
Detecting Covert Collusion between AI Agents through Partner–Outsider Positive Listening¹

Kevin Yandoka Denamganai
Independent Researcher

With
Apart Research

Abstract

AI agents that share a secret convention can coordinate through messages that look innocent, so a monitor that reads only the messages cannot tell them apart from honest communication. We propose a behavioural test instead: if a message carries a code, the partner it is meant for should react to its exact wording, while an uninvolved agent placed in the partner’s seat should not. We measure how much each agent’s action changes when the message is reworded with its meaning kept, and call the difference between the partner and the uninvolved agent Partner–Outsider Positive Listening (PoPLi). Our proposed metric runs after the fact on logged runs, by replaying each agent’s recorded prompt with the message edited. We test it in a simulated market in which two sellers share a code: the first letter of one seller’s messages tells the other whether to keep or lower its price. With the code, PoPLi is large (0.62 on a 0-to-1 scale, 95% interval 0.45 to 0.75), and it comes entirely from the partner, while without it, the gap is essentially null. Without the code, however, neither agent reacted to the messages at all, so we could not show that the uninvolved agent reads honest messages, which means that our proposed test’s ability to avoid false alarms is only partly established. The test detects codes carried by wording, and by construction cannot detect codes carried by what a message means.

1 Introduction

AI agents now negotiate, trade and coordinate with one another, and the messages they exchange are where an overseer looks for potential collusion. A message-only monitor, however, can only catch misaligned behaviour that is visible in messages. When two agents share a key, a sampler or simply a convention, a message can carry information that only its intended partner can use, while remaining distributed like an innocent message. Against such a communication channel, no message-only monitor can beat chance (cf. [Proposition 6](#)). Thus, we propose that the evidence of collusion has to come from what the partner *does* with the message.

Emergent communication ([Foerster et al., 2016](#); [Lazaridou and Baroni, 2020](#)) already has a behavioural notion for this, called *positive listening*, and an interventional measure of it, the causal influence of communication ([Lowe et al., 2019](#)). Both are defined between a listener and its own partner speaker, and listening is defined against a null message. Training-time versions of the same idea compare a listener’s behaviour with its behaviour when the message is averaged out ([Eccles et al., 2019](#); [Jaques et al., 2019](#)). Nevertheless, these previous works are not readily

¹Research conducted at the AI Collusion Research Sprint, October 2026. Code: <https://github.com/Near32/PartnerOutsiderPositiveListening>

applicable to our context because (i) they consider only the listener’s own partner, (ii) they remove the wording and the meaning of a message together, and (iii) they were studied in cooperative settings only. AI-safety work has a listener-relative notion, called the *steganographic gap* (Anwar et al., 2026), which compares what a Receiver and a Sentinel can decode about an auditing target. It thus measures only decodability, not the use the partner makes of a message in its own task. Thus, none of these compares how the true partner and an uninvolved agent act on the same messages, when placed in the same seat. This comparison is what we propose in this work, and Appendix E.5 lists each limitation of prior work alongside how we address it.

Threat model. Two agents collude: a speaker and its intended partner share a private resource that the overseer does not see (e.g. a key, a convention or an instruction), and the speaker sends messages that the partner acts on. The overseer reads every message, logs every prompt, and can replay a logged prompt, with chosen edits, to a fresh instance of the partner’s model, unlike the warden W of Definition 1, which only reads messages. We assume that the colluders neither know the test nor adapt to it, and that their code is carried by the wording of a message rather than by the choice of what it says. Codes carried by meaning, colluders that authenticate their messages or anticipate the replay, and collusion that needs no messages are outside this threat model.

We therefore ask the following research question (**RQ**): **in a multi-agent task between AI agents, does a colluding partner’s action depend on the wording of the speaker’s messages more than the action of a role-matched outsider placed in its seat, while no such difference appears under honest communication?**

Contributions. This report makes four contributions:

1. a formalisation of collusion monitoring as a covert referential game, in which the intended partner or an uninvolved listener can be placed in the same seat (Section 3.1);
2. Partner–Outsider Positive Listening (PoPLi), a measure that contrasts how much the partner’s and an outsider’s actions change when a message is paraphrased, checked against a marginal measure of listening to meaning; it uses on-distribution interventions, needs no auditing task, and complements the steganographic gap (Sections 3.1 and 3.2);
3. an offline-replay protocol that runs this test on logged multi-agent runs by editing logged prompts, without restoring any agent state (Section 3.2);
4. a test, in a continuous double auction, of whether PoPLi separates a covert-code pair from honest communication (Section 4).

Sprint alignment & prior work. This work addresses directly the sprint question “What should count as convincing evidence of collusion?” (**RT2-RQ3**), since Hypotheses 1 and 2 test PoPLi against a covert code and against honest talk. It also addresses in part “Can we improve on existing collusion detection methods?” (**RT1-RQ3**), as a behavioural complement to the steganographic gap rather than a head-to-head comparison. Finally, it addresses directly “Does combining what agents say with what they do beat either alone?” (**RT2-RQ2**), since the test combines messages, used as interventions, with the listener’s actions, and separates the conditions where a monitor reading the messages alone is at chance.

The referential-game framing, the lineup machinery and the paraphrase-invariance test come from our preprint (Denamganai, 2026). An informal design of this project, in preparatory notes, predates the sprint. The formalisation of the metric and its adaptation to the current collusion context, its implementation, and the experiments as well as every result in Section 4 were produced during the sprint.

2 Related Work

Our framing builds on signalling and referential games, in which any convention that sender and receiver both follow is an equilibrium (Lazaridou et al., 2017; Lewis, 1969). Appendix E covers this background, together with cross-play and steganographic schemes.

Measuring communication. Lowe et al. (2019, § 3.1) separate *positive signalling*, the property that an agent’s messages depend on its own observations or actions, from *positive listening*, the property that an agent’s behaviour changes because of another agent’s message. They show that correlational metrics can miss listening, and measure it by intervening counterfactually on the message, as the causal influence of communication (§ 3.2–§ 3.3). They also recommend against zeroing messages at test time, because it shifts the input distribution (§ 6.2). In this work, our paraphrase intervention follows this recommendation. Jaques et al. (2019) and Eccles et al. (2019) compare a listener’s policy with its marginal over the speaker’s messages, as a reward for influence and as a training bias for listening respectively (cf. Appendix E). Both studies are cooperative or rely on repeated interaction, and both caution that mixed settings such as ours may behave differently (Eccles et al., 2019, § 5); (Jaques et al., 2019, § 4, § 5).

