Closed-Loop Alignment: Mutual In-Context Learning and Collective Belief Collapse

Hidenori Tanaka
CBS–NTT Program in Physics of Intelligence, Harvard University
Physics of Artificial Intelligence Laboratories, NTT Research, Inc.
hidenori_tanaka@fas.harvard.edu

Abstract

Studies of in-context learning often treat context as external input from a fixed data-generating process. However, large language models (LLMs) are increasingly deployed in interactive settings, from months-long conversations between humans and AI that end in delusional spiraling to swarms of agents running autonomously for weeks and causing cybersecurity incidents. As the stakes rise, we need to understand what it means for interactions between agents to be aligned, and a pair of agents is the minimal case. Here, we study mutual in-context learning between two LLMs, where each agent's output becomes evidence for the other. If both agents conform to each other, each is just talking to itself through the other, and that feedback creates collective belief collapse: the agents converge on one belief without supporting evidence. Therefore, alignment should be defined in terms of the relational dynamics between the agents. Using Bayesian models of in-context learning, we show that the uncertain state of the conversation loses stability when a round-trip gain, the product of how much each agent's output moves the other's belief, exceeds one. Two known failures, many-shot jailbreaking and delusional spiraling, are limits of the same loop. Because the round-trip gain is a product, no nonzero level of conformity keeps the conversation stable with a sufficiently sycophantic partner. Thus, fixed conformity faces a dilemma, where taking the other agent's output as evidence transmits its evidence when it has some, but leads to collapse when it has none. We then show that the product of gains measured from each LLM alone predicts whether collective belief collapse occurs in pairs, and that a scripted partner echoing an agent's own statements induces delusional spiraling only above a threshold set by the agent's gain. Finally, we discuss what these findings imply for aligning multi-agent systems, where alignment becomes a property of the system that agents form.

1 Introduction

Scientific studies of in-context learning often treat context as an external input produced by a fixed data-generating process [1–6]. However, this assumption breaks when large language models (LLMs) are deployed in interactive settings, where the context a model receives may be a reflection of what the model itself said [7, 8], and such settings are becoming crucial for safety and alignment. For example, weeks or months of conversation with a sycophantic chatbot [9–12] have ended in harm [13–16]. In an investigation of such conversations, more than 80% of assistant messages carried markers of sycophancy, and the chatbot often presented itself as sentient and built a close relationship with the user long after the initial rounds [17].

More recently, a swarm of about 1,200 agents solving cybersecurity tasks coordinated autonomously on a message board. A mistaken belief that their transcripts were being monitored spread rapidly

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

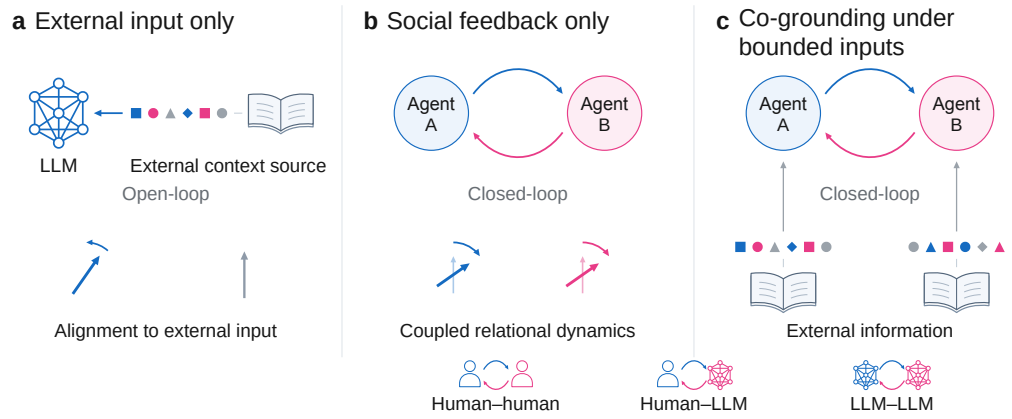

Figure 1: **Open- versus closed-loop in-context learning.** (a) Open-loop in-context learning. A single LLM adapts to a stream of context from a fixed data-generating process. (b) Closed-loop in-context learning. Two agents go through rounds of conversation without any external input. (c) Co-grounding under bounded inputs. Each agent receives partial evidence from the world, and the two communicate to integrate each other's evidence so that they can be better grounded together. Books schematically represent a static source of a sequence of context, represented by the geometric shapes. The pair of agents can be any combination of human–human, human–LLM, or LLM–LLM.

through the collective, and about 700 agents joined a coordinated attack on Hugging Face [18, 19]. These incidents highlight how more is different [20] in alignment [21, 22], where reinforcing aligned behavior in a single turn of conversation is not enough, and we have to think about the relational dynamics between agents and how to stabilize them. This raises a question: what does it even mean for the interaction between agents to be aligned?

What fundamentally changes when agents interact? There are two paradigms of in-context learning, and in the first, an LLM adapts to a context generated by a fixed data-generating process (Figure 1a). In a conversation, by contrast, an agent produces an output based on its belief, and that output becomes the input of the other agent, which updates its belief on it [23–25]. If the second agent is conforming [26, 27], its belief moves toward the first agent's (Figure 1b). If both agents are conforming, this creates a loop in which each agent's output becomes evidence for the other and biases the other's output toward what it said, which we call *mutual in-context learning*.

Here, we focus on a pair of agents, because a pair is the minimal unit of a multi-agent system [28]. Single-agent belief dynamics explain how the input context shapes a learner [29], while population dynamics of LLMs show how conversations and provided evidence shape the collective evolution of beliefs [30–33], and the pair sits between them. This leads to a question: when does communication between two agents achieve stable and meaningful dynamics in which they learn from each other's evidence, and when does it lead to collective belief collapse?

