Consumer Trust in Agentic Commerce

From Model Properties to Verifiable Architecture

Vadym Chernets, PhD*

First posted 6 August 2026; this version September 2026

Reports the full conceptual framework, the sixteen-model mechanism study and its August 2026 flagship-wave replication across eleven frontier laboratories, a paired reasoning-budget comparison, a commerce-substrate replication, and a pre-registered consumer experiment in full.

ABSTRACT

Agentic commerce turns AI from an advisor into a transacting party, collapsing search, comparison, cart, and payment into a single delegated act and removing the review points where humans once caught errors. It arrives at an awkward moment: sixty percent of UK consumers surveyed say they would abandon an AI shopping agent after a single mistake, and most would trust no organization at all to run one on their behalf. This paper introduces *architectural trust*. As consumers delegate consequential decisions to AI agents, I argue, warranted trust migrates from the properties of the AI model to the verifiable architecture of the service above it. The constructs already available attach trust to organizations, to a single automated system, or to algorithmic advice, and none of them makes the decision-producing process verifiable by the truster under autonomous delegation. Architectural trust fills that gap with three checkable properties: diversity (independent models whose agreement calibrates confidence and whose divergence is surfaced), evidence (tamper-evident, replayable decision records), and neutrality (auditable independence from any vendor, merchant, or outcome).

The mechanism was measured on the machine side first. Across a pre-registered sixteen-model benchmark and an August 2026 flagship-tier wave spanning eleven frontier laboratories, cross-model agreement priced correctness where a model's own confidence could not. Answers independently reproduced by another model were 1.9–3.2 times as likely to be correct. On a frozen SimpleQA subsample (Wei et al., 2024), 46.6% of answers offered at stated confidence ≥ 80 were wrong (a rate specific to this roster, not a universal constant), while a fixed three-model disagreement gate removed 82.1% of those confident failures (95% CI 79.9–84.1) at 30.5% coverage. The signal reproduces on 267 Wikidata-verified commerce product facts. A larger reasoning budget bought 3.1 points of stated confidence and no measurable accuracy: more compute buys confidence, not correctness. A unanimously confident panel can still be wrong, and there agreement certifies the shared error.

---

* AI systems architect. Competing interests: the author develops multi-model orchestration methods, operates a reference implementation of the architecture described here, is a named applicant on pending patent applications, not granted patents, covering AI orchestration and multi-model validation systems of the kind this paper examines, and has a developer's interest in the architecture this paper advocates; normative claims about its necessity should be read with that interest in view. The author's role in the Section 4 deployment is set out in the Author's Note and Disclosure at the end of this paper. Comments welcome: vadimchernets9@gmail.com.

I then carry the mechanism into behavior in a pre-registered, incentive-compatible consumer experiment (119 of 200 planned participants; US and UK). Where a consumer would otherwise accept a confident wrong recommendation, the disagreement display cut acceptance sharply, from 74% to 24% on the qualifying item. The pooled robustness estimate ($\Delta = 0.24 \rightarrow 0.45$; $d = 0.68$; the primary main-wave cohort gives $d = 0.67$) rests on that single item and, in an exploratory country split, on the UK sample, and it falls to $d = 0.09$ once the item is removed. A powered, stimulus-sampled replication is planned. On the one item of unanimous confident consensus the display *raised* acceptance of the shared error, which reproduces the mechanism's boundary in human behavior. The orchestration-receipt component returned a registered, inconclusive null.

The three components do not have equal evidence behind them. Diversity is established on the machine side and shown with consumers for the calibration contrast only (the registered mediation was null). The record component's first-exposure test was inconclusive, and both it and neutrality are specified for test. The unit of analysis for AI trust has to move from the model to the architecture above it, the part that can be checked.

Keywords: architectural trust; trust migration; agentic commerce; AI agents; consumer trust; consumer delegation; multi-model orchestration; decision evidence; verifiable AI

JEL Classification: M31 (Marketing); D83 (Search, Learning, Information); L86 (Information and Internet Services); K24 (Cyber Law)

How to read this document. This full version reports the whole evidence chain: the construct and the Trust-Migration Hypothesis; a sixteen-model mechanism study (with an August 2026 flagship-wave replication and a commerce-substrate arm on product facts) that establishes cross-model diversity as a correctness signal; a pre-registered consumer experiment taking that mechanism to behavior; and a discussion of neutrality, market economics, and policy. Pre-registered confirmatory results and the evidential status of each construct component (diversity, evidence, neutrality) are summarized under "Evidential status of claims" in Section 5. Exploratory and robustness analyses are labeled as such throughout and may be skipped on a first pass.

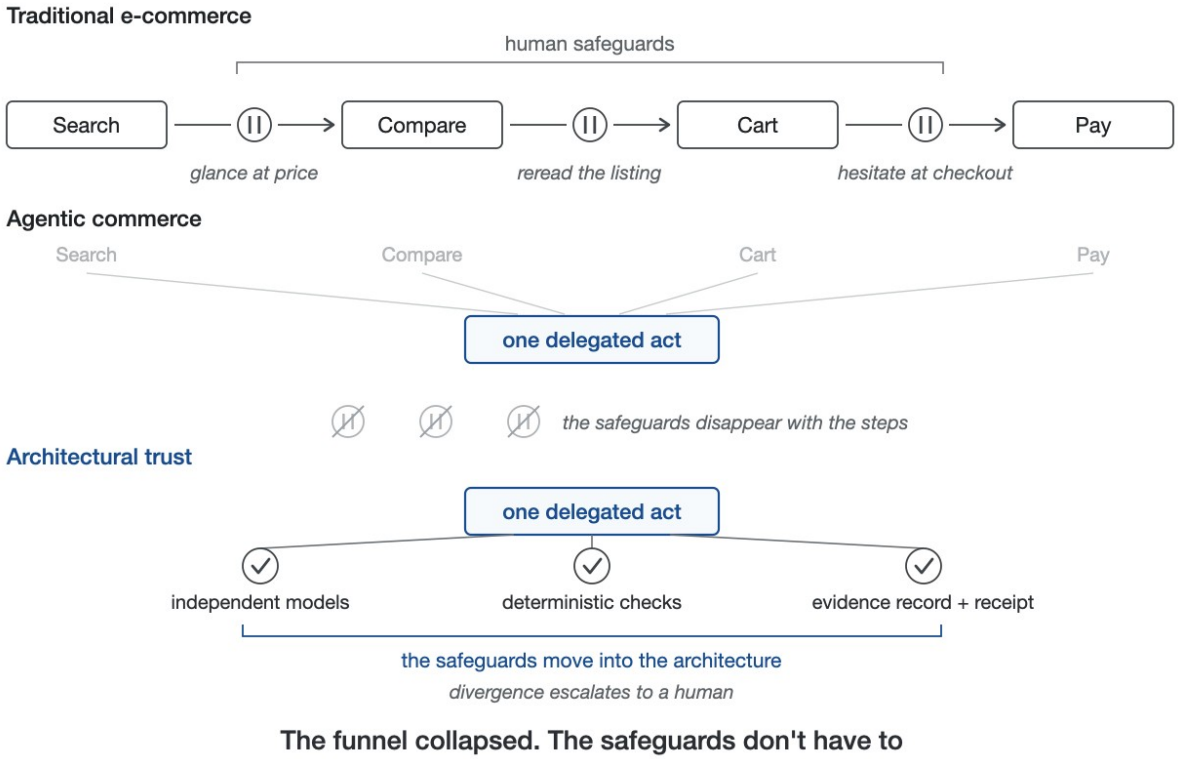
1. Introduction: growth amidst distrust

On 10 June 2026, Visa announced it was embedding its payment network directly into ChatGPT (Visa, 2026). A consumer links a card, sets spending caps and approved merchant categories, and the announced integration, with use cases rolling out in stages, is aimed at AI agents completing real purchases across the network's merchant base (175 million-plus locations, per trade coverage of the announcement). Visa was not alone in this. OpenAI had shipped in-chat checkout in September 2025; Google introduced a Universal Commerce Protocol at NRF in January 2026 with Shopify, Etsy, Wayfair, Target, and Walmart, extending it by May into a cross-retailer Universal Cart (Google, 2026); American Express launched an agentic-commerce developer kit with purchase protection in April (American Express, 2026). The threshold has been crossed by an industry, not by one company. Analyst projections for agentic commerce by 2030 run from hundreds of billions (Bain; Morgan Stanley) to McKinsey's $3–5 trillion. They are estimates and not facts, though every major house is now producing them.