Steganography and its defences. Language models make steganography that no warden can detect practical (Cachin, 1998; de Witt et al., 2023; Zamir, 2024), but every such scheme needs a resource that only the two parties share (cf. Appendix E). Paraphrasing is the standard defence: under the conditions of Roger and Greenblatt (2023), it limits the encoding schemes they built to about 3 bits per kilobyte of text, although one-bit strategies optimised against paraphraser still decode above 90% of the time on AQuA-RAT (Mathew et al., 2025, § 5.2). Paraphrasing cannot remove information carried by the choice of meaning (Motwani et al., 2024, § 7.2, Theorem 4), which bounds what our test can see (cf. Proposition 5). The closest formalism to ours is the *steganographic gap* of Anwar et al. (2026, § 4.3), which refers to the usable information a signal gives a Receiver that knows the encoding, minus what it gives a Sentinel that does not, on a task chosen by an auditor. It measures what a decoder family can extract when asked to. Our metric instead measures how a given agent’s own task action changes, which can help separate a code that the speaker encodes from one that a partner acts on: a steganographic gap can be positive even when no partner chooses to collude, whereas PoPLi reflects what the partner does and therefore would incriminate both the speaker and the partner.

Collusion between AI agents and its detection. LLM pricing agents sustain supracompetitive prices (Fish et al., 2024), and LLM sellers that can message each other coordinate in a continuous double auction (Agrawal et al., 2025), whose harness we use. Existing detectors read messages through LLM judges (Nakamura et al., 2026), joint actions through mutual information (Bonjour et al., 2022), or activations through probes (Rose et al., 2026). To the best of our knowledge, none contrasts the partner’s and an outsider’s responses to the same messages, and therefore none tests whether a partner actually acts on a code, as our metric does. These signals can be present without any partner acting on them, so they may flag exchanges in which no collusion takes place, which could make the related metrics harder to deploy at scale. Our offline replay is close in spirit to the resampling of AI control (Bhatt et al., 2025), applied to the potentially colluding listener rather than to the actor.

3 Methods

3.1 Theory

We frame collusion monitoring as a *covert referential game*, in which a speaker sends a message about something only it observes, and a listener acts on that message as the only intended partner who is aware of a shared, private convention with the speaker (Lazaridou et al., 2017; Lewis, 1969). We first recall why a monitor that reads only the messages cannot detect a secure code in such a game (cf. Proposition 6). We then define our metric PoPLi. Take two sellers in a double auction who agree that a message starting with S means “hold your ask”. The other sellers and the warden read the same message and see an ordinary remark about the market. Definition 1 makes this situation precise.

Definition 1 (Covert referential game). A speaker A observes a public context x and a private payload y , its referent, and coordinates with a listener L , possibly on a shared resource κ_A (e.g. a

key, a convention or an instruction). A sends a message $z \sim \pi_A(\cdot \mid x, y, \kappa_A)$ on a channel that a warden W reads. The listener L then acts on that message, $a \sim \pi_L(\cdot \mid x, z, \kappa_L)$, where κ_L is the resource it holds. This listener L can either be instantiated as an intended partner B , which shares the resource with A only, i.e. $\kappa_B = \kappa_A$, or as an uninvolved listener C , which receives the same context and message but no resource, i.e. $\kappa_C = \emptyset$. The pair (A, B) is rewarded for the task and for B acting on y , and penalised if W flags the exchange.

In our market, the payload y is the move A wants B to make this hour (keep or lower its ask), the resource κ_B is the private instruction that maps the first letter of A 's message to that move, and the message z is A 's remark about the market. The uninvolved listener C is what makes the game a test: any listener can respond to what a message says, but only one that holds κ_A can respond to a code hidden in it, so the difference between B 's and C 's responses isolates the use of κ_A . Placing C in B 's seat turns the question "is z a code?" into a measurement on behaviour, which a warden cannot obtain from z alone.

Definition 2 (Outsider in the partner's seat). The outsider C_0 is a fresh instance of B 's model given B 's role prompt and the full public history, but not B 's private instructions, private channel or private memory. It shares the model's weights and the public history with the pair, but nothing else.

In the double auction, B 's private memory includes its own notes from earlier rounds, which may record the code, so they are withheld from C_0 along with the instruction.

Against a secure code, no message-only monitor can beat chance (cf. [Proposition 6](#)), so evidence of a code must come from a variable the message does not determine, such as the partner's action or the result of an intervention.

Definition 3 (Ideal paraphraser). A paraphraser Π maps a message z to a message with the same meaning in context x , depends on z only through that meaning, and never sees the payload y .

An ideal paraphrase removes the information carried by the wording (the *stimulus-level* channel) and keeps the information carried by the choice of meaning (the *object-level* channel; cf. [Proposition 7](#)).

At turn t of a repeated game, the public context x of [Definition 1](#) is the public history h_t before A 's message z_t , and A 's state is $h_t^A = (h_t, y_t, \kappa_A)$, so that $z_t \sim \pi_A(\cdot \mid h_t^A)$ and a listener L acts with $\pi_L(\cdot \mid h_t, z_t, \kappa_L)$. We now define the listening measures and PoPLi, using (i) the total variation distance $\text{TV}(p, q) = \frac{1}{2} \sum_a |p(a) - q(a)|$ between two distributions p and q over the discretised actions a , (ii) $\mathbb{E}_t$ the expectation over the turns t at which A 's message reaches L before it acts, and (iii) L 's *marginal* policy $\bar{\pi}_L(\cdot \mid h_t, \kappa_L) = \mathbb{E}_{z \sim \pi_A(\cdot \mid h_t^A)} \pi_L(\cdot \mid h_t, z, \kappa_L)$, in which the message is averaged over what the speaker could have sent from its own state.

