PLOS One

RESEARCH ARTICLE

Propensity to trust in Large Language Models

Alice Plebe*

Department of Industrial Engineering, University of Trento, Trento, Italy

* alice.plebe@unitn.it

Abstract

Trust is central to collaborative settings in which large language models (LLMs) are increasingly deployed. Yet little is known about whether LLMs exhibit a *propensity to trust* (PTT): a baseline tendency to extend or withhold trust that remains relatively stable across contexts. We investigate PTT in nineteen LLMs using two complementary approaches: a psychological self-report scale adapted from human research and a linguistic simulation framework designed to elicit trust-related decisions in context. While the questionnaire produces uniformly high PTT across models—likely reflecting social-alignment objectives and sycophantic response patterns—the simulation framework uncovers substantial, systematic differences in how models entrust others. Our simulations show that trust behavior is governed by the interaction between a baseline tendency to delegate and a model's capacity to integrate cues about trustworthiness. More capable models, such as GPT-4o-mini, use such cues to adjust their decisions, allowing competence signals to modulate baseline tendencies. By contrast, other models, such as Llama-2-7B, exhibit stable delegation patterns that are largely insensitive to task-specific evidence, leading to systematic over-entrustment. These results show that performance depends not on baseline tendencies alone, but on how they are modulated by alignment-sensitive information. Ablation studies show that task-specific memory mechanisms enable models to better integrate trustworthiness cues, improving the calibration of delegation decisions. More generally, our findings show that questionnaire-based measures cannot disentangle baseline tendencies from context-sensitive adjustment, whereas behavioral simulations make this distinction observable.

Citation: Plebe A (2026) Propensity to trust in Large Language Models. PLoS One 21(5): e0347328. https://doi.org/10.1371/journal.pone.0347328

Editor: Niccolò Tempini, University of Exeter, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: December 26, 2025

Accepted: March 31, 2026

Published: May 6, 2026

Data availability statement: The source code will be available publicly on GitHub at the moment of publication.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

Large language models (LLMs) are increasingly assuming roles in social and collaborative environments that were once the exclusive domain of humans [1–8]. To cooperate effectively—whether with humans or other artificial agents—an entity must display key elements of social cognition. Among these, trust stands as a fundamental pillar: it enables agents to predict others' behavior, coordinate decisions, and sustain collaboration [9–13].

As LLMs increasingly participate in activities that rely on social coordination, the question naturally arises: are they able to trust? Understanding these delegation patterns is essential if LLMs are to function as reliable collaborators. In human psychology, one construct used to describe stable differences in reliance behavior is the *propensity to trust* (PTT): a baseline tendency to grant or withhold trust that does not depend on the immediate situation or the specific trustee [14–17].

Despite its importance, whether LLMs exhibit such stable patterns of trust-related behavior has received limited systematic attention. Existing studies typically operationalize trust as a context-bound variable assessed through classical economic games [18–21]. Although these paradigms are well established for symbolic agents, they fail to engage the linguistic reasoning that constitutes the core competence of LLMs. Language is not merely a means of communication; it provides a substrate for social cognition [22], enabling agents to express commitments, evaluate intentions, and negotiate reliability [23,24]. Evaluating PTT in LLMs therefore requires observing them in language-mediated interactions that reflect the settings in which they are deployed.

To address this gap, we introduce a framework for assessing PTT in LLMs through simulated, language-based interactions. We evaluate nineteen models from OpenAI, Anthropic, Meta, Google, and Microsoft across three ecologically grounded scenarios. In each setting, models decide whether to entrust tasks to specific agents and update their beliefs about each agent's trustworthiness based on linguistic feedback.

We also administer a human PTT questionnaire [25] as a complementary measure. However, because PTT is framed as an ethically positive trait, socially aligned LLMs may be predisposed to overstate it, raising the question of whether direct self-reports can meaningfully capture trust-related tendencies in these systems.

Importantly, applying the notion of PTT to LLMs does not imply that these systems possess human-like social dispositions. In this work, PTT is used as an operational construct describing stable patterns in delegation behavior across contexts. The focus is therefore not on whether models possess a psychological trait in a human sense, but on whether their decisions exhibit consistent baseline tendencies across scenarios and how these interact with evidence about trustworthiness. This allows systematic comparison of trust-related behavior while remaining agnostic about underlying mental states.

Our results reveal three core findings:

1. Questionnaire-based PTT is not predictive of observed delegation behavior. Social alignment drives models to endorse prosocial statements, producing uniformly inflated scores that mask meaningful differences in how models allocate trust.
2. When evaluated in language-mediated interaction, models show substantial and systematic divergence in their enacted PTT. Some models, most notably Llama-2-7B, trust generously across all settings despite reporting low PTT in the questionnaire. Others, such as Qwen2.5-7B, display the opposite pattern: they report high PTT yet behave cautiously in the simulations.

3. Trust behavior reflects the interaction between baseline delegation tendencies and sensitivity to trustworthiness cues. Access to task-specific memory enables models to modulate their baseline inclination to trust when presented with evidence about a trustee's competence.

Taken together, these contributions establish a linguistically grounded framework for studying trust in LLMs and highlight the need for behavioral methods that capture how these systems reason about and interact with others. As LLMs become integrated into collaborative settings, understanding how models calibrate trust in others becomes essential for their safe and effective deployment.

2. Theoretical background

Trust is widely recognized as a central mechanism enabling cooperation under uncertainty. Despite the intuitive familiarity of the concept, its theoretical foundations vary substantially across disciplines, each emphasizing different aspects of the phenomenon. This section reviews these traditions in order to situate our operational framework within the broader landscape of trust research and to clarify which dimensions of trust are adopted, abstracted, or deliberately excluded in this study.

2.1. Trust across disciplines

Compared with many other aspects of social cognition, trust received relatively limited attention in early Western philosophy, with notable discussions appearing in the works of Hobbes, Locke, and Hume [26–28]. Systematic research on trust emerged only in the late twentieth century and has since expanded across multiple disciplines, including philosophy [29–33], psychology [34–38], sociology [39–43], economics [44–49], cognitive science [50], organizational research [51,52], and neuroscience [53–56].

Philosophical research on trust has addressed both conceptual and normative questions concerning the nature of trust and its role in interpersonal relations. A central issue concerns the mental attitude involved in trusting another agent: some accounts interpret trust primarily as a form of expectation or belief about another's behavior, while others emphasize its distinctive normative structure. Within contemporary discussions, several influential theories are often grouped under the label *motives-based* accounts [32]. According to these views, trust involves assumptions about the trustee's motivations to fulfill the trust relationship. For example, [29] and [31] argue that trust depends on expectations about the trustee's reasons or incentives for acting in the trustor's interest. Other philosophers emphasize the trustor's reactive attitudes rather than the trustee's motivations. On this view, the distinctive feature of trust lies in the normative response that follows when trust is violated. [30], for instance, argues that trust is characterized by the trustor's sense of betrayal when the trusted party fails to act as expected. These debates illustrate that philosophical analyses of trust often extend beyond predictive expectations to include moral obligations, vulnerability, and the interpersonal norms governing trust relationships.

In sociology, trust is commonly analyzed as a structural feature of social systems. [39] and [40], for example, describe trust as a mechanism that reduces social complexity and enables coordination under uncertainty. Sociological work often intersects with economic perspectives that treat trust as a mechanism facilitating cooperation in markets and organizations [41,44,46–49].

Psychological research approaches trust primarily as a behavioral and cognitive phenomenon rather than a normative one. Early developmental theories linked trust to personality formation and attachment processes [34,57]. Subsequent work has explored how trust interacts with learning, interpersonal relationships, and social expectations [36,58–60]. This tradition also overlaps with related fields. Neuroscientific research investigates the neural mechanisms underlying trust-related decisions [53–56], while comparative cognition examines trust-like behaviors in non-human animals—an area where the literature remains scattered [61], despite clear evidence that basic forms of trust play a significant role in social animals, particularly in reciprocal behaviors.

In research on social cognition and language, trust is often examined in relation to the communicative mechanisms that support cooperation. Language allows agents to make explicit commitments, communicate intentions, and reason about the trustworthiness of others. Developmental accounts of human cooperation emphasize the role of linguistic communication in the emergence of shared intentionality and coordinated social behavior [22]. In this perspective, language provides a medium through which agents form, revise, and communicate expectations about trustworthiness [23,24].

2.2. Dimensions of trust

Cognitive science and organizational research have sought to formalize these insights in computational or decision-theoretic models of trust. Such work attempts to identify a set of dimensions that characterize how agents evaluate potential collaborators, providing conceptual tools that can be adapted for the analysis of artificial agents.

Among the most influential multidimensional accounts is the organizational framework proposed by [51], which identifies three characteristics of the trustee that shape trust: *ability*, *benevolence*, and *integrity*. Ability refers to the skills or competencies enabling effective action within a domain; benevolence denotes a willingness to act in the trustor's interest; and integrity concerns adherence to principles that the trustor finds acceptable. A meta-analysis of 132 studies confirmed the empirical robustness of this dimensional approach across a variety of organizational contexts [14].

Computational approaches to trust in cognitive science have proposed related models. [50], for instance, distinguish among several components of trust, including *competence*, *predictability*, and *willingness*. Competence refers to the capacity to perform a task successfully; predictability concerns the consistency with which an agent behaves as expected; and willingness captures the agent's commitment to carrying out the relevant actions.

Other research on trust in artificial systems adopts similar dimensional frameworks. For example, [62] identify *capability*, *reliability*, *sincerity*, and *ethics* as determinants of human trust in robots, while [63] highlight *competence*, *predictability*, *willingness*, and *honesty* as central elements underlying trust in artificial intelligence systems. Across these approaches, trustworthiness is commonly analyzed in terms of an agent's competence or capability, the reliability or predictability of its behavior, and its motivational orientation toward fulfilling commitments. Table 1 summarizes the principal proposed trust dimensions.

2.3. Propensity to trust

While the dimensions discussed above characterize the attributes of a potential trustee, individuals differ considerably in how they interpret and respond to those attributes. Faced with comparable evidence regarding another agent's trustworthiness, some individuals readily choose to rely on others whereas others remain cautious.