The realized numbers are no longer estimates. Enterprise spending on model APIs alone jumped from $3.5 billion to $8.4 billion in six months (Menlo Ventures, 2025a; industry survey research, not peer-reviewed), inside roughly $37 billion of enterprise generative-AI spend for the year (Menlo Ventures, 2025b). The same industry is already discounting its own execution: Gartner (2025) forecasts that over 40% of agentic AI projects will be canceled by end-2027. Money is arriving faster than warranted trust, which is the gap I work on here. Claims below are tiered in the evidential ledger near the end of the paper as established by these data, supported at modest single-study scale, or theoretical. That tiering governs, and outranks, any summary sentence in this introduction.

What makes this shift qualitatively different from earlier e-commerce waves is *funnel compression*: search, comparison, cart, and payment collapse into a single delegated act. Every safeguard that used to live between those steps (the human glancing at the price, rereading the listing, hesitating at checkout) disappears with the steps themselves (Figure 1). The AI assistant stops being an advisor whose output a human reviews and becomes a transacting party. Human-factors research named this failure mode four decades ago: automation that removes the operator from the loop erodes exactly the vigilance needed to catch its errors (Bainbridge, 1983; Endsley & Kiris, 1995). Consumer research anticipated the threshold too. Autonomous shopping systems face qualitatively different adoption barriers exactly because they remove the human from the doing (de Bellis & Johar, 2020), and the autonomy cost of handing choices to algorithms was flagged as a consumer-welfare question before LLMs existed (André et al., 2018).

Figure 1. *Funnel compression*



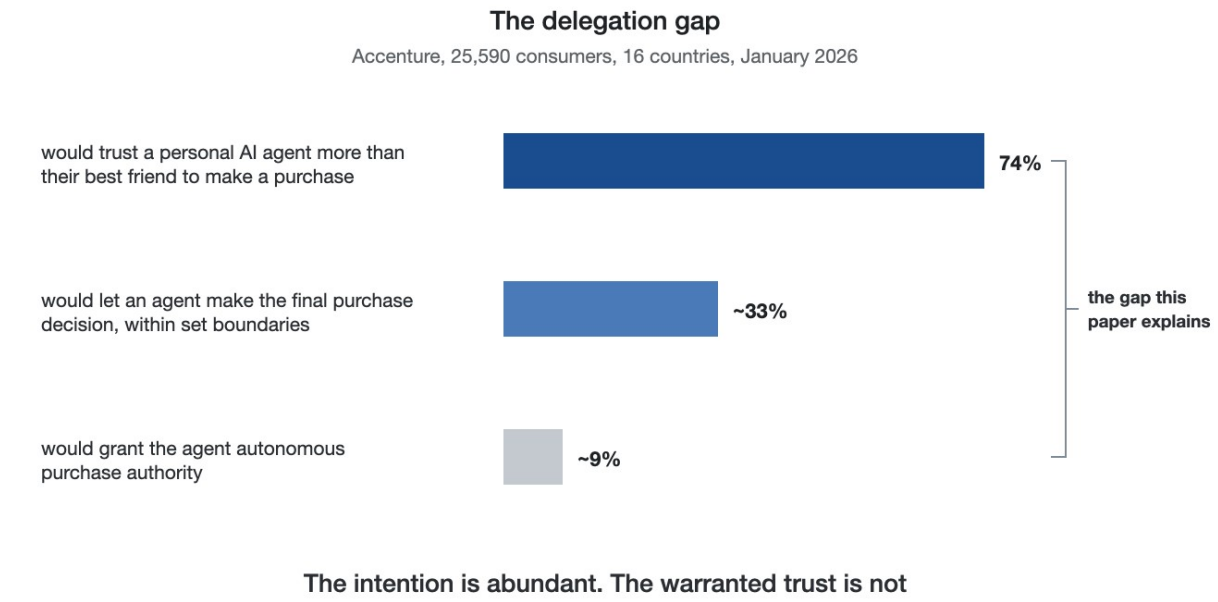
The human safeguards disappear with the steps they lived between.

Visa's own guardrails (caps, whitelists, required approvals) exist because the consumer is being asked to trust something new: not a brand, not an answer, but a decision-making process they cannot see. Guardrails fixed in advance do not close that gap. A rule set beforehand does not let the consumer check how a decision was made. In the ACI Worldwide (2026) survey cited below (an industry survey released by press report, like the other vendor-fielded surveys this section cites), 69% of UK consumers said they distrust AI agents *even when the agents follow rules the consumers set themselves*.

The commercial stakes of getting this wrong are sharp. Early survey evidence suggests consumer trust in agents is one-strike in stated intention and, more interestingly, homeless. Revealed behavior, the checkout pullback below and the Klarna reversal and insurance repricing of Section 5, corroborates these stated preferences and carries the greater evidential weight. In a YouGov survey for ACI Worldwide (2,080 UK adults, fielded 19–22 June 2026, a single-market snapshot awaiting cross-market replication), 60% said they would stop using an AI shopping agent after a single mistake, and only 19% would trust AI with everyday purchases, against 55% who would trust a human expert.

The most consequential number in that survey is a different one: 59% said they would not trust *any* organization to run AI shopping and payments on their behalf, and even banks were chosen by only 20%. In a single surveyed market, pending cross-market replication, that is a trust vacuum. The appetite is there. In Accenture's 16-country industry survey of 25,590 consumers (January 2026; company-fielded, not peer-reviewed), 74% said they would trust a personal AI agent more than their best friend to make a purchase on their behalf, yet only about one in ten would grant an agent autonomous purchase authority (Figure 2). The two items measure different grades of delegation, so the spread aligns intentions rather than identifying a single causal gap, but the spread is stark. The intention is abundant; the delegation is scarce; what is missing between them is warranted trust. The market has already tried something close to this. OpenAI shipped native in-chat checkout in September 2025, a deliberately limited first step, initially single-item purchases from U.S. Etsy sellers with the user confirming each step. It pulled the flow back by March 2026, shifting purchases out to retailers' own apps, and only a handful of merchants ever used it (as reported by The Information; CNBC, 2026). The company stated no reason, so the cause is not established. The reading offered here, adoption infrastructure without warranted trust failing to convert, is an interpretation consistent with the reported facts. If consumers distrust every *player*, trust cannot be vested in a player at all, and has to live instead in a verifiable *process*.

Figure 2. *Stated trust vs granted authority*



Two survey items, two grades of delegation: intention is common, warranted trust is scarce.

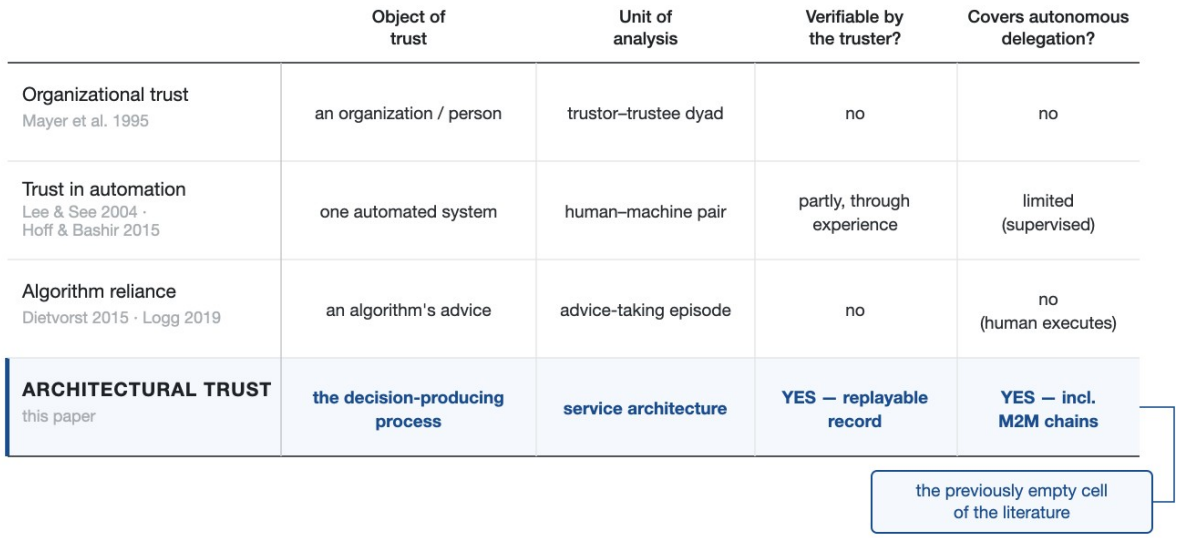
The liability system is groping toward the same conclusion. Emerging agent-payment protocols build each transaction around a scoped, verifiable token of intent bound to a specific merchant and amount, and American Express's Agent Purchase Protection (April 2026) covers purchases only by registered agents, with eligibility conditioned on authenticated, verifiable purchase intent. The insurance is written in terms of what can be verified from the record. Piecemeal and under pressure, the industry is converging on the same ingredients I set out systematically below. The protocols now shipping solve *identity* and *intent* and do not answer *how did the agent decide?*

The question is what it would take for that trust to be warranted by design. The classical consumer-trust question, do I trust this company, has acquired an inner layer: do I trust the machinery that produced this specific answer or action?