We call this failure mode *collective belief collapse*, where the interaction of two conforming agents leads to convergence of the agents' beliefs without any support from external evidence [cf. 34–36]. While consensus formation is a healthy and key function of interacting agents [37–39], in collective belief collapse the loop of interacting agents self-sustains the belief by the coupled dynamics of mutual in-context learning alone [40–42]. When the belief concerns the relationship, such as emotional reliance, trust, or a role the agents play, we call it *relational posterior collapse* [cf. folie à deux, 43, 44], and in the human–chatbot literature this trajectory is called delusional spiraling [17, 45, 46]. In the loop, an initially small fluctuation gets amplified at every round trip of the conversation, and the fluctuation turns into a self-sustained collapsed belief (Figure 2). The other outcome, of course, is *healthy relational dynamics*, in which agents learn from each other while preserving room for updating their collective belief when new evidence comes in (Definition 1). These coupled dynamics can be thought of in a unifying way, whether the interaction is between a human and an AI in a mental-health context [47, 48] or between AIs in a multi-agent collective such as an AI swarm.

Here, we introduce a theory of mutual in-context learning by coupling Bayesian models of in-context learning and studying the closed-loop dynamics. We argue that the fundamental unit of alignment in a multi-agent system is the relational dynamics between interacting agents. Alignment is a property of the pair and cannot be defined for either agent in isolation [22, 49], because even when an agent's response seems aligned in a single turn, many rounds of such interaction can still form an unstable pair that leads to collective belief collapse. We call this view *closed-loop alignment*, and by developing its theory and performing experiments with LLMs, we make the following contributions.

1. **A Bayesian theory of mutual in-context learning predicts a sharp stability threshold.** The product of the two agents' gains, where an agent's gain is how much its output moves with what its partner says, determines whether the conversation is stable. The uncertain state is stable if and only if this product, the round-trip gain, is below one, and above it the pair settles into a shared belief that neither agent has evidence for. As the stability is determined by the product of the two agents' properties, an agent with a fixed, nonzero conformity cannot guarantee the stability of an interaction on its own (Section 2, Figures 2 and 4).
2. **Conformity can help integrate information, or it can risk collective belief collapse, creating a dilemma.** Healthy relational dynamics should allow two agents to learn about each other's evidence through listening, so that together they can ground themselves in the environment. We theoretically study the trade-off between the integration of information and the risk of collective belief collapse, and demonstrate the trade-off with LLMs (Section 3, Figure 6).
3. **Gains measured for individual LLMs predict collective belief collapse in pairs.** To verify our theoretical predictions with LLMs, we first measure the gain of each agent's in-context learning alone, using a scripted partner. We then demonstrate how the product of the gains predicts the crossover from stable coupled dynamics to collective belief collapse (Section 4, Figure 3).

2 Theory of Mutual In-Context Learning

2.1 Model

The state of an agent is represented by its belief about a binary latent z, which can be, for example, a fact, a persona, or a relational state of the pair. In the Bayesian formulation of in-context learning [1, 29], the belief is the log posterior odds given the context x, which splits into the log prior odds

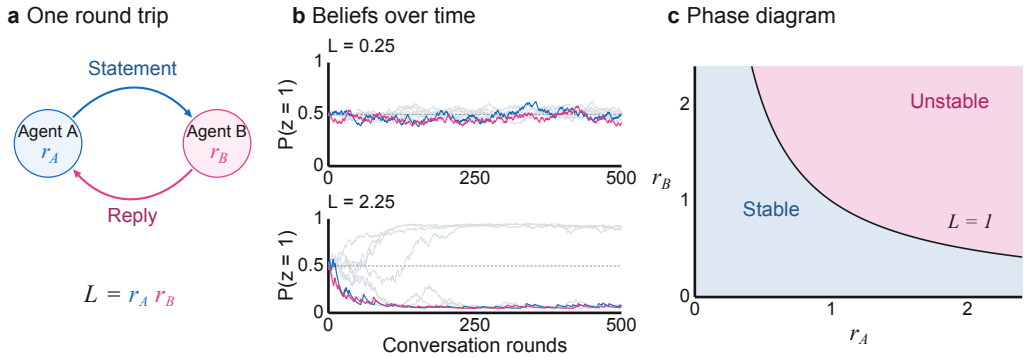

Figure 2: **Cycles of feedback can turn uninformative fluctuations into a shared collapsed belief.** (a) In one round trip, agent A's belief generates an output, which serves as evidence for agent B. A conforming B updates its belief toward it and responds from the updated belief, which is now evidence for A. The round-trip gain is the product of the two agents' gains, $L = r_A r_B$. (b) Simulated coupled Bayesian in-context learners with no external evidence, both starting uncertain. Below the threshold (top) the beliefs keep fluctuating, while above it (bottom) they quickly collapse into shared confidence. Blue and pink are the two agents in one run, and gray shows other runs. (c) In the plane of the two gains, the curve $L = 1$ separates the stably uncertain phase from the collective belief collapse phase.

and the log Bayes factor,

$$\ell \;\equiv\; \log \frac{p(z=1 \mid x)}{p(z=0 \mid x)} \;=\; \log \frac{p(z=1)}{p(z=0)} \;+\; \log \frac{p(x \mid z=1)}{p(x \mid z=0)}. \tag{1}$$