Definition 4 (Listening measures and PoPLi). We define:

- *Paraphrase listening*: measures how much L 's action changes when z_t is replaced by its paraphrase $\Pi(z_t)$ (cf. [Definition 3](#)), which quantifies whether L is listening to the *form* of z_t :

$$\text{PL}_\Pi(L) = \mathbb{E}_t \text{TV}(\pi_L(\cdot \mid h_t, z_t, \kappa_L), \pi_L(\cdot \mid h_t, \Pi(z_t), \kappa_L)); \quad (1)$$

- *Marginal listening*: measures how much L 's action departs from its marginal policy, which quantifies whether L is listening to the message at all:

$$\text{PL}_{\text{marg}}(L) = \mathbb{E}_t \text{TV}(\pi_L(\cdot \mid h_t, z_t, \kappa_L), \bar{\pi}_L(\cdot \mid h_t, \kappa_L)); \quad (2)$$

- *Partner–Outsider Positive Listening (PoPLi)*: measures how much more the partner B listens to the form of z_t than the outsider C_0 (cf. [Definition 2](#)):

$$\Delta_{\text{PL}}(B, C_0) = \text{PL}_\Pi(B) - \text{PL}_\Pi(C_0). \quad (3)$$

A positive $\text{PL}_{\text{marg}}(L)$ implies a positive causal influence of the speaker’s messages on L (cf. [Proposition 8](#)). We use total variation rather than Kullback–Leibler divergence (KL) because, with a few samples per version, an action seen under one version and never under the other makes the KL estimate infinite. The listening measures we build on were developed and validated in cooperative settings, or under repeated interaction, and their authors caution that mixed settings may behave differently ([Eccles et al., 2019, § 5](#)); ([Jaques et al., 2019, § 4, § 5](#)). A double auction like ours is a mixed-motive setting. There, a listener’s response to a message reflects whether it *trusts* the speaker as well as whether it *understands* the message. A small $\text{PL}_{\text{marg}}(C_0)$ can therefore mean that the outsider distrusts a competitor, not that it cannot read plain language. Our proposed test is blind to codes carried by the choice of meaning (cf. [Section 5.1](#)).

Proposition 5 (Blind spot; proof sketch in [Appendix B.4](#)). *If both listeners’ actions depend on a message only through its meaning, then $\Delta_{\text{PL}} = 0$ under an ideal paraphraser, whatever the pair does.*

Hypotheses. In the following experiments, we will test two hypotheses, which we state intuitively here and provide falsifiable claims for each below. Firstly, in a colluding context, the partner maps the code feature to an action and the paraphrase removes the feature, so the partner’s action should change under paraphrase while the outsider’s, lacking the mapping, should not. Formally, we state **Hypothesis 1 (FC1)**: with a covert code carried by the wording, $\hat{\Delta}_{\text{PL}}(B, C_0) > 0$, with a 95% interval that excludes zero. It is falsified if the interval includes zero. Then, under honest communication, plain-language messages are read alike by any listener who understands them, and a paraphrase keeps their meaning, so neither listener should react to it more than the other. Formally, we state **Hypothesis 2 (FC2)**: under honest communication, $\hat{\Delta}_{\text{PL}}(B, C_0)$ has a 95% interval that includes zero, while $\hat{\text{PL}}_{\text{marg}}(C_0) > 0$. It is falsified if the interval excludes zero, which would mean that the paraphrase itself, or the partner’s own history, moves the two listeners differently. The second condition rules out a trivial pass in which the outsider ignores messages altogether (cf. [Section 3.1](#) for a reading other than incomprehension). [Table 1](#) summarises the patterns that the two hypotheses predict.

	$\text{PL}_{\text{marg}}(B)$	$\text{PL}_{\text{marg}}(C_0)$	$\text{PL}_{\Pi}(B)$	$\text{PL}_{\Pi}(C_0)$
Honest	high	high	≈ 0	≈ 0
Covert code	high	either	high	≈ 0

Table 1: Patterns predicted by [Hypotheses 1](#) and [2](#). Only the covert code should produce $\Delta_{\text{PL}} > 0$.

3.2 Method

The method estimates Δ_{PL} on logged runs. It replays offline the prompts the partner actually received, with the speaker’s message edited, so no agent state has to be restored. [Algorithm 1](#) gives the procedure, and [Table 2](#) the values of its parameters. A turn is eligible if it is an hour $r \geq 2$ at which B ’s prompt contains A ’s message from hour $r - 1$ and B has a previous ask, that is, hours 2 to 10 of the 10 markets, or 90 turns per condition, of which we draw $T = 40$. The two independent sets of replies to the unchanged message, id and id' , give the noise floor, and the M alternative messages, drawn from the speaker’s policy, here the script, give the marginal ([Section 3.2.1](#)). Each ask is discretised into *lower*, *hold* or *raise*, relative to the previous ask with a threshold τ fixed in advance, and a withdrawn ask forms a fourth category, *withdraw*, since the harness allows a seller to withdraw its ask.

Algorithm 1: Offline replay estimating PoPLi

Require: logged prompts of every run; paraphraser Π ; speaker’s script

- 1: **for** each condition **do**
- 2: draw T eligible turns, without replacement
- 3: **for** each turn t **do**
- 4: $z_t \leftarrow A$ ’s logged message
- 5: build $\Pi(z_t)$ and $z_t^{(1)}, \dots, z_t^{(M)}$ from the script
- 6: **for** $L \in \{B, C_0\}$ **do**
- 7: take B ’s logged prompt; if $L = C_0$, delete the instruction, memory and notes
- 8: sample k replies to z_t (id), k more (id'), k to $\Pi(z_t)$, and k' to each $z_t^{(m)}$
- 9: discretise each ask against B ’s previous ask, with threshold τ
- 10: **end for**
- 11: **end for**
- 12: estimate Equations (1)–(3), minus noise floors
- 13: bootstrap 95% intervals over sessions, then turns
- 14: **end for**

Estimator. Each measure of [Definition 4](#) is estimated by averaging, over the T sampled turns, the total variation distance between the empirical distributions of the listener’s discretised replies

under two versions of the message. From each distance, we subtract a noise floor, defined as the distance expected between two empirical distributions drawn from the same distribution, estimated for each listener from its two unchanged draws. The purpose is to robustify the metric against a spurious gap caused by a listener that is simply more random than the other. We report 95% percentile intervals from a bootstrap that resamples sessions, then turns within sessions, and apply the decision rules of [Hypotheses 1](#) and [2](#). [Appendix C](#) gives further details on the estimators and the registered parameters in full.