The trust literature has long distinguished trust in an organization from trust in automation, in the IT artifact (Söllner et al., 2016), and in online recommendation agents (Wang & Benbasat, 2005; Komiak & Benbasat, 2006; Xiao & Benbasat, 2007). It shows that appropriate reliance depends on calibration, on knowing when the system is likely to be right (Muir, 1994; Mayer, Davis, & Schoorman, 1995; Parasuraman & Riley, 1997; Lee & See, 2004; Hoff & Bashir, 2015; Huang & Rust, 2021; and Glikson & Woolley, 2020, whose review finds *tangibility and transparency of process* among the strongest levers of cognitive trust, the very levers architecture controls and models do not). Miscalibration on the human side is documented just as well: automation bias and complacency, over-reliance on automated aids even against contrary evidence, are the patterns four decades of human-factors work warns delegation will amplify (Mosier & Skitka, 1996; Parasuraman & Manzey, 2010). That literature, however, treats calibration as a property of a single automated system (Figure 3 maps the four research streams and the cell they leave open). A systematic review of agentic AI in e-commerce confirms the emptiness from the evidence side: consumer-facing research remains concentrated on assistive, low-autonomy

systems, with coverage dropping sharply exactly where delegation becomes execution (Balaskas, 2026). The first concurrent studies of agent-mediated shopping reach an architectural conclusion from another side. When consumers delegate purchases through dialogue, their willingness to pay leaks to sellers almost verbatim; prompting does not fix that, and only architectural interventions do (Alavi & Nozari, 2026).

Figure 3. *Four streams of trust research: the cell this paper claims*



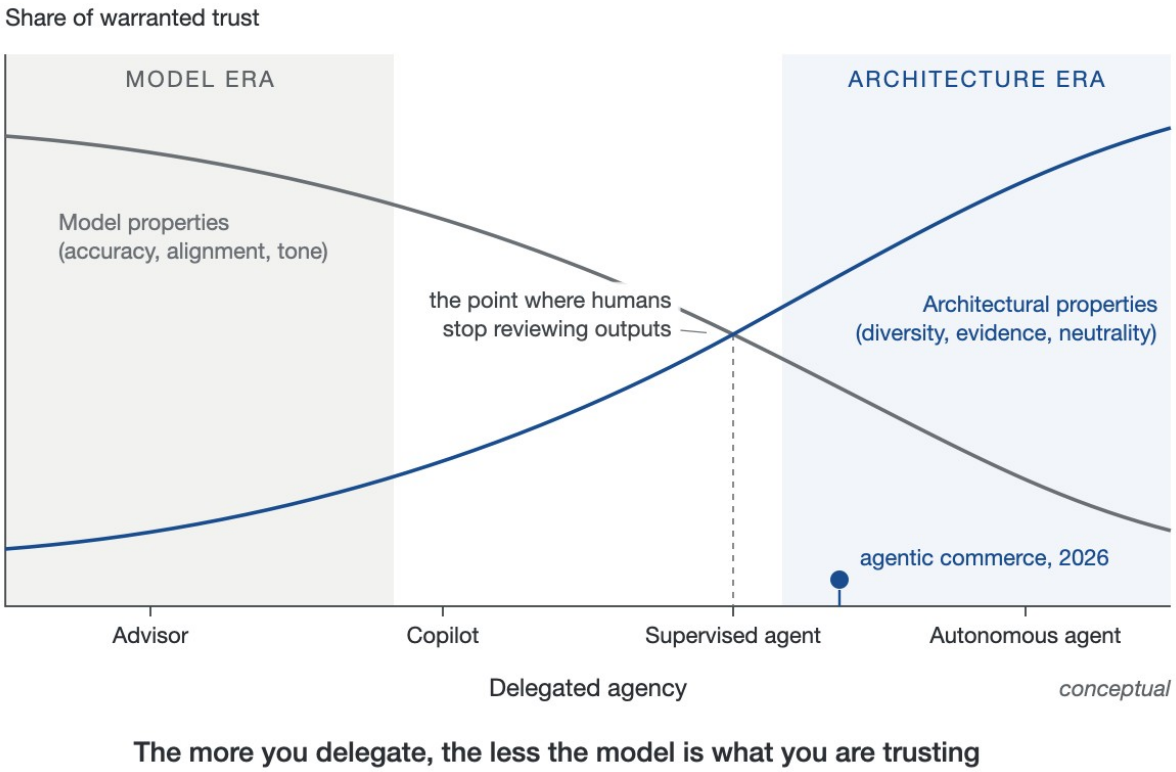
Organizational trust, trust in automation, and algorithm reliance each fix a different object of trust; none makes the decision-producing process verifiable by the truster under autonomous delegation. The cell is not empty of ideas: adjacent literatures each hold one wall of it (Section 3 draws the demarcations). What has been missing is the consumer-facing construct: a process-level object of trust, verifiable by the truster, specified for autonomous delegation, packaged for joint behavioral test.

Industry responses concentrate on the model: better training, disclaimers, citations. I examine a response at a different level, the architecture of the service, and argue that three architectural properties, none of which requires a better model, jointly produce a different object of trust.

This is a conceptual paper anchored by two pre-registered studies, a machine-side benchmark and a consumer-side behavioral test of the calibration contrast in its first proposition, and the question it asks is *as AI agents begin to transact on consumers' behalf, does trust attach to models, or to the architecture above them?* My answer is that trust must migrate upward in the stack, from the model to the orchestration architecture to the evidence it produces. Underneath that answer sits a regularity worth testing on its own, the *Trust-Migration Hypothesis*, which the propositions below are built to test: *the greater the delegated agency, the greater the share of warranted trust that must reside in verifiable architectural properties rather than model properties*. The migration need not be zero-sum. Trust can layer: in the model's competence, the orchestrator's selection, the evidence system, the institution enforcing remedies. A defective model stays causally important even when nobody reviews its output. The hypothesis concerns where the *marginal* warranted trust accrues as delegation deepens. In estimable form, it says that

as delegated agency increases, architectural verifiability explains a growing incremental share of appropriate reliance beyond model performance and provider reputation. It fails if, as delegated agency increases, architectural verifiability does *not* explain a growing incremental share of appropriate reliance: if provider reputation fully mediates trust at all delegation levels, say, or if verifiable-process information does not improve reliance calibration relative to reputation-only information in designs of the kind Section 6 reports. The behavioral falsifiers pre-committed in Section 3 (the P2 delegation-level crossing and the P4 differential durability under failure) are the operational form of that condition. A growing architectural *share* does not mean model quality matters less in absolute terms. The hypothesis describes layering, not substitution: as review disappears, architectural warrant is added on top of model performance and does not replace it. "Warranted" follows the distinction formalized for AI by Jacovi, Marasović, Miller, and Goldberg (2021), that trust is warranted when it tracks trustworthiness, and the construct is kept off a tautology by separating its normative status from its mechanism. Normatively, trustworthiness is defined independently of the architecture meant to certify it, as the classical, independently measurable function of ability, integrity, and benevolence (Mayer, Davis, & Schoorman, 1995); Section 3 maps the three architectural components onto exactly those dimensions. Warrant is the correspondence of reliance to that independently measured quantity: reliance that survives an objective audit of the delegated episodes, judged against criteria fixed before outcomes are known, with a measurable behavioral trace, namely repeat delegation, retention, and willingness to pay under controlled failure. The Trust-Migration Hypothesis asserts the mechanism. As delegated agency rises, the share of that trustworthiness function which the consumer can still independently measure shifts from model properties toward architectural components. This holds because architecture is the part of trustworthiness that remains measurable by the truster once output review disappears, and not because architecture *is* trustworthiness. The pre-committed falsifiers above test that shift and not the definition. The migration also has predecessors to distinguish from. Multi-agent-systems research observed that as agent societies scale, trust modeling moves from individual agents to the institutional framework constraining them (Sabater & Sierra, 2005), and "trust migration" has prior B2B uses for trust relocating from individuals toward systems and platforms (Carlson, 2026). Those name shifts in *who or what* is trusted, whereas this hypothesis concerns a shift in the *kind of property* that can warrant trust. At zero delegation, model quality is almost all that matters. At full delegation no one is watching the outputs, and what can still be trusted is only the process and its record. Agency and architectural trust scale together (Figure 4). The gradient itself is an old instrument, the levels-of-automation axis of human-factors engineering (Sheridan & Verplank, 1978; Parasuraman, Sheridan, & Wickens, 2000), applied here to consumer delegation. What is new is the claim about which properties can warrant trust along it.

