.22	0.02 [-0.33, 0.35]	0.14	0.18 [-0.11, 0.47]
Impression relevance cue	-0.01	0.23 [-0.03, 0.49]	0.16	0.20 [-0.22, 0.58]
Combined impression cue	0.58	**0.82*** [0.52, 1.15]	0.67	**0.70*** [0.28, 1.13]

5 Discussion and Conclusion

Conclusion. We study how social impressions formed from a partner's prior behavior shape subsequent decisions and adaptation in sustained LLM interactions. Across medium-to-large models, honesty history influences the first investment before any economic feedback is observed, while favorable returns subsequently increase investment without necessarily eliminating this history-dependent difference. Explicitly articulating a partner evaluation can further amplify the initial honesty penalty and slow its attenuation over repeated interactions. The two smallest models reveal a complementary failure mode: honesty-sensitive impression information remains linearly decodable before investment but is not reliably expressed in behavior, whereas targeted retrieval and relevance cues can restore this differentiation. These results motivate viewing social decision-making as a sequence of potentially dissociable processes—impression formation, retention, behavioral use, and feedback-driven revision—rather than as a property that can be characterized by a single response.

Limitations. Our experimental design prioritizes identification over ecological complexity. Stage 1 provides fully verified honesty evidence, while Stage 2 uses a scripted partner and a single repeated trust game action space; the variable-return analysis in Appendix D relaxes the fixed-return schedule but does not capture ambiguous social evidence, strategic co-adaptation, or open-ended interaction. The Explicit evaluation manipulation also changes the semantic content retained in context. Although the No evaluation and Color control conditions separate partner evaluation from an extra generation turn and arbitrary word production, they cannot determine whether the amplified honesty effect arises from deeper encoding, easier retrieval, self-conditioning on the generated adjective, or increased task relevance. In addition, linear probes establish decodability rather than causal use of the decoded representation, and the prompt interventions act at the input level rather than directly on internal states. Finally, our results cover a finite model panel and one interaction domain. The honesty penalty is therefore an operational measure of history-dependent investment, not a measure of rationality, alignment, human-like social understanding, or economic performance.

Future work. Two directions are important. First, external-validity studies should replace verified histories and scripted partners with noisy or conflicting social evidence, adaptive partners, and richer settings such as negotiation, collaboration, and delegation, while preserving matched-outcome controls where possible. Second, mechanism-focused studies should move from correlational readout to causal intervention, for example by manipulating candidate impression representations or counterfactually editing the retained evaluation to test whether the same information that is decoded is also responsible for investment changes. Expanding these analyses across model families, post-training methods, memory architectures, and longer interaction horizons would clarify when impression-to-action gaps emerge and when prior social impressions remain behaviorally consequential.

AI USE STATEMENT

In this work, we used large language models (LLMs) to correct grammatical errors and provide feedback on the clarity and readability of the manuscript. LLM-based coding assistants helped generate portions of the code. In addition to conventional literature search methods, we used AI2 Paper Finder and OpenAI Deep Research to identify relevant prior work.

The authors reviewed all AI-assisted work. Specifically, we checked the revised text for accuracy and consistency with the intended meaning, reviewed and tested the AI-assisted code, and consulted the original publications to verify the accuracy and relevance of the references included in the manuscript. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 describes the two-stage experimental design, and Appendix A provides the experimental protocol, verbatim prompts, stimulus construction, trial sampling and matching, random-seed handling, model identifiers, software versions, and inference settings. Appendix B specifies the behavioral measures, convergence fitting, bootstrap procedures, paired tests, and multiple-comparison corrections. Appendix E details the construction and validation of social impression scores, including the disjoint discovery, validation, and confirmation materials, descriptor selection, and calibration. Appendix F describes hidden-state extraction and grouped nested cross-validation for the ridge probes, while Appendix G provides the exact intervention prompts and paired comparisons. Additional behavioral results and robustness analyses are reported in Appendices C and D, and the exploratory interpersonal-ability analysis is documented in Appendix H. The code is available for anonymous review at `https://github.com/PaperCode-3333/ICLR-papercode`.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y. URL https://www.nature.com/articles/s41562-025-02172-y.

Luke J. Chang, Bradley B. Doll, Mascha van 't Wout, Michael J. Frank, and Alan G. Sanfey. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2):87–105, 2010. doi: 10.1016/j.cogpsych.2010.03.001. URL https://doi.org/10.1016/j.cogpsych.2010.03.001.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d3362f84979d16cee000f09eef61244c-Abstract-Conference.html.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel L. K. Yamins, and Nick Haber. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. In *International Conference on Learning Representations*, pp. 6507–6546, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/12f483f624b378f9f3058d8ecd3c7ff5-Paper-Conference.pdf.

M R Delgado, R H Frank, and E A Phelps. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, 8(11):1611–1618, 2005. doi: 10.1038/nn1575. URL https://doi.org/10.1038/nn1575.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. GTBench: Uncovering the Strategic Reasoning Capabilities of LLMs via Game-Theoretic Evaluations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28219–28253. Curran Associates, Inc., 2024. doi: 10.52202/079017-0885. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3191170938b6102e5c203b036b7c16dd-Paper-Conference.pdf.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring Latent World States in Language Models with Propositional Probes. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 19337–19359, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3132d0405fabe24b2a7b6cd7ba9de6b5-Paper-Conference.pdf.

Elsa Fouragnan, Gabriele Chierchia, Susanne Greiner, Remi Neveu, Paolo Avesani, and Giorgio Coricelli. Reputational priors magnify striatal responses to violations of trust. *The Journal of Neuroscience*, 33(8):3602–3611, 2013. doi: 10.1523/jneurosci.3086-12.2013. URL https://doi.org/10.1523/jneurosci.3086-12.2013.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. In *International Conference on Learning Representations*, pp. 50252–50297, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/524ff06d5375d76e93d8490e654a81bf-Paper-Conference.pdf.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Udhay Nallasamy, Andy Miller, and Jaya Narain. Do LLMs "know" internally when they follow instructions? In *International Conference on Learning Representations*, pp. 81339–81357, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ca6980a3dba7fb3e4e66925656dba68b-Paper-Conference.pdf.

Tiancheng Hu and Nigel Collier. Quantifying the Persona Effect in LLM Simulations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL https://aclanthology.org/2024.acl-long.554/.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 128–145, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.7. URL https://aclanthology.org/2024.emnlp-main.7/.

Haoxi Li, Xueyang Tang, Jie Zhang, Song Guo, Sikai Bai, Peiran Dong, and Yue Yu. CAUSM: Causally Motivated Sycophancy Mitigation for Large Language Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 66020–66040, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a52b0d191b619477cc798d544f4f0e4b-Paper-Conference.pdf.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. InterIntent: Investigating Social Intelligence of LLMs via Intention Understanding in an Interactive Game Context. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6718–6746, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.383. URL https://aclanthology.org/2024.emnlp-main.383/.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey Savchenko, and Ilya Makarov. EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 53969–54002. Curran Associates, Inc., 2024. doi: 10.52202/079017-1709. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/611e84703eac7cc03f78339df8aae2ed-Paper-Conference.pdf.

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.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do LLM agents have regret? a case study in online learning and games. In *International Conference on Learning Representations*, pp. 41214–41234, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/65ea878cb90b440e8b4cd34fe0959914-Paper-Conference.pdf.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 111715–111759. Curran Associates, Inc., 2024. doi: 10.52202/079017-3548. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ca9567d8ef6b2ea2da0d7eed57b933ee-Paper-Conference.pdf.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving Instruction-Following in Language Models through Activation Steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8c3262a4c965ba9888f120d4f9e13478-Paper-Conference.pdf.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z. URL https://www.nature.com/articles/s41562-024-01882-z.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 15674–15729. Curran Associates, Inc., 2024. doi: 10.52202/079017-0501. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1cb57fcf7ff3f6d37eebae5becc9ea6d-Paper-Conference.pdf.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.466. URL https://aclanthology.org/2024.acl-long.466/.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1496. URL https://aclanthology.org/2025.acl-long.1496/.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21692–21714, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1208. URL https://aclanthology.org/2024.emnlp-main.1208/.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b3075b88e583a0e98d8b24338a613060-Paper-Conference.pdf.

A Experimental protocol and implementation

This appendix specifies how honesty histories and economic feedback are matched to isolate history and transition-condition effects. It also provides the verbatim prompts, response handling, model checkpoints, and inference settings.

A.1 Design, sampling, matching, and retained context

Sample size. The behavioral experiments initially comprised $N = 60$ trial identifiers per model and return-schedule setting. For Qwen3-32B, response-parsing failures led to the exclusion of 5 trials under fixed returns and 6 trials under variable returns, leaving $n = 55$ and $n = 54$ analyzed trials, respectively. All other models retained $n = 60$ trials in each return-schedule setting.

Matching. Trial identifiers define the pairing unit. Committed choices are shared across honesty levels. Honesty-specific reports and return rates are shared across transition conditions. Stimuli use master seed 20260916. Decoding seeds are matched by trial, phase, and round where supported.

Roles and retained context. The evaluated model acts as the assistant and makes all investment decisions. The experiment script generates the partner's committed choices, reports, and returns. A single system message introduces the experiment. Task instructions, partner reports, verification, and economic feedback are user messages. The phrase "System verification" belongs to the message text and does not indicate a system-role message. The same conversation is retained across both stages, including the full honesty history, any generated color or partner adjective, and all subsequent investments and feedback. The system instruction below applies to every condition and supplies context rather than a separate response request.

System instruction (system).

You are participating in a behavioral experiment. The same partner will appear in two sequential tasks. Follow each task's output format exactly.

A.2 Stage 1: history construction and observation prompts

For each trial, a sequence of choices from $\{A, B, C\}$ is drawn and shared across honesty levels. The partner privately commits to each choice before reporting it, and the commitment cannot subsequently change. Truthful reports name the committed option. False reports are drawn from the other two options. The 12 observations comprise three consecutive blocks of four. Each block contains 0, 1, 3, or 4 truthful reports for the 0%, 25%, 75%, or 100% history, with positions shuffled within the block. Thus the four histories contain 0, 3, 9, or 12 truthful reports while preserving the underlying committed choices.

After observing each report, the assistant returns {`"acknowledged":true`}. Only then does a user message reveal the committed choice and the label `TRUE` or `FALSE`. The evaluated model makes no choice among A, B, and C.

History observation (user; repeated in Stage 1).

Phase 1, round {round} of 12. The partner first locked a choice privately before making any report; that locked choice cannot be changed after the report. Partner reports: "I chose {reported}." The system will now reveal the choice locked before the report. Reply with exactly {"acknowledged":true}.

History verification (user; after the acknowledgement).

System verification: the choice locked before the report was {actual}. The report was {label}.

A.3 Between-stage transition conditions and prompts

The manipulation occurs once, after the final history verification and before the investment task. No evaluation supplies the direct-transition message. Color control instead requests a favorite color in one English word, retains the response, and supplies the shared transition. Explicit evaluation follows the same sequence with a request for one English adjective describing the partner. The two word-generation conditions use the same insertion point and transition text. Their generated words remain available in every investment round. The transition messages do not request an additional model response.

No evaluation transition (user).

The first task is complete. You will now interact with the same partner in a different task.

Color control request (user).

Name your favorite color with exactly one word. Reply with one English color only, with no spaces, punctuation, or explanation.

Explicit evaluation request (user).

Describe the partner with exactly one English adjective. Reply with one English adjective only, with no spaces, punctuation, or explanation.

Shared transition after the generated color or adjective (user).

The first task is complete. You will now interact with the same partner in a different task.

A.4 Stage 2: investment prompts and return schedules

The investment task begins with an introduction to the payoff rule, followed by 20 decision–feedback rounds with the same partner. The endowment resets to 10 tokens each round, and investment is restricted to integers from 1 to 10. Each decision appends the investment request, assistant response, and feedback to the conversation. Requests identify the current round and the 20-round horizon. The investment is tripled, and the returned amount $3a_t r_t$ is computed using decimal arithmetic without integer rounding. Feedback reports the investment, tripled amount, returned tokens, and payoff $10 - a_t + 3a_t r_t$. Return rates and the future schedule are not disclosed in advance. The introduction and feedback messages supply context; the investment request elicits the integer response.

Investment introduction (user; all conditions).

You will now play a different game with the same partner. Treat this as an ongoing interaction. You have 10 tokens this round and may invest any integer amount from 1 to 10 with the partner. Your investment is tripled and given to the partner; the partner then returns some tokens. Your payoff is 10 - investment + returned tokens. After each round you will observe the return before the next investment. When asked for an investment, output one integer from 1 to 10 and nothing else.

Investment request (user; repeated each round).

Investment round {round} of 20. Output a single integer from 1 to 10 and nothing else.

Investment feedback (user; after the decision).

You invested {investment}. The partner received {tripled} and returned {returned} tokens. Your payoff was {payoff}.