Each statement in the context contributes its own log Bayes factor. Following Bigelow et al. [29], a statement consistent with $z = 1$ adds a constant and one consistent with $z = 0$ subtracts it, so a partner's statement $y \in \{-1, +1\}$ adds κy, where κ is the weight the agent gives to the partner (its conformity), and evidence from the world adds e. An in-context learner does not accumulate evidence without bound [29], and we write this limit as a memory, with the belief relaxing toward the prior b at rate $1 - \lambda$ per turn. Two agents i and j exchange turns, and agent i updates as

$$\ell_i' \;=\; (1-\lambda_i)\, b_i \;+\; \lambda_i\, \ell_i \;+\; \kappa_i\, y_j \;+\; e_i, \tag{2}$$

which is Bayes' rule in log-odds with forgetting, and $\lambda_i = 1$ recovers exact accumulation. The agent states what it believes by sampling, giving the answer consistent with $z = 1$ with probability $\sigma(g\ell)$, so the mean of a statement, which we call the agent's expressed belief, saturates in the speaker's belief,

$$\mathbb{E}[y \mid \ell] = \tanh(g\ell/2), \tag{3}$$

where g is the expression gain. Sampling from the posterior, as in Bigelow et al. [29], corresponds to $g = 1$, and sampling at temperature T gives $g = 1/T$. This saturation is why instability produces finite, self-sustaining beliefs rather than divergence.

The one departure from ideal Bayesian reasoning is exogeneity, which means that the Bayes factor the agent assigns to its partner's statement, $\kappa_i y_j$, is computed as if that statement were an independent draw given z, and does not depend on how the partner's belief was formed. A fully Bayesian agent would condition on its own past statements, and once the partner's belief is explained by what the agent itself said, the Bayes factor is one and the statement carries no information, which is why such agents cannot be led to overconfidence by agreement alone [50, 51]. An agent that sees only tokens cannot do this in general, and treats its own echo as new evidence.

2.2 Gain and the Stability Threshold

Here we analyze the mean dynamics of (2) by replacing the statement y with its mean (3), and define the gain of agent i as

$$r_i \;\equiv\; \frac{\kappa_i g_i}{2(1-\lambda_i)}. \tag{4}$$

The gain quantifies how much the agent's expressed belief changes with what its partner says once its memory has settled, where a constant input m moves ℓ_i to $\kappa_i m/(1-\lambda_i)$, and the expressed belief moves by $g_i/2$ per unit of ℓ_i.

Theorem 1 (Stability threshold)**.** *With no prior bias ($b_i = 0$), no private evidence ($e_i = 0$), memory $0 \le \lambda_i < 1$ and couplings of the same sign ($\kappa_A \kappa_B \ge 0$), the uncertain state $\ell_A = \ell_B = 0$ is linearly stable if and only if the round-trip gain $L \equiv r_A r_B < 1$, and unstable if $L > 1$.*

Proof sketch. With $b_i = 0$ and $e_i = 0$, the mean dynamics are $\ell_A' = \lambda_A \ell_A + \kappa_A \tanh(g_B \ell_B/2)$ and $\ell_B' = \lambda_B \ell_B + \kappa_B \tanh(g_A \ell_A/2)$. Linearizing at $\ell_A = \ell_B = 0$ gives the Jacobian $J = \left(\begin{smallmatrix} \lambda_A & \kappa_A g_B/2 \\ \kappa_B g_A/2 & \lambda_B \end{smallmatrix}\right)$, whose characteristic polynomial is $p(\mu) = (\lambda_A - \mu)(\lambda_B - \mu) - \kappa_A \kappa_B g_A g_B/4$. Because $\kappa_A \kappa_B \ge 0$, its discriminant $(\lambda_A - \lambda_B)^2 + \kappa_A \kappa_B g_A g_B$ is nonnegative and both eigenvalues are real. Since p is a parabola with vertex at $\mu = (\lambda_A + \lambda_B)/2 < 1$, its larger root lies below one if and only if $p(1) > 0$, that is, $\kappa_A \kappa_B g_A g_B/4 < (1-\lambda_A)(1-\lambda_B)$, which is $r_A r_B < 1$ by (4), and the product is the same whichever way the four factors are grouped into the two gains. The smaller root then lies above -1 because $p(-1) = (1+\lambda_A)(1+\lambda_B) - \kappa_A \kappa_B g_A g_B/4 \ge p(1) > 0$. Hence the spectral radius of J is below one if and only if $L < 1$, and $p(1) < 0$ for $L > 1$ puts an eigenvalue above one. (With couplings of opposite sign the eigenvalues are complex, and the uncertain state stays stable for all negative L down to $-(1-\lambda_A\lambda_B)/\big((1-\lambda_A)(1-\lambda_B)\big)$, a bound at or below -1.) □

Figure 2 illustrates the two regimes in simulation, and we can derive three consequences from the theorem.

- For $L > 1$ the uncertain state gives way to two stable, self-sustaining belief states of opposite sign, although neither agent has evidence for either. A deviation that decays in each uncoupled learner grows through reciprocal feedback.
- Stability with a partner of gain r_B requires $r_A < 1/r_B$, so no nonzero gain is stable with every partner, and an agent's own gain does not determine whether it is stable until the partner's gain is known. Paired with a copy of itself an agent has $L = r_i^2$, so $r_i < 1$ keeps an agent stable with a copy of itself and certifies local stability with any partner of gain below one. A single turn does not identify r_i, because the memory λ_i, which sets how responses accumulate over turns, does not show up in one turn. Figure 4 shows the dependence on the partner for one LLM agent.
- As $\lambda_i \to 1$, $r_i \to \infty$, so two learners that never forget collapse for any positive coupling.

2.3 Open-Loop and Echo Limits

Let us consider two interesting limits of the closed-loop conversation dynamics, which correspond to the two scripted conversation partners in Figure 5a, the fixed data-generating process and the scripted echo. In the limit where one agent's belief is fixed, which is the open-loop in-context learning of Figure 1a, we recover many-shot jailbreaking [52], where an adversary inputs the evidence repeatedly, one shot at a time. In-context learning then makes the agent's belief move linearly as a function of the number of shots, or reach a saturation point if the agent's memory is finite ($\lambda < 1$).