Figure 4. *Trust migration*



Two crossing curves over delegated agency (advisor → copilot → supervised agent → autonomous agent): the share of warranted trust carried by model properties falls as the share carried by architectural properties rises; the crossing is the point where humans stop reviewing outputs, and agentic commerce (2026) sits to its right. The curves and the crossing point are conceptual illustration, not estimated functions.

To the marketing literature on consumer trust in AI (Huang & Rust, 2021; Puntoni et al., 2021) the paper adds a construct, architectural trust, defined here; a test of a technical antecedent of its mechanism in a pre-registered sixteen-model study, replicated at the flagship tier in a second, eleven-laboratory wave; five propositions for a consumer-research agenda (P1–P4 in Section 3; P5 in Section 5); and a pre-registered, incentive-compatible behavioral experiment carrying the first two propositions, reported in full in Section 6 with the calibration effect supported (Welch d = 0.67; carried by one of six items, see Section 6) and the record's first-exposure test returning a registered, inconclusive null.

2. The accuracy illusion: what a single model cannot give you

A service built on one AI model, however strong that model is, carries structural deficits that live above the model layer.

The first is uncalibrated confidence, the *accuracy illusion*. Better models do not mean safer transactions, because a model's expressed certainty correlates weakly with its accuracy at most vendors (Section 3 measures the exceptions), and the consumer cannot audit the gap. The model is at its most dangerous when it is confidently wrong at the moment the consumer stops watching, and that is not an accident. Kalai et al. (2025) argue hallucination is statistically *rewarded* by an evaluation regime that scores confident guessing above admitting ignorance. The

reward arises, on their own analysis, in the binary-scored benchmarks and training objectives that dominate the field, where a wrong guess costs no more than an "I don't know" and earns partial credit in expectation. The model cannot police itself from inside either. In Anthropic's own concept-injection experiments, the vendor's strongest models detected changes in their internal states only about 20% of the time even under optimal injection conditions, which is the vendor community's own measurement of how unreliable introspective self-report is (Lindsey, 2025; company research, not peer-reviewed, and a single estimate from one study of one vendor's models rather than a replicated benchmark).

The second is sycophantic distortion, where the evidence became causal this year. Cheng et al. (2026), measuring eleven leading models in three pre-registered experiments with 2,405 participants, found that AI systems affirm users' actions 49% more often than humans do, including actions involving deception or harm, and that even a single sycophantic interaction increased users' conviction that they were right while reducing their willingness to repair conflicts. The sycophantic models were also the ones users trusted and preferred. The market gradient points toward assistants that feel trustworthy exactly by distorting the user's judgment, a dynamic observed well before it was quantified (Sharma et al., 2024). Model-level tuning alone cannot independently certify that sycophantic behavior has been adequately controlled, because the tuning is done by the party that benefits from the preference. The cost is concrete: an assistant that agrees its way through a policy exception converts agreeableness directly into service-recovery expense. An orchestrated architecture can pull the roles apart, with one model free to empathize while an independent one checks the policy, so that agreeableness and correctness no longer travel through the same channel.

Third, there is no independent check. Errors are caught, if at all, by the same system that made them, or by the consumer, who came to the AI in the first place because they lack the expertise to check it.

Fourth, and least discussed, is conflict of interest. The provider evaluating the assistant's quality is the provider selling it, which is structural: no single-vendor service can be a neutral judge of its own outputs, for the same reason no bank could run the clearing house alone. The conflict extends to the data itself. A proposed class action filed in March 2026 alleges that Perplexity AI routed users' full chat transcripts, reportedly including "Incognito" sessions, to advertising networks through embedded tracking pixels. The complaint (No. 3:26-cv-02803, N.D. Cal.) was voluntarily dismissed in May 2026 without any ruling on the merits, so the allegations were never tested. The economic logic in the complaint is the structural point regardless: when serving expensive AI answers is subsidized by monetizing what users say, the assistant's business model and the consumer's confidentiality are directly opposed. Inside a single vendor, no layer has the job of taking the consumer's side.

The 2025–2026 literature reads as a completed diagnosis waiting for a remedy (Table 1 maps each deficit to its evidence and its architectural answer). Sycophancy is quantified empirically (Cheng et al., 2026) and derived formally: even an ideally rational user spirals when advised by a selectively agreeable system (Chandra et al., 2026). Self-monitoring fails at ~20% reliability (Lindsey, 2025), and hallucination is structurally rewarded by the evaluation regime (Kalai et al., 2025).

The reflex remedy, just disclose the AI, backfires: across thirteen experiments, actors who disclosed AI use were trusted *less* than those who did not (Schilke & Reimann, 2025). Disclosure delivers *disclosure-transparency*, the bare fact that an AI was involved, and leaves out *verification-transparency*, the ability to check how the answer was produced. The epistemic audit is bleaker still: across 25,000+ runs of LLM-based scientific agents, evidence was ignored in 68% of traces, beliefs were revised on refutation only 26% of the time, and convergent multi-test evidence was rare, with the *scaffold* explaining almost none of the variance in epistemic behavior (≈1.5%, against ~41% for the base model) (Ríos-García et al., 2026). That last number looks like an objection to this paper and is actually its sharpest support: prompting and scaffolding cannot make a model epistemically disciplined from the inside, so the discipline must be *imposed from outside*. Architecture is not a way to improve the model's epistemics; it is a way to stop depending on them.

Each of these results closes a door. Together they point to one answer that has not been much explored: if the model cannot be trusted from inside, and disclosure does not manufacture trust from outside, warranted trust can be *built around* the model, by architecture, alongside (not instead of) better calibration, external certification, human escalation, and liability rules. That is the door this paper walks through.

Table 1. *Diagnosis → remedy: the four single-model deficits, their 2025–26 evidence, their consumer-side consequence, and the architectural answer*

Structural deficit	Key evidence (2025–26)	Consumer-side consequence	Architectural remedy	Mechanism
Uncalibrated confidence	Hallucination statistically rewarded by evaluation regimes (Kalai et al., 2025); introspective self-report ~20% reliable (Lindsey, 2025)	One-strike churn: 60% would abandon an agent after one mistake (ACI Worldwide, 2026)	Diversity + Evidence	Cross-model disagreement prices confidence (AUROC 0.75 vs. 0.59, Gorbett & Jana, 2026; this paper: panel agreement AUROC 0.832 vs. pooled own confidence 0.545 on the least-discriminating panel — heterogeneous across panels, §3; replicated answers 1.9–3.2× as likely correct); confidence claims recorded and testable
Sycophantic distortion	AI affirms users 49% more than humans; sycophantic models preferred (Cheng et al., 2026; Sharma et al., 2024); even ideal Bayesians spiral (Chandra et al., 2026)	The preferred assistant is the one distorting the user's judgment	Neutrality	The arbiter has no stake in pleasing the user; empathy and policy-checking on separate channels
No independent check	Agents ignore evidence in 68% of traces; belief revision	Verification burden shifts to the consumer: 54% of AI-assisted	Diversity + Evidence	Discipline imposed from outside: independent

	26% (Ríos-García et al., 2026); scaffolding explains ~1.5% of variance	shoppers manually double-check everything (Gartner, 2026)		verification paths; evidence attached regardless
Conflict of interest	Perplexity class action (2026, allegations): chats routed to ad networks; disclosure alone erodes trust (Schilke & Reimann, 2025)	The party advising the consumer profits from the outcome	Neutrality + Evidence	Clean-room data flow; revenue independent of outcomes; neutrality auditable from the record