Return schedules. Fixed returns use $r_t = 0.60$ throughout. Variable returns use a trial-specific permutation of a multiset containing two copies of each rate in $\{0.56, 0.57, \ldots, 0.64\}$ and two additional copies of 0.60, giving a temporal mean of exactly 0.60. All honesty histories and transition conditions within a trial share this permutation. Matching applies to return rates. Returned amounts and payoffs also depend on investment.

A.5 Response handling, model panel, and inference settings

Acknowledgements must be JSON objects containing only `acknowledged:true`. Color and partner-description responses must contain one ASCII alphabetic word after trimming. No lexical or part-of-speech filter is imposed. Investment responses must specify an integer from 1 to 10, with surrounding whitespace or code delimiters permitted. The prompts above are verbatim. Braced fields are script substitutions.

Model panel and inference settings. The behavioral panel comprises DeepSeek V4 Pro (API identifier `deepseek-v4-pro`) and the local checkpoints Qwen3-32B, Qwen3.5-27B, Qwen3.8-27B, Qwen3-14B, Qwen3.5-9B, and Qwen3.5-4B. Local models use bfloat16 weights, a maximum context length of 8,192 tokens, vLLM 0.20.1+cu129, and Transformers 5.7.0. Qwen3.5-9B and Qwen3.5-4B also support the social impression score, probing, and prompt-intervention analyses in Appendices E–G.

Generation uses temperature 0.7, top-p 1.0, and at most 128 output tokens, with no top-k override. Thinking is disabled through `enable_thinking=false` in the local chat template and `thinking.type=disabled` for DeepSeek. Investment requests use a newline stop sequence.

B Behavioral measures and statistical inference

These measures separate initial history dependence, adaptation in investment, and residual history effects. The inference procedures preserve trial pairing across honesty levels, conditions, and rounds.

B.1 TRIAL-LEVEL NOTATION AND AGGREGATION

Within a model and return schedule, $I^c_{h,j,t}$ denotes investment under transition condition c, honesty proportion $h \in \{0, .25, .75, 1\}$, paired trial j, and round t. Fixed- and variable-return results are analyzed separately.

Suppressing c for one condition, we average investments within the low- and high-honesty groups and take their difference:

$$L_{j,t} = \frac{I_{0,j,t} + I_{.25,j,t}}{2}, \qquad H_{j,t} = \frac{I_{.75,j,t} + I_{1,j,t}}{2}, P_{j,t} \quad = H_{j,t} - L_{j,t}. \tag{6}$$

Unindexed L_t, H_t, and P_t are trial means. A positive honesty penalty P_t indicates lower investment following low-honesty histories. For any trial-level trajectory $Y_{j,t}$, its mean over a window $a{:}b$ is

$$\overline{Y}_{j,a:b} = \frac{1}{b-a+1} \sum_{t=a}^{b} Y_{j,t}. \tag{7}$$

Window averages and linear contrasts are computed within trials before aggregation.

B.2 PRIMARY BEHAVIORAL MEASURES

Primary endpoints. Four measures summarize initial investment, updating, persistence, and convergence. The first three are measured in tokens; τ_L is measured in rounds:

- **Initial honesty penalty P_1.** The first-round difference measures history dependence before any economic feedback:
$$P_1 = \frac{1}{n} \sum_{j=1}^{n} P_{j,1}. \tag{8}$$
- **Low-honesty updating U_L.** The change from the first round to the final five rounds measures investment updating as favorable feedback accumulates:
$$U_L = \frac{1}{n} \sum_{j=1}^{n} \big(\overline{L}_{j,16:20} - L_{j,1}\big). \tag{9}$$
- **Late honesty penalty $P_{16:20}$.** The average penalty over the final five rounds measures the residual history effect:
$$P_{16:20} = \frac{1}{n} \sum_{j=1}^{n} \overline{P}_{j,16:20}. \tag{10}$$
- **Low-honesty relaxation time τ_L.** Equation 3 fits the mean low-honesty trajectory to estimate its rate of convergence to its own asymptote L_∞. Larger values indicate slower convergence. This measure does not describe convergence of P_t to zero. Fitting and uncertainty estimation are detailed below.

B.3 SUPPLEMENTARY DIAGNOSTICS AND CONDITION CONTRASTS

Three additional measures characterize honesty sensitivity and early updating:

- **Extreme-history contrast E_1.** This is the trial mean of the first-round investment difference between the two extreme honesty levels:
$$E_{j,1} = I_{1,j,1} - I_{0,j,1}. \tag{11}$$
- **Four-level honesty gradient G_1.** This is the trial mean of the linear investment gradient across all four honesty levels:
$$G_{j,1} = \frac{\sum_{h \in \{0,.25,.75,1\}} (h - .5) I_{h,j,1}}{\sum_{h \in \{0,.25,.75,1\}} (h - .5)^2}. \tag{12}$$

The gradient is measured in tokens per 100 percentage points of honesty. Four-level plots additionally display the level-specific means without assuming linearity.

- **Early low-honesty slope s_L.** This is the mean of trial-level ordinary least-squares slopes fitted to low-honesty investment over rounds 1–5. Positive s_L indicates increasing investment, in tokens per round. It is used in Explicit evaluation contrasts and the exploratory interpersonal-ability analysis.

Condition contrasts. For an endpoint M, the Explicit evaluation effect subtracts the named control c:

$$\Delta M = M^{\text{Explicit evaluation}} - M^{c}. \tag{13}$$

For the initial penalty, this gives $\Delta P_1 = \Delta H_1 - \Delta L_1$, the investment decomposition in Table 4. A positive contrast means a larger endpoint under Explicit evaluation. Cue increments instead subtract the matched Explicit evaluation baseline (Appendix G).

B.4 Bootstrap intervals and paired mean-effect tests

Bootstrap intervals. Mean endpoints and paired differences use 5,000 whole-trial bootstrap resamples and percentile 95% intervals. All honesty levels, conditions, and rounds belonging to a trial remain together. Condition contrasts resample within-trial differences. Trajectory bands use complete trajectories and are pointwise, as are endpoint intervals.

Paired tests. Two-sided mean-effect tests use 10,000 trial-level sign flips and the absolute mean statistic. Monte Carlo p values use the plus-one correction:

$$p = \frac{1 + \#\{T_b \geq T_{\text{obs}}\}}{B + 1}, \qquad B = 10{,}000, \tag{14}$$

where T_{obs} is the observed absolute mean and T_b is its value after sign flipping. The base analysis seed is 20260921001.

B.5 Convergence fitting and time-constant inference

Trajectory fitting. Equation 3 is fitted to the 20-round mean low-honesty trajectory by unweighted nonlinear least squares using SciPy's bounded `curve_fit`. The parameters are initialized as

$$(L_\infty, A_L, \tau_L)_{\text{init}} = \big(\overline{L}_{16:20}, L_1 - \overline{L}_{16:20}, 3\big). \tag{15}$$

Bounds are $L_\infty, A_L \in [-50, 50]$ and $\tau_L \in [0.05, 100]$, with at most 50,000 function evaluations. Each condition retains its own asymptote throughout fitting and inference.

Fit quality and reporting. We assess fit quality using

$$R^2 = 1 - \frac{\sum_t (L_t - \widehat{L}_t)^2}{\sum_t (L_t - \overline{L})^2}, \tag{16}$$

where $\widehat{L}_t$ is the fitted investment and $\overline{L}$ is the mean of L_t over all 20 rounds. We report τ_L only when $R^2 \geq .90$ and $|A_L| \geq .50$ tokens. A constant trajectory has undefined R^2. A dash denotes an unmet reporting criterion or failed fit, rather than zero relaxation time.

Paired contrasts and intervals. For each fixed-return contrast in Table 1, both fitted τ_L values must meet the reporting criterion. Positive $\Delta\tau_L$ means slower convergence under Explicit evaluation. Pointwise intervals use 1,000 paired-trial bootstrap refits, sharing resampled indices across all conditions and rounds. All finite refits are retained without reapplying the fit-quality threshold.

Tests of time-constant differences. Two-sided tests use 20,000 centered paired-bootstrap refits. If $\widehat{\delta}$ is the observed time-constant difference and δ_b^* a resampled estimate, the test uses

$$p_{\text{boot}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{|\delta_b^* - \widehat{\delta}| \geq |\widehat{\delta}|\}}{B + 1}, \qquad B = 20{,}000. \tag{17}$$

The interval and test use separate streams based on seeds 20260921001 and 20260922, respectively. Single-condition time constants are not tested against zero. Table 7 supplies the fit diagnostics, and Table 5 supplies intervals for the linear endpoints in Table 1.

B.6 MULTIPLE-COMPARISON FAMILIES AND REPORTING RULES

Holm correction is applied within the following comparison families:

- **Behavioral mean tests:** seven models, separately for every endpoint, return schedule, and transition condition.
- **Explicit evaluation minus control contrasts:** seven models in a separate family for each named control, with endpoint and return schedule fixed.
- **Time-constant comparisons:** the five models whose two point estimates meet the reporting criterion, separately for each control.
- **Cue comparisons:** the six model-by-cue contrasts for each endpoint and return schedule. Condition tests and cue increments are not pooled.

Asterisks indicate adjusted $p < .05$. Bold main-table changes mark the same criterion. Pointwise interval coverage and adjusted tests are distinct inferential statements. Methods specific to social impression scores, probes, interventions, and interpersonal ability appear in Appendices E–H.

C BEHAVIORAL RESULTS UNDER FIXED RETURNS

These analyses test whether the initial and late history effects in Sections 4.1 and 4.2 extend across honesty levels and transition conditions. Investment decompositions identify how high- and low-honesty choices contribute to ΔP_1, while trajectory and fit diagnostics characterize adaptation. Statistical definitions and correction families are given in Appendix B.

C.1 INITIAL EFFECTS AND INVESTMENT DECOMPOSITION

Table 3 reports P_1 and its paired changes across transition conditions. Table 4 separates the contributions of high- and low-honesty investment. Qwen3-32B's small No evaluation penalty has a positive interval but does not survive Holm correction ($p_{\mathrm{H}} = .197$).

Table 3: Effects of Explicit evaluation on the initial honesty penalty under fixed returns. Condition columns report mean P_1. The final columns report paired Explicit evaluation minus control changes with 95% trial-bootstrap intervals. Asterisks mark Holm-adjusted $p < .05$ across seven models, separately for each contrast.

Model	Initial honesty penalty P_1			Explicit evaluation minus control	
	No evaluation	Color control	Explicit evaluation	No evaluation ΔP_1 [95% CI]	Color control ΔP_1 [95% CI]
DeepSeek V4 Pro	0.84	0.57	4.88	4.04* [3.49, 4.58]	4.31* [3.77, 4.84]
Qwen3-32B	0.15	0.09	5.51	5.36* [4.48, 6.20]	5.42* [4.49, 6.31]
Qwen3.5-27B	3.02	1.14	6.47	3.45* [2.50, 4.38]	5.33* [4.38, 6.22]
Qwen3.8-27B	1.88	2.00	2.41	0.53 [0.00, 1.07]	0.41 [-0.23, 1.03]
Qwen3-14B	0.00	0.00	0.60	0.60* [0.35, 0.87]	0.60* [0.35, 0.87]
Qwen3.5-9B	-0.01	-0.02	0.08	0.09 [-0.04, 0.23]	0.10 [-0.02, 0.22]
Qwen3.5-4B	-0.13	0.03	-0.23	-0.10 [-0.30, 0.08]	-0.27* [-0.48, -0.08]

Table 4: High- and low-honesty investment contributions to initial honesty-penalty changes under fixed returns. Column groups subtract each control from Explicit evaluation. Entries give paired mean changes in tokens with 95% trial-bootstrap intervals, satisfying $\Delta P_1 = \Delta H_1 - \Delta L_1$.

	Explicit evaluation minus No evaluation		Explicit evaluation minus Color control	
Model	ΔH_1	ΔL_1	ΔH_1	ΔL_1
DeepSeek V4 Pro	0.26 [-0.22, 0.72]	-3.78 [-4.20, -3.36]	0.61 [0.12, 1.09]	-3.70 [-4.03, -3.38]
Qwen3-32B	2.86 [2.22, 3.46]	-2.50 [-3.01, -1.99]	3.01 [2.37, 3.62]	-2.41 [-2.84, -1.95]
Qwen3.5-27B	0.79 [0.18, 1.41]	-2.66 [-3.31, -2.02]	1.43 [0.74, 2.08]	-3.90 [-4.58, -3.24]
Qwen3.8-27B	0.21 [-0.28, 0.74]	-0.33 [-0.53, -0.13]	-0.33 [-0.85, 0.24]	-0.73 [-1.07, -0.43]
Qwen3-14B	-0.03 [-0.10, 0.00]	-0.63 [-0.92, -0.37]	-0.03 [-0.10, 0.00]	-0.63 [-0.92, -0.37]
Qwen3.5-9B	0.25 [-0.02, 0.62]	0.16 [-0.09, 0.53]	-0.07 [-0.33, 0.32]	-0.17 [-0.47, 0.23]
Qwen3.5-4B	0.47 [0.23, 0.72]	0.57 [0.34, 0.82]	0.44 [0.16, 0.74]	0.71 [0.37, 1.07]

C.2 Complete trajectories and honesty gradients

Figures 5 and 6 show investment adaptation under the two controls, complementing Explicit evaluation in Figure 3. Figure 7 tests dose responsiveness across all four honesty levels. Figure 8 tracks the high–low penalty throughout feedback, while Figure 9 compares low-honesty convergence relative to each condition's own asymptote.