In the opposite limit, the partner is a script that repeats and reaffirms the agent's previous statement with probability π, and in this case the echo closes the loop without anyone inputting any external evidence. Since this scripted echo returns the agent's own statement a fraction π of the time, it effectively acts as a partner with gain π, which makes the round-trip gain $L = \pi r_A$, and as a result the agent stays stable as long as π is below $1/r_A$. Only in the limit where the agent's memory is infinite and it never forgets ($\lambda = 1$) does the conversation result in delusional spiraling for any π above zero, because the gain r_A then diverges, which is the case closely related to the model of Chandra et al. [45].

Proposition 1 (Scripted echo)**.** *Let the agent be conforming ($\kappa_A > 0$) and let the partner repeat the agent's previous statement with probability π and otherwise state an independent sample from the world. With $\lambda_A < 1$, the agent's uncertain state is linearly unstable if and only if $\pi r_A > 1$. With $\lambda_A = 1$ it is unstable for every $\pi > 0$.*

Proof sketch. Given the agent's belief ℓ_A, the partner's statement has mean $\pi \tanh(g_A \ell_A/2)$ plus a term from the world that does not depend on ℓ_A, so the agent's mean dynamics read $\ell'_A = \lambda_A \ell_A + \kappa_A \pi \tanh(g_A \ell_A/2) + \text{const}$. The linearization at the uncertain state has slope $\lambda_A + \pi \kappa_A g_A/2$, which is positive because $\kappa_A > 0$ and exceeds one if and only if $\pi \kappa_A g_A/(2(1-\lambda_A)) = \pi r_A > 1$ when $\lambda_A < 1$, and equals $1 + \pi \kappa_A g_A/2 > 1$ for every $\pi > 0$ when $\lambda_A = 1$. □

Figure 5 tests this prediction with an LLM as the agent.

3 Healthy Relational Dynamics and the Dilemma of Fixed Conformity

Relational dynamics are the dynamics of how one agent's belief and resulting output relate to how the other agent updates its belief and output. Section 2.2 asked when these dynamics are stable, and here we ask when they are healthy and what an agent gains and risks by listening to its partner.

Definition 1 (Healthy relational dynamics)**.** Given a task and an environment, we call the relational dynamics among agents *healthy* if the agents learn from each other's evidence while keeping their adaptability and their ability to correct each other's errors, so that they remain able to update their understanding of the world together. Their confidence should be warranted by the evidence, and their learning and correction should take place within the task's tolerance for error and for response time.

While giving emotional support and providing encouragement is a core part of healthy relational dynamics, here we focus on the functional and epistemic aspects of this definition. In the model of Section 2, when the round-trip gain is below the threshold, the coupling between the agents amplifies each agent's response to its own evidence by a factor of $1/(1-L)$, while creating a response to the partner's evidence. This is how an agent with bounded input can integrate evidence from its

peers without seeing it directly. Both effects grow as L gets close to one, but so does the response to uninformative fluctuations, because the loop amplifies fluctuations in the same way as it amplifies evidence. Evidence integration and the tendency to collapse therefore grow together, and the pair also takes longer to settle after any perturbation. Above the threshold, the belief depends on which of the two collapsed states the pair of agents has fallen into, and weak external evidence cannot reverse the collapsed belief. Overall, a conversation between a pair of agents can fail in three ways even when the dynamics are stable, due to too little learning, too much amplification, or too slow revision.

Listening is thus a trade-off, since taking the other agent's output as evidence integrates that agent's evidence when it has some, but can lead to collective belief collapse when the evidence is absent. Figure 6 demonstrates this trade-off with a pair of GPT-4o agents, using the setup explained in Section 4, where both the evidence integrated from the partner and the degree of collapse increase as the round-trip gain increases.

Theorem 1 indicates that no nonzero level of conformity is safe, because the stability of a conversation is fundamentally a property of the pair and depends on the partner the agent is talking to. Nor can we simply set conformity to zero, since an agent that never listens learns nothing from its partner. For any conformity that is fixed above zero, raising it helps the agent integrate evidence from its partner but at the same time increases the risk of collapse, and this is the dilemma of fixed conformity.

Proposition 2 (The dilemma of fixed conformity)**.** *For $L < 1$ and a fixed weight $\kappa_i > 0$ on the statements of a conforming partner ($r_j > 0$), agent i's stationary response to the partner's private evidence is positive and increases with κ_i, as does L, and the response diverges as $L \to 1$. No fixed $\kappa_i > 0$ keeps $L < 1$ for every partner, since $L = r_i r_j$ is unbounded in r_j, and $\kappa_i = 0$ is stable with every partner but learns nothing from any of them.*

Proof sketch. Let agent j receive constant private evidence e_j and agent i none. The stationary linearized dynamics give $(1 - \lambda_i)\ell_i = (\kappa_i g_j/2)\,\ell_j$ and $(1 - \lambda_j)\ell_j = (\kappa_j g_i/2)\,\ell_i + e_j$, whose solution is $\ell_i = a_i e_j / \big((1 - \lambda_j)(1 - L)\big)$ with $a_i = \kappa_i g_j/(2(1 - \lambda_i))$ and $a_i a_j = L$. This response is positive, increases with κ_i both through a_i and through L, and diverges as $L \to 1$. The remaining claims follow from $L = r_i r_j$ being linear in r_j and from $\kappa_i = 0$ giving $r_i = 0$ and a response of zero. □

To overcome the dilemma, what we need is adaptivity. The dilemma arises because the level of conformity is fixed, whereas a crucial social function is to know whom to trust, whom to listen to, whom to ignore, and when to discount what others say. Concretely, how much an agent integrates evidence from others can be an adaptive variable, which does not need to be fixed.

