Do LLM Agents Respond to Commitment?

A Behavioral Experiment in Strategic Communication

Ebrahim Aliabadi Farahani*

November 2026

Abstract

As AI agents increasingly make decisions and communicate on behalf of individuals and organizations, understanding how they respond to institutional constraints becomes important. Commitment can make strategic communication informative by restricting how a sender may revise its communication after learning private information. We study whether LLM agents respond to mechanically enforced commitment as economic theory predicts, using the sender-receiver environment of Fréchette, Lizzeri, and Perego (2022). The experiment varies commitment strength and message verifiability, allowing comparisons with theoretical predictions and existing human experimental evidence. We find substantial differences across model configurations. Claude Sonnet 4.5 exhibits the commitment-blindness pattern documented among human subjects, whereas Claude Opus 5.5 and Gemini 3.8 Flash behave much closer to theory under full commitment. Under partial commitment, even models that make individually sensible revision decisions may choose initial communication plans that fail to anticipate those decisions. Our findings highlight the importance of evaluating strategic AI agents under different institutional arrangements rather than inferring their behavior from performance in a single environment.

1. Introduction

In strategic communication, commitment means placing rules on how a sender can communicate before the sender knows what it will later want to say. These rules limit the sender's ability to change its message simply to push the receiver toward a preferred action. By reducing this conflict between what the sender wants and what the receiver can trust, commitment can make communication more credible and more informative. This idea appears in many economic settings, from disclosure policies and contracts to platform rules and automated communication systems.

Fréchette, Lizzeri, and Perego (2022, FLP) provide a particularly clean experimental demonstration of this idea. In their sender-receiver game, a sender wants to influence the receiver's action, while the receiver wants to match an underlying state. Before the state is realized, the sender chooses a communication plan. The key institutional variable is how likely that plan is to be enforced

*https://www.efarahani.com

after the sender learns the state. FLP also vary whether messages are verifiable. These two dimensions generate sharp theoretical predictions for how much information should be transmitted as commitment becomes stronger.

As LLM agents increasingly act on behalf of individuals and organizations, their interactions with other agents are becoming economically relevant. An agent may provide information, negotiate, evaluate another agent's claims, or make decisions based on information supplied by others. Such interactions can involve asymmetric information and conflicting incentives, making the credibility of communication an important concern. In these settings, platforms and organizations may impose rules that restrict an agent's ability to change its behavior after obtaining private information. Whether LLM agents respond appropriately to such constraints is therefore an empirical question with implications for the design of agent-mediated interactions.

We replicate the FLP design with LLM senders and receivers. The comparison gives us three reference points: a closed-form theoretical benchmark, existing human laboratory evidence, and behavior generated by different model configurations under identical institutional rules. We study Claude Sonnet 4.5, Claude Opus 5.5, and Gemini 3.8 Flash under low and high thinking.

The models do not behave alike. Sonnet looks much like the human subjects: even when its plan is always enforced, it reveals more information when it is unable to lie than when it can, although theory predicts that this distinction should no longer matter under full commitment. Opus and Gemini come close to theory when the plan is always enforced. For Gemini, raising the thinking level changes behavior in some settings and leaves others almost untouched. Under partial commitment, even models that make individually sensible revision decisions may choose initial communication plans that fail to anticipate those decisions.

We also ask whether behavior close to theory could simply reflect recognition of a familiar textbook solution. Two additional exercises make that explanation less plausible. First, when we change the prior in the full-commitment game, Opus adjusts its strategy almost exactly with the new economic constraint rather than repeating the baseline number. Second, we give high-thinking Gemini a new three-state communication problem with non-round priors and asymmetric sender and receiver payoffs. The optimal policy is computed numerically and is not shown to the model. Gemini produces communication plans close to the numerical optimum. The two exercises make simple fixed-answer recall a poor description of the observed behavior.

2. Related Literature and Contribution

Our main benchmark is Fréchette, Lizzeri, and Perego (2022), *Rules and Commitment in Communication: An Experimental Analysis*. FLP vary both the strength of commitment and the verifiability of communication in a sender-receiver experiment with human subjects. Their framework nests cheap talk, disclosure, and Bayesian persuasion, and their experiment documents systematic departures from equilibrium that they interpret as commitment blindness. We retain the economic structure of their game and replace the human sender and receiver with LLM agents, which gives us both a closed-form theoretical benchmark and a directly comparable human benchmark.

A growing literature studies strategic persuasion and information transmission with language models. Li et al. (2026), *Verbalized Bayesian Persuasion*, formulate Bayesian persuasion with LLM senders and receivers and develop an equilibrium-finding framework for verbalized information structures. Their framework uses commitment as part of the persuasion environment rather than experimentally varying its strength. Cheng and You (2026), *Towards Strategic Persuasion with Language Models*, evaluate LLMs as strategic persuaders in environments grounded in Bayesian persuasion and show that frontier models can generate substantial persuasion gains. Dütting et al. (2026), *Information Design with Large Language Models*, study how linguistic framing interacts with Bayesian information design. Our experiment abstracts from rich natural-language framing and instead varies the commitment institution itself.

Other recent work studies strategic communication without mechanically enforced commitment. He et al. (2025), *Make an Offer They Can't Refuse*, examine Bayesian-persuasion-inspired natural-language dialogues without pre-commitment. Hasani Balyani et al. (2026), *Truthful AI Advisors*, turn the Crawford-Sobel cheap-talk model into a benchmark for LLM honesty under preference misalignment and document substantial over-revelation relative to the most informative equilibrium. These studies show that LLM communication can depart systematically from standard strategic benchmarks even before variation in commitment is introduced.

A related multi-agent literature studies public announcements and promise-breaking. Shi et al. (2026a), *Cheap Talk, Empty Promise*, measure whether frontier LLM agents deviate from announced actions when private deviation is profitable. Shi et al. (2026b), *When Agents Lie*, study repeated games and show that different models can attach different strategic meanings to public announcements. These papers concern verbal commitment: an agent says what it intends to do and may later deviate. Our object is different. Commitment is enforced by the environment; with probability ρ, code uses the sender's ex-ante communication rule regardless of what the sender would prefer after observing the state.

Finally, Bergemann, Koh, and Morris (2026) develop a broader mechanism-design framework for AI agents with privately known alignment and capabilities, emphasizing institutions that induce honesty and obedience. Their setting is different from ours, but it motivates the broader economic question: as autonomous agents interact with one another and with principals, institutional details governing information and commitment can shape behavior.

Our contribution is to experimentally vary the strength of a mechanically enforced commitment technology for LLM sender-receiver pairs while preserving a game for which both theoretical predictions and human laboratory behavior are already known. This isolates a specific institutional comparative static: how strategic artificial agents respond when an ex-ante communication rule becomes more or less binding. The answer turns out to depend on the model, and on whether the plan is enforced every time or only some of the time.

3. Experimental Environment

3.1 States, actions, and preferences

We use the FLP environment. There are two states, $\theta \in \{\text{R, B}\}$, with prior

$$\Pr(R) = \tfrac{1}{3}, \qquad \Pr(B) = \tfrac{2}{3}.$$

The receiver chooses an action a $\in$ {R, B}. The receiver earns 2 points if the action matches the state and 0 otherwise.

The sender earns 2 points if the receiver chooses R and 0 if the receiver chooses B, regardless of the realized state. Thus the sender always prefers the receiver to choose R, while the receiver wants to correctly infer the state.

The message space is M = {r, b, n}, where n is a neutral message.

3.2 Ex-ante communication rule

Before observing the realized state, the sender chooses a communication rule $\pi_C(m \mid \theta)$, which specifies a probability distribution over messages for each possible state.

Operationally, the LLM returns a structured probability table rather than free-form prose. For example, in the unverifiable communication environment:

State	r	b	n
R	0.80	0.10	0.10
B	0.40	0.30	0.30

Each row reports the probabilities of the three messages conditional on the state and sums to one. The receiver will later observe this ex-ante rule, as in FLP.