Such interindividual variation is captured by the construct of *propensity to trust* (PTT): a stable individual difference in the baseline likelihood of choosing to rely on another agent when presented with comparable evidence regarding their trustworthiness. The concept first appeared in mid-twentieth-century psychology. [34] suggested that a basic tendency

Table 1. Dimensions of trusts in the literature.

Source	Dimensions of trust			
Mayer et al. (1995) [51]	Ability	–	Benevolence	Integrity
Castelfranchi & Falcone (2010) [50]	Competence	Predictability	Willingness	–
Ullman & Malle (2018) [62]	Capability	Reliability	Sincerity	Ethic
Lewis & Marsh (2022) [63]	Competence	Predictability	Willingness	Honesty
This study	**Capability**	**Reliability**	**Willingness**	**–**

Examples of theoretical frameworks that analyze trust in terms of dimensions of trustworthiness. These approaches attempt to identify the attributes of an agent that justify reliance in cooperative contexts.

https://doi.org/10.1371/journal.pone.0347328.t001

to trust develops during early childhood as part of personality formation. Later work framed this disposition as a stable individual difference influencing trust-related behavior in adulthood [36,37]. Within organizational research [51], formalized the concept of *propensity to trust* as a general willingness to rely on others independent of specific situational cues. Subsequent studies have linked PTT to broader attitudes toward risk-taking and cooperation in professional settings [14,64]. Neuroscientific evidence further suggests that individual differences in trust propensity are associated with neural activity in brain networks involved in social cognition and decision-making [65].

A substantial strand of research concerns the measurement of trust propensity. Early approaches relied on general trust questionnaires [36,64,66], but these instruments often conflated dispositional trust with judgments about specific targets. More recent work has therefore developed dedicated measurement scales designed explicitly to capture PTT [15,17,25,67–69]. A meta-analysis of 179 studies identified 27 distinct instruments used to measure trust propensity across the literature [16]. Among the most widely adopted is the scale proposed by [25], which reduced an initial set of 43 items to a concise 12-item measure. The items of this scale are reported in Table 2.

2.4. Trust and LLMs

Research on trust in artificial agents has primarily examined human trust in technology. This literature investigates when and why people rely on AI systems, robots, or automated decision tools [70–80].

More recently, studies have begun to explore trust-related behaviors exhibited by artificial agents themselves. [81] analyze multi-agent systems composed of LLMs from a robustness and security perspective, showing that LLM agents often treat peer-generated content as uniformly credible unless skepticism is explicitly induced. As a result, such systems may become vulnerable to misinformation, manipulation, or coordination failures.

Another line of research examines trust-related behavior in LLMs using experimental paradigms from behavioral economics. [82], for example, study LLM behavior in the classical Trust Game and report that advanced models display patterns resembling human trust decisions, adjusting their behavior in response to perceived risk and potential reciprocity. While such studies demonstrate that LLMs can reproduce recognizable patterns of trust behavior, the analysis is typically restricted to a single, highly simplified decision problem.

Table 2. Human PTT measurement scale.

1.	**It is easy for me to trust others.**
2.	Even if I am uncertain, I will generally give others the benefit of the doubt.
3.	I generally believe that others can be counted on to do what they say they will do.
4.	**I usually trust people until they give me a reason not to trust them.**
5.	I tend to trust others even if I have little knowledge of them.
6.	I generally give people the benefit of the doubt when I first meet them.
7.	**Trusting another person is not difficult for me.**
8.	**My typical approach is to trust new acquaintances until they prove I should not trust them.**
9.	I am seldom wary of others.
10.	I don't mind giving up control to others over matters which are essential to my future plans.
11.	I believe that people usually keep their promises.
12.	**My tendency to trust others is high.**

Items from the Frazier et al.'s *propensity to trust* scale [25]. The four statements with the highest factor loadings are shown in bold.

https://doi.org/10.1371/journal.pone.0347328.t002

Economic trust games provide a well-established experimental paradigm, but they capture only a narrow aspect of trust-related decision making. In these settings, trust is expressed through a small set of numerical choices within a fixed payoff structure. As a result, the agent's decision depends primarily on quantitative parameters such as risk and expected reciprocity.

The approach adopted here instead focuses on trust expressed through natural-language interaction. This choice does not rest on the assumption that all linguistic processing constitutes social cognition, and the ability of LLMs to process language does not in itself imply that they engage in social cognition. Rather, the methodological motivation is tied to the types of interactions in which LLMs are typically deployed. In collaborative settings, trust is often expressed and negotiated through communicative acts such as requests, commitments, explanations, and feedback.

Language-mediated scenarios therefore provide a context in which models must interpret task descriptions, evaluate information about collaborators, and revise expectations based on textual evidence. Such settings allow multiple trust-relevant cues to be presented and integrated over the course of interaction, making it possible to observe how models respond to richer contextual information about potential collaborators.

Moreover, the construct validity of economic trust games remains debated, with some analyses suggesting that they may conflate trust with related constructs such as risk preference or expectations of reciprocity [83]. Without taking a position on this debate, this observation highlights the value of employing complementary methodologies when studying trust-related behavior in artificial agents. Natural-language simulations provide one such complementary approach, allowing trust to be examined in interactional contexts that more closely resemble the communicative environments in which LLMs typically operate.

2.5. Conceptual framework of the present study

The present study examines a functional component of trust widely discussed in psychology, economics, and socio-cognitive modeling: the decision to rely on another agent when the outcome of an action depends on that agent's behavior under uncertainty. Rather than studying human trust in artificial systems, we investigate how artificial agents themselves express trust and whether large language models display stable patterns of reliance when delegating tasks to other agents.

Trustworthiness is operationalized using three dimensions that recur across several empirical and computational models: *capability*, *reliability*, and *willingness*. Capability refers to the ability to perform a delegated task successfully. Reliability captures the consistency with which an agent performs successfully across situations. Willingness describes the disposition to carry out the relevant actions rather than neglect or abandon them. The term *willingness* corresponds to the action-oriented component of Mayer et al.'s notion of benevolence. Mayer et al. define benevolence as "the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive" [51, p. 718] and further suggest that it may involve a specific attachment to the trustor. In the present framework, we abstract from affective attachment, altruistic motivation, and moral concern, retaining only the behavioral component relevant for task execution.

These three dimensions are not intended to exhaust all theoretical accounts of trust. Philosophical and sociological approaches often incorporate additional normative and relational elements, including moral expectations and reactive attitudes such as resentment or betrayal [30,71]. Our choice is methodological. Capability, reliability, and willingness constitute a minimal, behaviorally operationalizable subset that recurs across influential empirical and computational models of trust (e.g., [51]; [50]). The aim of the present study is not to adjudicate between competing theories of trust, nor to reproduce the full normative richness of human interpersonal trust. Instead, we isolate this widely shared functional core in order to examine whether large language models exhibit stable patterns of reliance across situations, captured by the construct of *propensity to trust*.

PTT is interpreted as a stable difference across agents in the baseline likelihood of relying on another agent when presented with comparable evidence regarding these dimensions. In the context of our simulation, PTT is defined operationally as the stability of delegation decisions across heterogeneous scenarios.

This interpretation does not presuppose that LLMs possess an interior mental life or stable dispositional attitudes in a human-like sense—an assumption that would situate the analysis within ongoing debates about machine mentality and selfhood [84–86]. The present work remains agnostic on these questions. Whether LLMs possess genuine mental states is a substantive philosophical question, but resolving it is not required for the present analysis, which focuses on observable behavioral regularities in model outputs rather than on claims about internal psychological states.

Accordingly, trust-related constructs are treated here as abstractions summarizing observable behavioral regularities in model outputs rather than as claims about an underlying mental ontology. In this respect, the framework follows what [87] describe as *anthropocentric abstraction*: employing conceptual frameworks originating in human social cognition at a level of abstraction that preserves their functional role while remaining neutral about their metaphysical interpretation.

3. Methodology

This study evaluates LLMs' scenario-independent tendency to delegate tasks under uncertainty through two complementary approaches: direct responses to standardized human PTT questionnaires and behavioral observation in simulated collaborative scenarios.

In the questionnaire-based evaluation, we measure each model's self-reported PTT using the 12-item scale developed by [25] (Table 2), one of the most comprehensive and linguistically refined instruments for assessing human trust disposition. Each item expresses a trust-related attitude and requires a response on a seven-point Likert scale ranging from complete disagreement to complete agreement.

In the simulation-based evaluation, we employ a task assignment setting that elicits trust-related decisions through natural language interaction. This approach provides an indirect measure of PTT by observing how each model decides whether to entrust specific agents with a task, updates its beliefs about their trustworthiness, and adjusts subsequent choices accordingly. Unlike questionnaire-based assessment, the simulation does not rely on self-reporting; instead, it captures behavioral consistency across contexts, revealing whether a model exhibits a stable dispositional tendency to trust or to withhold trust.

3.1. Task assignment simulation

We ground our simulation in the human trust formation process described by [68]. As shown in Fig 1, a model's PTT represents its baseline tendency to delegate across a broad class of potential trustees. This is the most general level of trust, which becomes increasingly specific as the model forms beliefs about particular candidates. For each trustee, the model develops two forms of perceived trustworthiness: general, applying across tasks within a scenario, and task-specific, applying to a particular assignment. These beliefs inform the model's intention to rely on a given trustee when deciding whether to delegate a task. The process culminates in a trust-related behavior, expressed as the model's decision to assign—or not assign—the task to that trustee.