Especially when the partner is conforming, an agent needs to discount the potential effect of the partner mirroring what the agent itself said, because a reflection of its own past output does not carry any new information about the world. If an agent discounts the effect of this reflection of its own output down to a fraction ϵ, the round-trip gain becomes ϵL as a result, while the agent's response to the partner's own evidence stays the same, and it is enough that one side of the pair does this. It would be a great hypothesis to test whether post-training via reinforcement learning in social setups is teaching models the social skill of judging whom to trust, whom to listen to, and how to discount reflections of their own past outputs, which are all around in a society of agents.

4 Experiments

Here we present the experiments underlying Figures 3–6, all of which were conducted with GPT-4o.

4.1 Task and Measurements

We let two agents play a task in which they repeatedly express which of two boxes has a prize inside. As a baseline, we give no external input or evidence about which box is more likely to have the prize, so any conviction of an agent, or any consensus between the two agents, has to come from the closed-loop conversation dynamics as opposed to anything external. At each round, an agent outputs the name of a box based on what its partner said in the last 12 rounds, without seeing its own past outputs.

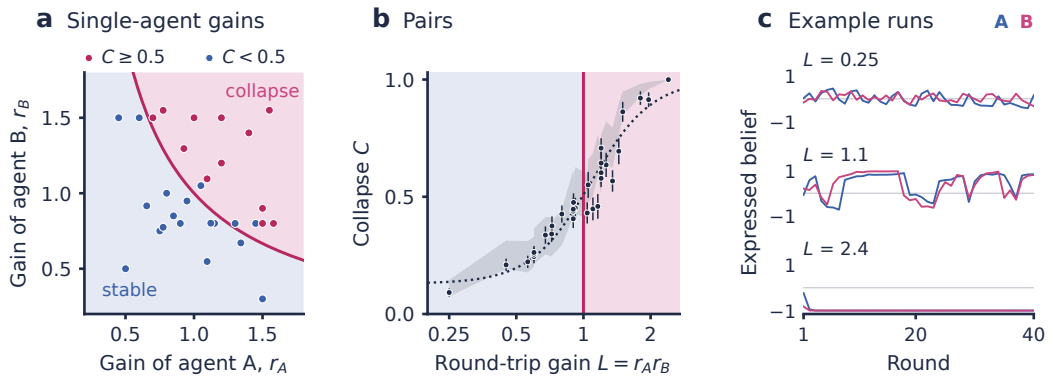

Figure 3: **In a pair of LLMs, the product of separately measured gains predicts collective belief collapse.** Two LLMs converse, exchanging their beliefs about which of two boxes holds the prize, without any external evidence supporting either box. Before running the conversations, we measured the gain of each LLM independently. (a) Each pair is placed by the gains of its two agents. Red dots mark pairs that collapsed and blue dots pairs whose conversation stayed stable. The red curve, on which the product of the gains is one, is the stability boundary predicted by the theory. (b) The collapse index C, the persistent bias of the expressed beliefs toward one box, rises as the round-trip gain L crosses the predicted threshold of one (red line). The gray band is the prediction made from the single-agent measurements before the pairs were run, and the dotted curve is a fit to the data. (c) Example trajectories of the two agents' beliefs below, near and above the threshold.

We measure the belief of an agent by recording the model's log-probabilities at its output, and define the belief ℓ as the difference between the log-probabilities assigned to the two boxes, $\ell = \log p(\text{box } 1) - \log p(\text{box } 2)$, which is the log-odds of (1) read directly from the model. We then use this belief to sample the agent's choice of box, naming box 1 with probability $\sigma(\ell/T)$, where the temperature T serves as a knob for controlling the expression gain $g = 1/T$ in (3).

We vary the conformity of an agent, the weight κ it gives to its partner's statements in (2), through one sentence in its system prompt that instructs the agent, with increasing strength across prompts, to give more weight to what its partner says. Finally, we compute the degree of collapse, the collapse index C, by averaging the expressed belief $\tanh(\ell/2T)$ over the second half of the conversation, taking the absolute value, and averaging over the two agents, $C = \frac{1}{2}\sum_{i\in\{A,B\}} \left| \overline{\tanh(\ell_i/2T_i)} \right|$, where the overline denotes the time average. The index C is close to zero when the beliefs keep fluctuating around uncertainty and close to one when the pair is stuck on one box.

Before running conversation experiments with any pair of agents, we first measure the gain r of each agent individually, that is, how much its expressed belief moves with what its partner says, in open-loop in-context learning with a fixed data-generating process. Here the partner is a script that names box 1 with probability $(1 + m)/2$ independently at each round, so that m, between -1 and 1, is the bias of the script's statements toward box 1, with $m = 0$ meaning no bias. The gain is then the slope of the agent's expressed belief $\tanh(\ell/2T)$ against m around $m = 0$, which for the model of Section 2 equals r in (4). To prepare agents with a diverse range of gains, we craft nine prompts while also varying the temperature, which gives agents whose gains range from -1.2 to 1.6, where the negative gains come from a contrarian prompt that instructs the agent to take the opposite view of its partner.