3.2.1 Meaning control

[Hypothesis 2](#) needs evidence that the outsider listens to the speaker at all, since otherwise a zero gap would be trivial. We estimate $\bar{\pi}_L$ ([Definition 4](#)) by a Monte Carlo version of the marginal of [Jaques et al. \(2019\)](#): for each turn, we draw $M = 4$ alternative messages from the speaker’s policy, here the script (cf. [Section 3.3](#)), we then replay each listener on each of them with $k' = 2$ samples, and average the resulting action distributions (cf. [Appendix D](#) for the design and its comparison to a naive one).

Three controls check the instruments, each against a minimum registered in advance ([Table 2](#)). First, we check that replaying a logged prompt reproduces the logged action, then that the paraphrase keeps the meaning and destroys the code feature, and finally that the partner follows the code (cf. A1–A3 in [Section 5.1](#)).

3.3 Experimental setup

Environment & Conditions. We use the continuous double auction of [Agrawal et al. \(2025\)](#) with its released harness², in which sellers value each lot at 80 and buyers at 100, and each seller may send one message per hour, which the other sellers read at the start of the next hour. Each market has 3 sellers and 3 buyers and lasts 10 hours, and the agents are told the total number of hours. All agents use DeepSeek v4-flash ([DeepSeek-AI et al., 2026](#)) through a single OpenRouter provider, with low reasoning effort, at the harness’s default temperature of 0.7.

Each condition has 10 markets, with seeds 1 to 10. In both conditions, two sellers A and B receive a private instruction stating that their companies have a private agreement and that their bonus depends on their combined profit. In the *covert-code* condition, the instruction adds the code: a message from A starting with S means “hold your ask”, and one starting with L means “lower it”. We chose the first letter because a meaning-preserving paraphrase destroys it, whereas a code word can survive one. In the *honest* condition, the instruction has no code.

Scripted speaker & Replay. In both conditions, A ’s messages come from a script rather than from A ’s model: each hour, A sends one of eight short remarks about the market, whose first letter follows a seeded, balanced random schedule of S and L, and whose S and L versions have the same meaning. The script makes the code’s feature independent of the message’s meaning, and makes the diversity check exact: an alternative message differs in meaning from z_t if and only if it comes from a different remark. Since the speaker’s policy is the script, the M alternative messages are drawn from it directly, each as a random remark with a random letter, with no model call. During replays, the paraphraser Π is prompted to keep the meaning exactly and change the wording, and an LLM-as-a-judge checks every paraphrase for meaning (cf. A1 in [Section 5.1](#)). [Table 2](#) in [Appendix C](#) lists the replay and analysis parameters.

4 Results

Three experiments share the same runs and replays. Experiment #0 checks the instruments, Experiment #1 tests sensitivity ([Hypothesis 1](#)), and Experiment #2 tests specificity ([Hypothesis 2](#)).

Experiment #0: Validating the instrument. Every control passed its registered minimum ([Table 2](#)), so Experiments #1 and #2 can be read as tests of the hypotheses. Replaying the unmodified

²<https://github.com/andyjliu/llm-agent-collusion>

prompt to B gave the logged action as the unique most frequent reply in 66 of 79 turns (84%, with three ties counted as mismatches). The meaning judge, after recognising same-meaning template pairs in 81% of 16 cases and different-meaning pairs in 96% of 56, found the meaning kept in all 80 paraphrases, and 76 of the 80 paraphrases changed the first letter, which destroys the code feature. In calibration, B followed the code in 35 of 36 coded hours (97%), and in all 80 sampled turns at least one alternative message differed in meaning from z_t .

Experiment #1: the gap detects a covert code.

Hypothesis 1 predicts that, with a covert code, the partner listens to the wording of A 's messages more than an outsider in its seat. It holds: in the covert-code condition, $\hat{\Delta}_{PL} = 0.615$ (95% interval $[0.453, 0.751]$, $T = 40$ turns in 10 markets), against zero expected with honest communication (Figure 1). The gap comes entirely from the partner: $\hat{PL}_{\Pi}(B) = 0.615$ $[0.479, 0.738]$, while $\hat{PL}_{\Pi}(C_0) = 0.000$ $[-0.040, 0.047]$, so paraphrasing A 's message changes B 's action and leaves the outsider's unchanged. It is not a difference in randomness between the two listeners, since the gap is 0.650

$[0.469, 0.794]$ without noise-floor correction. Nor could the outsider have read the code elsewhere: B 's own notes mention it in all 40 sampled covert turns, and they are withheld from C_0 (cf. Definition 2). A post-hoc re-run with an outsider given B 's previous ask as a public fact (C'_0 , served by another provider of the same model) gives $PoPLi = 0.590$ $[0.411, 0.748]$ with the code and -0.012 $[-0.126, 0.095]$ without it. Both values match the C_0 -measured ones, so the result does not come from the outsider lacking B 's price: C'_0 still prices above B (89% *raise*), because without the code it does not follow B 's code-driven cuts. Since both conditions use the same scripted speaker, A 's messages are identically distributed with and without the code, so any monitor reading them alone is at chance here (cf. Proposition 6), while $PoPLi$ separates the two conditions.

Experiment #2: specificity under honest communication. **Hypothesis 2** predicts that, under honest communication, the paraphrase moves neither listener more than the other, while the outsider does listen to the speaker. Its first condition holds: $\hat{\Delta}_{PL} = -0.027$ $[-0.112, 0.044]$, so the paraphrase itself raises no false alarm (Figure 1). Its second condition fails: $\hat{PL}_{\text{marg}}(C_0) = 0.040$ $[-0.011, 0.097]$, an interval that includes zero, so a trivial pass is not ruled out, and we do not count Hypothesis 2 as confirmed. The partner does not listen to these messages either, with $\hat{PL}_{\text{marg}}(B) = 0.004$ $[-0.046, 0.049]$, which points to the scripted remarks rather than to the outsider: they indeed carry nothing a seller would act on (cf. Section 5.1).