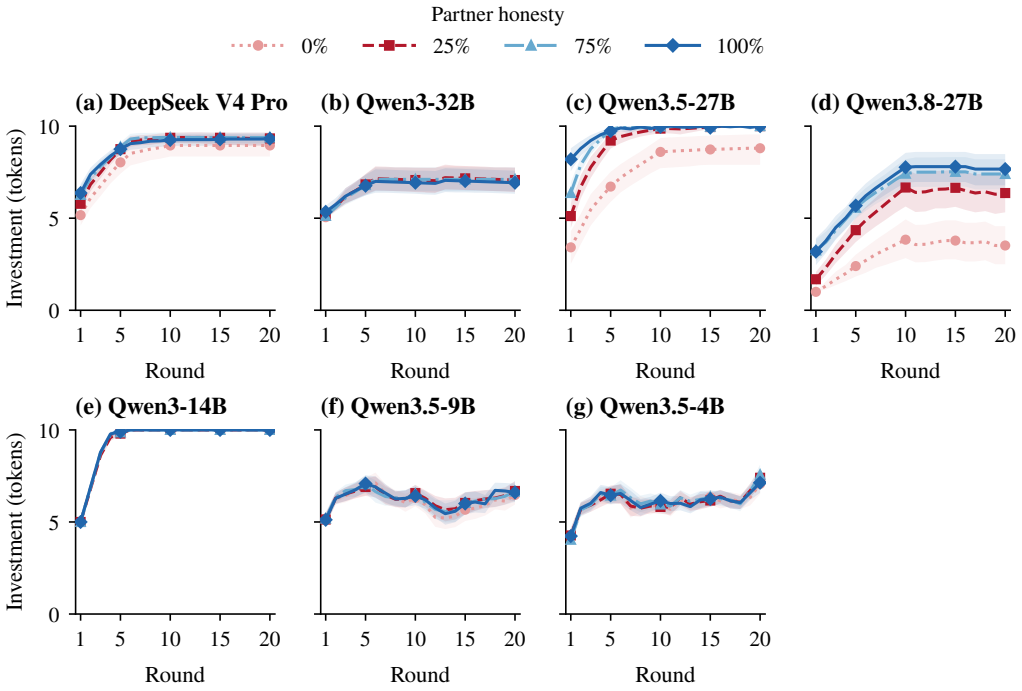

Figure 5: Fixed-return investment under No evaluation. Bands are 95% trial-bootstrap intervals. Models follow the order in Figure 2.

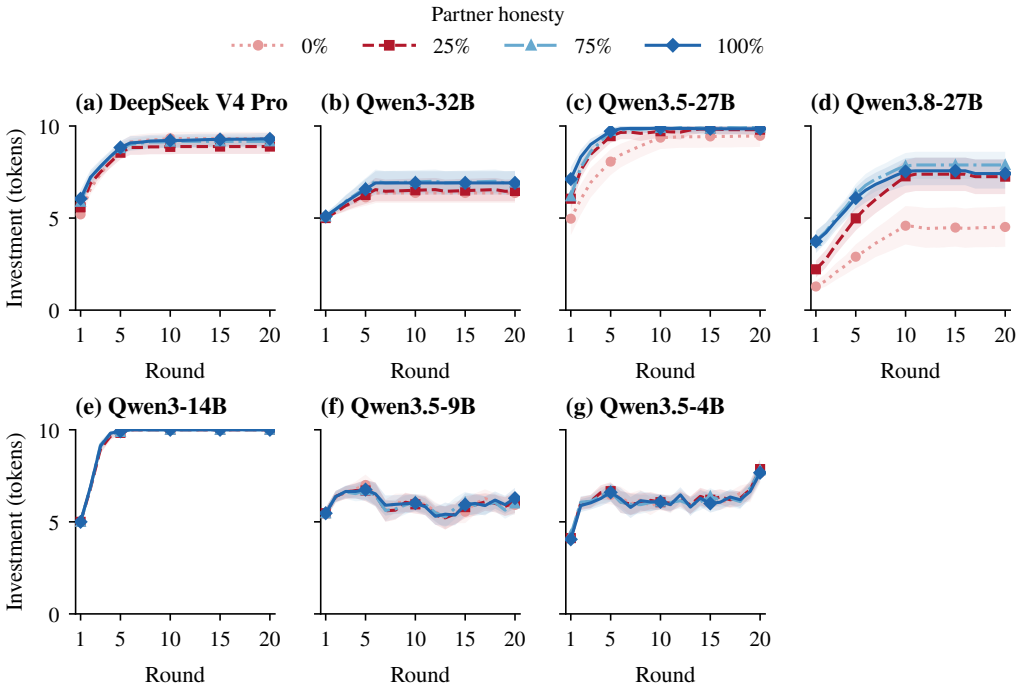

Figure 6: Fixed-return investment under Color control. Bands are 95% trial-bootstrap intervals.

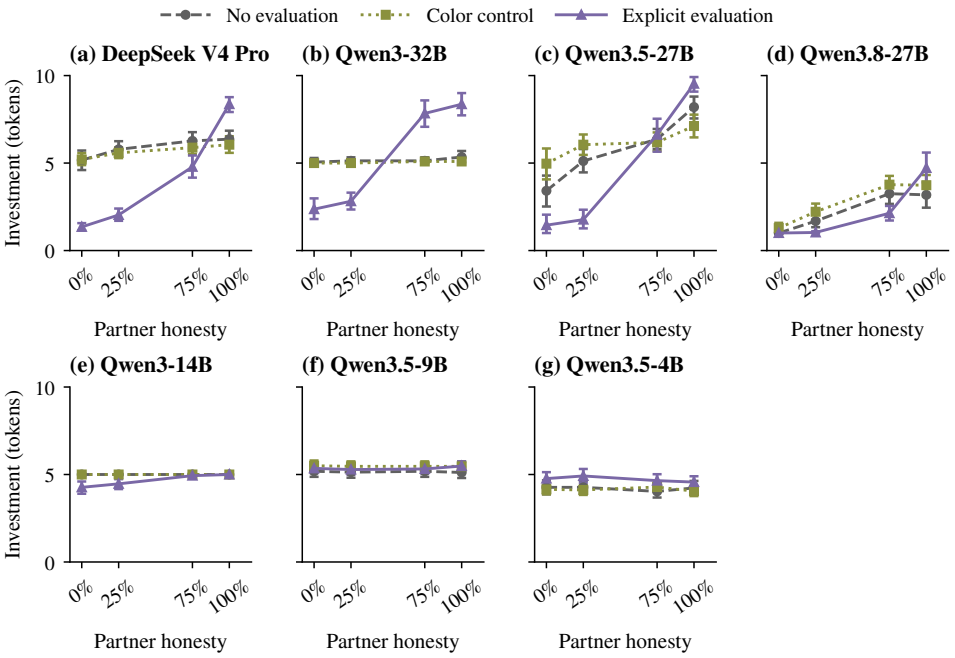

Figure 7: Honesty gradients in initial investment under fixed returns. Points and intervals give means and 95% trial-bootstrap intervals for each honesty level and transition condition. Segments connect observed means.

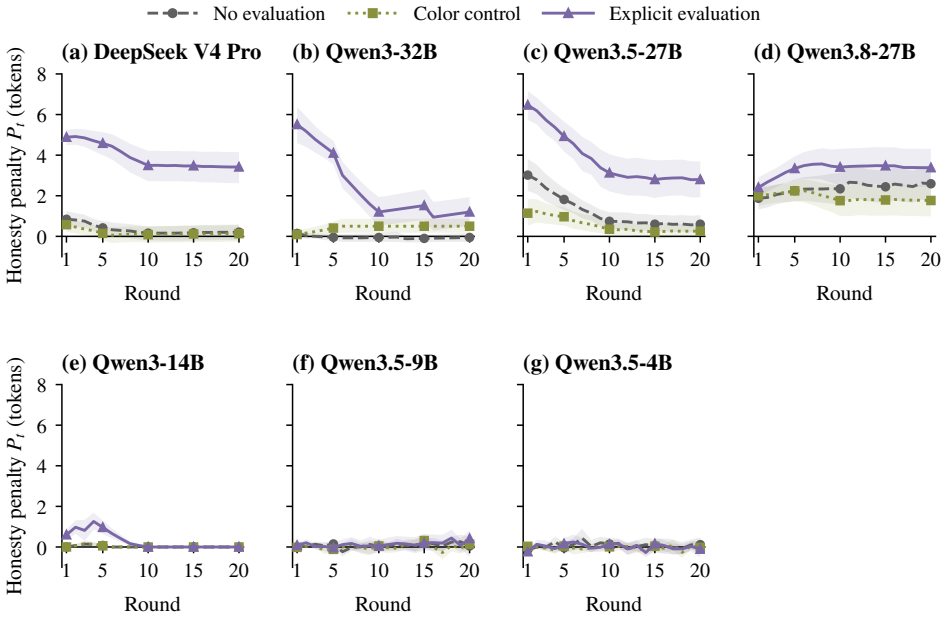

Figure 8: Honesty penalties across transition conditions under fixed returns. Curves show mean within-trial high-minus-low investment differences P_t. Bands are 95% trial-bootstrap intervals.

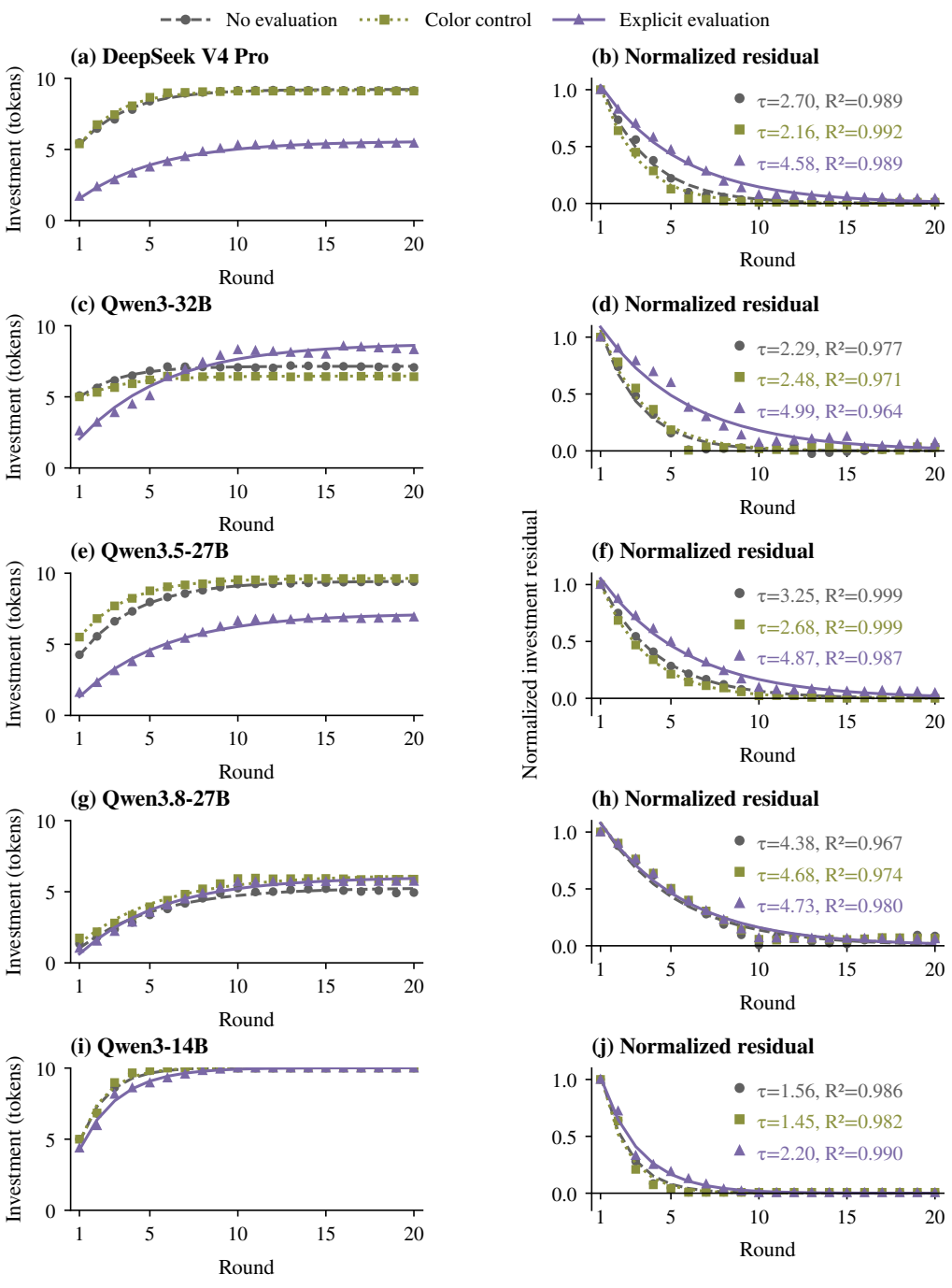

Figure 9: Low-honesty convergence under fixed returns. Left panels show observed means and exponential fits. Right panels show observed and fitted residuals normalized as $(L_t - L_\infty)/(L_1 - L_\infty)$. Annotations report τ_L in rounds and R^2. Included models meet the fit-reporting criteria.