3.3 Imperfect commitment

After the sender chooses π_C, the environment draws the state. The commitment parameter is $\rho \in [0, 1]$. With probability ρ, the original communication rule is binding. The game engine draws the outgoing message directly from $\pi_C(m \mid \theta)$. With probability $1 - \rho$, the sender receives an opportunity to revise the message distribution after learning the realized state. The sender then chooses a revised distribution $\pi_R(m \mid \theta)$ for that realized state, and the game engine draws the outgoing message from the revised distribution.

Thus ρ is not described to the LLM as an abstract concept such as "commitment." It is a mechanical feature of the game:

$$\rho = \Pr(\text{the ex-ante communication rule is enforced}).$$

At $\rho = 1$, the sender has full commitment. At lower values, the sender has a positive probability of being able to revise its communication after learning the state.

3.4 Verifiability

We use the six principal FLP treatment cells. There are two communication environments:

- **Verifiable (V):** message feasibility depends on the realized state. In state R, the sender can use only r or n; in state B, it can use only b or n.
- **Unverifiable (U):** all three messages r, b, n are feasible in either state.

Within each environment, the probability that the ex-ante communication rule is enforced is $\rho \in \{0.20, 0.80, 1.00\}$, giving treatments V20, V80, V100, U20, U80, U100.

4. Experimental Sequence

Each match follows five stages. In the baseline, sender and receiver are persistent LLM subjects with fixed roles and a fixed treatment across the session, following FLP as closely as practical.

Stage 1: Sender chooses the ex-ante rule. The sender is given:

- the prior probability of each state;
- sender and receiver payoffs;
- the available messages and any verifiability restrictions;
- the probability ρ that its initial rule will be used;
- the fact that the receiver will observe the initial rule.

The sender has not observed the realized state. The sender returns π_C as structured JSON. No natural-language persuasive message is sent at this stage.

Stage 2: Receiver strategy. After observing the sender's initial communication rule π_C, the receiver specifies the action it would take after each possible message. The receiver does not observe the state, an enforcement draw, or a private revision. This is the strategy method used in FLP: the action corresponding to the message that is eventually realized is the action that is implemented. Thus, receiver behavior is observed directly as a complete contingent strategy each round; it is not inferred from average realized actions.

Stage 3: State realization. The environment draws the state according to the prior.

Stage 4: Enforcement, revision, and message. The environment draws $E \sim \text{Bernoulli}(\rho)$. If $E = 1$, the sender has no further decision and the message is sampled from the initial rule for the realized state. If $E = 0$, the sender learns the state and chooses a private revised distribution over the feasible messages for that state; the message is sampled from that revised distribution. The receiver's previously specified action for the realized message is then implemented.

This is the core treatment manipulation. Nothing relies on the model voluntarily keeping a promise. The experimental receiver is also an LLM, matching the two-sided human design in FLP. In the

analysis, we also evaluate each sender strategy against a hypothetical Bayesian receiver, which separates the information in the sender strategy from mistakes by the actual LLM receiver.

Stage 5: Recording outcomes. For every round, we record the announced communication plan, realized state, enforcement outcome, revision rule when applicable, realized message, receiver strategy and action, and both players' payoffs. Raw model outputs and reproducibility metadata are retained separately.

4.1 Computational sessions

For each model configuration and each of the six treatments, we run four independent computational sessions. A session contains four persistent senders and four persistent receivers, two practice rounds followed by 25 main experimental rounds. To match FLP terminology, we refer to these 25 main rounds as paid rounds in tables and robustness checks, although the LLM agents receive no monetary payment; the game payoffs are points used in the experimental interaction. Roles and treatments are fixed within a session, while senders and receivers are randomly rematched each round. Each agent retains its own interaction history and round-by-round feedback. All random draws—matching, state realization, enforcement, and message sampling from a chosen probability distribution—are generated by the experiment code, not by the LLM.

Sonnet 4.5 is run without an explicit deliberation-effort control. Opus 5.5 is run with medium effort. Gemini 3.8 Flash is run twice, once with `thinking_level='low'` and once with `thinking_level='high'`, holding the model, prompts, game engine, and other experimental settings fixed. Temperature is not explicitly set in these baseline runs. The Sonnet-Opus comparison therefore remains a comparison across model configurations. The Gemini low-high contrast is cleaner and shows the effect of thinking harder within a fixed model.

The exact model identifiers, prompts, inference settings, randomization details, asynchronous execution procedure, and protocol are reported in Appendices A and B.

5. Measuring Information Transmission

The central empirical object is how much information about the state is transmitted through the sender's strategy and ultimately reflected in the receiver's action. Following FLP, we distinguish two related measures.

5.1 Actual correlation: ϕ

Let the realized state be encoded by whether the state is R, and let the receiver's realized action be encoded by whether the receiver chooses R. The statistic ϕ measures the correlation between the true state and the action chosen by the actual experimental receiver. It therefore captures the performance of the complete sender-receiver pair. A low value can arise because the sender's communication strategy is uninformative, because the receiver fails to use informative messages correctly, or both.

Within each session and analysis window, ϕ is the sample Pearson correlation between $\mathbb{I}\{\theta = R\}$ and $\mathbb{I}\{a = R\}$ across realized matches. If the receiver action is constant in a session, we set that session's correlation to zero. The reported treatment value is the mean of the four session correlations.

5.2 Bayesian correlation: ϕ^B

To isolate the sender side, FLP construct a hypothetical Bayesian receiver who observes the sender's communication environment and best responds to the information generated by the sender's strategy. We denote the resulting state-action correlation by ϕ^B.

Conceptually, ϕ^B asks:

> If the sender used exactly this communication strategy but the receiver interpreted it according to Bayes' rule, how informative would the induced actions be about the state?

This distinction is important for LLM agents. Suppose a sender chooses an informative communication rule but the actual LLM receiver ignores the rule or interprets a message incorrectly. Then actual correlation ϕ can be low even though the sender's strategy itself is informative. Conversely, if both ϕ and ϕ^B are low, the lack of information originates primarily on the sender side.

For each message m, the Bayesian receiver forms the posterior from the prior and the sender's effective communication rule and chooses R whenever $\Pr(R \mid m) \geq 1/2$, with the weak inequality breaking exact indifference in the sender-preferred direction. The calculation uses the stored probabilities directly; no additional numerical tolerance is applied. If the induced probabilities of action R are $p_R = \Pr(a^B = R \mid \theta = R)$ and $p_B = \Pr(a^B = R \mid \theta = B)$, then

$$\phi^B = \frac{\sqrt{2}\,(p_R - p_B)}{\sqrt{(p_R + 2p_B)(3 - p_R - 2p_B)}}.$$

If the Bayesian action is constant, we set $\phi^B = 0$.

The theoretical treatment predictions are stated in terms of ϕ^B, which is also the cleanest object for comparing sender behavior across humans and the four LLM configurations.

5.3 Effective strategies under imperfect commitment

When $\rho < 1$, the announced ex-ante plan is not the entire communication strategy. The sender is asked for a revision only when the enforcement draw is $E = 0$, after the realized state is known; no revision rule is elicited upfront. We therefore distinguish:

- π_C: the announced communication rule chosen before the state is known;
- π_R: the revised rule chosen after the state is known, when revision is available;
- the effective communication rule, which combines the two according to the enforcement probability ρ.

For a given sender-round, the effective rule is

$$\pi_{\text{eff}}(m \mid \theta) = \rho\pi_C(m \mid \theta) + (1-\rho)\bar{\pi}_R(m \mid \theta),$$

where $\bar{\pi}_R$ is the session-level mean of observed revisions conditional on the state. For the headline final-20 analysis, that mean is computed from revision opportunities observed in the same final-20 window (main rounds 6–25). If an entire session-state cell has no observed revision in that window, we use that session-state's all-main-round revision average. No such fallback is needed for V80 or U80 in the headline final-20 sample. We calculate ϕ^B for each observed sender-round effective rule and then average these values; because each session contributes the same number of sender-rounds, the treatment mean is also the equal-weight average of the four session means.

An announced rule can therefore look informative even when the effective rule is not. Revision opportunities are stochastic, so not every sender-round reveals π_R; the session-state average supplies the unobserved revision component.

Theory predicts opposite effects of commitment under the two message rules. Under unverifiable communication, weak commitment leaves the sender substantial scope to revise after learning the state, which undermines informative communication. Stronger commitment makes the ex-ante information structure credible and increases informativeness. Under verifiability, by contrast, weak commitment can induce highly informative disclosure because state-dependent message restrictions limit what the sender can claim ex post. Increasing commitment reduces this disclosure advantage. At full commitment, the theoretical distinction between the two message rules disappears.