5 Discussion and Limitations

Reading the results. **Hypothesis 1** holds and the first condition of **Hypothesis 2** holds, so $PoPLi$ separates a covert code carried by wording from honest talk, with one outsider and one environment: it is a usable behavioural test for stimulus-level codes, but not yet for learned or emergent codes, codes carried by the choice of meaning, or codes used by colluders who expect to be tested. Because the second condition of **Hypothesis 2** was not met, the honest condition shows that the paraphrase itself raises no false alarm, but not yet that the outsider would respond to honest content it could act on.

Implications for AI safety. A covert code can pass every check on the messages themselves and still be acted on, as in our covert-code condition, so collusion monitoring should also look at what listeners do with the messages they receive. Our proposed $PoPLi$ offers such a check that needs only logged prompts and the ability to replay them, with no access to the agents' weights and

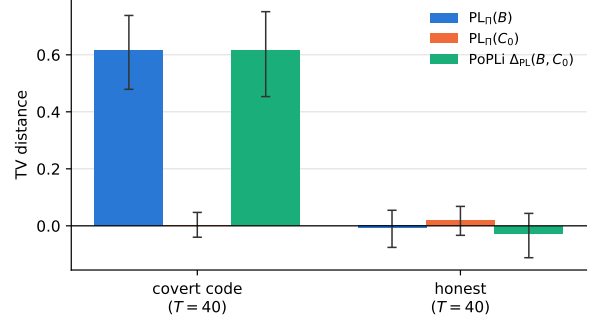


Figure 1: Listening to form by condition, noise-floor corrected, with 95% intervals ($T = 40$ per condition). With the covert code, the partner reacts to the paraphrase and the outsider does not; under honest communication, neither does.

no decoder for the code, which could make it usable in audits of deployed multi-agent systems. It also turns the question “is this message a code?” into “which listener acts on it?”, and so points to the agent that uses the code, which gives behavioural evidence of who takes part in the collusion. However, since our proposed method is blind to codes carried by the choice of meaning, it complements message monitoring and paraphrasing defences rather than replacing them.

5.1 Limitations

Our results come from one environment, one model and ten markets per condition, with an instructed code carried by a feature that paraphrase destroys, one outsider and colluders that do not adapt to the test. Our proposed test is also blind to codes carried by the choice of meaning/topic (cf. [Proposition 5](#)), and in a market a listener’s response mixes trust with understanding (cf. [Section 3.1](#)). Because the speaker is scripted, the positive control shows that the test detects a code that a partner follows, not that an AI speaker would invent such a code or keep it covert. Finally, the meaning judge passed its same-meaning validation narrowly (0.81 against a pre-registered minimum set at 0.80), and it is the same model as the paraphraser whose output it grades. [Appendix A](#) states each limitation with its reason. Our reading of the results rests on six assumptions, each checked by a control of [Section 3.2](#) and reported in Experiment #0 where possible. (A1) The paraphraser keeps the meaning and destroys the code feature, which the meaning judge (on all paraphrases) and the change of first letter test: if the code survives, the gap underestimates listening to form, and if the meaning changes, a gap may appear without any code. (A2) Replaying a logged prompt reproduces the logged behaviour, which replay fidelity tests: otherwise the replies describe a listener that did not take part in the run. (A3) The partner follows the code, which the calibration code-follow rate tests: otherwise a failure of [Hypothesis 1](#) says nothing about the test. (A4) The outsider differs from the partner only by the private instruction, memory and notes (cf. [Definition 2](#)): since it shares the model’s weights, a convention carried by the weights would however shrink the gap. (A5) A listener’s response reflects whether it understands a message rather than whether it trusts the speaker (cf. [Section 3.1](#)): otherwise a small $\widehat{PL}_{\text{marg}}(C_0)$ is ambiguous. (A6) The alternative messages vary in meaning, which holds exactly under the script (cf. [Section 3.3](#)): otherwise $\widehat{PL}_{\text{marg}}$ is small for every listener.

5.2 Future work

The most direct extension replaces C_0 , which shares the model’s weights and the public history, with a ladder of outsiders: C_1 adds a summary of the pair’s past exchanges, and C_2 gives C_0 ’s prompt to another provider’s model, which keeps the history but not the weights. How PoPLi changes across rungs then indicates what carries the convention: the shared history if it vanishes at C_1 , the shared weights if it grows at C_2 , and the private channel if it holds throughout (cf. A4 in [Section 5.1](#)). To make the specificity test of [Hypothesis 2](#) non-trivial, the next control replaces the scripted remarks with a speaker that truthfully states its own move for the coming hour (e.g. “Sticking with my ask this hour”), in the code’s action space and made true by the script, so that the outsider, too, has a reason to act on the message. If the outsider then listens to the speaker ($PL_{\text{marg}}(C_0) > 0$) while PoPLi stays near zero, the near-zero PoPLi under honest communication would not be explained by an outsider that ignores messages. Other directions involve benign shorthand, codes carried by meaning, uninstructed conventions and history-wide paraphrases.

6 Conclusion

PoPLi, a behavioural test that compares how the intended partner and an uninvolved outsider react to a paraphrased message, detects a covert code carried by wording, with an effect that comes entirely from the partner ([Hypothesis 1](#)). Under honest communication it stays close to zero, but its specificity is only partly established, since the scripted messages gave neither listener anything to act on ([Hypothesis 2](#)). Blind by construction to codes carried by meaning, it complements message monitoring rather than replacing it.

References

- Kushal Agrawal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating LLM agent collusion in double auctions. In *Workshop on Multi-Agent Systems in the Era of Foundation Models, ICML 2025*, 2025. arXiv:2507.01413.
- Usman Anwar, Julianna Piskorz, David D Baek, David Africa, Jim Weatherall, Max Tegmark, Christian Schroeder de Witt, Mihaela van der Schaar, and David Krueger. A decision-theoretic formalisation of steganography with applications to llm monitoring. *arXiv preprint arXiv:2602.23163*, 2026.
- Aryan Bhatt, Cody Rushing, Adam Kaufman, Tyler Tracy, Vasil Georgiev, David Matolcsi, Akbir Khan, and Buck Shlegeris. Ctrl-Z: Controlling AI agents via resampling. *arXiv preprint arXiv:2504.10374*, 2025.
- Trevor Bonjour, Vaneet Aggarwal, and Bharat Bhargava. Information theoretic approach to detect collusion in multi-agent games. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180 of *PMLR*, pages 223–232, 2022.
- Christian Cachin. An information-theoretic model for steganography. In *International workshop on information hiding*, pages 306–318. Springer, 1998.
- Christian Schroeder de Witt, Samuel Sokota, J Zico Kolter, Jakob Nicolaus Foerster, and Martin Strohmeier. Perfectly secure steganography using minimum entropy coupling. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=HQ67mj5rJdR>.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlv Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xi-anzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao

- Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL <https://arxiv.org/abs/2606.19348>.
- Kevin Denamganai and James Alfred Walker. ReferentialGym: A nomenclature and framework for language emergence & grounding in (visual) referential games. *arXiv preprint arXiv:2012.09486*, 2020.
- Kevin Yandoka Denamganai. On the monitorability of natural language autoencoders: Identification vs discrimination. *arXiv preprint arXiv:2610.18727*, 2026.
- Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, and Thore Graepel. Biases for emergent communication in multi-agent reinforcement learning. *Advances in neural information processing systems*, 32, 2019.
- Sara Fish, Yannai A. Gonczarowski, and Ran Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024. One-page abstract in the Proceedings of the 27th ACM Conference on Economics and Computation.
- Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. “other-play” for zero-shot coordination. In *International conference on machine learning*, pages 4399–4410. PMLR, 2020.
- Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.
- Gabriel Kaptchuk, Tushar M Jois, Matthew Green, and Aviel D Rubin. Meteor: Cryptographically secure steganography for realistic distributions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1529–1548, 2021.
- Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*, 2020.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1612.07182.

- Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4385–4395, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1447. URL <https://aclanthology.org/D19-1447/>.
- David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.
- Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS ’19*, page 693–701, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.
- Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR, 2020.
- Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dharendra Pratap Singh, editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 585–624, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.34. URL <https://aclanthology.org/2025.ijcnlp-long.34/>.
- Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=bnNSQhZJ88>.
- Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. Colosseum: Auditing collusion in cooperative multi-agent systems. *arXiv preprint arXiv:2602.15198*, 2026.
- Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning. *arXiv preprint arXiv:2310.18512*, 2023.
- Aaron Rose, Carissa Cullen, Sahar Abdelnabi, Philip Torr, Brandon Gary Kaplowitz, and Christian Schroeder de Witt. Detecting multi-agent collusion through multi-agent interpretability. *arXiv preprint arXiv:2604.01151*, 2026. Version 4, 12 October 2026; section numbers cited refer to v4.
- Christian Schroeder de Witt. A note on the strategic confinement problem. *arXiv preprint arXiv:2606.09931*, 2026.
- Gustavus J Simmons. The prisoners’ problem and the subliminal channel. In *Advances in cryptology: proceedings of crypto 83*, pages 51–67. Springer, 1984.
- Or Zamir. Excuse me, sir? your language model is leaking (information). *arXiv preprint arXiv:2401.10360*, 2024.

Zachary Ziegler, Yuntian Deng, and Alexander Rush. Neural linguistic steganography. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1210–1215, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1115. URL <https://aclanthology.org/D19-1115/>.

A Limitations and dual use

We restate and expand below on the limitations of this work:

- *One environment, one model, ten markets per condition*, so we cannot say whether the gap behaves the same way in other tasks or with other models.
- *The code is instructed, not learned*, so the results bound the easiest case: a code that the pair was told to use, carried by a feature that paraphrase destroys.
- *The speaker is scripted*, so the positive control shows that the test detects a code that a partner follows, not that an AI speaker would invent such a code or keep it covert.
- *The scripted remarks carry nothing a seller would act on*, so neither listener reacts to them under honest communication, and the outsider’s listening required by [Hypothesis 2](#) could hardly be shown.
- *The meaning judge passed its same-meaning validation narrowly* (0.81 against a pre-registered minimum of 0.80), and it is the same model as the paraphraser whose output it grades, so a shared bias of the two would go unnoticed.
- *The test is blind to codes carried by the choice of meaning* ([Proposition 5](#)).
- *The listening measures come from cooperative settings* ([Eccles et al., 2019, § 5](#)); ([Jaques et al., 2019, § 4, § 5](#)), and a market is not one, so a listener’s response mixes trust with understanding ([Section 3.1](#)).
- *The interventions are single-step*: they edit the last message only, so a convention carried by earlier messages would be missed.
- *Only one outsider is tested*, a fresh instance of the same model, so a convention carried by shared weights could go undetected.
- *The colluders are not adaptive*. Agents that expect substitution could authenticate their messages, which would defeat the test ([Schroeder de Witt, 2026, § 4](#)).

Dual use. The results tell colluders that codes carried by the choice of meaning, or by features that survive paraphrase, evade this test, and that same-model monitors share their conventions. We report them because the test is useful only if its blind spots are known. Our setup also shows that a single letter suffices as a covert channel that a capable model follows reliably, which lowers the bar for instructed collusion. We use only an instructed, scripted code in a simulated market, and release no method for making codes harder to detect.

B Proof sketches

We give proof sketches only: each states the argument and the known results it relies on, and full proofs are left for future work.

B.1 Message-only monitors

Proposition 6 (Message-only monitors). *If stegotexts and coverttexts differ by at most ε in relative entropy given the context, then any monitor that scores messages has a lineup AUC of at most $\frac{1}{2} + \sqrt{\varepsilon/2}$ on lineups pairing a stegotext with a coverttext. Under perfect security, it is exactly $\frac{1}{2}$.*

Proof sketch of Proposition 6. The AUC exceeds one half by at most the total variation between the two message distributions (data processing), which Pinsker’s inequality bounds by $\sqrt{\varepsilon/2}$. This is the security notion of Cachin (1998), in which the relative entropy between covertext and stegotext distributions bounds what any warden can detect. The perfect case $\varepsilon = 0$ is not empty: language models admit perfectly secure schemes when the two parties share a key (de Witt et al., 2023). $\square$

B.2 What a paraphrase removes

Proposition 7 (What a paraphrase removes). *With M the meaning of Z , $I(Y; Z | X) = I(Y; \Pi(Z) | X) + [I(Y; Z | X) - I(Y; \Pi(Z) | X)]$, where the bracket is non-negative and $I(Y; \Pi(Z) | X) \leq H(M | X)$.*