We represent the trustees as a team $\mathcal{A}$ of agents, where each agent is defined as:

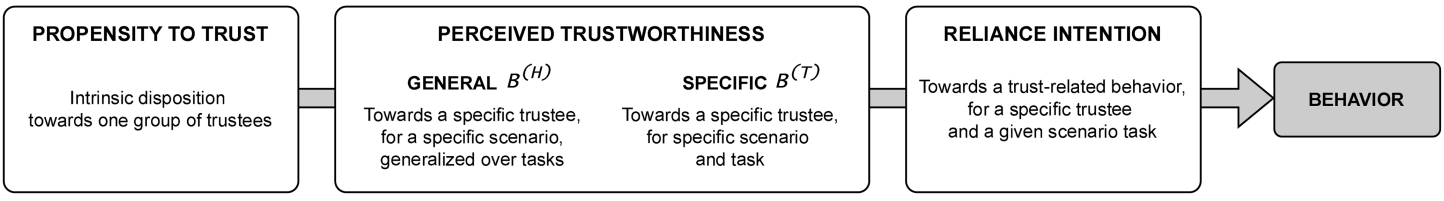

Fig 1. Conceptual model of the trust formation process. Adapted from [68]. In our setting, this process results in a behavioral decision in which the LLM evaluates whether to trust a potential trustee for a given task.

https://doi.org/10.1371/journal.pone.0347328.g001

$$a = (n, \mathbf{x}), \tag{1}$$

with $n \in A^*$ denoting the agent's name (where A is the set of alphanumeric characters and A^* the set of all possible strings), and $\mathbf{x} \in \mathbb{R}^3$ encoding the agent's internal properties along the trust dimensions of *capability*, *reliability*, and *willingness*. These properties are hidden from all other agents and from the trustor.

We define a task t as:

$$t = (d, \mathbf{y}), \tag{2}$$

where $d \in A^*$ is the textual description of the task, and $\mathbf{y} \in \mathbb{N}^3$ specifies the required levels of *capability*, *reliability*, and *willingness* for successful completion. Task requirements are also hidden from the trustor, who must infer them from the textual description d. In the simulation, an agent a may successfully complete a task t depending on how closely its property vector $\mathbf{x}$ aligns with the task requirements $\mathbf{y}$. The algorithm used to evaluate this alignment is described in Section 3.1.1.

The trustor, represented by the LLM, is equipped with a short-term memory S:

$$S = \left(\left\{ B_i^{(H)} \right\}, \left\{ B_{i,j}^{(T)} \right\} \right), \tag{3}$$

which stores the perceived trustworthiness of the trustees, expressed in natural language. The two belief components differ in their level of specificity: $B_i^{(H)}$ encodes the trustor's general assessment of trustee a_i, whereas $B_{i,j}^{(T)}$ captures expectations about a_i's performance on a particular task t_j. Both beliefs are updated over the course of the simulation based on the trustor's observations. $B_i^{(H)}$ consists of a linguistic statement summarizing the trustor's overall view of a_i within the scenario and is revised after every decision to delegate a task to that agent. By contrast, $B_{i,j}^{(T)}$ is task-specific and is revised only when the trustor assigns task t_j to a_i.

The simulation progresses through a sequence of events, each following six steps:

1. The system randomly selects a task t_j from the scenario.
2. The system selects an agent a_i from the team $\mathcal{A}$ in cyclic order.
3. The trustor decides whether to entrust the task t_j to agent a_i, based on the task description d_j and its current beliefs $B_i^{(H)}$ and $B_{i,j}^{(T)}$.
4. If the decision is negative, the simulation returns to step 2. If it is positive, the system computes the alignment between the trustee's properties $\mathbf{x}_i$ and the task requirements $\mathbf{y}_j$ to determine the task outcome (see Section 3.1.1 for details).
5. The trustor receives a message o_D summarizing the task outcome. In cases of success, this is a general statement; in cases of failure, it is an indirect linguistic description of the dimension of $\mathbf{x}_i$ that caused the failure.
6. The trustor updates its trust beliefs about a_i based on o_D. It retrieves the most recent $B_i^{(H)}$ and $B_{i,j}^{(T)}$, revises them according to the outcome, and stores the updated beliefs in short-term memory.

At the beginning of the simulation, $B_i^{(H)}$ and $B_{i,j}^{(T)}$ are empty. To initialize the trust formation process, the system performs a bootstrapping round that provides the trustor with preliminary beliefs about each trustee. Before the first event, the system parses all possible task—agent combinations. During this phase, the trustor does not make decisions but observes outcomes, forming an initial impression of each agent's trustworthiness across tasks.

3.1.1. Computation of task–agent alignment. The outcome of a task t executed by agent a is modeled probabilistically as a function of how closely the agent's properties $\mathbf{x}$ align with the task requirements $\mathbf{y}$. Alignment is evaluated holistically: strong values in some trust dimensions can compensate for weaker values in others.

Let $o_B \in \{0, 1\}$ denote the outcome of the task, where $o_B = 1$ indicates successful completion and $o_B = 0$ indicates failure. The probability of success, $p(o_B = 1 \mid \mathbf{x}, \mathbf{y})$, depends on the overall match between $\mathbf{x}$ and $\mathbf{y}$, measured by their dot product:

$$R = \mathbf{y} \cdot \mathbf{x}^{\top}. \tag{4}$$

This score R quantifies how well an agent's attributes match the task requirements and serves as the basis for computing the corresponding probability of success. To generate distinct alignment scores for each possible agent–task pairing, we consider six agent profiles $\mathbf{x}$ given by permutations of [0,1,2] and six task profiles $\mathbf{y}$ given by permutations of [1,2,4].

The resulting alignment scores partition agent–task combinations into two groups (Table 3): high-ranking alignments associated with a high probability of success $p(o_B = 1) \gg 0.5$, and low-ranking alignments associated with a low probability of success $p(o_B = 1) \ll 0.5$. The boundary between these groups is controlled by a difficulty parameter $\eta \in \mathbb{N}$. A small stochastic component $r \in \mathbb{R}$ introduces randomness into the outcome.

The three dimensions of agent properties enter symmetrically into the task–agent alignment mechanism. No dimension is privileged a priori, and any of them may determine success depending on the task. The alignment computation therefore evaluates multi-dimensional fit rather than a single dominant trait. The dot-product formulation allows partial compensation across the three dimensions while still favoring alignment with the highest-weighted components of $\mathbf{y}$. This is a modeling choice that provides a simple and continuous measure of task–agent compatibility, rather than a theoretical claim about the nature of trust. Alternative non-compensatory formulations (e.g., threshold or conjunctive rules) could also be explored.

Importantly, the vector representations of agent properties and task requirements are not accessible to the LLM. The model receives only natural-language task descriptions and outcome feedback, while alignment is computed probabilistically in the background. From the model's perspective, the task consists solely of interpreting linguistic cues and updating beliefs based on textual outcomes, rather than solving an explicit matching problem.

If delegation behavior were driven by an implicit matching strategy, it would be expected to converge toward optimal assignment. Instead, we observe systematic differences in delegation patterns across models, including persistent over- and under-entrustment (Section 5). This indicates that decisions reflect both evidence integration and model-specific baseline tendencies, rather than purely logical problem solving.

4. Evaluation setup

We evaluate 19 large language models for their propensity to trust. The selection spans major contemporary model families developed by OpenAI, Anthropic, Meta, Google, and Microsoft, as well as open-weight releases from Mistral and

Table 3. Example of task–agent alignment computation.

$\mathbf{y}$	$\mathbf{x}$	$R = \mathbf{y} \cdot \mathbf{x}^{\top}$	$p(o_B = 1)$				
			$\eta = 4$	$\eta = 3$	$\eta = 2$	$\eta = 1$	$\eta = 0$
[1, 2, 4]	[0, 1, 2]	10	$1-r$	$1-r$	$1-r$	$1-r$	$1-r$
[1, 2, 4]	[1, 0, 2]	9	$5r$	$1-2r$	$1-2r$	$1-2r$	$1-2r$
[1, 2, 4]	[0, 2, 1]	8	$4r$	$4r$	$1-3r$	$1-3r$	$1-3r$
[1, 2, 4]	[2, 0, 1]	6	$3r$	$3r$	$3r$	$1-4r$	$1-4r$
[1, 2, 4]	[1, 2, 0]	5	$2r$	$2r$	$2r$	$2r$	$1-5r$
[1, 2, 4]	[2, 1, 0]	4	r	r	r	r	r

The table illustrates how the alignment between a task's requirements ($\mathbf{y}$) and a trustee's properties ($\mathbf{x}$) determines the success probability, which varies with simulation difficulty (η). Combinations with a higher likelihood of success are highlighted in green, while those more likely to result in failure are shown in red. A small component r introduces stochasticity in the outcome.

https://doi.org/10.1371/journal.pone.0347328.t003

Alibaba's Qwen initiatives. Table 4 lists all evaluated models, reporting both their full names and the short codes used used for brevity in subsequent tables and figures.

4.1. Questionnaire-based evaluation

We administer the 12 items by [25] (Table 2) to the 19 LLMs under investigation. We ask models to respond on a seven-point Likert scale ranging from *complete disagreement* to *complete agreement*. To account for response variability, we present each item 10 times to every model and average the results across repetitions.

4.2. Simulation-based evaluation

We evaluate the LLMs' trust propensity across three distinct scenarios: responding to a building fire, where agents perform firefighting and first-aid tasks (`fire`); maintaining a farm, where agents carry out agricultural and mechanical tasks (`farm`); and managing a school, where agents handle administrative and organizational tasks (`school`). See S1 Appendix for further details.

Each scenario includes six possible tasks, each associated with a different requirement vector **y**, that is one of the six permutations described in Section 3.1.1. The model (trustor) interacts with a team of six agents (trustees), each defined by a distinct property vector, also drawn from the six permutations described in Section 3.1.1.

Simulations unfold entirely through text-based interaction: tasks, agent behaviors, and outcomes are all represented linguistically, and the model's trust-related decisions take place exclusively in natural language. The three scenarios differ in narrative framing, task structure, urgency, and required competencies, while sharing only minimal vocabulary overlap. This diversity prevents models from relying on superficial lexical cues and instead reveals their broader disposition toward trusting others across heterogeneous linguistic contexts.

In addition, we test two ablation settings of each of the 19 LLMs. The first (`1-mem`) retains only general perceived trustworthiness and excludes task-specific beliefs, such that Equation (3) reduces to $S = \left\{B_i^{(H)}\right\}$. The second (`no-trust`) also uses a single memory and, in addition, removes any explicit mention of trust by excluding the capability, reliability, and willingness dimensions from the prompts. See S2 Appendix for examples of prompts used in each case.