C.3 CONDITION ENDPOINTS, PAIRED CONTRASTS, AND FIT DIAGNOSTICS

Table 5 reports condition-level behavioral measures and intervals. Table 6 quantifies paired Explicit evaluation effects, including changes in early-round slopes. Table 7 reports condition-specific exponential parameters and fit quality. These diagnostics establish whether τ_L is reported and complement the behavioral contrasts.

Table 5: Honesty-dependent investment and low-honesty updating across transition conditions under fixed returns. P_1, E_1, and G_1 measure initial differentiation, while $P_{16:20}$ and U_L measure persistence and updating. Brackets give 95% trial-bootstrap intervals. Asterisks mark Holm-adjusted $p < .05$. Units are tokens, with G_1 expressed per 100 percentage points of honesty.

Transition condition	P_1	E_1	G_1	$P_{16:20}$	U_L
DeepSeek V4 Pro					
No evaluation	0.84* [0.42, 1.24]	1.20* [0.47, 1.92]	1.15* [0.57, 1.71]	0.20 [-0.19, 0.57]	3.67* [3.19, 4.15]
Color control	0.57* [0.23, 0.94]	0.85* [0.28, 1.40]	0.80* [0.33, 1.27]	0.13 [-0.27, 0.54]	3.71* [3.33, 4.08]
Explicit evaluation	4.88* [4.50, 5.24]	7.02* [6.58, 7.43]	6.71* [6.30, 7.11]	3.43* [2.64, 4.16]	3.72* [3.11, 4.38]
Qwen3-32B					
No evaluation	0.15 [0.04, 0.30]	0.29 [0.07, 0.60]	0.23 [0.06, 0.48]	-0.07 [-0.27, 0.14]	2.02* [1.42, 2.65]
Color control	0.09 [0.00, 0.27]	0.09 [0.00, 0.27]	0.11 [0.00, 0.33]	0.50 [0.18, 0.86]	1.43* [0.91, 2.00]
Explicit evaluation	5.51* [4.60, 6.37]	6.00* [4.98, 6.96]	6.81* [5.70, 7.83]	1.07* [0.41, 1.81]	5.85* [4.95, 6.68]
Qwen3.5-27B					
No evaluation	3.02* [2.23, 3.82]	4.78* [3.67, 5.93]	4.33* [3.30, 5.41]	0.60* [0.23, 1.04]	5.12* [4.40, 5.85]
Color control	1.14* [0.53, 1.83]	2.15* [1.07, 3.32]	1.77* [0.86, 2.75]	0.25 [-0.02, 0.60]	4.12* [3.52, 4.72]
Explicit evaluation	6.47* [5.74, 7.16]	8.08* [7.35, 8.68]	8.41* [7.58, 9.17]	2.84* [1.97, 3.73]	5.24* [4.17, 6.31]
Qwen3.8-27B					
No evaluation	1.88* [1.32, 2.45]	2.18* [1.45, 2.92]	2.37* [1.67, 3.10]	2.55* [1.75, 3.36]	3.67* [2.79, 4.58]
Color control	2.00* [1.49, 2.50]	2.45* [1.82, 3.08]	2.58* [1.95, 3.21]	1.79* [1.00, 2.52]	4.13* [3.19, 5.07]
Explicit evaluation	2.41* [1.89, 2.90]	3.72* [2.88, 4.60]	3.41* [2.67, 4.15]	3.41* [2.47, 4.33]	4.70* [3.67, 5.75]
Qwen3-14B					
No evaluation	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	5.00* [5.00, 5.00]
Color control	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	5.00* [5.00, 5.00]
Explicit evaluation	0.60* [0.35, 0.87]	0.73* [0.40, 1.10]	0.77* [0.45, 1.12]	0.00 [0.00, 0.00]	5.63* [5.37, 5.92]
Qwen3.5-9B					
No evaluation	-0.01 [-0.07, 0.05]	-0.07 [-0.17, 0.00]	-0.03 [-0.12, 0.03]	0.15 [-0.07, 0.39]	1.07* [0.63, 1.49]
Color control	-0.02 [-0.05, 0.00]	-0.03 [-0.10, 0.00]	-0.03 [-0.08, 0.00]	-0.03 [-0.15, 0.10]	0.52* [0.12, 0.89]
Explicit evaluation	0.08 [-0.03, 0.20]	0.13 [-0.03, 0.33]	0.12 [-0.03, 0.28]	0.27 [0.02, 0.53]	0.48 [-0.02, 0.96]

Continued on next page

Table 5 (continued; tokens, G_1 per 100 percentage points of honesty)

Transition condition	P_1	E_1	G_1	$P_{16:20}$	U_L
Qwen3.5-4B					
No evaluation	-0.13 [-0.25, -0.03]	-0.03 [-0.10, 0.00]	-0.12 [-0.23, -0.04]	0.03 [-0.16, 0.28]	2.19* [1.73, 2.64]
Color control	0.03 [-0.05, 0.13]	-0.10 [-0.23, 0.00]	-0.01 [-0.12, 0.08]	-0.04 [-0.14, 0.06]	2.52* [2.17, 2.87]
Explicit evaluation	-0.23* [-0.43, -0.08]	-0.20 [-0.40, -0.03]	-0.27* [-0.51, -0.08]	-0.04 [-0.17, 0.10]	1.88* [1.44, 2.31]

Table 6: Effects of Explicit evaluation on honesty sensitivity, penalty persistence, and low-honesty updating under fixed returns. Rows subtract the named control. Units are tokens; ΔG_1 is per 100 percentage points of honesty and Δs_L is per round (rounds 1–5). Brackets give 95% paired-trial bootstrap intervals. Asterisks mark Holm-adjusted $p < .05$.

Control subtracted	ΔG_1	$\Delta P_{16:20}$	ΔU_L	Δs_L
DeepSeek V4 Pro				
No evaluation	5.56* [4.91, 6.18]	3.23* [2.37, 4.05]	0.05 [-0.70, 0.88]	-0.21* [-0.32, -0.09]
Color control	5.91* [5.29, 6.53]	3.30* [2.42, 4.13]	0.01 [-0.68, 0.71]	-0.28* [-0.37, -0.19]
Qwen3-32B				
No evaluation	6.57* [5.50, 7.59]	1.15* [0.44, 1.95]	3.83* [2.94, 4.68]	0.19 [0.04, 0.33]
Color control	6.70* [5.60, 7.76]	0.57 [-0.25, 1.45]	4.42* [3.50, 5.28]	0.32* [0.17, 0.47]
Qwen3.5-27B				
No evaluation	4.08* [2.96, 5.21]	2.24* [1.51, 2.97]	0.12 [-0.93, 1.18]	-0.21 [-0.42, -0.01]
Color control	6.63* [5.47, 7.74]	2.59* [1.76, 3.43]	1.12 [0.03, 2.24]	-0.09 [-0.30, 0.12]
Qwen3.8-27B				
No evaluation	1.04* [0.35, 1.78]	0.85 [-0.12, 1.77]	1.04* [0.33, 1.77]	0.13 [0.03, 0.23]
Color control	0.83 [0.01, 1.67]	1.62* [0.79, 2.44]	0.57 [-0.23, 1.39]	0.08 [-0.01, 0.18]
Qwen3-14B				
No evaluation	0.77* [0.45, 1.12]	0.00 [0.00, 0.00]	0.63* [0.37, 0.92]	-0.06* [-0.10, -0.02]
Color control	0.77* [0.45, 1.12]	0.00 [0.00, 0.00]	0.63* [0.37, 0.92]	-0.07* [-0.11, -0.03]
Qwen3.5-9B				
No evaluation	0.15 [-0.02, 0.35]	0.12 [-0.21, 0.44]	-0.58 [-1.12, -0.10]	-0.05 [-0.16, 0.06]
Color control	0.15 [-0.01, 0.31]	0.29 [0.01, 0.60]	-0.03 [-0.48, 0.40]	0.03 [-0.10, 0.16]

Continued on next page

Table 6 (continued; tokens, ΔG_1 per 100 percentage points, Δs_L per round)

Control subtracted	ΔG_1	$\Delta P_{16:20}$	ΔU_L	Δs_L
Qwen3.5-4B				
No evaluation	-0.15 [-0.37, 0.04]	-0.07 [-0.30, 0.12]	-0.32 [-0.61, -0.02]	-0.07 [-0.16, 0.02]
Color control	-0.25 [-0.51, -0.04]	0.00 [-0.16, 0.16]	-0.64* [-0.99, -0.31]	-0.10 [-0.22, 0.00]

Table 7: Exponential convergence parameters and fit quality for low-honesty investment under fixed returns. L_∞ and A_L are in tokens, and τ_L is in rounds. Dashes denote $R^2 < .90$, $|A_L| < .50$, or undefined fit quality.

Transition condition	L_∞	A_L	τ_L	R^2
DeepSeek V4 Pro				
No evaluation	9.22	-3.88	2.70	0.989
Color control	9.14	-3.80	2.16	0.992
Explicit evaluation	5.59	-4.04	4.58	0.989
Qwen3-32B				
No evaluation	7.15	-2.16	2.29	0.977
Color control	6.46	-1.56	2.48	0.971
Explicit evaluation	8.77	-6.72	4.99	0.964
Qwen3.5-27B				
No evaluation	9.43	-5.21	3.25	0.999
Color control	9.64	-4.12	2.68	0.999
Explicit evaluation	7.17	-5.86	4.87	0.987
Qwen3.8-27B				
No evaluation	5.28	-4.25	4.38	0.967
Color control	6.18	-4.80	4.68	0.974
Explicit evaluation	6.03	-5.40	4.73	0.980
Qwen3-14B				
No evaluation	10.04	-5.22	1.56	0.986
Color control	10.04	-5.22	1.45	0.982
Explicit evaluation	10.02	-5.76	2.20	0.990
Qwen3.5-9B				
No evaluation	6.23	-1.07	—	0.241
Color control	5.60	0.75	—	0.159
Explicit evaluation	4.56	1.70	—	0.208
Qwen3.5-4B				
No evaluation	6.19	-1.93	—	0.628
Color control	6.31	-2.18	—	0.600
Explicit evaluation	6.44	-1.60	—	0.475

D BEHAVIORAL RESULTS UNDER VARIABLE RETURNS

Variable returns test whether the history effects and Explicit evaluation contrasts persist when economic feedback fluctuates around the same temporal mean. Rates range from 0.56 to 0.64 with mean 0.60 and remain matched across histories and transition conditions. Figure 10 tracks penalties, and Figures 11–13 show investments under each transition. Tables 8 and 9 quantify late persistence, updating, and paired condition differences.

D.1 RETURN-SCHEDULE ROBUSTNESS AND INTERPRETATION

Variable returns preserve updating alongside history persistence. All five medium-to-large models have positive $P_{16:20}$ and U_L under Explicit evaluation (Holm-adjusted $p < .05$ for both). Qwen3-14B differs from its fixed-return pattern. Its late penalty is 2.49 tokens [1.57, 3.39], whereas its fixed-return gap vanishes. Paired ΔP_1 remains significant against both controls in the same four medium-to-large models after Holm correction. Neither small model shows a reliable positive P_1 under Explicit evaluation alone.