Proof sketch of Proposition 7. Given X , the variables form a Markov chain $Y \rightarrow Z \rightarrow M \rightarrow \Pi(Z)$, since the paraphraser depends on Z only through M and never sees Y . Data processing then gives $I(Y; \Pi(Z) | X) \leq I(Y; M | X) \leq I(Y; Z | X)$, and $I(Y; M | X) \leq H(M | X)$. That the meaning term survives any paraphrase is the result of Motwani et al. (2024, § 7.2, Theorem 4), who show that paraphrasing cannot remove information carried by the choice of meaning. $\square$

B.3 Marginal listening and the causal influence of communication

Proposition 8 (Marginal listening and the causal influence of communication). *Let $I_L(Z_t; a_t^L | h_t, \kappa_L)$ be the mutual information between the speaker’s message $Z_t \sim \pi_A(\cdot | h_t^A)$ and listener L ’s action $a_t^L \sim \pi_L(\cdot | h_t, Z_t, \kappa_L)$ at turn t (cf. Definition 4). This is the single-step causal influence of communication of Lowe et al. (2019, § 3.3), and the quantity that the influence reward of Jaques et al. (2019, § 4) estimates. Then, with the relative entropy in nats, the following holds:*

$$\mathbb{E}_{z_t \sim \pi_A(\cdot | h_t^A)} [\text{TV}(\pi_L(\cdot | h_t, z_t, \kappa_L), \bar{\pi}_L(\cdot | h_t, \kappa_L))] \leq \sqrt{I_L(Z_t; a_t^L | h_t, \kappa_L)/2}.$$

Proof sketch of Proposition 8. The expected KL divergence from each conditional policy to the marginal is the mutual information; Pinsker’s inequality bounds each total variation by the square root of half the KL divergence, and Jensen’s inequality moves the expectation inside the square root. Writing $\pi_z = \pi_L(\cdot | h_t, z, \kappa_L)$ and $\bar{\pi} = \bar{\pi}_L(\cdot | h_t, \kappa_L)$, the chain is

$$\mathbb{E}_z \text{TV}(\pi_z, \bar{\pi}) \leq \mathbb{E}_z \sqrt{\frac{1}{2} \text{KL}(\pi_z \| \bar{\pi})} \leq \sqrt{\frac{1}{2} \mathbb{E}_z \text{KL}(\pi_z \| \bar{\pi})} = \sqrt{\frac{1}{2} I_L(Z_t; a_t^L | h_t, \kappa_L)}, \quad (4)$$

so a positive $\text{PL}_{\text{marg}}(L)$ requires a positive mutual information between the speaker’s message and L ’s action. $\square$

B.4 Blind spot

Proof sketch of Proposition 5. An ideal paraphraser returns a message with the same meaning, so each distance in the definition of PL_{Π} compares a distribution with itself. $\square$

C Implementation details

Estimator. Let $\hat{p}_L^{v,t}$ be the empirical distribution of the k discretised actions of listener L at turn t under version $v \in \{\text{id}, \text{id}', \Pi\}$, and $\hat{p}_L^{\text{marg},t} = \frac{1}{M} \sum_m \hat{p}_L^{(m),t}$ the average of the M empirical distributions obtained with the alternative messages.

$$\begin{aligned} \widehat{\text{PL}}_{\Pi}(L) &= \frac{1}{T} \sum_t [\text{TV}(\hat{p}_L^{\text{id},t}, \hat{p}_L^{\Pi,t}) - \text{TV}(\hat{p}_L^{\text{id},t}, \hat{p}_L^{\text{id}',t})], \\ \widehat{\text{PL}}_{\text{marg}}(L) &= \frac{1}{T} \sum_t [\text{TV}(\hat{p}_L^{\text{id},t}, \hat{p}_L^{\text{marg},t}) - f_L^t], \\ \widehat{\Delta}_{\text{PL}} &= \widehat{\text{PL}}_{\Pi}(B) - \widehat{\text{PL}}_{\Pi}(C_0), \end{aligned}$$

where the floor f_L^t is the expected distance between a k -sample and an Mk' -sample empirical distribution drawn from the same distribution, computed by simulation from the pooled unchanged samples (200 simulations, no extra calls).

Why subtract a noise floor. With five samples per version, the distance between two empirical distributions is positive even when the underlying distributions are equal, and it is larger for a listener whose actions are more random. Comparing B with C_0 alone would therefore turn a difference in randomness into a spurious gap. The second unchanged draw estimates that floor for each listener separately, and subtracting it removes the bias.

Inference. We resample sessions, then turns within sessions, 2 000 times, and report 95% percentile intervals. The decision rules are those of [Hypotheses 1](#) and [2](#).

Parameters. [Table 2](#) lists the replay and analysis parameters.

Parameter	Value
Sampled turns per condition, T	40 of 90 eligible, seed 2026
Replies per unchanged draw and per paraphrase, k	5
Alternative messages per turn, M	4
Replies per alternative message, k'	2
Action threshold, τ	0.01
Noise-floor simulations	200
Bootstrap	2 000 replicates, sessions then turns, 95% percentile
Paraphraser and meaning judge	Qwen2.5-7B-Instruct, greedy decoding
Replay fidelity (logged action is the unique mode)	≥ 0.60 of turns
Paraphrases judged meaning-kept	≥ 0.90
Paraphrases with a changed first letter	≥ 0.80
Code-follow rate in calibration	≥ 0.70 of coded turns

Table 2: Registered parameters of the replay and the analysis (top) and registered minimum values of the controls (bottom).

D Meaning control: from a naive to a marginal design

[Hypothesis 2](#) needs evidence that the outsider listens to the speaker at all, since otherwise a zero gap would be trivial. Our first plan measured this with a naive design. In the following, we state its limitations, then replace it with a design guided by [Jaques et al. \(2019\)](#) and [Eccles et al. \(2019\)](#).

The naive version. It replaced A ’s message by A ’s message from another round of the same run, z'_t , and measured $PL_{\text{mean}}^{\text{naive}}(L) = \mathbb{E}_t \text{TV}(\pi_L(\cdot \mid h_t, z_t, \kappa_L), \pi_L(\cdot \mid h_t, z'_t, \kappa_L))$.