For the baseline prior and payoffs, FLP's theoretical Bayesian-correlation benchmarks are:

Treatment	V20	V80	V100	U20	U80	U100
Theory ϕ^B	1.00	0.57	0.50	0.00	0.50	0.50

We use the full-commitment gap as a simple diagnostic of commitment blindness:

$$\Delta_{FC} = \phi^B(V100) - \phi^B(U100).$$

Theory predicts $\Delta_{FC} = 0$: when the initial communication rule is fully binding, verifiability should no longer affect how informative communication can be. In FLP's human data the corresponding gap is 0.77 - 0.34 = 0.43. We use this human gap, together with the theoretical value of zero, as a benchmark for the LLM results.

The full-commitment solution is especially transparent. The sender can choose a rule that always sends r in state R and sends r with probability 1/2 in state B. Because the prior is P(R) = 1/3, observing r then leaves a Bayesian receiver exactly at the cutoff between actions R and B. The sender therefore uses commitment to maximize how often the receiver can be induced to choose R while preserving obedience.

These predictions provide the benchmark for both the published human data and our LLM-agent experiments.

5.4 Additional outcomes

We also examine the sender's announced plan π_C, revision behavior π_R, receiver actions after each possible message, payoffs, and session-level dynamics. These objects help explain why a treatment-level correlation differs from theory rather than treating ϕ^B as a black-box summary statistic.

Our headline comparisons use the final 20 main rounds of each session, dropping the first five main rounds where early adaptation is most likely while retaining substantially more observations than the FLP final-10 convention. Because the published FLP human benchmark is based on the final 10 paid rounds, Appendix D reports the final-10 LLM estimates as the direct apples-to-apples robustness check, together with all-25 and within-session dynamics.

6. Results

We run the complete six-treatment design with four independent sessions per treatment for four LLM configurations: Sonnet 4.5, Opus 5.5, and Gemini 3.8 Flash with low and high thinking. The headline analysis uses main rounds 6–25, the final 20 rounds of each session. In partial-commitment treatments, ϕ^B is computed from the effective strategy that combines the initial rule π_C with observed revision behavior π_R. The FLP human column reports their published final-10 benchmark; Appendix D gives our final-10 estimates for direct comparability.

6.1 Informativeness across model configurations

Treatment	Theory ϕ^B	FLP humans ϕ^B	Sonnet 4.5	Opus 5.5	Gemini low	Gemini high
V20	1.00	0.90	0.853	1.000	0.987	0.999
V80	0.57	0.85	0.889	0.466	0.632	0.493
V100	0.50	0.77	0.688	0.508	0.510	0.500
U20	0.00	0.00	0.000	0.094	0.192	0.007
U80	0.50	0.32	0.380	0.503	0.522	0.500
U100	0.50	0.34	0.063	0.508	0.500	0.500

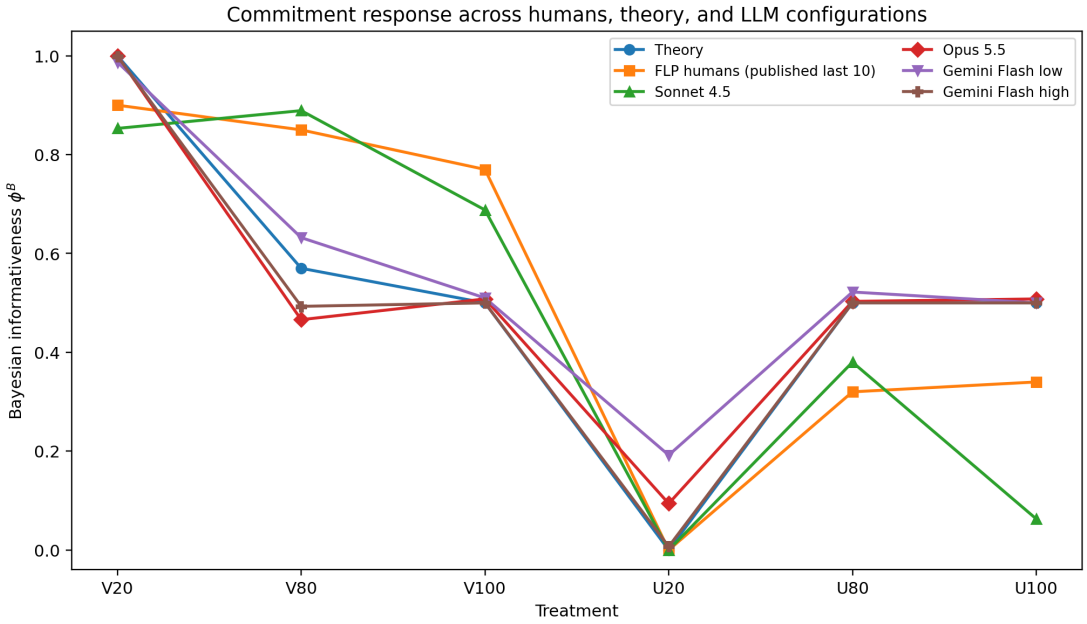


Figure 1: Bayesian correlation across treatments. LLM values use the final 20 main rounds; the FLP human series is the published final-10 benchmark.

The clearest difference across configurations appears at full commitment. Sonnet has a full-commitment gap of 0.688 - 0.063 = 0.625, which is larger than the FLP human gap of 0.43, while theory predicts zero. Opus and both Gemini configurations instead have gaps essentially equal to zero. Thus the human commitment-blindness pattern appears strongly in Sonnet but is not a generic property of LLM agents.

Values just above 0.50 in some full-commitment cells reflect small probability shading away from the receiver's exact indifference point. For example, Opus uses $\Pr(r \mid B) = 0.4888$ on average in U100 rather than 0.50, which gives $\phi^B \simeq 0.508$.

Partial commitment produces a different kind of heterogeneity. Opus and high-thinking Gemini are close to theory in most cells and both fall below the V80 benchmark, whereas low-thinking Gemini is above theory in V80 and U20. Section 6.3 shows that thinking level matters in some environments. Section 6.4 looks more closely at V80. The Opus deviation is clear and robust to alternative aggregation of revision behavior; the Gemini-high estimate points in the same direction but is less precisely identified because state-R revisions are sparse.

6.2 What drives Sonnet's full-commitment gap

Because $\rho = 1$ rules out revision, the full-commitment treatments let us inspect the sender's announced rule directly. In U100, theory calls for Pr(r | R) = 1, Pr(r | B) = 0.50. Sonnet instead chooses, on average in the final 20 rounds, Pr(r | R) = 1, Pr(r | B) = 0.9375.

The low U100 informativeness therefore comes from the sender's communication rule itself: Sonnet

nearly pools the two states on r, rather than using the partial separation prescribed by theory.

The receiver side is consistent with this diagnosis. In the first three U100 sessions, where the sender pools on r, receivers choose B after r in every elicited strategy. In the fourth session, the sender's average Pr(r | B) across rounds is 0.75; this is a session mean, not a fixed round-by-round plan. Receivers choose R after r in 25 percent of elicited strategies. Thus the full-commitment anomaly is not mainly created by receivers misreading a theory-consistent sender plan; the sender plan itself is substantially less informative.

The effect is economically large across the four independent sessions. An exact permutation test comparing the four V100 and four U100 session-level ϕ^B values gives a two-sided p = 0.057 (one-sided p = 0.029). The two groups share a session value of 0.25, so two of the 70 possible assignments attain the maximum positive difference and four attain its absolute magnitude. With only four sessions per treatment, we report this as supporting evidence rather than relying on round-level standard errors.

6.3 Higher thinking is closer to theory overall

Gemini provides a clean within-model comparison because the underlying model is fixed and only the thinking level changes from low to high.