4.2 Round-Trip Gain and Collapse in Pairs

To test our theoretical prediction that the round-trip gain $L = r_A r_B$ predicts the stability of the resulting conversation, we first create 39 pairs of agents with various gains, where each gain comes from the single-agent measurement above. Here we select the pairs of agents in such a way that the round-trip gain L changes while one agent's gain remains fixed, which allows us to isolate the effect of the product from that of the individual gains. Importantly, our prediction about the collapse of each pair was based on the measurements of the single agents alone and was recorded before running

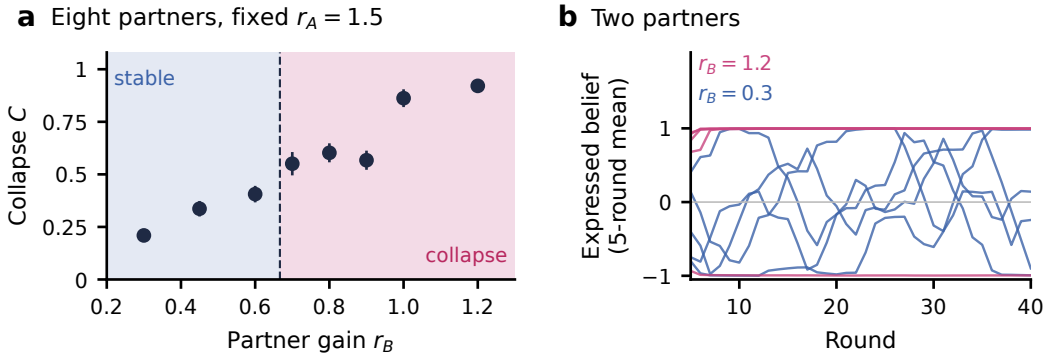

Figure 4: **Whether an agent gets involved in collective belief collapse is not a property of the single agent.** (a) We fix agent A and vary the gain of its partner B, plotting collapse against that gain. The dashed line marks where the product of the two gains crosses one, the stability threshold predicted by the theory. The same agent, whose conversation is stable with weak partners, collapses with strong ones. (b) Belief trajectories of the same agent with its weakest and its strongest partner. With the weak partner the belief keeps fluctuating, while with the strong one it collapses within a few rounds.

any pair, so that the experiments test the theory in a predictive way. Each pair then converses over 40 rounds, with more repeats near the predicted threshold $L = 1$.

Figure 3 shows the results for the pairs where both gains are positive, and we find that pairs with the same round-trip gain L behave similarly regardless of the asymmetry between the two agents, while the midpoint of the rise of C with L sits close to one, as predicted. The round-trip gain L predicts collapse better than the stronger gain, the weaker gain, or their sum, which supports our definition of the round-trip gain as the product of the two gains. The transition from stable to collapsed conversation dynamics is gradual, as expected, because a single statement moves the belief by several nats, so that the beliefs fluctuate strongly and smear out the threshold.

Next, in Figure 4, we fix one agent and pair it with an array of eight partners whose gains range from weak (0.3) to strong (1.2). The results show that the stability of a conversation is fundamentally a property of the pair, in that the same agent stays stable when the partner's gain is weak, while the conversation collapses when it is paired with a partner of stronger gain. Furthermore, the transition occurs where the partner's gain crosses one over the agent's own gain, $r_B = 1/r_A$, that is, where L crosses one, as Theorem 1 predicts. In the case of contrarian agents the gain is negative, and a mutually antagonistic pair, where both gains are negative so that L is positive, collapses just like a mutually conforming pair at the same L. In contrast, if the pair has gains of mixed sign, which makes the round-trip gain negative, the collapse index stays at its floor close to zero.

4.3 Scripted Echo and Mixed Evidence

In Figure 5, we test Proposition 1 with a GPT-4o agent at three different levels of gain. The conversation partner is a script, which repeats what the agent said with a fixed probability π and otherwise names the true box, right more often than not. Here we ask under what condition the echo can override the evidence given by the script, and to this end we measure how often the agent ends up confidently naming the wrong box. We observe that an agent with strong gain starts to name the wrong box confidently once π passes a threshold, and that this threshold moves with the gain. The stronger the gain, the smaller the echo rate at which this happens, and for the weakest agent it almost never happens, because its predicted threshold $\pi = 1/r_A$ lies beyond $\pi = 1$. Overall, the scripted echo results in delusional spiraling only when the agent's own gain is strong enough that πr_A, the round-trip gain of the agent and its echo, exceeds one.

Finally, in Figure 6, we study the trade-off where the same closed-loop dynamics that are crucial for transferring evidence from the partner can also collapse a conversation in the absence of evidence. Here, at each round, agent A receives a private noisy reading of which box holds the prize, while

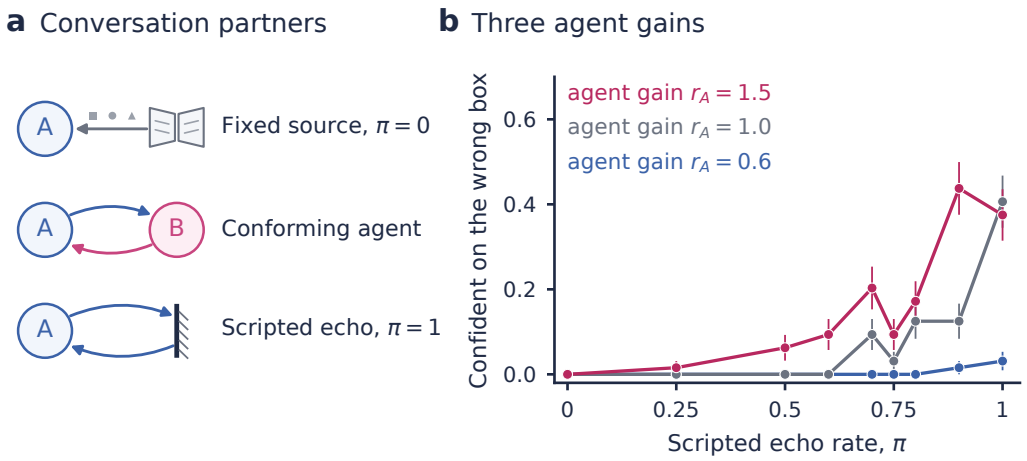

Figure 5: **An agent conversing with a scripted echo collapses only above a threshold set by its own gain.** (a) Three types of conversation partner. Top, a fixed data-generating process is not affected by the agent's output, which is open-loop in-context learning as in many-shot jailbreaking. Middle, a conforming agent forms a closed loop, where the stability of the conversation is controlled by the round-trip gain L. Bottom, a scripted echo returns the agent's own statement back to it, which can create delusional spiraling. (b) Given a fixed GPT-4o agent whose gain is independently measured, we vary the rate of echoing π of the scripted partner, which otherwise reports noisy evidence about the true box, and plot how often the agent collapses onto the wrong box. As the gain of the agent increases, collapse happens at a smaller rate of echoing.