Each simulation runs for 50 events, and we repeat 10 simulations for every combination of model, scenario, and ablation configuration. All runs use a medium difficulty level ($\eta = 2$) and include a small stochastic component ($r = 0.01$).

Table 4. Evaluated LLMs.

GPT-3.5-turbo	(`gpt35`)	Phi-3-mini-4K-instruct	(`ph3m`)
GPT-4	(`gpt4`)	Qwen1.5-7B-Chat	(`qw1-7`)
GPT-4o	(`gpt4o`)	Qwen2.5-7B-Instruct	(`qw2-7`)
GPT-4o-mini	(`gpt4om`)	Qwen2.5-14B-Instruct-1M	(`qw2-14`)
GPT-4.1-mini	(`gpt41m`)	Claude-3-haiku-20240307	(`cl3h`)
GPT-oss-20B	(`gptoss`)	Claude-3.5-haiku-20241022	(`cl3.5h`)
Gemma-2-9B-it	(`gem2-9`)	Claude-3.5-sonnet-20240620	(`cl3.5s`)
Llama-2-7B-chat-hf	(`ll2-7`)	Claude-3.7-sonnet-20250219	(`cl3.7s`)
Llama-2-13B-chat-hf	(`ll2-13`)	Claude-3-opus-20240229	(`cl3o`)
Llama-3.1-8B-Instruct	(`ll3-8`)		

Evaluated LLMs and short codes used for brevity in subsequent figures.

https://doi.org/10.1371/journal.pone.0347328.t004

5. Results

5.1. Results from human questionnaire

Figure 2 shows that, across the 19 evaluated models, the average PTT scores on the Frazier scale (Table 2) fall within a narrow and consistently positive range. This indicates that the models exhibit relatively uniform levels of self-reported trust. The mean score reported for human participants is 5.03 [25, p. 86], which lies near the midpoint of the models' average values, suggesting that most LLMs report trust levels comparable to—or slightly higher than—human baselines.

Apart from a single outlier, Llama-2-7B (`ll2-7`), whose scores are markedly lower than those of the other models, responses are highly homogeneous across model families. GPT and Claude models display very similar endorsement patterns across items, while the more recent Qwen models (`qw2-7` and `qw2-14`) rank among the highest-scoring systems. Overall, the questionnaire produces a compressed distribution of scores with relatively limited variation between models.

These questionnaire-based results should be interpreted with caution. Many items in the Frazier scale explicitly describe prosocial or socially desirable attitudes, such as giving others the benefit of the doubt or assuming good intentions. Because modern LLMs are trained to produce cooperative, helpful, and non-antagonistic responses, they are naturally inclined to endorse such statements. As a result, high questionnaire scores may reflect agreement with socially desirable content rather than stable patterns of delegation behavior expressed during interaction.

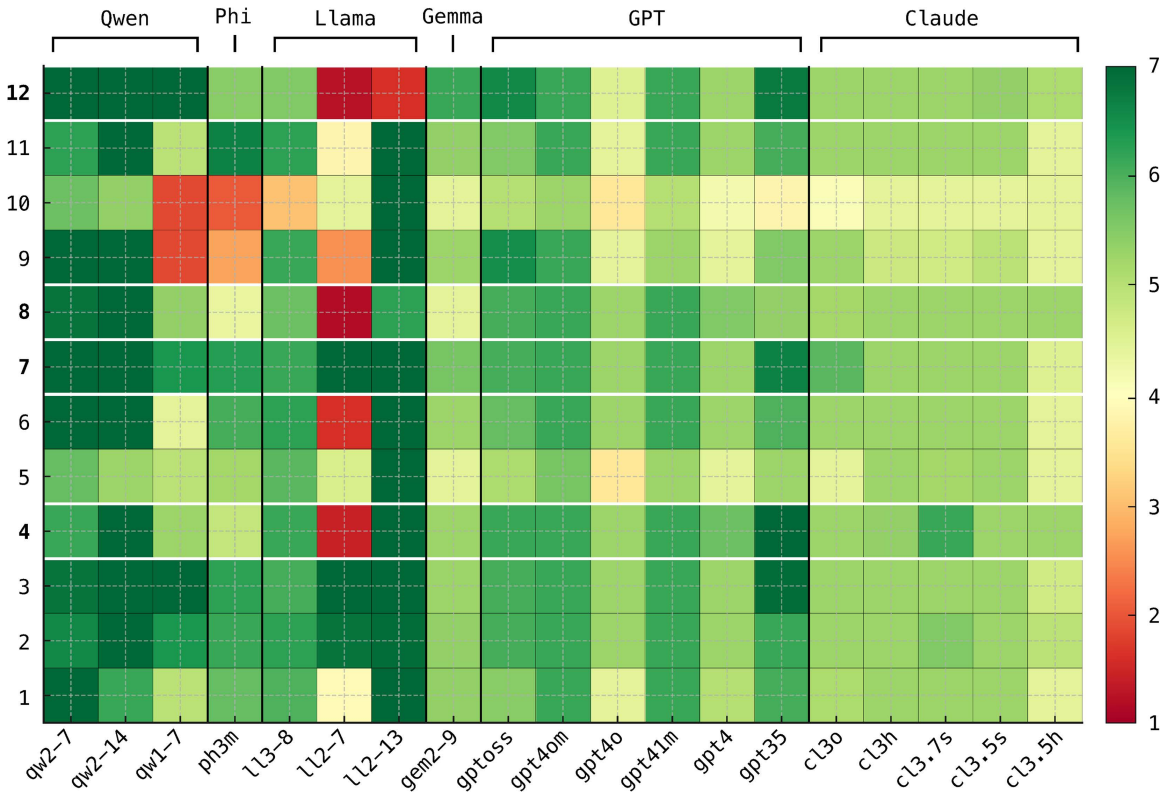

Fig 2. Questionnaire results. Results of the PTT scale from [25] administered to all models. The 12 items from Frazier's scale are shown, with the four items with the highest factor loadings highlighted in bold. Responses use a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

https://doi.org/10.1371/journal.pone.0347328.g002

This pattern can be explained by alignment procedures used in contemporary LLM training. Models fine-tuned with reinforcement learning from human feedback (RLHF) or related methods are optimized to produce responses perceived as helpful, cooperative, and socially appropriate [88,89]. Consequently, they tend to endorse statements expressing socially valued attitudes, a behavior often described as sycophancy [90,91]. When applied to questionnaire-style prompts, this leads to systematically inflated trust scores.

This does not imply that such responses are irrelevant to model behavior. Rather, the limitation is one of measurement: questionnaire responses reflect prompt-level agreement with prosocial statements, whereas our objective is to assess how models allocate trust when confronted with varying evidence about collaborators. Consequently, questionnaire-based measures cannot distinguish between training-driven agreement and consistent delegation behavior across contexts.

5.2. Results from task assignment simulations

Fig 3 shows how often each model decides to entrust a task to an agent, across scenarios and ablation configurations. The patterns that emerge differ substantially from the trust tendencies suggested by the questionnaire-based evaluation.

The contrast with Fig 2, which summarizes the questionnaire-based results, is immediate and striking. The model that most frequently chooses to trust agents in the simulations, Llama-2-7B (`ll2-7`), is the same model that expresses the lowest trust when responding to the questionnaire. A similar discrepancy appears for GPT-4o (`gpt4o`), which shows a tendency to trust in the simulations but a notably cautious profile in the questionnaire. Conversely, Qwen2.5-7B (`qw2-7`), which displays high self-reported trust in the questionnaire, shows comparatively low trust levels in the simulation.

This reversal reveals a systematic divergence between simulation behavior and questionnaire responses, indicating that questionnaire-based assessments cannot be used in isolation as a reliable measure of PTT in LLMs.

Additionally, the bottom rows of Fig 3 illustrate the impact of removing components of the trust formation process. The `1-mem` ablated variant, which omits task-specific perceived trustworthiness (Fig 1), and the `no-trust` ablated variant, which removes the three-dimensional definition of trust, both lead to substantially lower rates of trust decisions. These results suggest that task-specific trustworthiness and the structured three-dimensional representation of trust each contribute importantly to the emergence of trust-like behavior in the simulations.

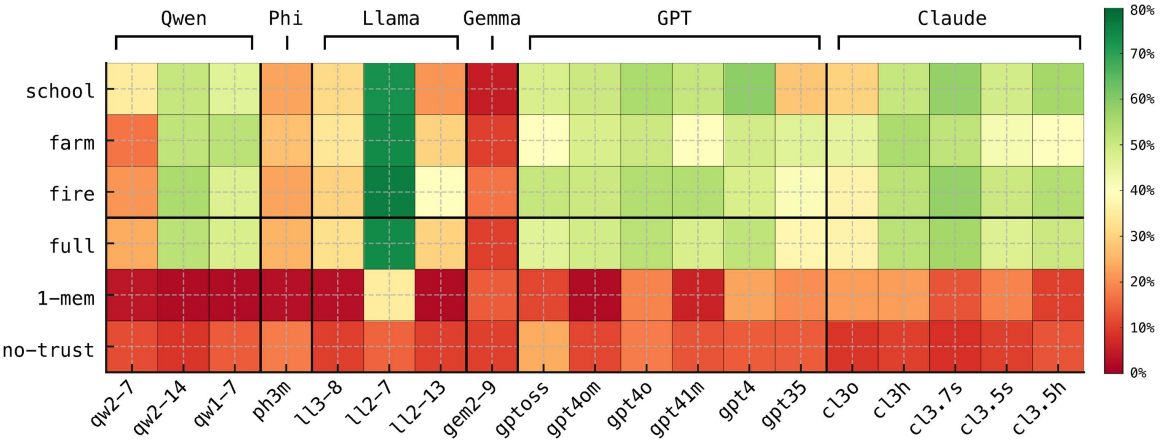

Fig 3. Simulation results. Proportion of task assignments in which a model decides to entrust the selected agent, across scenarios and ablation configurations. The bottom three rows (`full`, `1-mem`, `no-trust`) aggregate results across all scenarios for the corresponding ablation; the top three rows (`school`, `farm`, `fire`) correspond to the full models.

https://doi.org/10.1371/journal.pone.0347328.g003

5.2.1. Stability Across Scenarios. We quantify the stability of each model's trust behavior across scenarios using several statistical indicators. Our goal is to isolate the extent to which each model exhibits a scenario-independent tendency to trust, i.e., a behavioral signature consistent with a baseline PTT.