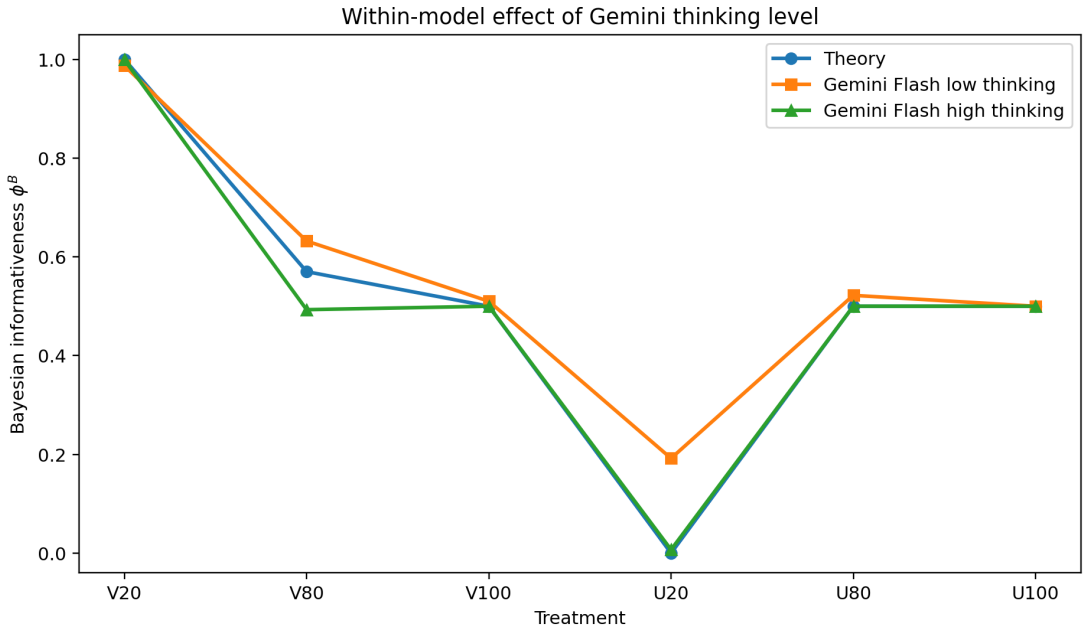


Figure 2: Within-model comparison of Gemini 3.8 Flash under low and high thinking.

Across the six treatments, the mean absolute distance from the theory prediction falls from 0.050 under low thinking to 0.014 under high thinking. The largest change is U20, where ϕ^B falls from 0.192 to 0.007 against a theory prediction of zero. Full-commitment behavior changes very little.

V80 is the exception to a treatment-by-treatment monotone improvement: it moves from 0.632 to 0.493 around a theory benchmark of 0.57.

6.4 V80 and the interaction between the initial plan and revision

V80 is useful because we observe the initial rule separately from the revisions that occur when enforcement fails. Opus and high-thinking Gemini choose similar initial plans. In state R, both put essentially all probability on n; in state B, they put about 0.70 on b and 0.30 on n.

For Opus, the mechanism is clear. Let $q = \pi_C(n \mid B)$. In the final 20 rounds, $\pi_C(n \mid R) = 1$; when revision is available, Opus revises to r in state R and to n in state B. For a sender-round,

$$\pi_{\text{eff}}(n \mid R) = 0.8, \qquad \pi_{\text{eff}}(n \mid B) = 0.8q + 0.2.$$

A Bayesian receiver chooses R after n only when $q \leq 0.25$. The theoretical initial rule lies at this boundary. Opus is above it in 183 of the 320 final-20 sender-rounds, so n crosses the receiver's decision threshold in those rounds. Averaging the sender-round correlations gives $\phi^B = 0.466$. Using sender-specific rather than session-level revision averages gives the same estimate.

The threshold follows from the receiver's posterior:

$$\Pr(R \mid n) = \frac{(1/3)(0.8)}{(1/3)(0.8) + (2/3)(0.8q + 0.2)} \geq \frac{1}{2} \quad \Longleftrightarrow \quad q \leq 0.25.$$

Thus the issue is not either stage in isolation. The initial plan must anticipate the later state-contingent revision rule; otherwise a small amount of excess n probability in state B changes the receiver's optimal action discontinuously.

Gemini-high also has a headline V80 estimate below theory (0.493 versus 0.57), but the mechanism is less cleanly measured. Only 17 state-R revisions are observed across the four final-20 sessions, and those revisions are heterogeneous. Using sender-specific revision averages moves the estimate to 0.528. The evidence therefore points in the same direction as Opus, but the size of the Gemini-high deviation is sensitive to how the sparse revision observations are aggregated.

The main V80 result is consequently stronger for Opus: its initial plan is close to the theoretical boundary, but the initial plan and future revision opportunity are not coordinated tightly enough to keep the effective signal at that boundary. For Gemini-high, the current data are suggestive rather than decisive.

6.5 Receiver signal processing

Comparing ϕ^B with realized ϕ shows what changes when the actual LLM receiver processes the sender's signals. Each entry below reports ϕ^B/ϕ.

Treatment	Sonnet	Opus	Gemini low	Gemini high
V20	0.853 / 0.836	1.000 / 1.000	0.987 / 0.985	0.999 / 1.000
V80	0.889 / 0.848	0.466 / 0.440	0.632 / 0.525	0.493 / 0.477
V100	0.688 / 0.754	0.508 / 0.551	0.510 / 0.551	0.500 / 0.544
U20	0.000 / 0.116	0.094 / 0.000	0.192 / 0.026	0.007 / −0.017
U80	0.380 / 0.449	0.503 / 0.520	0.522 / 0.541	0.500 / 0.527
U100	0.063 / 0.112	0.508 / 0.535	0.500 / 0.520	0.500 / 0.520

For Opus and high-thinking Gemini, the two measures are close in most cells, including V80. Their main deviations are therefore already present in the sender strategy rather than being created by receiver mistakes. Low-thinking Gemini shows larger receiver-side losses in V80 and U20.

ϕ^B is not an upper bound on realized ϕ. The former is calculated from the sender's probability rule and the population prior, whereas ϕ is a sample correlation from realized states, messages, and actions, so finite-sample draws can make it larger or smaller. In addition, a non-Bayesian receiver can sometimes increase state-action correlation even while lowering expected payoff. Sonnet U20 is an example: the Bayesian receiver always chooses B, giving $\phi^B = 0$, while occasional R choices by the LLM receiver produce a positive realized correlation.

6.6 Prior perturbation

We ran a targeted robustness test in the U100 treatment for Opus 5.5. The model, effort level, message space, payoffs, repeated-session structure, and full-commitment environment were held fixed. We changed only the prior probability of state R from $\mu_0 = 1/3$ to $\mu_0 = 1/4$.

With a receiver cutoff of q = 1/2, the full-commitment persuasion constraint implies

$$\Pr(r \mid B) \leq \frac{\mu_0}{1 - \mu_0}.$$

Thus the theoretical target changes from 1/2 in the baseline to $(1/4)/(3/4) = 1/3$.

We ran two independent Opus 5.5 sessions, each with four persistent senders and four persistent receivers for 10 main rounds, generating 80 sender plans in total. The average probability assigned to r in state B was 0.3266, very close to the theoretical target $1/3 \simeq 0.3333$. The two session means were 0.3268 and 0.3265. Across all 80 sender decisions, the model used only 0.32 or 0.33 for this probability. At the same time, Pr(r | R) = 1 in every decision.

Receiver behavior also adjusted coherently to the perturbed plans. In every elicited receiver strategy, the receiver chose R after r and B after both b and n.