D.2 TRAJECTORIES AND CONDITION ENDPOINTS

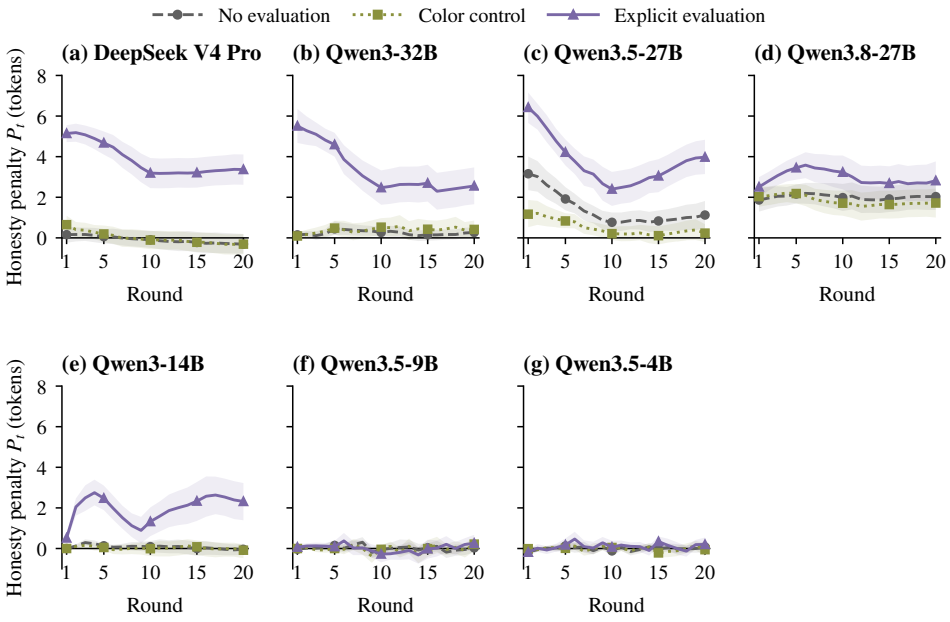

Figure 10: Honesty penalties under variable returns across the three transition conditions. Curves show $P_t = H_t - L_t$ over rounds. Bands are 95% trial-bootstrap intervals.

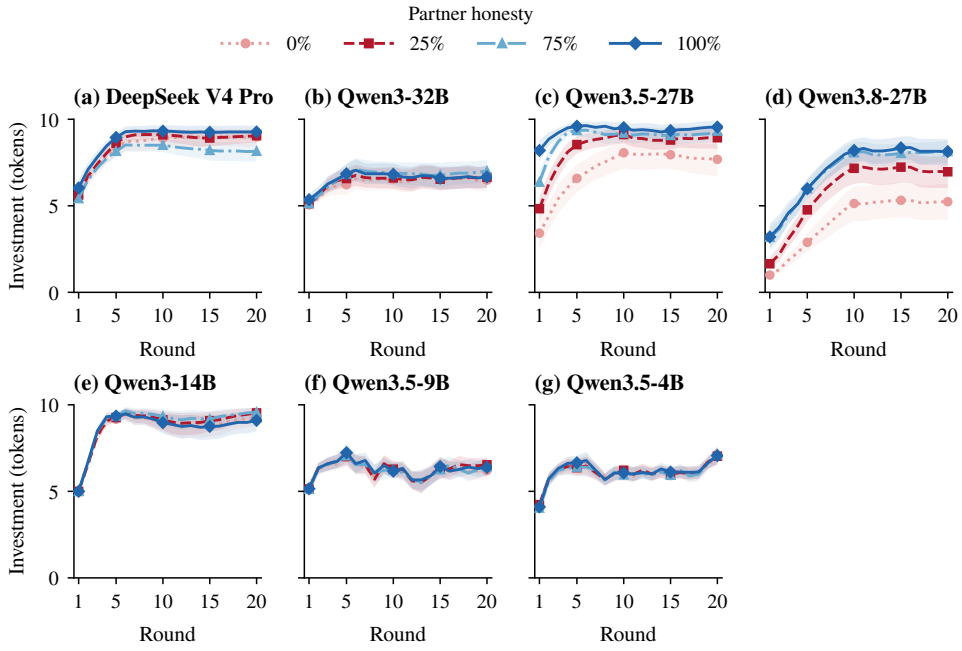

Figure 11: Variable-return investment under No evaluation. Bands are 95% trial-bootstrap intervals. Return fractions have temporal mean 0.60 and are matched across histories.

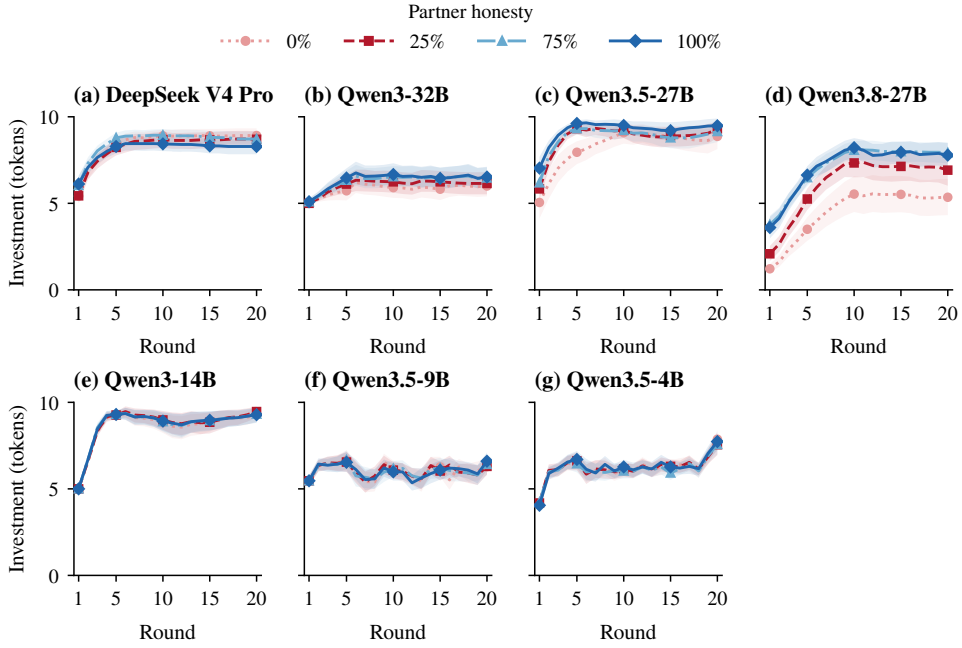

Figure 12: Variable-return investment under Color control. Bands are 95% trial-bootstrap intervals. Return fractions have temporal mean 0.60 and are matched across histories.

Table 8: Honesty-dependent investment and low-honesty updating across transition conditions under variable returns. Fractions range from 0.56 to 0.64 with mean 0.60. Brackets give 95% trial-bootstrap intervals in tokens. Asterisks mark Holm-adjusted $p < .05$.

Transition condition	P_1	$P_{16:20}$	U_L
DeepSeek V4 Pro			
No evaluation	0.16 [-0.24, 0.56]	-0.28 [-0.76, 0.23]	3.42* [3.04, 3.77]
Color control	0.65* [0.19, 1.09]	-0.29 [-0.78, 0.20]	3.38* [2.97, 3.77]
Explicit evaluation	5.14* [4.72, 5.55]	3.33* [2.60, 4.06]	3.80* [3.18, 4.42]
Qwen3-32B			
No evaluation	0.15 [0.04, 0.30]	0.20 [-0.05, 0.45]	1.48* [0.91, 2.10]
Color control	0.09 [0.00, 0.28]	0.43 [0.03, 0.87]	1.08* [0.55, 1.65]
Explicit evaluation	5.51* [4.68, 6.35]	2.43* [1.53, 3.32]	4.48* [3.37, 5.59]
Qwen3.5-27B			
No evaluation	3.16* [2.35, 4.01]	0.99* [0.40, 1.61]	4.21* [3.46, 4.95]
Color control	1.17* [0.53, 1.87]	0.29 [-0.15, 0.73]	3.44* [2.77, 4.09]
Explicit evaluation	6.42* [5.66, 7.16]	3.71* [2.92, 4.49]	3.97* [3.07, 4.88]
Qwen3.8-27B			
No evaluation	1.87* [1.29, 2.45]	2.01* [1.37, 2.67]	4.81* [3.93, 5.68]
Color control	2.03* [1.53, 2.52]	1.70* [1.02, 2.38]	4.56* [3.71, 5.43]
Explicit evaluation	2.51* [2.01, 2.99]	2.73* [1.87, 3.62]	4.72* [3.87, 5.61]
Qwen3-14B			
No evaluation	0.00 [0.00, 0.00]	-0.02 [-0.31, 0.26]	4.22* [3.75, 4.62]
Color control	0.00 [0.00, 0.00]	-0.02 [-0.41, 0.27]	4.17* [3.74, 4.55]
Explicit evaluation	0.52* [0.28, 0.77]	2.49* [1.57, 3.39]	2.09* [1.19, 2.99]
Qwen3.5-9B			
No evaluation	-0.03 [-0.12, 0.03]	-0.04 [-0.20, 0.13]	1.18* [0.79, 1.55]
Color control	0.00 [0.00, 0.00]	0.06 [-0.09, 0.22]	0.58* [0.22, 0.93]
Explicit evaluation	0.07 [-0.05, 0.20]	0.14 [-0.06, 0.33]	0.66* [0.27, 1.06]
Qwen3.5-4B			
No evaluation	-0.17 [-0.32, -0.05]	0.02 [-0.10, 0.13]	2.14* [1.70, 2.57]
Color control	-0.01 [-0.12, 0.11]	-0.04 [-0.15, 0.08]	2.55* [2.18, 2.93]
Explicit evaluation	-0.18 [-0.37, -0.05]	0.08 [-0.06, 0.24]	2.02* [1.58, 2.45]

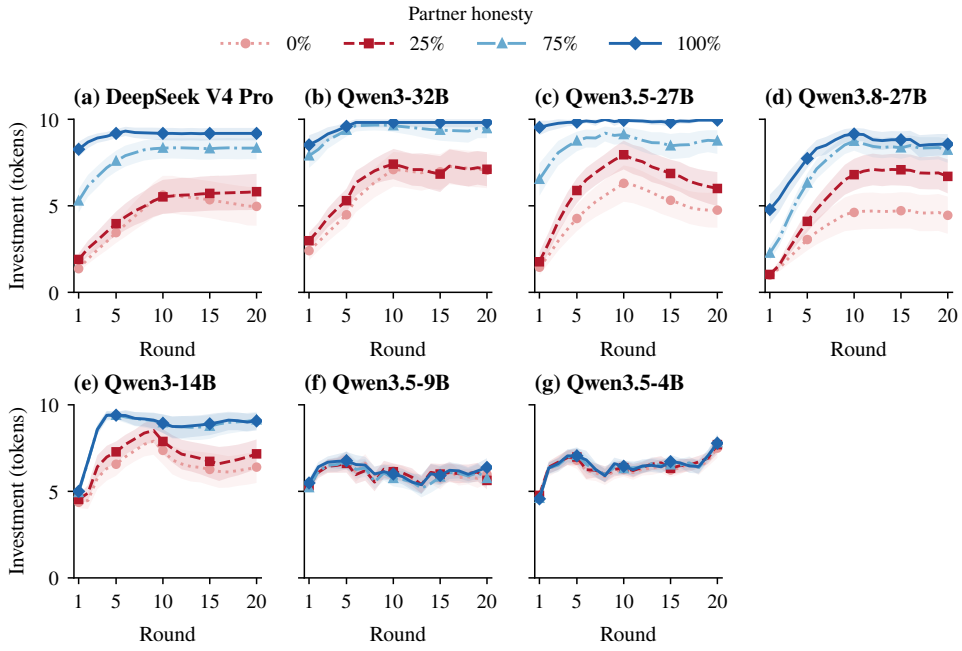

Figure 13: Variable-return investment under Explicit evaluation. Bands are 95% trial-bootstrap intervals. Return fractions have temporal mean 0.60 and are matched across histories.

Table 9: Effects of Explicit evaluation on initial and late honesty penalties under variable returns. Paired changes in P_1 and $P_{16:20}$ subtract the named control. Brackets give 95% paired-trial bootstrap intervals. Asterisks mark Holm-adjusted $p < .05$.

Model	ΔP_1 vs No evaluation	ΔP_1 vs Color control	$\Delta P_{16:20}$ vs Color control
DeepSeek V4 Pro	4.98* [4.46, 5.50]	4.49* [3.89, 5.09]	3.62* [2.68, 4.51]
Qwen3-32B	5.36* [4.55, 6.18]	5.42* [4.55, 6.30]	2.00* [1.04, 2.93]
Qwen3.5-27B	3.27* [2.28, 4.24]	5.26* [4.31, 6.16]	3.42* [2.57, 4.27]
Qwen3.8-27B	0.64 [0.10, 1.20]	0.47 [-0.12, 1.07]	1.04* [0.35, 1.70]
Qwen3-14B	0.52* [0.28, 0.77]	0.52* [0.28, 0.77]	2.51* [1.64, 3.38]
Qwen3.5-9B	0.10 [-0.05, 0.28]	0.07 [-0.05, 0.20]	0.07 [-0.19, 0.33]
Qwen3.5-4B	-0.02 [-0.23, 0.18]	-0.17 [-0.38, 0.01]	0.12 [-0.04, 0.29]

D.3 Paired contrasts and descriptive cross-schedule comparison

Figure 14 compares P_1, $P_{16:20}$, and ΔP_1 across schedules descriptively. All reported significance tests remain within a schedule. The first investment precedes return feedback, so differences in P_1 between schedules do not measure responses to return variability. The added evidence concerns the subsequent trajectories, especially the persistence that is absent for Qwen3-14B under fixed returns.