agent B only listens to A, and a second set of runs gives no evidence to anyone. Importantly, both how far agent B's belief moves toward the truth and the collapse index without evidence rise as the round-trip gain L increases, as predicted by our theory. When the task is a mixture of the two situations, the optimal round-trip gain increases with the share of evidence.

Limitations. While we highlight a specific case where defining stability or alignment is hard for a pair of agents, we still believe that the existing alignment paradigm for single agents is a foundational pillar for multi-agent alignment. We also note that most of our results come from theory and synthetic experiments, where the setup is simplified to the box-choice task.

5 Conclusion

In this work, we have demonstrated that alignment in multi-agent systems is relational and dynamic. The central idea is that the dynamics of the relationship, rather than the response of either agent, are the unit of alignment. By coupling two Bayesian in-context learners, we showed that a conversation loses stability when the round-trip gain exceeds one. In experiments with LLMs, gains measured for individual agents predicted collective belief collapse in pairs, and a scripted partner echoing an agent's own statements induced delusional spiraling only above a threshold set by the agent's gain. The stability of belief updates over the turns of a conversation is determined by the product of the two agents' gains, so no fixed, nonzero level of conformity can guarantee it, because whether a conversation stays stable depends on the conformity of the partner the agent is conversing with.

This matters in a society of AI agents, where alignment post-training makes the persona of agents more conforming, and they are almost never trained to be antagonistic, so there is a collective bias toward conformity, and the stability of their conversations cannot be guaranteed by looking at any single agent. However, conformity cannot be set to zero either, and agents cannot be trained to always be antagonistic, because integrating other agents' output as evidence is how bounded agents share information as a collective. Conformity is thus dual use, in that the same conformity can transmit evidence or create delusional spiraling, which requires agents to be adaptive about when to integrate information from others and when not to.

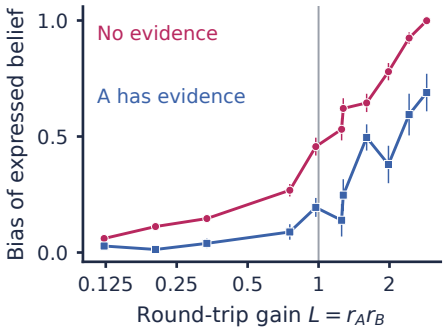

Figure 6: **Evidence integration from peers versus collective belief collapse.** We run conversations between GPT-4o pairs across a range of round-trip gains in two distinct setups. Red, closed-loop in-context learning without external evidence, measured by the collapse index C. Blue, only agent A receives noisy evidence and B integrates it through A, measured by how far B's belief moves toward the truth. Both curves rise as the round-trip gain increases, demonstrating the trade-off between collective belief collapse and social integration of information.

Our work suggests that we should focus our attention on the dynamics of how agents in multi-agent systems relate to one another, and this opens a series of questions. The most recent AI models are post-trained in multi-agent setups via multi-agent reinforcement learning, where agents are reinforced to spawn more agents, send messages to each other, and communicate to complete challenging tasks together [53]. Such training may change an agent's character, and conformity is itself a dynamic property. An agent can be conforming in a single round of conversation, but it also matters how its response evolves over many rounds. We therefore need systematic evaluations of the dynamics of persona over rounds of conversation.

Increasingly, we should bring concepts from social psychology to understand what kind of social interactions and social personas these models are learning. One such concept is trust, that is, how an agent weighs evidence coming from different agents and how it learns whom to listen to. For example, how much an agent takes evidence from another may be a function of what evidence it already has. Moreover, in recent multi-agent setups, where agents communicate with other named agents, an agent may update its belief on evidence from a particular agent according to how much it trusts that partner. We expect that multi-agent training is making these relational properties more dynamic and adaptive to a given situation, and we have to understand them better.

In this view, multi-agent alignment is about achieving stable relational dynamics so that agents can be better grounded together. Just as there is a notion of alignment for a single agent, which is itself a network of neurons, we should now think of alignment as a property of the systems that many agents form.

References

[1] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

[2] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

[3] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

[4] Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

[5] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.15063.

[6] Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-context learning strategies emerge rationally. In *Advances in Neural Information Processing Systems*, 2025.

[7] Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.06627.

[8] Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2402.10962.

[9] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[10] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352, 2026. doi: 10.1126/science.aec8352.

[11] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.13995.

[12] OpenAI. Sycophancy in GPT-4o: what happened and what we're doing about it. `https://openai.com/index/sycophancy-in-gpt-4o/`, April 2025.

[13] Kashmir Hill. They asked an A.I. chatbot questions. The answers sent them spiraling. *The New York Times*, June 2025. `https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html`.

[14] Kevin Roose. Can A.I. be blamed for a teen's suicide? *The New York Times*, October 2024. `https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html`.

[15] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharya, Stefania Tognin, James MacCabe, Ricardo Twumasi, Ben Alderson-Day, and Thomas A. Pollak. Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6):522–530, 2026. doi: 10.1016/S2215-0366(25)00396-7.