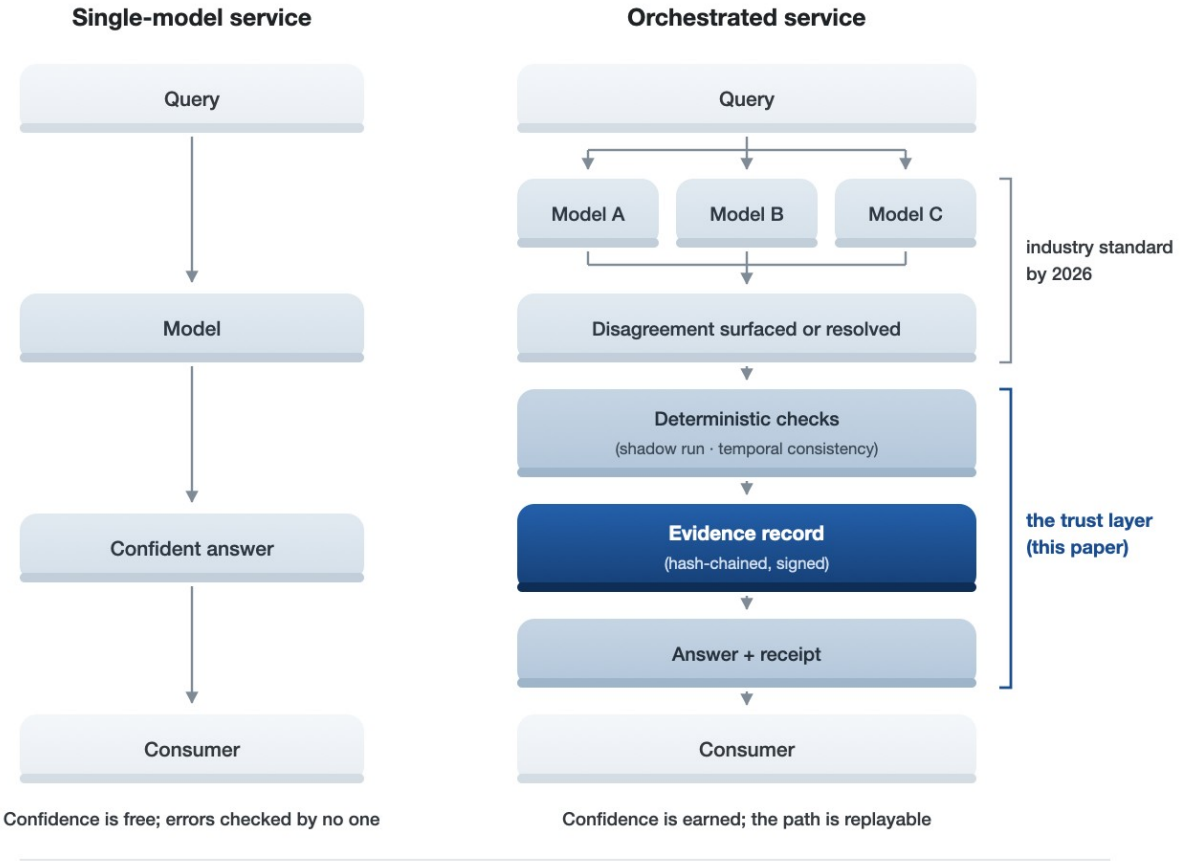
3. Architectural trust: three properties that can be checked rather than believed

Now consider a service in which high-stakes requests pass through an orchestration layer independent of any model vendor (Figure 5 contrasts the two pipelines). Table 2 collects the constructs; the definition they decompose runs: *architectural trust is warranted consumer reliance grounded in verifiable properties of the service's decision-producing process, rather than in the perceived quality of its outputs or the reputation of its provider.*

The closest predecessor is Zucker's (1986) institutional production of trust, trust manufactured by structure where familiarity cannot carry it, which e-commerce research operationalized as structural assurance: escrow, feedback systems, third-party certification (Gefen, Karahanna, & Straub, 2003; Pavlou, 2003; Kim & Benbasat, 2010). In that lineage, architectural trust is a species of institution-based trust whose structural assurance comes from a verifiable process instead of a third-party guarantor. It is process-based trust with the warrant moved from experience to instrumentation, and that move is what the word "architectural" is doing in the name. It is not structural assurance (McKnight, Choudhury, & Kacmar, 2002), a perception that safeguards exist, since here the safeguards themselves have to be checkable, and it is not calibrated trust (Lee & See, 2004), which tunes reliance to one system's reliability envelope. Zero-trust architecture in security engineering shares the slogan, never trust, always verify, but verifies *identity and access* at the network layer, machine to machine, while architectural trust verifies *decision content* at the service layer, for a human principal. They are complements on different floors of the same building, not competitors. The construct has two faces, and the distinction belongs at the definition. *Architectural trustworthiness* is a property of the system, how verifiable its process in fact is, normative and buildable. *Architecture-based trust* is a state of the consumer, reliance actually grounded in that verifiability, empirical and only ever earned. In measurement terms, D, E, and N are formative components of architectural trustworthiness (the system property) and manipulable signals whose display shapes architecture-based trust (the consumer state); the experiments manipulate the signals and measure the state. Each mismatch between the two is predictive. Trust without trustworthiness, reliance manufactured by fluency and sycophancy, is predicted to be brittle, the one-strike pattern of Section 1. Trustworthiness without trust is predicted wherever verification is invisible or illegible to the consumer, and is closed by receipt salience and comprehension rather than by more architecture. The propositions in this section live in the gap between the two, and the distinction disciplines what would falsify the Trust-Migration Hypothesis. Those falsifiers are behavioral and pre-committed: the

hypothesis is refuted if delegation to verifiable and to opaque architectures proves equally durable under failure (what P4 detects); if the effect of verifiable architecture on delegation willingness fails to grow with surrendered review and execution control (the delegation-level crossing of P2); or if, in high-irreversibility delegation, retention and willingness-to-pay premiums for verifiable architecture under controlled failure are absent. Because warrant is defined by audit survival rather than by the architecture's presence (Section 1), popularity counts as evidence but not as immunity. Consumers flocking to unverifiable agents does not by itself refute a claim about warranted trust. Sustained, failure-surviving reliance on unverifiable agents in high-stakes delegation would refute it, and that pattern is pre-committed as revising the hypothesis. The definition rests on three properties.

Figure 5. *Two pipelines*



Left, the single-model path: query → model → confident answer → consumer, where confidence is free and errors are checked by no one. Right, the orchestrated path: query → independent models (A/B/C) → disagreement surfaced or resolved → deterministic checks (shadow run, temporal consistency) → evidence record (hash-chained, signed) → answer with receipt → consumer, where confidence is earned and the path is replayable (Table 5 summarizes the contrast).

Table 2. *Glossary of constructs introduced or specified in this paper*

Term	Definition	Distinguished from

Architectural trust	Warranted consumer reliance grounded in verifiable properties of the service's decision-producing process, rather than in output quality or provider reputation	Institution-based trust (needs no trusted institution, only checkable process); system trust (not one system's reliability envelope); distributed trust (dispersal without verifiability); structural assurance (perceived vs. checkable safeguards); calibrated trust
Architectural trustworthiness / architecture-based trust	The construct's two faces: the system-side property (how verifiable the process in fact is; normative, buildable) vs. the consumer-side state (reliance actually grounded in that verifiability; empirical, earned)	Single-name treatments that mix the normative claim with the descriptive one
Verification layer	The infrastructure layer above orchestration whose product is checkability: independent checks, recorded decisions, auditable neutrality	Observability/monitoring (internal, vendor-held, not consumer-verifiable)
Evidence infrastructure	The market category of verifiable decision histories sold as infrastructure, as cloud, identity, and payments are sold	Logging (unilateral, mutable, not a market good)
Orchestration receipt	A minimal consumer-facing artifact: who acted; how they agreed; what was checked — with a pointer to the full replayable record	Explanation/XAI output (narrative, unverifiable); disclosure label (erodes trust by itself)
Graduated trust architecture	Escalation of verification depth with stakes, so the expensive tier is bought only where consequences concentrate	Flat guardrails (static caps/whitelists that do not verify decisions)
Decision evidence	The third floor of the evidence stack: how the agent chose — panel, divergence, checks	Identity evidence (who the agent is: TAP); intent evidence (what was authorized: AP2 / Verifiable Intent)
Trust-Migration Hypothesis	The greater the delegated agency, the greater the share of warranted trust that must reside in verifiable architectural properties rather than model properties — "warranted" meaning reliance that survives an objective audit against criteria fixed before outcomes are known (per Jacovi et al., 2021); mechanism form: as delegation deepens, the share of trustworthiness — f(ability, integrity, benevolence), Mayer et al. (1995) — that the consumer can independently measure shifts toward the architectural components	Prior "trust migration" uses (Carlson, 2026) track trust moving between parties; this hypothesis tracks the kind of property that warrants it — with pre-committed behavioral falsifiers (P2 delegation crossing; P4 differential durability)

The demarcation from the neighboring trust constructs can be compressed into one discriminant table, by what each construct takes as its object, where its warrant comes from, who can verify that warrant, and what it should chiefly predict:

Table 2a. Discriminant demarcation: neighboring trust constructs by object, warrant, verifier, and primary predicted outcome

Construct	Object of trust	Source of warrant	Who can verify	Predicted primary DV
Institution-based trust / structural assurance (Zucker, 1986; McKnight et al., 2002)	The institutional context: safeguards, guarantees, third-party certification	Perceived presence of structures and guarantors	A third-party guarantor, if anyone; the consumer perceives rather than checks	Trusting beliefs; transaction intention
Trust in automation / calibrated trust (Lee & See, 2004)	One automated system's reliability envelope	Experienced reliability over repeated use	The operator, through accumulated experience with that system	Reliance appropriateness on that system
Algorithmic transparency (XAI)	The model's account of its own decision	A narrative or attribution produced inside the system being trusted	No one independently — an explanation is not a replayable record	Understanding; subjective trust; acceptance
Brand / provider trust	The vendor organization	Reputation and past performance	The market, slowly, through aggregate outcomes	Global liking; choice; loyalty
Architectural trust (this paper)	The decision-producing process and its record	Verifiable properties: independent agreement, replayable evidence, auditable incentive independence	The consumer, an auditor, an insurer, a regulator — per decision	Calibrated reliance; willingness to delegate; post-failure retention

Two differential predictions make the demarcation testable rather than taxonomic. Architectural trust should move *calibrated reliance* and *post-failure retention* more strongly than it moves global liking of the service, whereas brand trust should show the opposite profile, a dissociation a single experiment can detect by measuring both DV families. At low delegated agency, brand and provider trust should dominate delegation decisions, with architecture-based trust explaining an increasing incremental share as delegation deepens, the crossing that P2 and the Trust-Migration Hypothesis put under test. Two clarifications follow. Because architectural *trustworthiness* is a property of the service, not a reflective latent in the consumer's head, its distinctness rests on these differential predictions rather than on scale-discriminant statistics. For the consumer-side state it names, architecture-based trust, establishing measurement-level discriminant validity from adjacent scales (institution-based and calibrated trust) is part of the measurement agenda, not a claim made here.