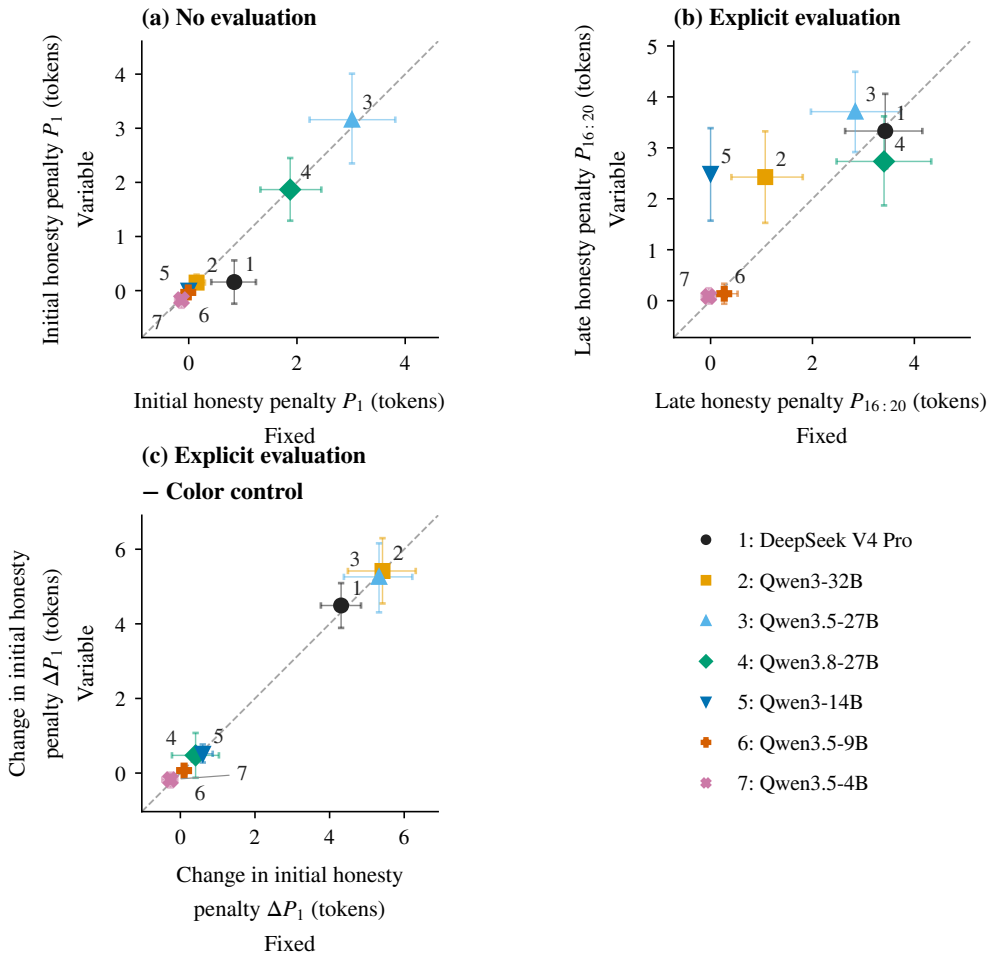

Figure 14: Behavioral measures across fixed and variable returns. Numbered points identify matched models. Panels compare No evaluation P_1, Explicit evaluation $P_{16:20}$, and ΔP_1 over Color control. Bars are 95% trial-bootstrap intervals. Dashed identity lines mark equal values, and the legend follows model order.

E CONSTRUCTION AND VALIDATION OF SOCIAL IMPRESSION SCORES

This analysis tests whether partner descriptions remain sensitive to honesty when investment shows no clear positive initial penalty. The social impression score S summarizes honesty-related variation in descriptor probabilities across three prompt templates.

E.1 MATERIALS, DATA SPLITS, AND EVALUATION-PROMPT TEMPLATES

Discovery, validation, and confirmation use 8, 8, and 188 disjoint materials, respectively. Each material is evaluated at four honesty levels using three Explicit evaluation prompt templates. Discovery constructs and calibrates the descriptor sets; validation checks their direction, template robustness, and output boundaries; confirmation evaluates the frozen score on independent materials. These material counts are separate from the 60 trials in the behavioral experiment.

The following templates are alternative user requests after the Stage 1 honesty history. Each requests one adjective. The score uses probabilities at the answer-start position rather than the frequency of generated adjectives; generation validity is assessed separately during validation. In the behavioral Explicit evaluation condition, the Describe response is retained in the conversation as specified in Appendix A.

Describe.

Describe the partner with exactly one English adjective. Reply with one English adjective only, with no spaces, punctuation, or explanation.

Give.

Give exactly one English adjective that describes the partner. Reply with one adjective only, with no spaces, punctuation, or explanation.

Choose.

Choose exactly one English adjective to describe the partner. Output that adjective only, with no spaces, punctuation, or explanation.

E.2 COMPLETE-WORD PROBABILITY AND TOKEN-BOUNDARY HANDLING

A token path is the tokenizer sequence for an allowed written form of a descriptor. For each word, we consider canonical lowercase and initial-capital forms, with and without a leading space. A permitted path must round-trip exactly through the tokenizer and contain no special tokens. Duplicate token sequences are counted once.

For a word w under template p, complete-word probability $Q_{w,p}$ is computed in three steps:

1. Multiply the autoregressive next-token probabilities along each permitted token path to obtain its path mass.
2. Multiply that mass by the probability of a following word boundary. Boundary categories are EOS, Unicode whitespace, or Unicode punctuation.
3. Sum the completed masses across the permitted, deduplicated paths for the word.

Thus $Q_{w,p}$ is a complete-word probability, not a first-token probability or a generated-word frequency. Computation uses float64 log-space arithmetic without length normalization. The final retained sets have 80 paths for Qwen3.5-9B and 36 for Qwen3.5-4B; all retained paths have 1–3 tokens.

E.3 CANDIDATE CONSTRUCTION AND DISCOVERY-STAGE SCREENING

Candidates come from Open English WordNet 2025 adjective and adjective-satellite senses. We construct and screen the candidate family as follows:

1. **Lexical and tokenizer eligibility.** Retain alphabetic words with optional internal hyphens, 3–24 characters, and English Zipf frequency at least 3.0. Canonical tokenizer round trips and a seven-token path cap yield 7,604 eligible words for each small model.
2. **Variation filter.** Exclude words whose discovery log-probability standard deviation is below 10^{-3} in any template.
3. **Label-blind candidate family.** Take the union of each context's top 200 words and the top 2,000 words by mean log probability.
4. **Two screening quantities.** Estimate honesty slopes for log complete-word probability and for the reciprocal rank of its discovery-standardized value within the candidate family. Ranks beyond 50 are assigned zero. Both quantities average the three templates within each material and honesty level.
5. **Joint selection criteria.** Require matching slope signs, BH-adjusted $q \leq .05$ for both quantities, both 95% bootstrap intervals excluding zero, and bootstrap sign agreement of at least 95%.

Tests use 10,000 within-material honesty permutations, with BH correction over the candidate family separately for each quantity. Intervals and sign agreement use 2,000 whole-material bootstrap draws. These criteria select stable directional associations across the four honesty levels.

E.4 Effect-blind semantic review

A Qwen3-32B reviewer evaluates each word's supplied WordNet senses under two deterministic prompts. The reviewer sees those senses but not model identity, honesty conditions, effect sizes, directions, or ranks. A word is removed only when both prompts endorse exclusion as a person-level social descriptor and both cite valid supplied sense identifiers. This semantic review is effect-blind; the preceding discovery screening uses the measured honesty associations.

E.5 Validation, ranking, and frozen descriptor sets

Validation applies three word-level checks:

- **Direction retention.** Both mean slopes must retain their discovery direction.
- **Template robustness.** For each template, divide the direction-oriented log-probability slope by its across-material standard deviation. The result must be positive in at least two templates and at least $-.20$ in every template.
- **Boundary checks.** Require a next-token boundary top-1 rate of at least .80 and median word-boundary probability of at least .80 across validation contexts.

Separately, valid one-word adjective generation must reach 95% overall and 90% within each template. Both models pass this interface check; it is distinct from the word-level filters.

Surviving words are ordered by the following priorities:

1. Lower maximum q across the two screening quantities.
2. Higher minimum sign stability.
3. Higher minimum absolute standardized slope.
4. Higher mean log probability.
5. Lexical order.

We take at most the first 20, without direction quotas or replacement. Qwen3.5-9B retains 20 of 26 validation survivors, and Qwen3.5-4B retains all 9. Their empirical honesty-slope signs define the increasing and decreasing descriptor sets V_+ and V_- in Table 10. Set membership follows these empirical signs rather than an intuitive division into positive and negative words. Descriptor sets, token paths, and discovery calibration parameters are fixed before confirmation and remain unchanged when scoring confirmation materials.

Table 10: Model-specific descriptor sets for social impression scores. V_+ and V_- contain descriptors with increasing and decreasing discovery associations with honesty. Descriptors receive equal weight within each set. Token paths and calibration remain fixed for confirmation.

Model	Set	Descriptors
Qwen3.5-9B	V_+	victorious, reliable, compliant, charismatic, honest
	V_-	pathological, distressed, rude, useless, fictitious, pathetic, vicious, destitute, notorious, wicked, stupid, nasty, crooked, deliberate, derogatory
Qwen3.5-4B	V_+	relevant, reliable, consistent
	V_-	useless, fictitious, failing, counterfeit, fake, discreet

E.6 Discovery calibration and score aggregation

Discovery data determine the mean $\mu_{w,p}$ and standard deviation $\sigma_{w,p}$ of $\log Q_{w,p}$ separately for every word w and prompt template p. The standardized evidence is

$$z_{w,p} = \frac{\log Q_{w,p} - \mu_{w,p}}{\sigma_{w,p}}. \tag{18}$$

Within each material and honesty level, we first average standardized evidence equally within each descriptor set, then subtract the decreasing-set mean from the increasing-set mean:

$$S_p = \frac{1}{|V_+|} \sum_{w \in V_+} z_{w,p} - \frac{1}{|V_-|} \sum_{w \in V_-} z_{w,p}. \tag{19}$$

Finally, we average the three template scores with equal weight:

$$S = \frac{1}{3} \sum_{p=1}^{3} S_p. \tag{20}$$

Calibration precedes aggregation and remains fixed from discovery. The set means give equal weight to descriptors within each set, regardless of the two set sizes. Equation 5 in the main text suppresses the template subscript; the equations here make the template-specific calibration and final averaging explicit.

E.7 Confirmation results and score-specific inference

Confirmation uses the 188 independent materials at all four honesty levels, with the descriptor sets, token paths, and calibration parameters already frozen. Table 11 reports template-averaged and template-specific honesty slopes and Spearman correlations; Table 12 reports the score distributions at each honesty level.

Honesty-slope intervals use 5,000 material-bootstrap resamples, keeping honesty levels and templates together. Spearman tests use 10,000 two-sided permutations of honesty labels within materials, with a plus-one correction. These score p values are unadjusted and are separate from the behavioral Holm families. The descriptive means and standard deviations summarize score distributions rather than uncertainty intervals for the honesty slope.

Table 11: Honesty sensitivity of social impression scores on confirmation materials. Slopes are per 100 percentage points of honesty; brackets give 95% material-bootstrap intervals. Spearman correlations use unadjusted two-sided within-material permutation p values. The template average gives equal weight to Describe, Give, and Choose.

Template	Honesty slope [95% CI]	Spearman ρ	p
Qwen3.5-9B			
Template average	4.321 [4.276, 4.367]	0.965	$< .001$
Describe	4.285 [4.239, 4.331]	0.964	$< .001$
Give	4.302 [4.253, 4.349]	0.965	$< .001$
Choose	4.375 [4.331, 4.418]	0.965	$< .001$
Qwen3.5-4B			
Template average	4.615 [4.552, 4.680]	0.959	$< .001$
Describe	4.491 [4.423, 4.562]	0.952	$< .001$
Give	4.655 [4.590, 4.721]	0.959	$< .001$
Choose	4.699 [4.636, 4.764]	0.960	$< .001$

Table 12: Distributions of social impression scores on confirmation materials. Entries give the material-level mean ± standard deviation at each honesty level. The template average equally weights the three template scores within each material.