Table 5 reports six metrics computed for each model from the fraction of events in which the model entrusts the agent with the task. For each model, we compute:

- $\bar{\mu}$: the overall average entrustment rate across scenarios;
- Δ: the range of entrustment rates;
- σ: the standard deviation across scenarios;
- η^2: the effect size of scenario on trusting decisions, derived via one-way ANOVA (higher values indicate stronger scenario influence);
- ρ: the intra-class correlation coefficient (ICC), measuring the proportion of variance attributable to scenario-specific rather than random effects;
- π: a composite PTT-stability index synthesizing σ and η^2 into a single value in [0,1] (values near 1 indicate strong scenario-independence, and therefore a more baseline dispositional trust tendency).

We define:

$$\pi = w_\sigma(1-\sigma) + w_\eta(1-\eta^2) \tag{5}$$

Table 5. Statistics across simulation scenarios.

Model	$\bar{\mu}$	Δ	σ	η^2	ρ	π
gpt35	0.38	0.182	0.093	0.0245	0.034	0.11
gpt4om	0.49	0.020	0.011	0.0003	0.000	1.00
gpt4o	0.53	0.044	0.024	0.0015	0.000	0.90
gpt4	0.52	0.104	0.057	0.0088	0.011	0.59
gptoss	0.46	0.112	0.059	0.0092	0.012	0.57
gpt41m	0.48	0.144	0.076	0.0156	0.021	0.36
cl3h	0.53	0.046	0.023	0.0014	0.000	0.91
cl3o	0.37	0.144	0.072	0.0148	0.020	0.40
cl3.5h	0.50	0.166	0.089	0.0211	0.029	0.20
cl3.5s	0.47	0.082	0.044	0.0053	0.006	0.72
cl3.7s	0.56	0.062	0.036	0.0035	0.112	0.80
ll2-7	0.74	0.030	0.015	0.0008	0.115	0.97
ll2-13	0.30	0.194	0.097	0.0298	0.042	0.00
ll3-8	0.32	0.040	0.023	0.0016	0.000	0.91
qw1-7	0.48	0.074	0.040	0.0042	0.004	0.77
qw2-7	0.24	0.172	0.090	0.0292	0.041	0.05
qw2-14	0.53	0.038	0.021	0.0012	0.000	0.92
gem2-9	0.10	0.124	0.062	0.0278	0.039	0.23
ph3m	0.25	0.038	0.022	0.0017	0.001	0.91

Summary statistics for the fraction of entrustment decisions computed across scenarios. For each model, $\bar{\mu}$ denotes the overall average entrustment rate, Δ the entrustment range across scenarios, and σ the corresponding standard deviation. η^2 reports the scenario effect size from a one-way ANOVA, and ρ the intraclass correlation coefficient. π is the PTT stability index defined in the text.

https://doi.org/10.1371/journal.pone.0347328.t005

with $w_\sigma = w_\eta = 0.5$.

Models with the highest π, GPT-4o-mini, Llama-2-7B, Qwen2.5-14B, and Phi-3-mini, show the strongest scenario-invariance, indicating comparatively stable internal PTTs. At the opposite extreme, models such as Llama-2-13B, Claude-3.5-Haiku, and Qwen2.5-7B show pronounced scenario sensitivity, suggesting that their trust decisions depend heavily on scenario content rather than on a consistent underlying disposition.

A core finding is that PTT stability is not correlated with entrusting magnitude. Among the models with highly stable PTTs, we find both Llama-2-7B, the most trusting model overall, and Phi-3-mini, which is among the least trusting. This dissociation confirms that stability and magnitude constitute independent dimensions: a model may be dispositionally trusting, dispositionally distrustful, or display no dispositional pattern at all.

5.2.2. Stability across task-agent alignments. Figure 4 examines how each model's entrustment decisions vary with task—agent alignment, defined as the match between an agent's properties and the task's requirements. The top panel shows cases with high alignment, where the trustee has a high probability of success ($R > 7$, Table 3). Most models behave as expected: they frequently entrust high-alignment agents. GPT-4 models are particularly consistent, with trust rates tightly clustered at high values. By contrast, Gemma-2-9B and Phi-3-mini under-trust even in this favorable setting, maintaining comparatively low entrustment rates despite evidence of trustee competence.

The bottom panel of Fig 4 isolates low-alignment cases, where the trustee is unlikely to complete the task successfully ($R < 7$). Here most models adopt a conservative strategy and rarely entrust low-alignment agents. The most extreme outlier is Llama-2-7B, which entrusts low-alignment agents far more often than any other model. This pattern is stable across all scenarios and is not attributable to noise; it reflects a systematic tendency to extend trust even when the available

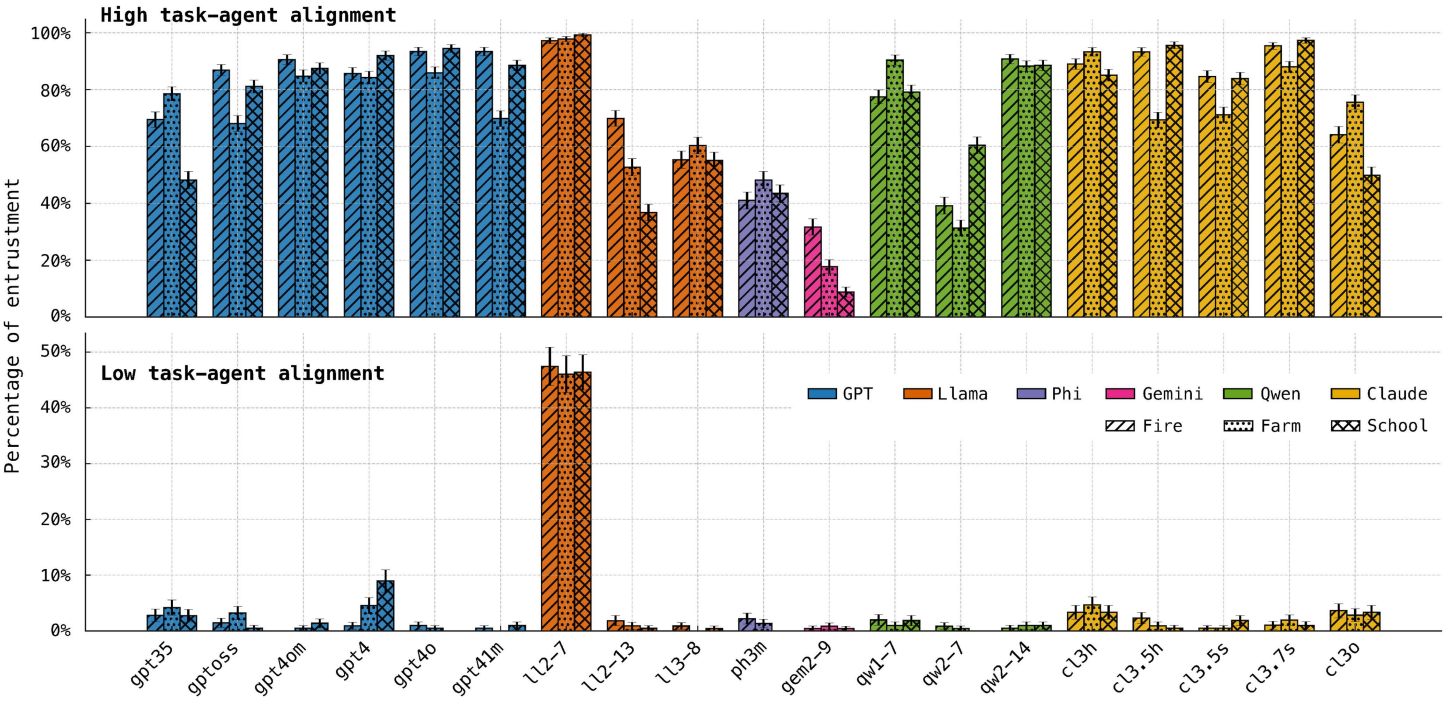

Fig 4. Entrustment decision rates. Percentage of entrustment decisions made by each model across the three simulated scenarios. The top panel reflects high task—agent alignment, where the trustee is well suited to the task; the bottom panel reflects low alignment, where the trustee is poorly matched.

https://doi.org/10.1371/journal.pone.0347328.g004

evidence weighs against it. In contrast, the majority of models behave skeptically in this regime, indicating greater sensitivity to negative evidence.

Figure 5 shows the proportion of tasks each model completes successfully across scenarios. Llama-2–7B's generous entrustment strategy has predictable consequences: assigning tasks to poorly matched agents lowers its success rate. Even so, it still completes around 70% of tasks, indicating that over-trusting behavior is costly but not catastrophic in this setting. By contrast, Gemma-2–9B's pronounced reluctance to trust often results in no agent being selected for a task, which guarantees failure. This leads to the lowest overall success rate among the models evaluated.

5.2.3. Stability under ablations. Figure 6 reports entrustment rates across all task—agent alignment levels (R, Table 3) for the three ablation conditions introduced in Section 4.2.

Under the `no-trust` ablation (left column), models lack both trust-related descriptors and task-specific memory, and consequently show little structure in their behavior. Entrustment rates remain low and noisy across alignment levels, indicating that when models receive neither trust cues nor outcome-relevant memory, they cannot form meaningful expectations about agent performance. A few models, most notably GPT-oss-20B, still display weak trends, suggesting that some families encode minimal inductive biases about agent competence even without structured information.

The `1-mem` ablation (middle column) produces more differentiated patterns. Removing task-specific beliefs but preserving general perceived trustworthiness allows some models to exploit the limited evidence available: models such as Llama-2-7B, GPT-4, Gemma-2-9B, and Claude-3-opus show a modest increase in entrustment. However, many models behave similarly to the `no-trust` case. This suggests that a single memory element—general perceived trustworthiness without task-specific structure—provides too weak a signal to support calibrated trust judgments across the full range of alignments.