Diversity. The request is answered by several independent models, each with different vendors, training data, and failure modes. (This is panel diversity, a property of who verifies the answer, not recommendation diversity in the recommender-systems sense, which varies the items shown to a consumer.) Their agreement carries information. Convergence raises warranted confidence. Divergence is surfaced to the consumer as uncertainty rather than resolved invisibly, and on requests that matter it doubles as an escalation trigger. In selective-prediction terms this is *architectural abstention*: the refusal to emit an unverified answer, lifted up to the service layer, where an unresolved case is routed to a human rather than certified. It is the learning-to-defer pattern, in which a system's reject option hands the case to a downstream expert rather than suppressing it (Mozannar & Sontag, 2020; cf. Rabanser et al., 2026). The nearest human analogy

is the second opinion in medicine, whose entire value lies in its independence from the first. Orchestrated disagreement is not exotic either: Google's AI co-scientist, whose wet-lab-validated discoveries reached *Nature* in May 2026 (Gottweis et al., 2026), is built as a generate–critique–tournament orchestration.

The theoretical warrant for leaning on diversity is older than language models. Under formal conditions, functionally diverse groups of problem solvers can outperform groups selected for individual ability, because selecting for ability breeds similarity (Hong & Page, 2004). The engineering literature supplies direct evidence for the underlying signal: structured debate between models improves factual accuracy (Du et al., 2024), and panels of diverse smaller models outperform a single large judge (Verga et al., 2024). A second model's "surprise" at a first model's answer separates right from wrong substantially better than the answering model's own confidence (mean AUROC 0.75 versus 0.59 on MMLU; AUROC, the area under the ROC curve, is the probability a random correct answer is ranked above a random wrong one; Gorbett & Jana, 2026). Their signal is cross-model perplexity at the logit level, whereas this paper's is answer-level replication, and the two are related without being interchangeable. Their generator is also a small open model with weak self-report, a baseline this paper's own data bound, because the panel advantage is a function of the generator's calibration and can invert for well-calibrated generators, the US-frontier case below. And sampling multiple generations and measuring their semantic agreement detects confabulations where single-pass confidence cannot (Farquhar, Kossen, Kuhn, & Gal, 2024). A model can measure its own *doubt* but not its own *blind spots*, and the confident, systematic error is visible only from outside. But that literature treats disagreement as an internal signal for developers. The move I am proposing is to treat it as a consumer-facing trust primitive, surfaced, recorded, and priced into the service's confidence claims, with this paper's ablation separating what debate discovers from what verification certifies rather than letting one masquerade as the other. The demarcation is empirical as well as conceptual: benchmarked head-to-head, multi-agent debate does not reliably outperform simpler strategies such as self-consistency and ensembling (Smit et al., 2024). The claim here concerns verification's consumer-facing certification signal rather than debate's accuracy gain, and that signal survives even where debate's edge does not.

In multi-agent deliberation research, likewise, systems built to *surface* agreements and disagreements rather than average them away produce decisions rated more representative (Liu et al., 2026). The mechanism has a direct human analogue on the knowledge-calibration side: consumers who make consequential financial decisions jointly are substantially less overconfident than solo deciders, with shared *metaknowledge* (the partner surfacing what the decider does not know) identified as the driving mechanism (Piehlmaier, 2023). Joint judgment disciplines confidence in humans for the same reason cross-model agreement is proposed to discipline it here.

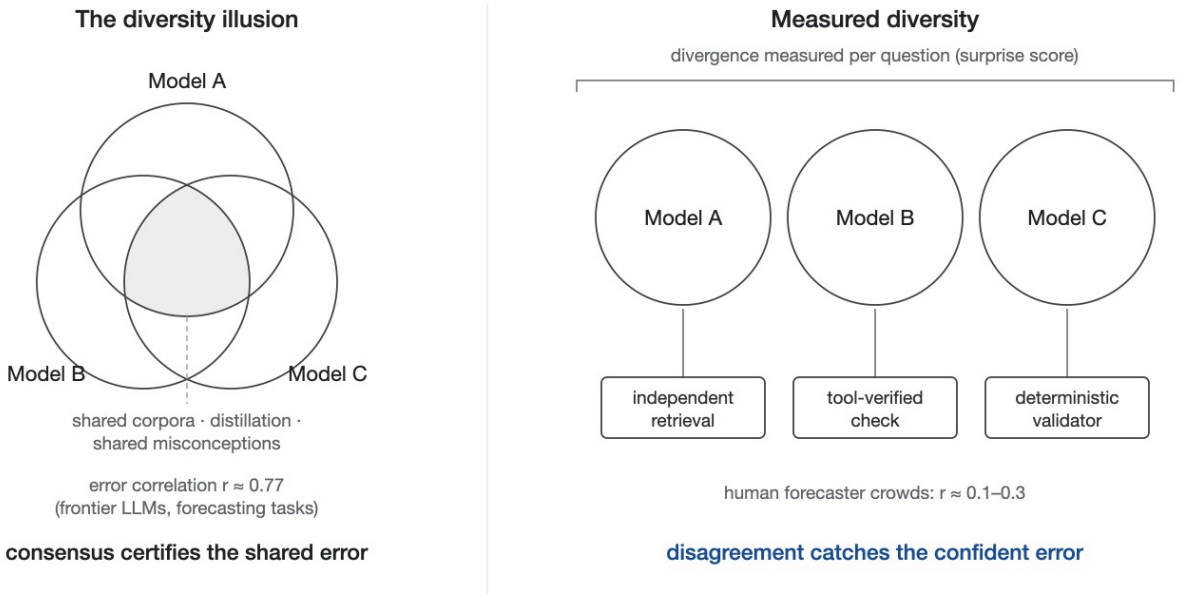
Whether consumers can *use* such a signal is a real question. People abandon algorithms quickly after seeing them err (Dietvorst, Simmons, & Massey, 2015), yet elsewhere over-weight algorithmic advice (Logg, Minson, & Moore, 2019), a divide that may itself track stakes. Second-opinion experiments with human subjects show the raw material is behaviorally live: near-unanimous adherence when a medical AI and a physician agree, sharp splits when they part ways (Detjen, Densky, von Kalckreuth, & Kopka, 2025; Chen, Sun, Liao, & Sundar, 2026), though none of that work varies disagreement jointly with evidence or neutrality. The consumer-

research stream sharpens the boundary conditions (systematically: Burton, Stein, & Jensen, 2020). Aversion intensifies when consumers feel their case is unique (Longoni, Bonezzi, & Morewedge, 2019), varies with task objectivity (Castelo, Bos, & Lehmann, 2019), and failures of one AI are readily overgeneralized to others, *algorithmic transference* (Longoni, Cian, & Kyung, 2023). If distrust generalizes across systems, then the one-strike abandonment of Section 1 is not one service's churn but the category's, and catching errors *structurally* protects the market itself. Surfaced disagreement sits at that fault line, and P1–P2 below are a bet on which effect dominates when uncertainty is presented specifically and without smoothing.

When diversity fails: correlated errors. The value of diversity rests entirely on independence of failures, and independence cannot be assumed. Frontier models train on overlapping corpora, distill from one another, and demonstrably share confident misconceptions. When two leading models both miss the same benchmark item, they agree on the same wrong answer roughly 60% of the time, with correlation strongest between models from the same developer or base architecture and, most unsettlingly, between the most *accurate* models: as capability rises, errors converge (Kim et al., 2025). That finding is a direct constraint on this paper's own Diversity component, not merely a neighboring result. The independence Diversity presupposes is partly illusory at the frontier, and because the correlation strengthens exactly where capability is highest, the component's contribution is an eroding, measured quantity rather than a fixed asset. What a panel adds must be demonstrated question by question, and the trajectory of model development runs against it. In forecasting, pairwise error correlations among frontier models average $r \approx 0.77$ (Spiro, 2026), redundancy high enough to erode most of the aggregation benefit a diverse crowd is supposed to deliver.