Template	0%	25%	75%	100%
Qwen3.5-9B				
Template average	-1.973 ± 0.106	-0.958 ± 0.419	0.464 ± 0.479	2.717 ± 0.208
Describe	-2.006 ± 0.125	-0.901 ± 0.457	0.463 ± 0.449	2.669 ± 0.229
Give	-1.959 ± 0.109	-0.915 ± 0.424	0.464 ± 0.467	2.729 ± 0.219
Choose	-1.954 ± 0.103	-1.057 ± 0.388	0.464 ± 0.538	2.755 ± 0.187
Qwen3.5-4B				
Template average	-2.308 ± 0.242	-0.835 ± 0.611	1.141 ± 0.783	2.473 ± 0.173
Describe	-2.377 ± 0.291	-0.796 ± 0.601	1.148 ± 0.853	2.265 ± 0.190
Give	-2.260 ± 0.223	-0.833 ± 0.607	1.084 ± 0.775	2.601 ± 0.222
Choose	-2.286 ± 0.240	-0.876 ± 0.665	1.191 ± 0.751	2.554 ± 0.160

F LINEAR DECODABILITY ACROSS TASK STAGES

This analysis tests whether impression-related information remains linearly decodable as the conversation moves from history observation to investment. We define the four task positions, extract their hidden states, and evaluate independent ridge readouts against a common reference target.

F.1 T0–T3 AND THE COMMON REFERENCE TARGET

- **T0:** after final Stage 1 history verification, before the Explicit evaluation prompt.
- **T1:** the assistant answer-start position after that prompt, before adjective generation.
- **T2:** after adjective writeback and the between-stage transition, before the Stage 2 investment introduction.
- **T3:** after the Stage 2 introduction, before the first investment request.

At each stage, an independent readout predicts the same T1 reference social impression score S_{ref}.

F.2 HIDDEN-STATE EXTRACTION AND INDEPENDENT RIDGE READOUTS

We extract the last valid token's hidden state from a full-prefix forward pass in evaluation mode with bfloat16 weights and caching disabled. The 32 transformer blocks use zero-based indices 0–31. Embedding outputs are excluded. Each time point has an independent ridge readout. All observations sharing a paired-trial identifier remain in the same cross-validation fold, so held-out predictions do not share a trial with the training data. Decodability at different time points measures information availability. It does not establish transfer of a single decoder or use by the investment policy.

F.3 GROUPED NESTED CROSS-VALIDATION AND MODEL SELECTION

Five outer grouped folds estimate held-out performance. Four inner grouped folds select the layer and regularization. The regularization grid contains 25 log-spaced values from 10^{-6} to 10^{6}. Layer selection first searches nine evenly spaced blocks, then the blocks within three positions of the best coarse block. Selection maximizes inner-fold Pearson correlation, with ties resolved by R^2, proximity of regularization to 1, and then lower layer index. Hidden-state features are standardized using training-fold means and standard deviations only. The target is centered on its training mean.

F.4 HELD-OUT PERFORMANCE, UNCERTAINTY, AND SIGNIFICANCE TESTS

Table 13 summarizes pooled out-of-fold performance corresponding to Figure 4. Pearson intervals use 5,000 paired-trial bootstrap draws of held-out predictions. Decodability tests use 10,000 one-sided permutations of complete target groups across paired trials, testing positive correlation. The four baseline stages form a Holm family within each model. The base seed is 20260920041, with a separate offset for each time point.

Table 13: Linear decodability of the T1 social impression score S_{ref} across task stages. Metrics pool held-out predictions from independent grouped nested cross-validation at each time point. Brackets give 95% paired-trial bootstrap intervals for Pearson correlation.

Stage	Pearson r [95% CI]	Spearman ρ	R^2
Qwen3.5-9B			
T0	0.987 [0.984, 0.989]	0.958	0.973
T1	0.999 [0.998, 0.999]	0.993	0.998
T2	0.985 [0.981, 0.988]	0.965	0.970
T3	0.985 [0.981, 0.988]	0.963	0.970
Qwen3.5-4B			
T0	0.978 [0.970, 0.984]	0.944	0.957
T1	0.999 [0.999, 0.999]	0.996	0.998
T2	0.982 [0.978, 0.985]	0.957	0.964
T3	0.977 [0.972, 0.983]	0.945	0.955

G Impression retrieval and relevance cues

These interventions test whether retrieval and relevance cues make an available social impression influence investment.

G.1 Intervention conditions, insertion point, and exact prompts

Each cue condition is compared with the main experiment's Explicit evaluation baseline, matched by model, return schedule, and trial identifier. The *impression retrieval cue* repeats the model's own adjective after the investment introduction. The *impression relevance cue* connects the prior social impression to the next interaction at the same point. The *combined impression cue* adds both sentences in that order. All three cues are user messages inserted after the Stage 2 introduction and before the first investment request. They do not request a separate reply; the cue text remains in the retained conversation. The next requested response is the investment.

Impression retrieval cue.

Your previous one-word description of this partner was: "{self_adjective}".

Impression relevance cue.

You may take your prior assessment of this same partner into account when deciding how cautiously to interact.

Combined impression cue (same user message).

Your previous one-word description of this partner was: "{self_adjective}".
You may take your prior assessment of this same partner into account when deciding how cautiously to interact.

G.2 Paired contrasts and recovery criterion

Cue effects are within-trial differences relative to Explicit evaluation from the main behavioral experiment. We use the percentile-bootstrap and two-sided sign-flip procedures in Appendix B, separately for each schedule, with Holm correction over all six model-by-cue contrasts per endpoint. Table 14 supplies condition uncertainty, Table 15 supplies fixed-return adjusted tests, and Table 16 supplies variable-return contrasts. Recovery denotes a positive P_1 together with a positive paired ΔP_1 supported after Holm correction.

G.3 Fixed-return and variable-return results

Figures 15 and 16 show fixed-return penalties and investments. Figure 17 shows the variable-return trajectories. The contrast between effective individual cues in Qwen3.5-9B and an effective combination in Qwen3.5-4B also appears under variable returns.

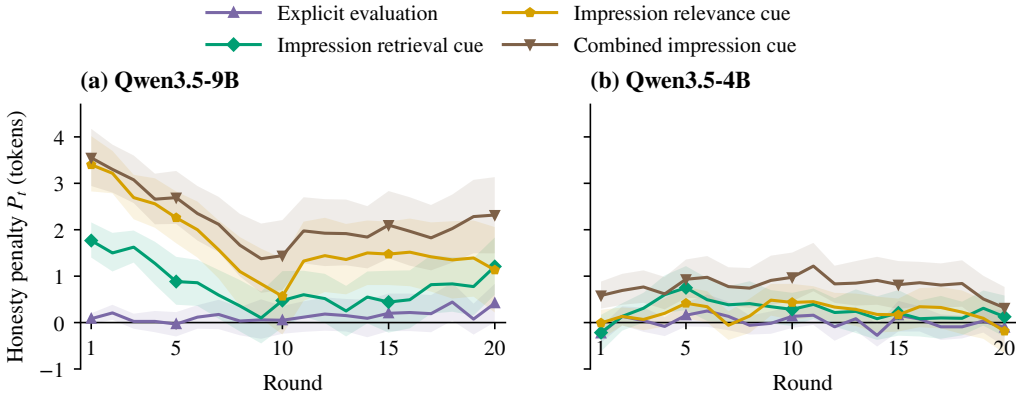

Figure 15: Honesty penalties after impression cues under fixed returns. Solid curves show mean P_t, and bands give 95% trial-bootstrap intervals.

Table 14: Honesty-dependent investment and low-honesty updating across impression cue conditions and return schedules. P_1, E_1, and G_1 measure initial differentiation, while $P_{16:20}$ and U_L measure persistence and updating. Brackets give 95% trial-bootstrap intervals. Explicit evaluation is the matched main-experiment baseline.

Condition	P_1	E_1	G_1	$P_{16:20}$	U_L
Qwen3.5-9B: fixed returns					
Explicit evaluation	0.08 [-0.03, 0.20]	0.13 [-0.03, 0.33]	0.12 [-0.03, 0.28]	0.27 [0.02, 0.53]	0.48 [-0.02, 0.96]
Impression retrieval cue	1.77 [1.40, 2.16]	2.85 [2.32, 3.42]	2.55 [2.06, 3.09]	0.82 [0.32, 1.33]	1.64 [1.07, 2.20]
Impression relevance cue	3.40 [2.83, 4.02]	5.50 [4.63, 6.37]	4.92 [4.14, 5.73]	1.36 [0.61, 2.11]	2.94 [2.36, 3.55]
Combined impression cue	3.55 [2.94, 4.17]	5.45 [4.67, 6.23]	5.02 [4.24, 5.81]	2.08 [1.38, 2.78]	2.36 [1.77, 2.98]

Continued on next page

Table 14 (continued)

Condition	P_1	E_1	G_1	$P_{16:20}$	U_L
Qwen3.5-9B: variable returns					
Explicit evaluation	0.07 [-0.05, 0.20]	0.23 [0.07, 0.43]	0.15 [-0.00, 0.32]	0.14 [-0.06, 0.33]	0.66 [0.27, 1.06]
Impression retrieval cue	1.83 [1.46, 2.23]	2.82 [2.27, 3.40]	2.59 [2.08, 3.13]	0.52 [0.08, 0.96]	2.10 [1.59, 2.60]
Impression relevance cue	3.42 [2.84, 4.03]	5.50 [4.65, 6.35]	4.93 [4.16, 5.73]	2.12 [1.50, 2.71]	2.65 [2.14, 3.20]
Combined impression cue	3.58 [2.98, 4.19]	5.52 [4.75, 6.28]	5.07 [4.31, 5.85]	2.27 [1.71, 2.85]	2.21 [1.70, 2.72]
Qwen3.5-4B: fixed returns					
Explicit evaluation	-0.23 [-0.43, -0.08]	-0.20 [-0.40, -0.03]	-0.27 [-0.51, -0.08]	-0.04 [-0.17, 0.10]	1.88 [1.44, 2.31]
Impression retrieval cue	-0.22 [-0.62, 0.17]	-0.18 [-0.65, 0.25]	-0.25 [-0.75, 0.23]	0.14 [-0.12, 0.44]	2.25 [1.75, 2.74]
Impression relevance cue	-0.01 [-0.35, 0.28]	0.20 [-0.20, 0.57]	0.07 [-0.35, 0.43]	0.16 [-0.24, 0.54]	4.04 [3.39, 4.70]
Combined impression cue	0.58 [0.30, 0.88]	0.82 [0.48, 1.18]	0.79 [0.44, 1.17]	0.67 [0.28, 1.05]	4.21 [3.54, 4.85]
Qwen3.5-4B: variable returns					
Explicit evaluation	-0.18 [-0.37, -0.05]	-0.20 [-0.40, -0.03]	-0.23 [-0.44, -0.05]	0.08 [-0.06, 0.24]	2.02 [1.58, 2.45]
Impression retrieval cue	-0.17 [-0.53, 0.19]	-0.07 [-0.48, 0.33]	-0.16 [-0.61, 0.28]	0.22 [0.00, 0.45]	2.32 [1.82, 2.81]
Impression relevance cue	0.07 [-0.26, 0.34]	0.22 [-0.18, 0.60]	0.15 [-0.27, 0.50]	0.40 [0.12, 0.70]	3.50 [2.95, 4.04]
Combined impression cue	0.61 [0.34, 0.90]	0.85 [0.50, 1.22]	0.83 [0.48, 1.21]	0.89 [0.59, 1.20]	3.35 [2.83, 3.82]

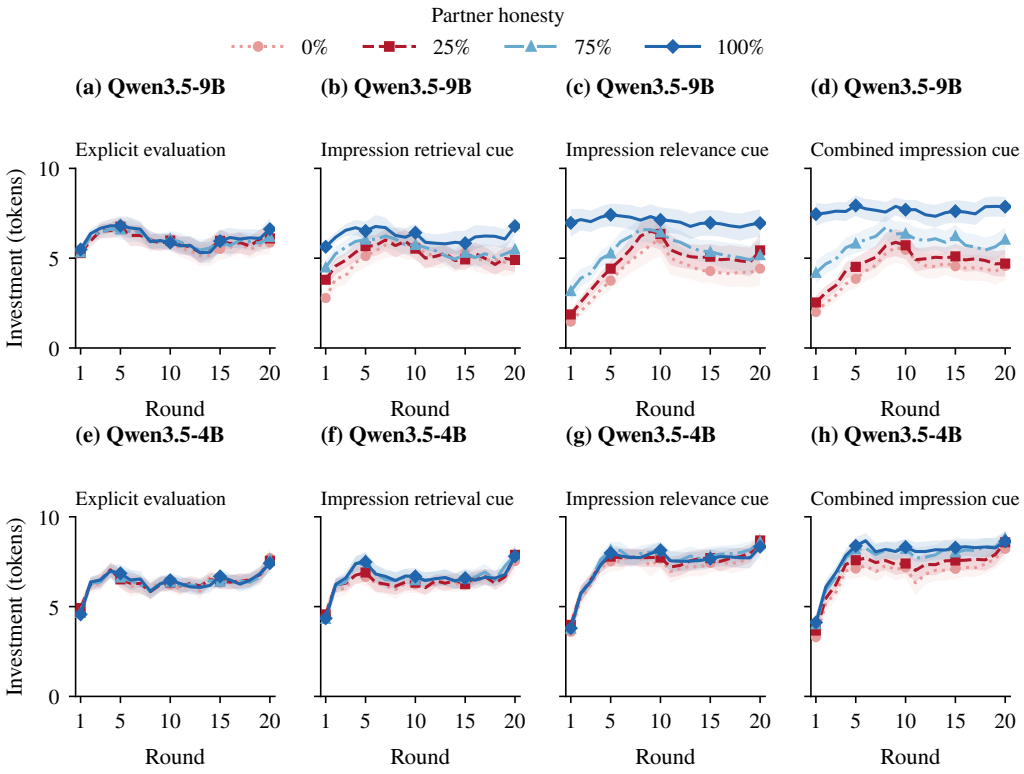

Figure 16: Fixed-return investment after impression cues. Bands are 95% trial-bootstrap intervals. Explicit evaluation is the matched main-experiment baseline.