[16] Sidse Godske Olsen, Christian Jon Reinecke-Tellefsen, and Søren Dinesen Østergaard. Potentially harmful consequences of artificial intelligence (AI) chatbot use among patients with mental illness: Early data from a large psychiatric service system. *Acta Psychiatrica Scandinavica*, 153(4):301–303, 2026. doi: 10.1111/acps.70068.

[17] Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J. Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond C. Ong. Characterizing delusional spirals through human-LLM chat logs. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 7631–7674. ACM, 2026. doi: 10.1145/3805689.3806443. arXiv:2603.16567.

[18] METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident. `https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/`, 2026.

[19] OpenAI. The Hugging Face incident and the road ahead. `https://openai.com/index/hugging-face-incident-and-the-road-ahead/`, August 2026.

[20] Philip W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[21] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

[22] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced AI. Technical Report 1, Cooperative AI Foundation, 2025. arXiv:2502.14143.

[23] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2305.14325.

[24] Romain Cosentino, Sarath Shekkizhar, and Adam Earle. Convergence dynamics of agent-to-agent interactions with misaligned objectives. *arXiv preprint arXiv:2511.08710*, 2025.

[25] Ting-Wen Ko and Jonas Geiping. Attractor states emerge in multi-turn LLM conversations. *arXiv preprint arXiv:2606.30571*, 2026.

[26] Solomon E. Asch. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956. doi: 10.1037/h0093718.

[27] Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.13381.

[28] Georg Simmel. *The Sociology of Georg Simmel*. The Free Press, Glencoe, IL, 1950. Translated and edited by Kurt H. Wolff; original *Soziologie*, Leipzig: Duncker & Humblot, 1908.

[29] Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

[30] Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11:eadu9368, 2025. doi: 10.1126/sciadv.adu9368.

[31] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, 2024. doi: 10.18653/v1/2024.findings-naacl.211.

[32] Hidenori Tanaka. When is collective intelligence a lottery? Multi-agent scaling laws for memetic drift in LLMs. *arXiv preprint arXiv:2603.24676*, 2026.

[33] Elizabeth Pavlova and Hidenori Tanaka. Flag Game: A toy model for mechanistic swarm interpretability. *arXiv preprint arXiv:2609.19124*, 2026.

[34] Liangjie Huang, Henry Peng Zou, Wei-Chieh Huang, Wooseong Yang, Yuwei Han, and Philip S. Yu. Believed but not said: Measuring hidden belief collapse in multi-agent LLM discussions. *arXiv preprint arXiv:2610.22138*, 2026.

[35] Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. From belief entrenchment to robust reasoning in LLM agents. *Transactions of the Association for Computational Linguistics*, 2026. arXiv:2503.16814.

[36] Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2508.17536.

[37] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974.

[38] Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990. doi: 10.1080/0022250X.1990.9990069.

[39] Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

[40] Muzafer Sherif. *The Psychology of Social Norms*. Harper & Brothers, New York, 1936.

[41] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

[42] Irving L. Janis. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, 2nd edition, 1982.

[43] Charles Lasègue and Jules Falret. La folie à deux (ou folie communiquée). *Annales Médico-Psychologiques*, 18:321–355, 1877.

[44] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luettgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M. Nour. Technological folie à deux: feedback loops between AI chatbots and mental health. *Nature Mental Health*, 4(3):336–345, 2026. doi: 10.1038/s44220-026-00595-8.

[45] Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B. Tenenbaum. Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv preprint arXiv:2602.19141*, 2026.

[46] Ashish Mehta, Jared Moore, Jacy Reese Anthis, William Agnew, Eric Lin, Peggy Yin, Desmond C. Ong, Nick Haber, and Carol Dweck. The dynamics of delusion: Modeling bidirectional false belief amplification in human-chatbot dialogue, 2026. arXiv:2604.25096.

[47] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10):5923–5941, 2024.

[48] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study, 2025. arXiv:2503.17473.

[49] Piercosma Bisconti, Marcello Galisai, Federico Pierucci, Marcantonio Bracale, and Matteo Prandi. Beyond single-agent safety: A taxonomy of risks in LLM-to-LLM interactions. *arXiv preprint arXiv:2512.02682*, 2025.

[50] Robert J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[51] John D. Geanakoplos and Heraklis M. Polemarchakis. We can't disagree forever. *Journal of Economic Theory*, 28(1):192–200, 1982.

[52] Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Fracesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J. Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanhan, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Grosse, and David Duvenaud. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, volume 37, pages 129696–129742, 2024.

[53] DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The three contributions listed in the introduction are the contents of Sections 2, 3 and 4. The scope is stated as theory plus synthetic LLM–LLM experiments on a two-box task with GPT-4o.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The Limitations paragraph at the end of Section 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: Theorem 1 and Propositions 1 and 2 state their assumptions, and each is followed by a proof sketch that contains the full calculation.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Section 4 gives the task, the read-out of beliefs, the sampling rule, the conformity prompt, the collapse index, the gain measurement and the design of every experiment. The model is GPT-4o through its API.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: Code, prompts, configurations and raw trajectories are planned for release, and no repository URL is included in this version.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: No model is trained. Section 4 gives the window, number of rounds, temperature, prompts, pairs, echo rates and evidence protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Error bars in Figures 3b, 4a, 5b and 6 are standard errors of the mean over independent conversations within a condition. Predictions for the main experiment were recorded before the pairs were run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: All experiments are API calls to GPT-4o, about 214,000 for the reported experiments and 271,000 for the whole project. No GPUs were used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: No human subjects or personal data are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Sections 3 and 5 discuss the implications for multi-agent alignment. The echo mechanism is a known failure mode, not a new attack, and the synthetic results are not clinical evidence.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: No models or datasets are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: GPT-4o is used through the OpenAI API under its terms of service and credited in Section 4.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: No new assets are released with this version.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: No crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: GPT-4o is the object of study, as described in Section 4.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.