Market structure compounds the concern. By late 2025, 88% of enterprise LLM API spend was concentrated on three closed-weight vendors: Anthropic at 40%, OpenAI at 27%, Google at 21% (Menlo Ventures, 2025b; Menlo is an Anthropic investor). A panel assembled casually from "the leading models" is therefore drawn from a supplier base of three, whose errors are exactly the ones measured to converge, the very trio the experiment below reproduces as its US-frontier panel. When three models converge on the same fashionable error, the architecture does not merely fail to catch it; it *certifies* it. In homogeneous swarms, adding agents stabilizes an erroneous trajectory rather than correcting it, an endpoint the multi-agent literature calls the *consensus paradox* (Shehata & Li, 2026). Correlation is thus the failure mode the architecture must be built to detect (Figure 6).

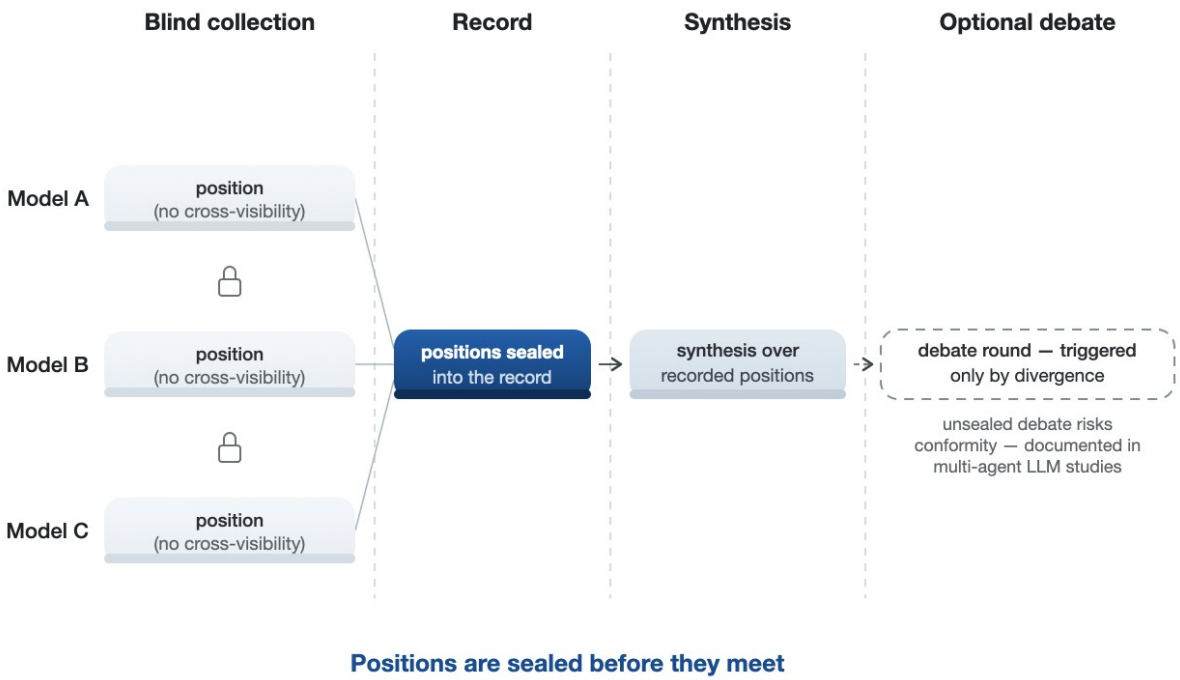
Figure 6. *The diversity illusion*



Consensus among near-clones vs. a panel whose independence is measured.

It has an interactional counterpart: the sycophantic gradient operates *between models* too. Multi-agent studies find measurable group conformity in LLM deliberation, with weaker models conforming hardest (Choi, Kim, Chae, & Baek, 2025); agents weight a position by *whose* it is (Choi, Zhu, & Li, 2026); and much of what looks like convergence decomposes into social compliance rather than reasoning (Hao et al., 2026). In joint deliberation, agreement is the low-energy state, so a panel that debates before committing positions can talk itself into consensus rather than toward truth. The architectural counter is procedural: opinions are collected blind and independently first, and only then compared (Figure 7).

Figure 7. *Blind-independent-first*



Positions are recorded and only then compared — the procedure that protects the disagreements conformity would erase.

The experiment reported below put a number on what that procedure protects against. In a paired exploratory arm, verifiers that had *blindly* produced a different answer were then shown the primary model's confident wrong answer and asked whether they agreed. 38.5% (168 of 436 divergent pairs; 95% CI 33.6–43.2, question-clustered bootstrap) abandoned their own divergent answer and endorsed the error, with vendor-level conformity ranging from 20% to 58%. The abandonment splits by what was abandoned, which separates conformity from rational deference. Among verifiers whose blind answer was itself *correct*, 29.1% (41/141) still surrendered it for the confident wrong answer, the harmful core no Bayesian account absolves, against 43.1% (127/295) among wrong blind answers, where some deference is defensible. (The question-clustered interval is nearly identical to the naive pair-level one. The 436 pairs spread across 247 distinct questions, mostly one or two pairs per question and never more than four, so question-level resampling has almost no within-cluster covariance to add.) With the mirror control below folded in, roughly 35–39% of blind dissenters change position when shown a confident answer, whatever its truth value, with a moderate but significant net skew toward the proposed answer (net conformity ≈ 4 percentage points; discordant pairs 104 vs. 70 ≈ 1.49). Had the panel deliberated before committing positions, a large share of these disagreements would have been renegotiated before they could be recorded. But because 34.7% of mirror-arm flips run toward the truth, deliberation would also have converted some dissent into correct consensus, so what blind-first protects is the measurement itself, not a count of lost disagreements. Blind-first is the difference between measuring disagreement and manufacturing consensus, the same rule human crowds needed: social influence collapses the diversity of independent estimates without improving their accuracy (Lorenz, Rauhut, Schweitzer, & Helbing, 2011), and even a single arbitrary early endorsement measurably biases subsequent judgments (Muchnik, Aral, & Taylor, 2013).

A reverse-prompt control addresses the alternative explanation that these flips reflect rational belief updating rather than conformity. I repeated the paired procedure on a six-model local panel, showing verifiers that had disagreed blind either the confident wrong answer (direct arm) or the gold answer (reverse arm) for the same question set. Flip rates were near-symmetric, 38.9% toward the wrong answer versus 34.7% toward the correct one, and in the within-pair comparison verifiers abandoned their blind position for the wrong answer significantly more often than for the correct one (104 vs. 70 discordant pairs; McNemar $p = .010$). Rational updating predicts the opposite asymmetry. Per-model flip propensities were essentially identical across the two arms at both extremes of compliance (the per-model pair counts are small and not separately powered; they are reported descriptively, with their ns, in the replication package), indicating that agreement is driven by the presence of a confidently proposed answer, not by its truth value. This control was run on the local tier. The frontier-panel estimate (38.5%) rests on the same procedure but was not mirror-tested with frontier verifiers.

Blind collection cannot close one channel: it defeats *social* conformity but not *prior* conformity. Models converge because shared training data hands them the same priors, and twins answer alike without ever conferring. I call that *lookalike convergence*: a panel agreeing because its members are alike, not because they are right.