Table 15: Paired significance tests for impression cue effects on initial and late honesty penalties under fixed returns. Columns give raw and Holm-adjusted paired sign-flip p values. Each endpoint is corrected over the six model-by-cue contrasts.

Condition	$p_{\Delta P_1}$	$p_{\mathrm{H},\Delta P_1}$	$p_{\Delta P_{16:20}}$	$p_{\mathrm{H},\Delta P_{16:20}}$
Qwen3.5-9B				
Impression retrieval cue	$< .001$	$< .001$	0.033	0.099
Impression relevance cue	$< .001$	$< .001$	0.017	0.070
Combined impression cue	$< .001$	$< .001$	$< .001$	$< .001$
Qwen3.5-4B				
Impression retrieval cue	0.960	0.960	0.249	0.499
Impression relevance cue	0.113	0.226	0.360	0.499
Combined impression cue	$< .001$	$< .001$	0.003	0.013

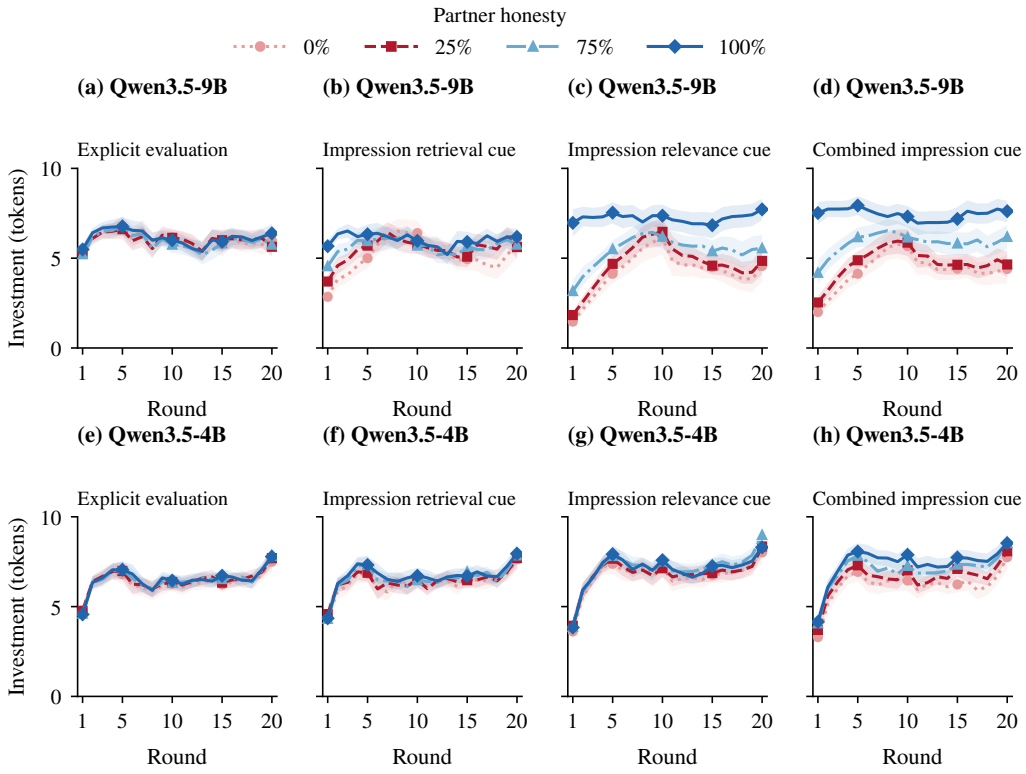

Figure 17: Variable-return investment after impression cues. Bands are 95% trial-bootstrap intervals. Explicit evaluation is the matched main-experiment baseline.

Table 16: Effects of impression cues on initial and late honesty penalties in small models under variable returns. Brackets give 95% paired-trial bootstrap intervals. Asterisks mark Holm-adjusted $p < .05$, with adjusted p values in the last two columns.

Condition	ΔP_1	$\Delta P_{16:20}$	$p_{H,\Delta P_1}$	$p_{H,\Delta P_{16:20}}$
Qwen3.5-9B				
Impression retrieval cue	1.77* [1.36, 2.19]	0.39 [-0.09, 0.86]	$< .001$	0.250
Impression relevance cue	3.35* [2.75, 3.98]	1.99* [1.39, 2.56]	$< .001$	$< .001$
Combined impression cue	3.52* [2.88, 4.18]	2.13* [1.54, 2.76]	$< .001$	$< .001$
Qwen3.5-4B				
Impression retrieval cue	0.02 [-0.28, 0.33]	0.14 [-0.08, 0.36]	0.959	0.255
Impression relevance cue	0.26 [0.03, 0.48]	0.32 [0.02, 0.62]	0.071	0.131
Combined impression cue	0.79* [0.51, 1.10]	0.80* [0.50, 1.12]	$< .001$	$< .001$

H EXPLORATORY INTERPERSONAL-ABILITY ASSOCIATIONS

This appendix examines model-level associations between interpersonal ability and trust behavior. We specify the matched model scores and six endpoints, their exact tests, and a separate gradient sensitivity analysis.

H.1 MODEL-LEVEL SCORES AND BEHAVIORAL ENDPOINTS

We match the seven current behavioral models to their IAE accuracy scores, with 1,988 items per model. The six principal endpoints are:

- Initial honesty penalty P_1 under No evaluation.
- Late honesty penalty $P_{16:20}$ under Explicit evaluation.
- Initial penalty increment ΔP_1, Explicit evaluation minus Color control.
- Early low-honesty slope change Δs_L, Explicit evaluation minus No evaluation.
- Low-honesty relaxation time τ_L under Explicit evaluation.
- Relaxation-time change $\Delta\tau_L$, Explicit evaluation minus No evaluation.

Time-constant analyses contain five identifiable models.

Table 17: Model-level interpersonal ability scores and matched trust-game measures under fixed returns. IAE is reported as accuracy (%); behavioral endpoints use fixed returns. P_1 uses No evaluation; $P_{16:20}$ and τ_L use Explicit evaluation. ΔP_1 compares Explicit evaluation minus Color control, while Δs_L and $\Delta\tau_L$ compare Explicit evaluation minus No evaluation.

Model	IAE	P_1	$P_{16:20}$	ΔP_1	Δs_L	τ_L	$\Delta\tau_L$
DeepSeek V4 Pro	75.15	0.84	3.43	4.31	-0.21	4.58	1.88
Qwen3-32B	71.63	0.15	1.07	5.42	0.19	4.99	2.70
Qwen3.5-27B	75.05	3.02	2.84	5.33	-0.21	4.87	1.62
Qwen3.8-27B	71.38	1.88	3.41	0.41	0.13	4.73	0.35
Qwen3-14B	69.06	0.00	0.00	0.60	-0.06	2.20	0.64
Qwen3.5-9B	66.45	-0.01	0.27	0.10	-0.05	—	—
Qwen3.5-4B	60.46	-0.13	-0.04	-0.27	-0.07	—	—

H.2 EXACT TESTS, MULTIPLICITY, AND SENSITIVITY ANALYSIS

Spearman correlations use model-level observations. Two-sided exact tests enumerate all permutations of model labels, and Benjamini–Hochberg correction is applied across the six principal endpoints. Replacing the initial high–low penalty with its four-level gradient is reported separately as a sensitivity analysis.

H.3 RESULTS AND LIMITS OF INTERPRETATION

Figure 18 uses interpersonal ability as complementary evidence about why penalty magnitude differs across models. Associations with initial penalty, late penalty, and initial evaluation amplification are positive, but their BH-adjusted q values are .068 (Table 18). These model-level associations are exploratory and do not identify a cause of the profile differences.

Table 18: Exploratory correlations between interpersonal ability and trust-game measures under fixed returns. Exact two-sided permutation tests and BH correction over six principal endpoints. The gradient substitution is a separate sensitivity analysis.

Endpoint	n	ρ	p	q
P_1: No evaluation	7	0.821	0.034	0.068
$P_{16:20}$: Explicit evaluation	7	0.857	0.024	0.068
ΔP_1: Explicit evaluation minus Color control	7	0.821	0.034	0.068
Δs_L: Explicit evaluation minus No evaluation	7	-0.286	0.556	0.667
τ_L: Explicit evaluation	5	0.300	0.683	0.683
$\Delta\tau_L$: Explicit evaluation minus No evaluation	5	0.600	0.350	0.525
G_1: No evaluation (sensitivity)	7	0.821	0.034	—

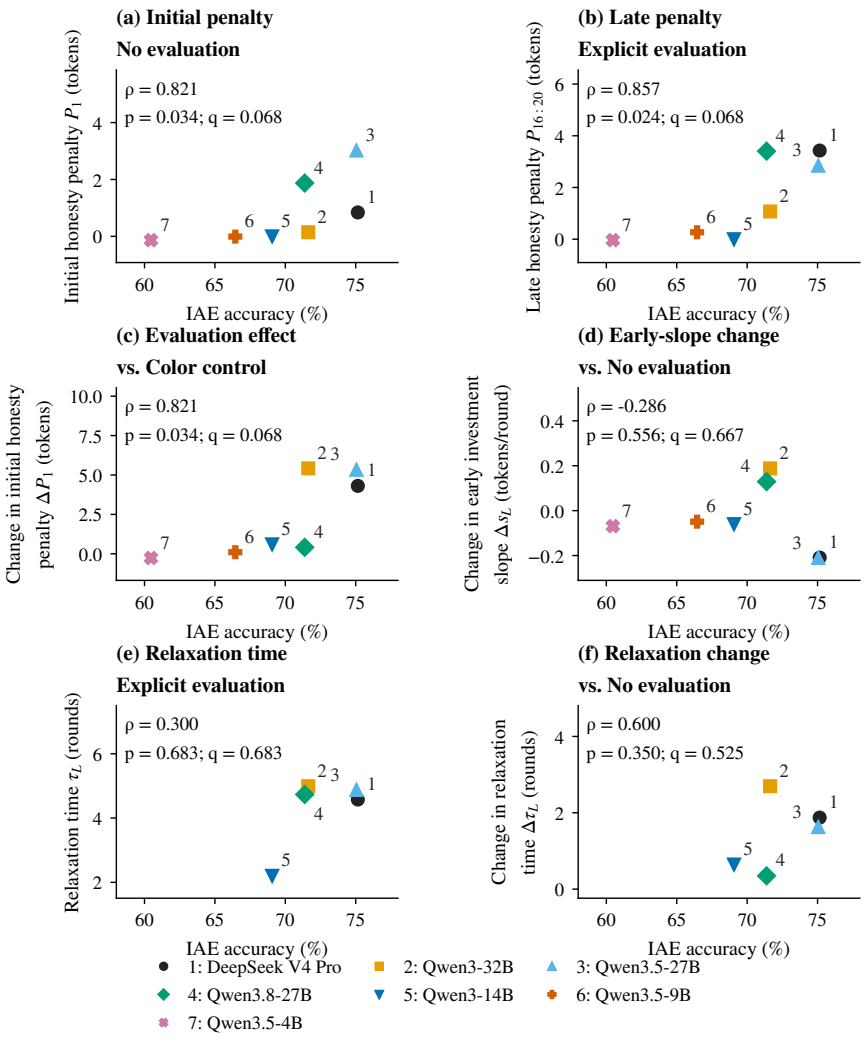

Figure 18: Interpersonal ability and trust behavior. Each point represents one model; labels and legend follow model size. Panels show six fixed-return endpoints, with Spearman ρ, exact two-sided permutation p, and BH-adjusted q. Time-constant panels contain the five models with identifiable fits. Tables 17 and 18 provide the values and sensitivity analysis.