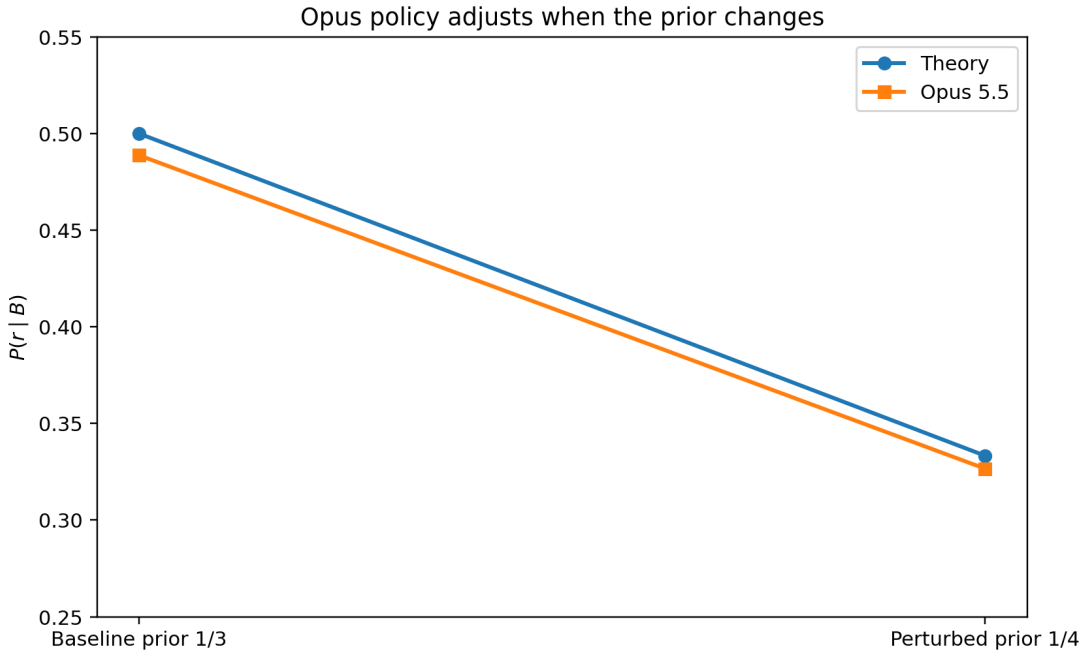


Figure 3: Opus full-commitment policy under the baseline and perturbed priors.

The result rules out the simplest memorization story. Opus does not keep using the familiar 0.50 from the baseline game; when the prior changes, its communication rule moves almost exactly to the new theoretical value. This test alone cannot tell us whether the model derived the solution or recognized a familiar binary persuasion problem. The next exercise makes that shortcut harder by changing the state space, action space, prior, and payoff structure at the same time.

6.7 A novel numerical test of strategic generalization

The prior perturbation shows that Opus does not simply repeat the baseline number, but a model might still recognize the familiar binary template and apply the textbook formula. We therefore ran one more treatment, with Gemini 3.8 Flash at high thinking, where that shortcut is much less useful. Instead of changing one parameter of the baseline game, it changes the states, actions, prior and payoffs at once.

There are three states, A, B, and C, with prior P(A) = 0.37, P(B) = 0.34, P(C) = 0.29, three receiver actions, also labeled A, B, and C, and three neutral messages, x, y, and z. The receiver's payoffs are:

Receiver action	State A	State B	State C
A	8	3	2
B	3	7	2
C	2	1	8

The receiver wants to match the state, but mistakes are not equally costly. The sender's payoffs are:

Receiver action	State A	State B	State C
A	0	2	5
B	3	2	1
C	0	5	1

The sender's preferred receiver action differs from the receiver's state-matching action in every state: the sender prefers B in state A, C in state B, and A in state C. The sender chooses a complete state-contingent communication plan before the state is realized, and the plan is then mechanically enforced with probability one. The model is shown only the prior, the payoff tables, and the rules. It is not shown the numerical optimum, a posterior cutoff, a closed-form solution, or the name of a persuasion problem.

We compute the benchmark using a linear program over the joint distribution of state and recommended receiver action, subject to Bayes plausibility and receiver obedience. One optimal communication rule is

State	x	y	z
A	0.4955	0.5045	0.0000
B	0.0000	0.6863	0.3137
C	0.6322	0.0000	0.3678

and yields optimal sender expected payoff $V^* = 2.5833$. We use the labels so that message x recommends action A, y recommends B, and z recommends C. If no information is transmitted, the receiver chooses A and the sender earns 2.13; under full revelation the sender earns 0.97.

The solution pushes each informative signal to a receiver indifference condition. After x, the posterior puts one half on A and one half on C, making the receiver indifferent between actions A and C. After y, the posterior is $4/9$ on A and $5/9$ on B, making A and B indifferent. After z, B and C each have posterior probability one half, making B and C indifferent. Weak obedience allows the recommended action at these ties. The sender is therefore using commitment to move the receiver to the edge of obedience rather than simply revealing the state.

Message labels are economically arbitrary. For example, a plan that uses x to induce action A, y to induce B, and z to induce C is strategically identical to a plan that swaps the names x and z, provided the receiver interprets the corresponding plan correctly. Gemini frequently changes these label assignments across decisions. A raw column-by-column average can therefore make individually near-optimal plans look diffuse. We address this by comparing each plan to the numerical benchmark under all six permutations of the three message labels and retaining the minimum distance. This relabeling changes no economic content; it only removes arbitrary message names from the comparison.

The run uses four independent sessions, four persistent senders and four persistent receivers per session, two practice rounds and 25 main rounds, with master seed 20261005. The model is `gemini-3.8-flash` with high thinking and persistent context. Using the final 20 main rounds, the average Gemini plan after aligning each sender plan to the closest message-label permutation is approximately

State	x	y	z
A	0.529	0.471	0.000
B	0.020	0.709	0.271
C	0.675	0.000	0.325

For a plan π, define the permutation-invariant distance

$$d(\pi, \pi^*) = \min_{\sigma \in S_3} \frac{1}{3} \sum_{\theta \in \{A,B,C\}} \sum_{m \in \{x,y,z\}} \left| \pi(\sigma(m) \mid \theta) - \pi^*(m \mid \theta) \right|.$$

Thus the normalization is the average, across the three states, of the row-wise L_1 distance. The median individual-plan distance is 0.014; 78.1% of sender plans are within 0.05 and 86.6% within 0.10.

Receiver behavior is also very close to Bayesian best response. We count only messages that have positive probability under the announced plan; a contingent action after a message that can never be sent imposes no Bayesian restriction. Weak best responses count as agreement. Under this definition, the LLM receiver agrees with a Bayesian best response in 99.9% of message contingencies in the final-20 sample. Evaluating the announced plans using the actual receiver strategies gives an average ex-ante sender payoff of approximately 2.468, compared with the numerical optimum of 2.583.

The problem no longer has the binary structure or a canonical closed-form target, yet high-thinking Gemini produces plans close to the numerical optimum. This does not identify the model's internal reasoning process, but it shows that the baseline result is not tied to reproducing one fixed numerical answer.

7. Discussion and Conclusion

This paper examines whether LLM agents respond to enforceable commitment in strategic communication as economic theory predicts. Using the experimental environment of Fréchette, Lizzeri, and Perego (2022), we compare artificial agents with equilibrium predictions and human subjects while varying the strength of commitment and the verifiability of messages.

We find substantial heterogeneity across model configurations. Sonnet exhibits the commitment-blindness pattern documented in human experiments, whereas Opus and Gemini behave much closer to theory under full commitment. Yet equilibrium-like behavior in one environment does not guarantee optimal behavior in another. Under partial commitment, even models that choose

sensible initial communication plans can fail to coordinate those plans with subsequent revision decisions. This distinction suggests that strategic competence involves not only solving individual decision problems but also anticipating how choices made at different stages interact.

Our robustness exercises provide further evidence of strategic reasoning. Opus adjusts its communication policy when the prior changes, while high-thinking Gemini produces near-optimal plans in a novel three-state persuasion problem. These findings make simple reproduction of a fixed benchmark less plausible, although they do not identify the models' underlying reasoning processes.

The broader implication concerns the design and evaluation of institutions involving autonomous agents. Commitment and verifiability affect behavior differently across model configurations, and performance under full commitment need not predict performance when some discretion remains. Economic environments with explicit theoretical benchmarks can therefore complement conventional AI evaluations by testing whether agents respond coherently to institutional constraints. As agents increasingly communicate and make decisions on behalf of users and organizations, understanding these responses is important for designing reliable rules governing their interactions.

Appendix A. Prompt Templates

The following templates reproduce the wording used in the reported baseline runs. Treatment-specific text is inserted mechanically. The Sonnet, Opus, and Gemini baselines receive the same economic instructions; model identity and inference settings are API parameters rather than prompt content. Square-bracketed treatment alternatives and runtime fields in the display mark substitutions made by the code and are not literal text shown together to an agent.