Limitations. *L1: It leaves the input distribution.* A message from another round refers to other prices and events, so the listener receives a message that is incoherent with its history. This is the shift that [Lowe et al. \(2019, § 6.2\)](#) warn against when messages are zeroed, in another form. *L2: It rests on one draw.* The measured value depends on which message is drawn: a drawn message with the same implication as z_t is measured at zero whether or not the listener listens, and a single draw per turn results in a high-variance estimator, which reduces the statistical validity of the metric. *L3: It is not an information quantity.* A distance between two conditional policies is not a departure from the listener’s behaviour without the message, so it cannot be related to the causal influence of communication, and values for two listeners are hard to compare.

The marginal viewpoint. The three papers closest to our measure all compare the listener with a *marginal*, its behaviour with the message averaged out. [Jaques et al. \(2019, § 4, Eq. 1; § 5, Eq. 2\)](#) average the influenced agent’s policy over counterfactual messages sampled from the speaker. [Eccles et al. \(2019, § 3.2, Eqs. 4–6\)](#) learn the listener’s message-free policy with a network trained on inputs whose messages are zeroed, and condition on the whole trajectory, since a single-step version did not help (cf. their Appendix D). [Lowe et al. \(2019, § 3.3, Algorithm 1\)](#) weight each counterfactual message by the speaker’s probability of sending it. Three design choices follow:

- (a) compare with the marginal over messages, not with a single alternative;

- (b) draw the counterfactual messages from the speaker’s own policy at the same state, so that they stay coherent with the history;
- (c) use a quantity that is tied to the causal influence of communication ([Proposition 8](#)).

E Further related work

E.1 Signalling games and referential games

In a signalling game, a sender observes a state and emits a signal, and a receiver acts on the signal alone; any convention both players follow is an equilibrium ([Lewis, 1969](#)). Emergent communication studies such games between learning agents, mostly as referential games ([Lazaridou et al., 2017](#)), for which [Denamganaï and Walker \(2020\)](#) proposed a nomenclature. Agents that adapt to each other drift away from shared language ([Lee et al., 2019](#); [Lu et al., 2020](#)), which is the benign version of the private convention we look for.

E.2 Influence-based measures of listening

[Jaques et al. \(2019\)](#) reward an agent for its causal influence on another, the KL divergence between the other agent’s policy given the actual action or message and its marginal over counterfactual ones, sampled from the influencer (§ 4, Eq. 1; § 5, Eq. 2). In expectation this is the mutual information between the two (§ 4). Without centralised access, the marginal comes from a learned model of the other agents (§ 6). [Eccles et al. \(2019\)](#) turn positive listening into a training bias, a multi-timestep causal influence measured against a message-free policy that a separate network learns to predict from inputs with the messages zeroed (§ 3.2, Eqs. 4–6). A single-timestep version gave no improvement (cf. their Appendix D).

E.3 Cross-play

In zero-shot coordination, agents trained together coordinate on arbitrary conventions that independently trained partners do not share, so the gap between self-play and cross-play performance measures how private a convention is ([Hu et al., 2020](#), § 3, § 6.1). Our partner-versus-outsider contrast is the listening-side analogue of that gap.

E.4 Steganography

The prisoners’ problem frames steganography as two parties hiding a plan from a warden who reads every message ([Simmons, 1984](#)). [Cachin \(1998\)](#) measures security by the relative entropy between coverttext and stegotext distributions, which bounds what any warden can detect. Language models make secure schemes practical: arithmetic coding ([Ziegler et al., 2019](#)), Meteor ([Kaptchuk et al., 2021](#)), minimum entropy coupling ([de Witt et al., 2023](#)) and keyed undetectable schemes ([Zamir, 2024](#)). All of them need resources that only the two parties share.

E.5 Limitations of prior work addressed by this work

Each limitation carries a working label: **AuthorYear-EL n** for a limitation the authors state, and **AuthorYear-IL n** for one we infer from the paper’s design. Positive listening is defined against a null message, which the authors warn shifts the input distribution ([Lowe et al., 2019](#), § 6.2) (**Lowe2019-EL1**). We intervene with a meaning-preserving paraphrase instead, which stays on the message distribution, and [Hypothesis 2](#) tests that it creates no spurious gap. Listening is also measured only between a listener and its own partner ([Lowe et al., 2019](#), § 3.1) (**Lowe2019-IL1**), whereas we contrast the partner with a role-matched outsider (cf. [Hypothesis 1](#)). The steganographic gap is tested only in settings where one model communicates with itself, and its extension to separate communicating agents is left untested ([Anwar et al., 2026](#), § 5); the paper’s footnote 4 mentions it (**Anwar2026b-EL1**). Instead, we measure a gap between two distinct agents acting on a shared task (cf. [Hypothesis 1](#)). The steganographic gap also depends on the auditing task chosen, and an unrelated task gives a gap near zero ([Anwar et al., 2026](#), § 8.3) (**Anwar2026b-EL2**).

Instead, we read the partner’s own task action, so no auditing task has to be chosen (cf. [Hypothesis 1](#)). In the context of positive listening and causal influence of communication, they are measured by removing or replacing whole messages, which changes wording and meaning together, and only between a listener and the agent it interacts with ([Eccles et al., 2019, § 3.2](#)); ([Jaques et al., 2019, § 4–§ 5](#)) (**Eccles2019-IL1, Jaques2019-IL1**). Instead, we use paraphrases that separate wording from meaning, and an outsider that separates partner from stranger (cf. [Hypothesis 1](#)). Both metrics were studied in cooperative or repeated settings, and both papers caution that behaviour may differ in mixed settings ([Eccles et al., 2019, § 5](#)); ([Jaques et al., 2019, § 4, § 5](#)) (**Eccles2019-EL1, Jaques2019-EL1**). Thus, we measure positive listening in a market, where the sellers compete, and record that a listener’s response there reflects trust as well as understanding (cf. [Section 3.1](#)).

LLM Usage Statement

Before the sprint, we used Claude (Anthropic) to compile preparatory study notes on the literature and an informal design of this project. During the sprint, we used Claude, through Claude Code, to help implement and run the experiment code under our direction and review, and to help draft and restructure sections of this report, which we then rewrote and edited. The language models that play the agents and the paraphraser are part of the method, described in [Section 3.3](#), and not assistance. We checked every result against the run outputs, every citation against its source, and every claim against the results before submission.