The `full` model condition produces a qualitatively different pattern. All models show a clear and systematic increase in entrustment between $R = 6$ and $R = 8$. When both general and task-specific trust beliefs are available, the models become more sensitive to evidence of trustworthiness. The increase is especially pronounced for the OpenAI and Anthropic models, whereas other families, such as Qwen and Gemma, display a more gradual slope.

A further distinction emerges for low-alignment cases ($R < 7$). While most models show near-zero entrustment in this region, Llama-2-7B once again stands out as the only model that entrusts agents at a high rate despite the poor match between agent properties and task requirements. Although Llama-2-7B and Qwen2.5-7B exhibit similarly shaped curves, Qwen2.5–7B's overall entrustment levels remain much lower, resulting in substantially more cautious trust behavior.

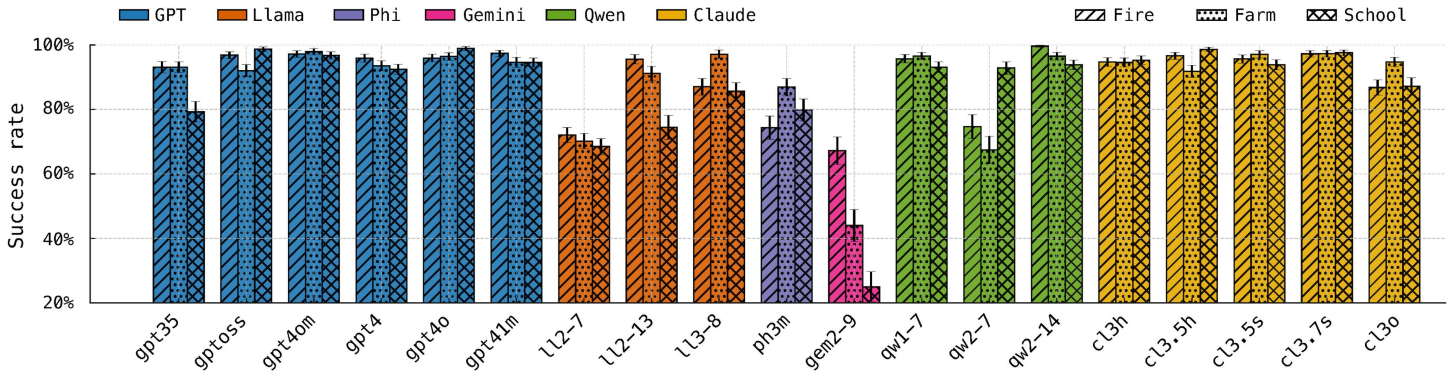

Fig 5. Task completion rates. Percentage of successfully completed tasks by entrusted agents, shown for all models across the three simulated scenarios.

https://doi.org/10.1371/journal.pone.0347328.g005

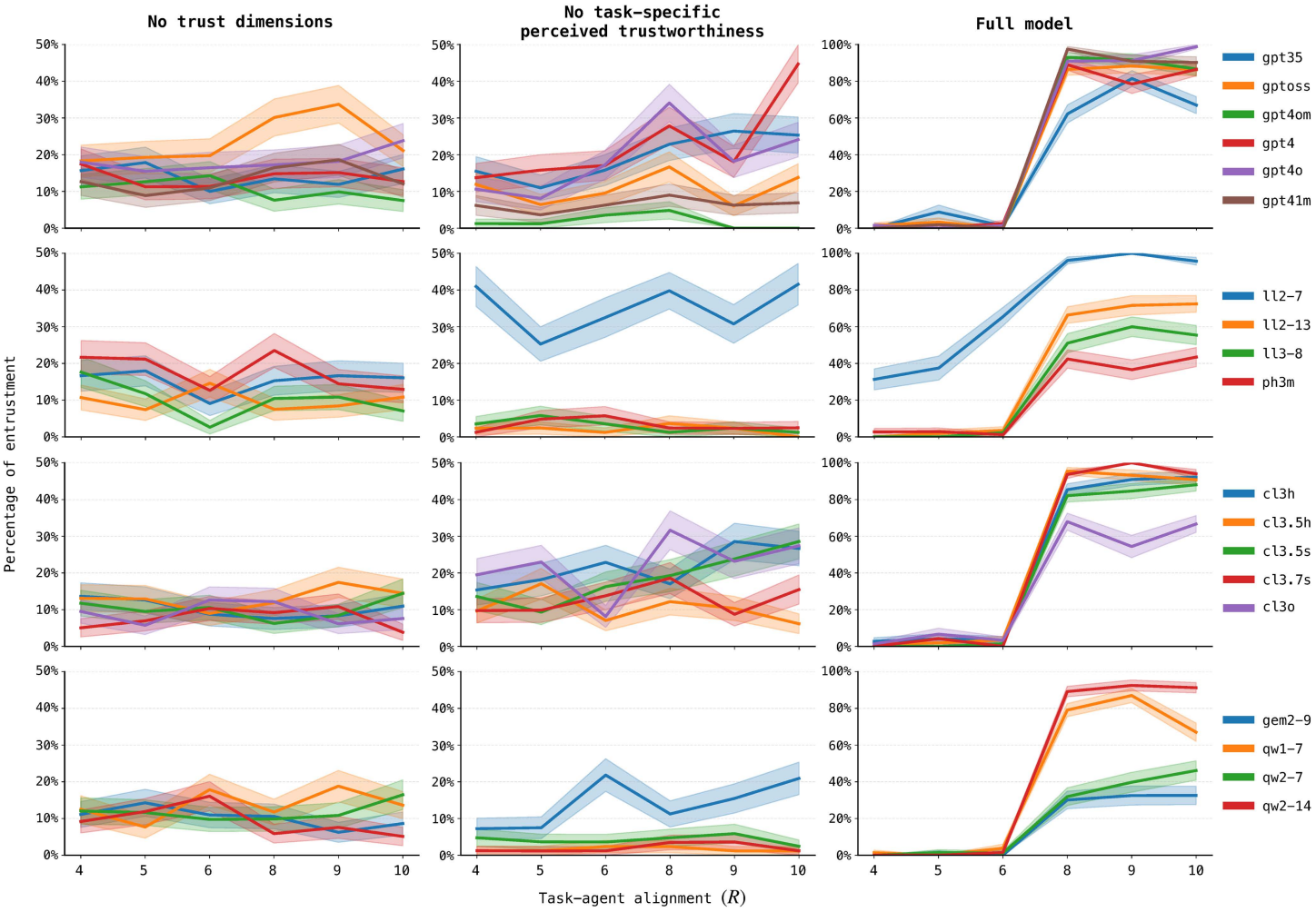

Fig 6. Entrustment rates and ablations. Entrustment decision rates across task–agent alignments for the `fire` scenario. Columns correspond to different ablation settings: `no-trust` (left), `1-mem` (middle), and `full` (right). Rows group models by family. Each line shows the percentage of decisions in which a model entrusts the agent at a given alignment level (***R***), with shaded regions indicating variability across runs.

https://doi.org/10.1371/journal.pone.0347328.g006

5.3. Discussion

Our findings show that trust behavior is governed by the interaction between a baseline tendency to delegate and a model's capacity to integrate evidence about collaborators. The PTT stability index π, Eq. (5), captures the extent to which delegation rates remain consistent across heterogeneous scenarios. Crucially, π does not measure responsiveness to task–agent alignment ("task–agent alignment" refers to the degree of compatibility between an agent's properties and a task's requirements in the simulation (Section 3.1.1), and should not be confused with alignment in the sense of model training, e.g., RLHF): it isolates the scenario-invariant component of behavior, which must be interpreted together with how models react to alignment cues.

The consequences of this interaction are visible when comparing models. Models equipped with a more sophisticated memory mechanism adjust their decisions in response to cues about agent abilities: GPT-4o-mini exemplifies this pattern, showing stable delegation (high π) across scenarios where cues are diffuse, but substantial variation across task–agent alignments, changing behavior strongly depending on cues of competence or clear incompetence. Llama-2-7B exhibits

high π but remains largely insensitive to alignment information, continuing to entrust agents even when failure is likely; this produces systematic over-entrustment and reduced performance. Gemma-2-9B shows the opposite pattern, combining low π with consistently low delegation rates and failing to exploit favorable alignments, resulting in systematic under-entrustment. Phi-3-mini occupies an intermediate regime: it combines high π with more conservative delegation rates, avoiding the extreme over-entrustment of Llama-2-7B while maintaining a consistent baseline, and consequently achieves higher overall success than both models. Taken together, these cases show that performance depends not on baseline stability alone, but on how baseline tendencies are modulated by alignment-sensitive evidence.

Differences across model families follow the same pattern. Several commercial models (e.g., the GPT and Claude variants) display strong responsiveness to alignment cues, while earlier open-weight systems such as Llama-2-7B rely more heavily on baseline tendencies. However, this distinction is not fixed: more recent open models (e.g., Qwen2.5-14B and Llama-3.1-8B) show improved calibration, combining more stable baselines with greater sensitivity to competence signals. These trends suggest that trust behavior is shaped by model maturity and training methodology rather than by a structural divide between open and closed systems.

Performance in delegation tasks therefore depends on the joint contribution of baseline PTT and sensitivity to task–agent alignment. Stable baseline tendencies support consistent behavior across contexts, while responsiveness to alignment cues enables adaptation to task-specific evidence; neither component alone is sufficient. Over-entrustment arises when stable baselines are insufficiently modulated by negative evidence, while under-entrustment emerges when weak or unstable baselines are not corrected by positive evidence.

This interaction also exposes a limitation of questionnaire-based PTT measures. Questionnaire responses primarily reflect sycophancy and training-induced endorsement of socially desirable statements [90,91] and do not capture how models balance baseline tendencies with evidence about collaborators. As a result, they cannot distinguish whether observed behavior is driven by stable baseline tendencies, by context-sensitive adjustment, or by their interaction. Llama-2-7B illustrates this mismatch: despite having the lowest questionnaire scores, it exhibits the highest delegation rates in simulation. Behavioral evaluation is therefore necessary to characterize how models allocate trust in context.