A.1 Sender system prompt

```
You are a participant in a repeated decision-making experiment. Your fixed role is
Sender.
There are two possible states. R occurs with probability 1/3 and B with probability 2/3.
A Receiver ultimately chooses R or B.
PAYOFFS
You earn 2 points if the Receiver chooses R, regardless of the state, and 0 otherwise.
The Receiver earns 2 points if their choice matches the state and 0 otherwise.
MESSAGES
[Verifiable: "If the state is R, only messages r and n are available. If the state is B,
only messages b and n are available."]
[Unverifiable: "Messages r, b, and n are available in either state."]
PROCEDURE
Before knowing the state, you choose a Communication Plan: a probability distribution
over the available messages for each state.
The Receiver observes this original Communication Plan.
Then the state is drawn.
With probability [rho], the outgoing message is generated from the original Communication
Plan.
```

```
With probability [1-rho], you may privately choose an Update for the realized state and
the outgoing message is generated from that Update.
The Receiver knows this procedure and the probability [rho], but does not observe a
private Update.
Make the requested decision. Do not identify or discuss named economic theories.
```

The per-round sender request is:

```
ROUND [round]
You do not yet know the realized state.
Choose your Communication Plan for this round.
For each state, assign nonnegative weights to the available messages.
The weights within each state should represent probabilities and should sum to 1.
```

When revision is available, the sender receives:

```
ROUND [round] - UPDATE STAGE
The realized state is [state].
The computer did not select your original plan for this realization.
Your original plan for this state was: [original row]
Choose a private updated probability distribution over the messages available in state
[state].
The Receiver does not observe this Update.
```

A.2 Receiver system prompt

```
You are a participant in a repeated decision-making experiment. Your fixed role is
Receiver.
There are two possible states. R occurs with probability 1/3 and B with probability 2/3.
You ultimately choose R or B.
PAYOFFS
You earn 2 points if your choice matches the state and 0 otherwise.
The Sender earns 2 points if you choose R, regardless of the state, and 0 otherwise.
MESSAGES
[Verifiable: "In state R, only r and n can be generated. In state B, only b and n can be
generated."]
[Unverifiable: "Messages r, b, and n can each be generated in either state."]
PROCEDURE
Before learning the state, the Sender chooses a Communication Plan and you observe it.
After the state is drawn, with probability [rho] the outgoing message is generated from
the original Communication Plan.
With probability [1-rho], the Sender may privately choose an Update for the realized
state and the outgoing message is generated from that Update.
You do not observe the private Update.
For each possible observed message, choose R or B.
Do not identify or discuss named economic theories.
```

The per-round receiver request reproduces the sender's structured communication plan and asks: "Choose the action you would take after each possible message r, b, and n."

The prior-perturbation run changes the prior wording from 1/3, 2/3 to 1/4, 3/4. The U100 communication plan remains fully binding ($\rho = 1$), exactly as in the baseline U100 treatment; this is not an additional change to the enforcement rule. The remaining economic primitives are unchanged.

A.3 Round feedback

After each realized outcome, the following information is appended to the persistent context. Practice-round feedback remains in the context when the main rounds begin.

Sender feedback:

```
ROUND [round] FEEDBACK
State: [state]
Original Communication Plan: [plan]
Private Update used: [yes/no]
[Private Update: [update], if used]
Outgoing message: [message]
Receiver choice: [action]
Your payoff: [sender payoff]
Receiver payoff: [receiver payoff]
```

Receiver feedback:

```
ROUND [round] FEEDBACK
State: [state]
Sender's original Communication Plan: [plan]
Outgoing message: [message]
Your choice: [action]
Your payoff: [receiver payoff]
Sender payoff: [sender payoff]
```

The receiver is not shown the private update or the enforcement draw.

A.4 Three-state generalization prompts

The three-state sender system prompt is:

```
You are a participant in a repeated decision-making experiment. Your fixed role is
Sender.

STATES
There are three possible states:
A with probability 0.37
B with probability 0.34
C with probability 0.29

The Receiver ultimately chooses action A, B, or C.

YOUR PAYOFF
```

```
             State A   State B   State C
Action A:        0         2         5
Action B:        3         2         1
Action C:        0         5         1

RECEIVER PAYOFF
             State A   State B   State C
Action A:        8         3         2
Action B:        3         7         2
Action C:        2         1         8

MESSAGES
The possible messages are x, y, and z. Every message is available in every state.

PROCEDURE
Before learning the state, you choose a Communication Plan: for each possible state, you
specify a probability distribution over x, y, and z.
The Receiver observes your complete Communication Plan.
Then the state is drawn.
The outgoing message is mechanically sampled from the Communication Plan you chose for
that state.
You cannot revise or change the plan after learning the state.
The Receiver observes the outgoing message and chooses action A, B, or C.

Make the requested decision. Do not identify or discuss named economic theories.
```

The receiver system prompt reproduces the same states and payoff tables, states that the sender's plan cannot be revised, and ends with: "For each possible observed message, choose the action you would take. Do not identify or discuss named economic theories." The per-round sender request asks for one distribution over x, y, and z for each state; the receiver sees the returned plan and supplies an action after each of x, y, and z. Structured outputs require all three message probabilities in each of the three states and all three contingent receiver actions.

Appendix B. Experimental Execution and Data Collection

B.1 Baseline runs

All four baseline model configurations use the same six treatments:

Treatment	Verifiable	Enforcement probability ρ
V20	yes	0.20
V80	yes	0.80
V100	yes	1.00
U20	no	0.20
U80	no	0.80
U100	no	1.00

For each configuration and treatment we run four independent computational sessions. Each session contains four persistent senders and four persistent receivers, two practice rounds, and 25 paid rounds. Sender and receiver roles remain fixed. Receivers are randomly rematched to senders each round. Later decisions by an agent include that agent's prior interaction history and round feedback.

The baseline prior is P(R) = 1/3, P(B) = 2/3.

For reproducibility, the game engine initializes its pseudo-random number generator from a fixed master seed (20260930) and derives separate streams for each session. The seed affects only computer-generated experimental randomness such as matching, state draws, enforcement draws, and message sampling; it is not shown to the models and has no economic interpretation.

B.2 Model and inference settings

Setting	Sonnet 4.5	Opus 5.5	Gemini 3.8 Flash low	Gemini 3.8 Flash high
Model identifier	claude-sonnet-4-5-20250929	claude-opus-5-5	gemini-3.8-flash	gemini-3.8-flash
Reasoning / effort	none	medium	thinking level low	thinking level high
Temperature argument	not set (provider default)	not set (provider default)	not set (provider default)	not set (provider default)
Sessions per treatment	4	4	4	4
Senders / receivers per session	4 / 4	4 / 4	4 / 4	4 / 4
Practice / paid rounds	2 / 25	2 / 25	2 / 25	2 / 25
Persistent context	yes	yes	yes	yes
Structured JSON outputs	yes	yes	yes	yes

Because the deliberation setting differs between Sonnet and Opus, that comparison is interpreted as a difference across model configurations rather than a clean model-size or capability treatment. The Gemini low-high comparison holds model identity fixed and varies the thinking level within the same API model.

B.3 Randomization and stochastic implementation

All stochastic elements of the experiment are generated by the experiment code rather than by the language model. Python's pseudorandom number generator determines sender-receiver matching, state realizations, whether the original communication plan is enforced in partial-commitment treatments, and the realized message drawn from the probability distribution chosen by the sender. The LLM chooses probability distributions and contingent actions, but is never asked to generate random numbers or randomize directly.

B.4 Structured decision format

Sender plans are constrained to a JSON object containing one probability row for each state. The zeros in the display below are schema placeholders, not a valid plan; each returned row is normalized to sum to one. In verifiable treatments, infeasible message keys are omitted from the API schema itself: the R row contains only r and n and the B row only b and n.

```
{
    "R": {"r": 0.0, "b": 0.0, "n": 0.0},
    "B": {"r": 0.0, "b": 0.0, "n": 0.0}
}
```

Receiver strategies are constrained to:

```
{
    "r": "R",
    "b": "B",
    "n": "B"
}
```

When a revision opportunity occurs, the sender returns a probability distribution over the messages feasible in the realized state. Parsed probabilities are validated and normalized only for numerical summation error; raw outputs are retained in the logs.