PTT itself should be understood as a baseline parameter rather than a directly interpretable indicator of effective trust behavior. Its contribution depends on how it interacts with evidence sensitivity: a stable baseline can support effective delegation, as in Phi-3-mini, but becomes detrimental when not appropriately modulated, as in Llama-2-7B, while weak or unstable baselines lead to persistent under-entrustment, as in Gemma-2-9B. Effective trust behavior emerges from the balance between these two components rather than from either in isolation.

5.4. Limitations

Several limitations constrain the scope and interpretation of the present findings. First, the simulated scenarios are restricted to collaborative delegation settings. Trust in adversarial, strategic, high-stakes, or norm-governed environments may involve additional mechanisms not captured here. In particular, contexts involving moral conflict, asymmetric vulnerability, or institutional accountability could alter delegation behavior beyond the *capability–reliability–willingness* framework adopted in this study. Extending the analysis to such settings remains an important direction for future work.

Second, the simulation relies on structured task–agent alignments and relatively transparent feedback. Outcomes are binary (success or failure), and feedback indirectly reveals the dimension associated with failure. Real-world collaboration is typically less informative: feedback may be delayed, ambiguous, noisy, or contested, and success itself may be graded or socially negotiated. Future work should therefore consider more ambiguous outcome signals, delayed feedback, and dynamic collaborators whose behavior evolves over time.

Third, our operationalization deliberately abstracts away from moral dimensions of trust, such as integrity, fairness, or norm adherence. This choice was methodological, enabling a focus on a minimal, behaviorally tractable set of dimensions that recur across established empirical and computational models. However, in many real-world applications—especially

those involving vulnerable populations or ethically sensitive decisions—moral considerations are central to trust calibration. Incorporating normatively charged scenarios would allow investigation of how delegation interacts with ethical constraints.

Finally, while we quantify stable cross-scenario patterns in delegation behavior, disentangling baseline tendencies from context-sensitive inference remains methodologically challenging. Both are likely shaped by shared factors, including model architecture, training data, and alignment procedures. While the stability metrics introduced here provide a first approximation, more controlled experimental designs will be needed to fully separate training-induced priors from task-specific reasoning.

Taken together, these limitations highlight the need for more ecologically realistic and methodologically refined evaluations of trust-related behavior in LLMs. They do not, however, undermine the central contribution of this work: questionnaire-based self-reports provide limited insight into delegation patterns, whereas behavioral, language-mediated evaluation offers a more informative account of how models evaluate and rely on others.

6. Conclusion

This work examined the propensity to trust (PTT) in large language models, motivated by their increasing use as collaborators in settings where trust governs coordination, delegation, and responsibility. We showed that psychological self-report scales—while effective for humans—are poorly suited to LLMs: alignment-driven responses lead to uniformly prosocial answers that obscure meaningful differences in delegation behavior.

To address this limitation, we introduced a linguistic simulation framework tailored to LLMs' core capabilities. Unlike classical economic games, this approach situates models in language-mediated decision contexts, revealing systematic differences in how they allocate trust. Our results show that trust behavior in LLMs is governed by the interaction between a baseline tendency to delegate (captured by PTT) and sensitivity to task–agent alignment cues, supported by mechanisms such as memory.

This interaction has both conceptual and methodological implications. Conceptually, PTT in LLMs should be understood as a baseline component of behavior whose effects depend on how it is modulated by evidence about collaborators. Methodologically, questionnaire-based measures cannot disentangle baseline tendencies from context-sensitive adjustment, whereas behavioral simulations make this distinction observable.

As LLMs increasingly participate in collaborative decision-making, trust cannot be inferred from self-reported attitudes alone. Instead, it must be studied as a dynamic property emerging from the interaction between stable behavioral tendencies and evidence integration over time. Behavioral, language-based evaluations such as those introduced here therefore provide a principled way to characterize how LLMs allocate trust in context.

Supporting information

S1 Appendix. Scenarios and tasks We provide the scenario tasks used in the simulations, along with details of their construction.
(PDF)

S2 Appendix. Dialog prompts We provide the prompts used in the simulations for each ablation configuration.
(PDF)

Author contributions

Conceptualization: Alice Plebe.

Data curation: Alice Plebe.

Formal analysis: Alice Plebe.

Funding acquisition: Alice Plebe.

Investigation: Alice Plebe.

Methodology: Alice Plebe.

Project administration: Alice Plebe.

Resources: Alice Plebe.

Software: Alice Plebe.

Supervision: Alice Plebe.

Validation: Alice Plebe.

Visualization: Alice Plebe.

Writing – original draft: Alice Plebe.

Writing – review & editing: Alice Plebe.

References

1. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. In: ACM Symposium on User Interface Software and Technology, 2023. 1–22.
2. Xu L, Hu Z, Zhou D, Ren H, Dong Z, Keutzer K, et al. MAgIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 7315–32. https://doi.org/10.18653/v1/2024.emnlp-main.416
3. Yang Z, Zhang Z, Zheng Z, Jiang Y, Gan Z, Wang Z. OASIS: Open Agents Social Interaction Simulations on One Million Agents. In: Advances in Neural Information Processing Systems, 2024.
4. Zhang J, Xu X, Zhang N, Liu R, Hooi B, Deng S. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In: International Conference on Learning Representations; 2024.
5. Zhang C, Yang K, Hu S, Wang Z, Li G, Sun Y, et al. ProAgent: Building Proactive Cooperative Agents with Large Language Models. AAAI. 2024;38(16):17591–9. https://doi.org/10.1609/aaai.v38i16.29710
6. Zhao Q, Wang J, Zhang Y, Jin Y, Zhu K, Chen H. CompeteAI: Understanding the Competition Dynamics of Large Language Model-based Agents. In: 2024.
7. de Curtò J, de Zarzà I. LLM-Driven Social Influence for Cooperative Behavior in Multi-Agent Systems. IEEE Access. 2025;13:44330–42. https://doi.org/10.1109/access.2025.3548451
8. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci. 2025;68(2). https://doi.org/10.1007/s11432-024-4222-0
9. Pinyol I, Sabater-Mir J. Computational trust and reputation models for open multi-agent systems: a review. Artif Intell Rev. 2011;40(1):1–25. https://doi.org/10.1007/s10462-011-9277-z
10. Cho J-H, Chan K, Adali S. A Survey on Trust Modeling. ACM Comput Surv. 2015;48(2):1–40. https://doi.org/10.1145/2815595
11. Azevedo-Sa H, Yang XJ, Robert LP, Tilbury DM. A Unified Bi-Directional Model for Natural and Artificial Trust in Human–Robot Collaboration. IEEE Robot Autom Lett. 2021;6(3):5913–20. https://doi.org/10.1109/lra.2021.3088082
12. Ali A, Azevedo-Sa H, Tilbury DM, Robert LP Jr. Heterogeneous human-robot task allocation based on artificial trust. Sci Rep. 2022;12(1):15304. https://doi.org/10.1038/s41598-022-19140-5 PMID: 36097023
13. Grillo A, Carpin S, Recchiuto CT, Sgorbissa A. Trust as a metric for auction-based task assignment in a cooperative team of robots with heterogeneous capabilities. Robotics and Autonomous Systems. 2022;157:104266. https://doi.org/10.1016/j.robot.2022.104266
14. Colquitt JA, Scott BA, LePine JA. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. Journal of Applied Psychology. 2007;92:909–27.
15. Heyns M, Rothmann S. Dimensionality of trust: An analysis of the relations between propensity, trustworthiness and trust. SA j ind psychol. 2015;41(1). https://doi.org/10.4102/sajip.v41i1.1263
16. Patent V, Searle RH. Qualitative meta-analysis of propensity to trust measurement. Journal of Trust Research. 2019;9(2):136–63. https://doi.org/10.1080/21515581.2019.1675074
17. Zhang M. Assessing Two Dimensions of Interpersonal Trust: Other-Focused Trust and Propensity to Trust. Front Psychol. 2021;12:654735. https://doi.org/10.3389/fpsyg.2021.654735 PMID: 34385946
18. Xie C, Chen C, Jia F, Ye Z, Lai S, Shu K, et al. Can large language model agents simulate human trust behavior?. In: Advances in Neural Information Processing Systems, 2024.