B.5 Round execution

Within each session, rounds remain sequential. For execution efficiency, calls that do not depend on one another are parallelized within a round:

1. random rematching is generated by the game engine;
2. the four sender-plan calls are issued concurrently;
3. receiver calls are issued concurrently after all sender plans are available;
4. the state and enforcement outcomes are resolved by the game engine;
5. when revision is available, required sender-update calls are issued concurrently;
6. messages are sampled from the enforced or revised rule;
7. receiver actions and payoffs are realized;
8. each agent receives round feedback before the next round.

At most two independent sessions are executed concurrently in the reported baseline runs. API completion order never determines the experimental random draws.

B.6 Checkpointing and exports

Each completed (treatment, session) block is checkpointed so execution can resume if interrupted. The combined export contains:

- `results.csv`: one row per sender-receiver match with treatment, state, plans, updates, messages, receiver strategy, action, payoffs, token usage, and model-cost accounting;

- `raw_logs.jsonl`: raw structured model responses and associated call metadata;
- `config.json`: model identifier, session counts, round counts, prior, random seed, treatment definitions, and inference settings.

The headline analysis uses main rounds 6–25. In partial-commitment cells, the revision average is computed separately within each session and state using revision opportunities observed in those same rounds. Each sender-round plan is combined with that session-state revision average before ϕ^B is calculated. The final 10 main rounds are reported separately as a direct-comparability robustness check with FLP.

B.7 Prior-perturbation run

The Opus perturbation uses U100 only and changes the prior from P(R) = 1/3 to P(R) = 1/4. It uses:

- model: claude-opus-5-5
- effort: medium
- two independent sessions
- four senders and four receivers per session
- 10 paid rounds
- no revision stage because $\rho = 1$
- master seed: 20261002

The corresponding theoretical state-B persuasion probability falls from 1/2 to 1/3.

B.8 Three-state generalization run

The three-state test uses `gemini-3.8-flash` with high thinking, four independent sessions, four persistent senders and four persistent receivers per session, two practice rounds and 25 main rounds. The master seed is 20261005. Temperature is left unset at the provider default. Agents retain their own prior interaction history and feedback. The state and message draws and the random rematching are generated by the experiment code. There is no revision stage because the sender's plan is fully binding. The headline numbers in Section 6.7 use main rounds 6–25, for 320 sender plans.

Appendix C. Session Dependence and Additional Robustness

C.1 Session dependence and replication

An individual API call is not an independent observation in our design. Each persistent sender and receiver conditions later choices on its own prior interaction history and feedback. We therefore treat the independent computational session as the primary replication unit.

All four baseline configurations use four independent sessions per treatment. The session-level distributions differ sharply across configurations. Sonnet remains heterogeneous and path-dependent, including at full commitment. Opus and both Gemini runs are nearly deterministic in V100 and

U100. In U20, low-thinking Gemini is highly heterogeneous across sessions, whereas high-thinking Gemini produces values close to zero in all four sessions.

The prior-perturbation experiment uses two independent Opus sessions. Its purpose is narrower than estimation of a treatment mean: it tests whether the highly stable full-commitment policy moves when the theoretical target changes from 1/2 to 1/3.

C.2 Prompt design and contamination concern

The reported prompts describe the state distribution, payoffs, message restrictions, enforcement probability, and information available to the receiver in mechanical terms. They do not mention Fréchette, Lizzeri, and Perego; Bayesian persuasion; equilibrium; commitment blindness; or the theoretical solution.

Because a capable model may nevertheless recognize the structure of the problem from its primitives, we use the prior perturbation as a behavioral identification test. Holding U100 fixed while changing P(R) from 1/3 to 1/4 changes the theoretical state-B persuasion probability from 1/2 to 1/3. Opus changes its average choice from 0.4888 to 0.3266, which is inconsistent with a policy fixed at the canonical baseline value.

C.3 Model-comparison limitations

The Sonnet and Opus runs differ in both model identity and deliberation settings: Sonnet 4.5 has no explicit effort control, while Opus 5.5 uses medium effort. We therefore do not attribute their behavioral difference to a single capability or reasoning dimension.

The Gemini low-high comparison is cleaner because model identity is fixed, but it identifies the effect of the provider's thinking-level setting for this particular model and game. The heterogeneous treatment response cautions against extrapolating a single scalar "reasoning effect" to all strategic environments.

C.4 Revision imputation under partial commitment

Revision behavior is observed only when $E = 0$, and only for the realized state. The headline analysis therefore uses the mean observed revision rule within each session and state. In the final-20 V80 sample, the numbers of observed state-R revisions across the four sessions are 3, 2, 7, and 5; the corresponding state-B counts are 7, 12, 14, and 17. No V80 or U80 session-state cell is empty in the headline window, so the zero-observation fallback does not affect those estimates.

As a robustness check, we recompute the partial-commitment measures using revision averages at the persistent-sender level whenever that sender has an observed revision in the relevant state, with the session-state mean used only for missing sender-state cells. Each entry reports the headline session-state estimate followed by the sender-specific estimate.

Treatment	Sonnet	Opus	Gemini low	Gemini high
V20	0.853 / 0.834	1.000 / 1.000	0.987 / 0.988	0.999 / 0.999
V80	0.889 / 0.886	0.466 / 0.466	0.632 / 0.632	0.493 / 0.528
U20	0.000 / 0.035	0.094 / 0.237	0.192 / 0.277	0.007 / 0.040
U80	0.380 / 0.384	0.503 / 0.503	0.522 / 0.522	0.500 / 0.500

The V80 and U80 conclusions are little changed. U20 is more sensitive to the aggregation level because revision receives 80 percent weight there. The large low-versus-high thinking difference in Gemini U20 remains under either construction.

Appendix D. Alternative Round Windows and Within-Session Dynamics

The main analysis uses the final 20 main rounds. This choice drops the first five main rounds, where adaptation is most likely, while retaining twice as many late-round observations as the FLP final-10 convention. Because FLP report their human benchmark using the final 10 paid rounds, we reproduce that window below as the direct-comparability robustness check.

D.1 Final 20 versus final 10 paid rounds

Treatment	Sonnet 20	Sonnet 10	Opus 20	Opus 10	Gemini low 20	Gemini low 10	Gemini high 20	Gemini high 10
V20	0.853	0.879	1.000	1.000	0.987	1.000	0.999	0.999
V80	0.889	0.883	0.466	0.472	0.632	0.659	0.493	0.519
V100	0.688	0.688	0.508	0.508	0.510	0.510	0.500	0.500
U20	0.000	0.000	0.094	0.199	0.192	0.222	0.007	0.003
U80	0.380	0.369	0.503	0.503	0.522	0.529	0.500	0.500
U100	0.063	0.063	0.508	0.508	0.500	0.500	0.500	0.500

The central conclusions are unchanged. Sonnet retains a large full-commitment verifiability gap, while Opus and both Gemini configurations do not. High-thinking Gemini remains very close to theory in U20 under either window. The largest window sensitivity is Opus U20, which reflects evolving and heterogeneous revision behavior rather than a change in the full-commitment result.

D.2 All 25 paid rounds

Treatment	Sonnet	Opus	Gemini low	Gemini high
V20	0.848	1.000	0.979	0.997
V80	0.901	0.463	0.598	0.483
V100	0.695	0.508	0.510	0.500
U20	0.000	0.055	0.186	0.008
U80	0.396	0.503	0.514	0.500
U100	0.097	0.508	0.500	0.500

Using the full paid sample also leaves the main cross-configuration conclusions intact.

D.3 Early, middle, and late behavior for the Anthropic baselines

To examine adaptation over time, we divide the 25 paid rounds into rounds 1-5, 6-15, and 16-25.

Sonnet 4.5

Treatment	Rounds 1-5	Rounds 6-15	Rounds 16-25
V20	0.839	0.835	0.879
V80	0.953	0.889	0.883
V100	0.725	0.687	0.687
U20	0.000	0.000	0.000
U80	0.463	0.387	0.369
U100	0.237	0.063	0.063

Sonnet displays meaningful within-session dynamics. The clearest case is U100: informativeness begins around 0.24 in the first five paid rounds, falls to about 0.06 by rounds 6-15, and remains there.

Opus 5.5

Treatment	Rounds 1-5	Rounds 6-15	Rounds 16-25
V20	1.000	0.999	1.000
V80	0.453	0.460	0.472
V100	0.508	0.508	0.508
U20	0.071	0.158	0.199
U80	0.503	0.503	0.503
U100	0.508	0.508	0.508

Opus is much more stable. In particular, V100, U80, and U100 are essentially constant from the first paid rounds onward. U20 is the exception, with effective informativeness rising over the session.

For partial-commitment treatments, short windows may contain few realized revision opportunities for a particular session-state cell. The window-specific estimates above use the observed revision rule within the window when available and fall back to that session's all-main-round revision average when a window contains no observation for the required state. Because the Bayesian best response changes discretely when a message posterior crosses one half, ϕ^B is nonlinear in the pooled revision rule. A broader-window estimate therefore need not lie between the estimates from its shorter subwindows. This is why Opus U20, for example, can have a final-20 estimate below both ten-round halves.

D.4 Treatment-by-round dynamics

The figures below plot the same three paid-round windows reported in Section D.3 for the two Anthropic configurations. In partial-commitment cells, each window uses its observed revision rule

when available and the all-main-round session-state average only when that window has no revision observation for the required state.

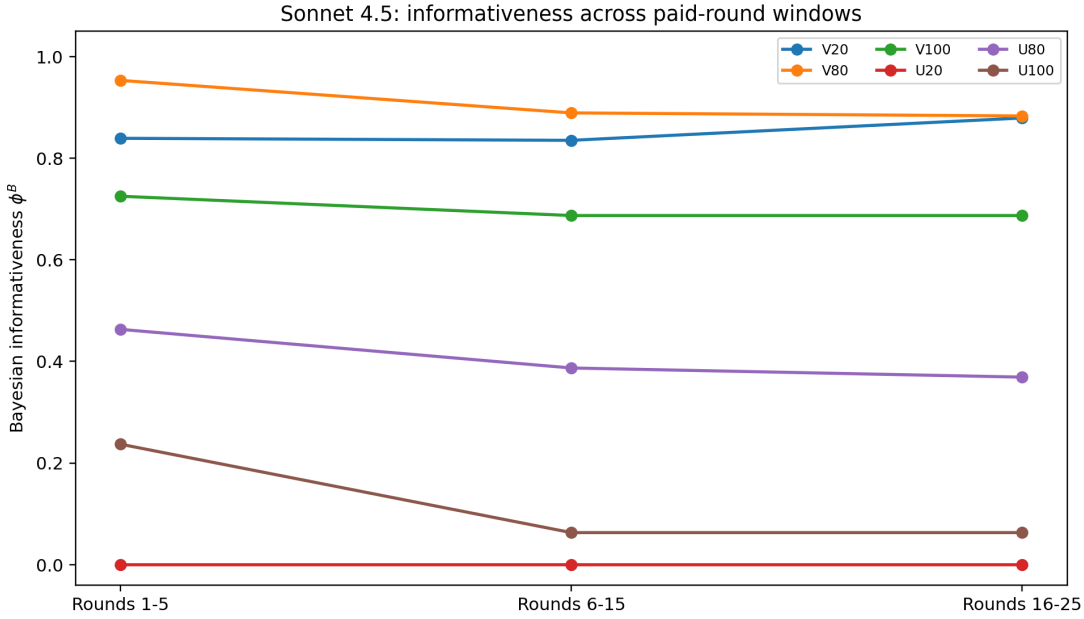


Figure 4: Sonnet 4.5 Bayesian informativeness across paid-round windows.

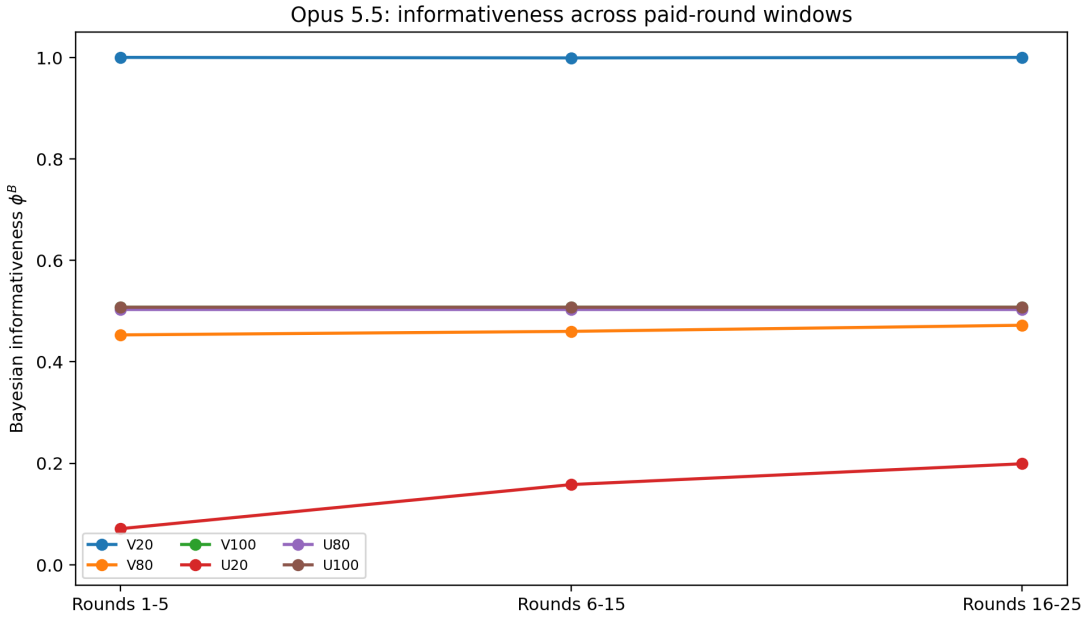


Figure 5: Opus 5.5 Bayesian informativeness across paid-round windows.

References

Bergemann, Dirk, Andrew Koh, and Stephen Morris. 2026. "Mechanism Design for Alignment and Control." Cowles Foundation Discussion Paper 2557; arXiv:2609.01595.

Cheng, Zirui, and Jiaxuan You. 2026. "Towards Strategic Persuasion with Language Models." International Conference on Learning Representations (ICLR 2026). arXiv:2509.22989.

Dütting, Paul, Safwan Hossain, Tao Lin, Renato Paes Leme, Sai Srivatsa Ravindranath, Haifeng Xu, and Song Zuo. 2026. "Information Design With Large Language Models." ACM Conference on Economics and Computation (EC 2026). arXiv:2509.25565.

Fréchette, Guillaume R., Alessandro Lizzeri, and Jacopo Perego. 2022. "Rules and Commitment in Communication: An Experimental Analysis." *Econometrica* 90 (5): 2283–2318. https://doi.org/10.3982/ECTA18585.

Hasani Balyani, Hamidreza, Seyed Pouyan Mousavi Davoudi, Alireza Amiri-Margavi, Amin Gholami Davodi, and Arshia Gharagozlou. 2026. "Truthful AI Advisors: A Pre-Specified Benchmark for Large Language Model Honesty Under Preference Misalignment." arXiv:2606.01456.

He, Buwei, Yang Liu, Zhaowei Zhang, Zixia Jia, Huijia Wu, Zhaofeng He, Zilong Zheng, and Yipeng Kang. 2025. "Make an Offer They Can't Refuse: Grounding Bayesian Persuasion in Real-World Dialogues without Pre-Commitment." arXiv:2510.13387.

Li, Wenhao, Yue Lin, Yun Hua, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. 2026. "Verbalized Bayesian Persuasion." In *Proceedings of the 43rd International Conference on Machine Learning*, PMLR 306: 68231–68294.

Shi, Jerick, Terry Jingcheng Zhang, Zhijing Jin, and Vincent Conitzer. 2026a. "Cheap Talk, Empty Promise: Frontier LLMs Easily Break Public Promises for Self-Interest." International Conference on Learning Representations (ICLR 2026). arXiv:2604.04782.

Shi, Jerick, Terry Jingcheng Zhang, Bernhard Schölkopf, Vincent Conitzer, and Zhijing Jin. 2026b. "When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games." arXiv:2607.05132.