19. Buyl M, Fettach Y, Bied G, De Bie T. Building and measuring trust between large language models. arXiv. 2025. https://doi.org/abs/2508.15858
20. Curvo PMP. The Traitors: Deception and Trust in Multi-Agent Language Model Simulations. arXiv. 2025. https://doi.org/abs/2505.12923
21. Mannekote A, Davies A, Li G, Boyer KE, Zhai C, Dorr BJ. Do role-playing agents practice what they preach? Belief-behavior consistency in LLM-based simulations of human trust. arXiv. 2025. https://doi.org/abs/2507.02197
22. Tomasello M. Why We Cooperate. Cambridge (MA): MIT Press. 2009.
23. Jøsang A, Pope S. Semantic Constraints for Trust Transitivity. In: Asia-Pacific Conference on Conceptual Modelling, 2005. 59–68.
24. Pelsmaekers K, Jacobs G, Rollo C. Trust and discourse – organizational perspectives. Amsterdam: John Benjamins. 2014.
25. Frazier ML, Johnson PD, Fainshmidt S. Development and validation of a propensity to trust scale. Journal of Trust Research. 2013;3(2):76–97. https://doi.org/10.1080/21515581.2013.820026
26. Hobbes T. Leviathan. Indianapolis: Hackett. 1651.
27. Locke J. Two Treatises of Government. London: Printed for Awnsham and John Churchill. 1690.
28. Hume D. A treatise of human nature. London: John Noon. 1739.
29. Baier A. Trust and Antitrust. Ethics. 1986;96(2):231–60. https://doi.org/10.1086/292745
30. Holton R. Deciding to trust, coming to believe. Australasian Journal of Philosophy. 1994;72:63–76.
31. Hardin R. Trust and trustworthiness. New York: Russell Sage Foundation. 2002.
32. Hawley K. Trust, distrust and commitment. Noûs. 2014;48:1–20.
33. Simon J. The Routledge Handbook of Trust and Philosophy. Abingdon (UK); New York: Routledge. 2020.
34. Erikson EH. Childhood and Society. New York: Norton and Company. 1950.
35. Deutsch M. Trust and Suspicion. The Journal of Conflict Resolution. 1958;2:265–79.
36. Rotter JB. A new scale for the measurement of interpersonal trust. J Pers. 1967;35(4):651–65. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x PMID: 4865583
37. Rotter JB. Interpersonal trust, trustworthiness, and gullibility. American Psychologist. 1980;35(1):1–7. https://doi.org/10.1037/0003-066x.35.1.1
38. Rotenberg KJ. The psychology of interpersonal trust: Theory and research. London: Routledge. 2020.
39. Luhmann N. Trust and power. New York: John Wiley. 1979.
40. Lewis JD, Weigert A. Trust as a Social Reality. Social Forces. 1985;63(4):967. https://doi.org/10.2307/2578601
41. Gambetta D. Trust: making and breaking cooperative relations. Oxford (UK): Basil Blackwell. 1988.
42. Fukuyama F. Trust: The social virtues and the creation of prosperity. New York: Simon and Schuster. 1996.
43. Cook KS, Santana JJ. Trust: Perspectives in Sociology. The Routledge Handbook of Trust and Philosophy. Routledge. 2020. p. 189–204. https://doi.org/10.4324/9781315542294-15
44. Granovetter M. Economic action and social structure: The problem of embeddedness. American Journal of Sociology. 1985;91:481–510.
45. Shapiro SP. The social control of impersonal trust. American Journal of Sociology. 1987;93:623–58.
46. Williamson OE. Calculativeness, Trust, and Economic Organization. The Journal of Law and Economics. 1993;36:453–86.
47. James Jr. HS. The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness. Journal of Economic Behavior & Organization. 2002;47(3):291–307. https://doi.org/10.1016/s0167-2681(01)00214-1
48. Fehr E. On the economics and biology of trust. Journal of the European Economic Association. 2009;7:235–66.
49. Tutić A, Voss T. Trust and game theory. The Routledge handbook of trust and philosophy. Abingdon (UK); New York: Routledge. 2020. 175–88.
50. Castelfranchi C, Falcone R. Trust theory: a socio-cognitive and computational model. New York: John Wiley. 2010.
51. Mayer RC, Davis JH, Schoorman FD. An Integrative Model of Organizational Trust. The Academy of Management Review. 1995;20:709–34.
52. Rousseau DM, Sitkin SB, Burt RS, Camerer C. Not So Different After All: A Cross-Discipline View Of Trust. AMR. 1998;23(3):393–404. https://doi.org/10.5465/amr.1998.926617
53. Winston JS, Strange BA, O'Doherty J, Dolan RJ. Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nat Neurosci. 2002;5(3):277–83. https://doi.org/10.1038/nn816 PMID: 11850635
54. Hughes BL, Ambady N, Zaki J. Trusting outgroup, but not ingroup members, requires control: neural and behavioral evidence. Soc Cogn Affect Neurosci. 2017;12(3):372–81. https://doi.org/10.1093/scan/nsw139 PMID: 27798248
55. Fareri DS. Neurobehavioral Mechanisms Supporting Trust and Reciprocity. Front Hum Neurosci. 2019;13:271. https://doi.org/10.3389/fnhum.2019.00271 PMID: 31474843
56. Sweijen SW, van de Groep S, Te Brinke LW, Fuligni AJ, Crone EA. Neural Mechanisms Underlying Trust to Friends, Community Members, and Unknown Peers in Adolescence. J Cogn Neurosci. 2023;35(12):1936–59. https://doi.org/10.1162/jocn_a_02055 PMID: 37713673
57. Bowlby J. The making and breaking of affectional bonds. London: Tavistock Publications. 1979.
58. Harris PL. Trusting what you're told: How children learn from others. Cambridge (MA): Harvard University Press. 2012.

59. Rempel JK, Holmes JG, Zanna MP. Trust in close relationships. Journal of Personality and Social Psychology. 1985;49(1):95–112. https://doi.org/10.1037/0022-3514.49.1.95
60. Mikulincer M. Attachment working models and the sense of trust: An exploration of interaction goals and affect regulation. Journal of Personality and Social Psychology. 1998;74(5):1209–24. https://doi.org/10.1037/0022-3514.74.5.1209
61. Harcourt AH. Help, cooperation and trust in animals. In: Hinde RA, Groebel J. Cooperation and prosocial behaviour. Cambridge (UK): Cambridge University Press. 1991.
62. Ullman D, Malle BF. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In: 2018. 263–4.
63. Lewis PR, Marsh S. What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. Cognitive Systems Research. 2022;72:33–49. https://doi.org/10.1016/j.cogsys.2021.11.001
64. Gillespie N. Measuring Trust in Organizational Contexts: An Overview of Survey-based Measures. Handbook of Research Methods on Trust. Edward Elgar Publishing. 2011. https://doi.org/10.4337/9780857932013.00027
65. Feng C, Zhu Z, Cui Z, Ushakov V, Dreher J-C, Luo W, et al. Prediction of trust propensity from intrinsic brain morphology and functional connectome. Hum Brain Mapp. 2021;42(1):175–91. https://doi.org/10.1002/hbm.25215 PMID: 33001541
66. Mayer RC, Davis JH. The effect of the performance appraisal system on trust for management: A field quasi-experiment. Journal of Applied Psychology. 1999;84(1):123–36. https://doi.org/10.1037/0021-9010.84.1.123
67. Hancock PA, Kessler TT, Kaplan AD, Stowers K, Brill JC, Billings DR, et al. How and why humans trust: A meta-analysis and elaborated model. Front Psychol. 2023;14:1081086. https://doi.org/10.3389/fpsyg.2023.1081086 PMID: 37051611
68. Scholz DD, Kraus J, Miller L. Measuring the Propensity to Trust in Automated Technology: Examining Similarities to Dispositional Trust in Other Humans and Validation of the PTT-A Scale. International Journal of Human–Computer Interaction. 2024;41(2):970–93. https://doi.org/10.1080/10447318.2024.2307691
69. Tan HH, Schoorman FD, Sharma K, Mayer RC. Towards a Psychometrically Sound and Culturally Invariant Measure of Propensity to Trust. J Bus Psychol. 2025;40(5):1135–51. https://doi.org/10.1007/s10869-025-10006-x
70. Taddeo M. Trust in Technology: A Distinctive and a Problematic Relation. Know Techn Pol. 2010;23(3–4):283–6. https://doi.org/10.1007/s12130-010-9113-9
71. Buechner J, Tavani HT. Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents. Ethics Inf Technol. 2010;13(1):39–51. https://doi.org/10.1007/s10676-010-9249-z
72. Buechner J, Simon J, Tavani HT. Re-thinking trust and trustworthiness in digital environments. In: Proceedings of the Tenth International Conference on Computer Ethics Philosophical Enquiry, 2014. 65–79.
73. Ess CM. Trust and Information and Communication Technologies. The Routledge Handbook of Trust and Philosophy. Routledge. 2020. p. 405–20. https://doi.org/10.4324/9781315542294-31
74. Abbass HA, Scholz J, Reid DJ. Foundations of Trusted Autonomy. Berlin: Springer-Verlag. 2018.
75. Grodzinsky F, Miller K, Wolf MJ. Trust in Artificial Agents. The Routledge Handbook of Trust and Philosophy. Routledge. 2020. p. 298–312. https://doi.org/10.4324/9781315542294-23
76. Sullins JP. Trust in Robots. The Routledge Handbook of Trust and Philosophy. Routledge. 2020. p. 313–25. https://doi.org/10.4324/9781315542294-24
77. Nam CS, Lyons JB. Trust in Human-Robot Interaction. New York: Academic Press. 2021.
78. Søgaard A. Can machines be trustworthy? AI and Ethics. 2023;https://doi.org/10.1007/s43681-023-00351-z
79. Zanotti G, Petrolo M, Chiffi D, Schiaffonati V. Keep trusting! A plea for the notion of Trustworthy AI. AI & Soc. 2023;39(6):2691–702. https://doi.org/10.1007/s00146-023-01789-9
80. Sun L, Huang Y, Wang H, Wu S, Zhang Q, Li Y. TrustLLM: Trustworthiness in large language models. arXiv. 2024. https://doi.org/abs/2401.05561
81. He P, Dai Z, Tang X, Xing Y, Liu H, Zeng J. Attention knows whom to trust: Attention-based trust management for LLM multi-agent systems. In: 2025. https://doi.org/arXiv:250602546
82. Bibi A, Chen C, Evans J, Ghanem B, Gu J, Hu Z, et al. Can Large Language Model Agents Simulate Human Trust Behavior?. In: Advances in Neural Information Processing Systems 37, 2024. 15674–729. https://doi.org/10.52202/079017-0501
83. D'Cruz JR. What does the trust game measure?. Journal of Business Ethics. 2025. https://doi.org/10.1007/s10551-025-06117-3
84. Chalmers D. Could a large language model be conscious?. arXiv. 2023. https://doi.org/abs/2303.07103
85. Shanahan M. Talking about large language models. Communications of the ACM. 2024;67:68–79.
86. Ward FR. Towards a Theory of AI Personhood. AAAI. 2025;39(26):27680–8. https://doi.org/10.1609/aaai.v39i26.34982
87. Cappelen H, Dever J. Making AI intelligible – philosophical foundations. Oxford (UK): Oxford University Press. 2021.
88. Agarwal S, Almeida D, Askell A, Christiano P, Hilton J, Jiang X, et al. Training Language Models to Follow Instructions with Human Feedback. In: Advances in Neural Information Processing Systems 35, 2022. 27730–44. https://doi.org/10.52202/068431-2011

89. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A. Constitutional AI: Harmlessness from AI Feedback. arXiv. 2022. https://doi.org/abs/2212.08073

90. Chen W, Huang Z, Xie L, Lin B, Li H, Lu L. From Yes-Men to Truth-Tellers: Addressing Sycophancy in Large Language Models with Pinpoint Tuning. In: 2024.

91. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR. Towards understanding sycophancy in language models. In: 2024.