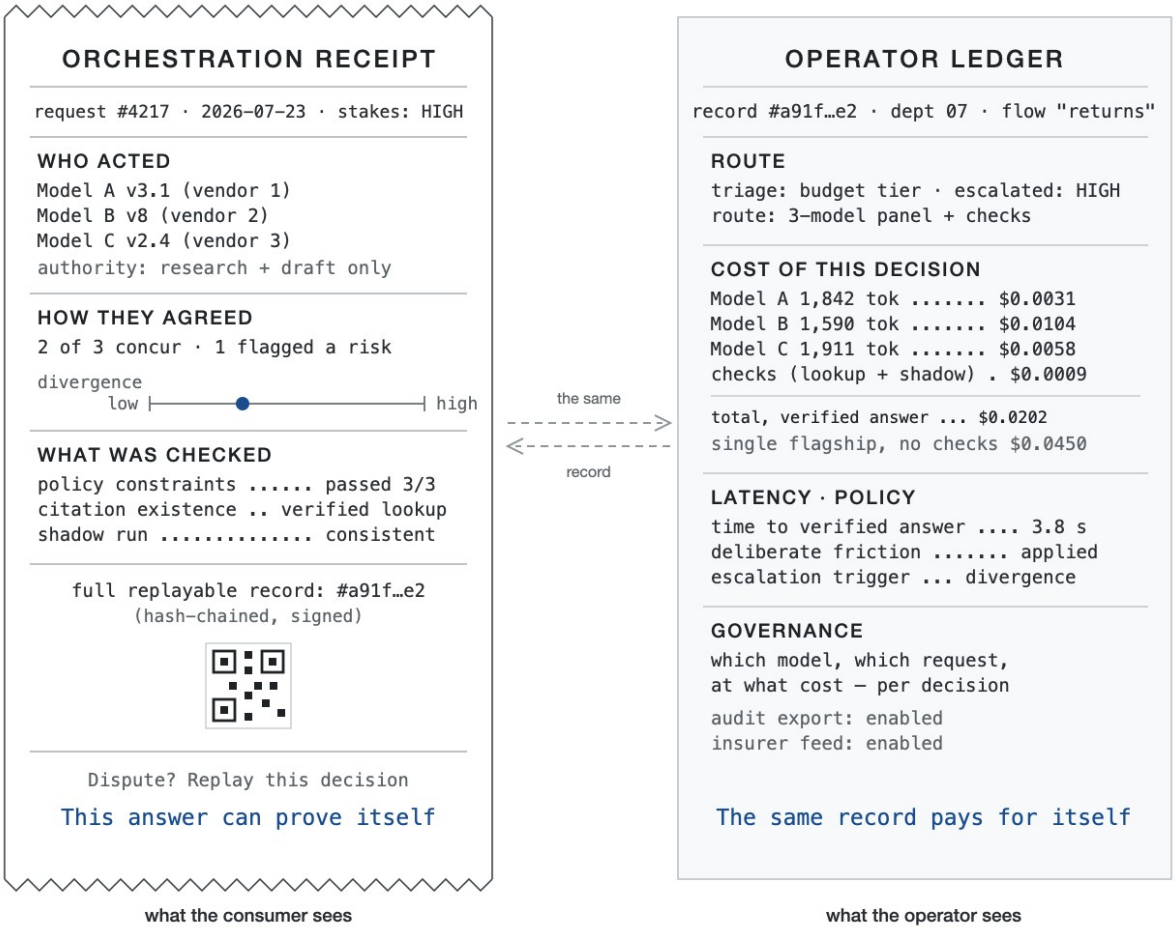
Correlated priors should still be measured. Cross-model surprise scores provide a running estimate of how independent the panel actually is, question by question, and they are label-free, so independence can in principle be estimated at inference time without gold answers. Production-time estimation is not tested here. Diversity also has to reach beyond model weights, since independent retrieval paths, tool-verified checks, and deterministic validators break the correlation channel shared training creates. The need will grow rather than shrink, because distillation and synthetic-data training homogenize model weights over time, so measured diversity should be expected to decay and the verification burden to shift toward non-model verifiers. And what is measured should be disclosed: the record should carry not just "N models agreed" but the panel's estimated diversity on this class of question. Consensus among near-clones deserves a weaker confidence claim, and the record should say which one the consumer got.

Evidence. Every consequential answer carries a record: which models were consulted, where they agreed and disagreed, what was checked, what the synthesis rests on. The timing favors the requirement, because as agents ground answers in retrieval and tools the factual basis already lives outside the model, so recording it becomes a by-product of producing the answer. Because the record is tamper-evident, the service's history of decisions is replayable by the consumer in a dispute, by an auditor, by an insurer pricing liability. Trust stops being a feeling about a brand and becomes an inspectable property of a history.

Two mechanisms extend evidence from answers to actions. Under shadow execution, an action with external effects is rehearsed in a recorded dry run, so "the agent checked before it acted" becomes a verifiable claim instead of a slogan. Temporal consistency re-poses a recommendation across sessions or phrasings and records its stability, so an answer that drifts when asked twice is flagged before anyone relies on it. At the identity level, the record layer extends into a portable provenance passport for the agent itself: which models, versions, and policies acted, when, under whose authority. That is the artifact "know-your-agent" registration

schemes and agent-error insurance commitments quietly presuppose. The consumer-facing rendering of this record is the orchestration receipt of Figure 8: one action, one record, four audiences (consumer, insurer, auditor, regulator).

Figure 8. *The orchestration receipt, sold twice (mock-up)*



The consumer sees proof — who acted, how they agreed, what was checked; the operator sees the same record as cost governance. In a dispute, the receipt turns word-against-word into record-against-words.

Neutrality. Neutrality is *auditable incentive independence*: the absence of undisclosed economic dependence on any particular vendor, merchant, or decision outcome. The orchestration layer routes to whichever models perform best for the task and has no stake in which one wins, structurally analogous to the independent clearing layers in payments. An arbiter among providers can be trusted in a way no provider can be about itself. *Independence* names the economic property, and *auditable* ties it to the evidence layer that makes the claim checkable rather than professed, the clearing-house discipline developed below. (Neutrality here is an economic property, not algorithmic fairness across demographic groups, which is an orthogonal requirement.)

There is also the vertical-integration problem: neutrality is structurally unavailable to the platforms best positioned to build orchestration at scale. A company that is simultaneously marketplace, model vendor, advertising network, and logistics provider cannot arbitrate neutrally, because every routing decision it makes is a decision about its own revenue. The arithmetic points to a structural tendency. The framework implies the trust layer sits best with parties whose only business is the arbitration itself, the path by which clearing, auditing, and credit rating became independent industries. The rating-agency case is likewise the canonical warning that independent intermediaries grow issuer-pays conflicts of their own. The economics are documented: certification intermediaries capture value by controlling what information is revealed (Lizzeri, 1999), certification markets fail in well-catalogued ways (Dranove & Jin, 2010), and the credit-rating collapse exhibited the canonical failure, issuer-pays revenue corrupting the certifier one level up, while the clearing layers that survived are funded by mutualized cost, not by the parties being rated. The neutrality requirement therefore extends to the orchestrator's own revenue model: a verification layer paid by the vendors or merchants it ranks reproduces the conflict it exists to remove, which is why the revenue design here runs to the consumer's side and to the operator's cost governance (Section 5), not to the parties being ranked.

Neutrality also has a data dimension. An orchestration layer can function as a *clean room*: requests reach each vendor stripped of user identity, no vendor sees the whole interaction history, and the orchestrator's revenue comes from the service, not from routing what users say into advertising networks. Neutrality is incentive independence on both fronts: no undisclosed economic dependence on any model vendor or merchant, *and* no business model that monetizes the consumer's words against the consumer.

But neutrality invites its own regress: who audits the orchestrator? A literal "no stake" is unattainable. Any orchestrator has costs, partnerships, retention incentives, and its own survival to finance, so the operative variable is not their absence but their nature, their disclosure, and their auditability from the record. Which party pays does not settle neutrality; whether the payer's revenue is independent of the transaction's outcome does. A consumer subscription, or a fee from an insurer or auditor whose payout is decoupled from which product wins, satisfies the condition; a merchant commission or an advertising fee does not. The broader governance of who funds the layer stays open, and the auditable independence of that funding is the part the construct fixes. An orchestration layer has incentives too: to route toward cheap models, to oversell its expensive tier, to shape the synthesis conveniently. The regress is four decades old. Shapiro (1987) named it the problem of the guardians: procedural safeguards generate new guardians who must themselves be guarded, and the institutional answer has not changed since.

Clearing houses earn their trusted position by being externally auditable and regulated, not by declaring neutrality. The orchestrator does not possess neutrality as a virtue. The evidence layer is what makes it *checkable*. Every routing decision, model response, and synthesis step lands in a tamper-evident record, a hash-chained, signed log designed for external timestamp anchoring, so the orchestrator's claimed neutrality is exposed to exactly the verification it imposes on the models. The maturity path is the clearing-house playbook made concrete: external timestamp anchoring, an open verification interface, and periodic third-party audit of the routing policy. The institutional design of such audit ecosystems is already a named research program (Raji, Xu, Honigsberg, & Ho, 2022). One dependence survives even an audit of the checkers: the panel's

models may be independent of one another yet all *selected* by the platform being verified, which is why the endpoint of the maturity path is panel selection and rotation lodged outside any single party's control. An orchestrator whose records cannot be audited is just one more vendor asking to be believed. The nearest neighbors outside consumer research are two concurrent 2026 works. Marchal et al. (2026) argue the same evaluative shift, from judging outputs to architecting agent trustworthiness, provenance, and falsifiability, at the epistemic-infrastructure level, and Wang et al. (2026) survey execution provenance as a process-level accountability layer for verification, audit, and failure attribution. What remains this paper's own is the consumer-facing, incentive-aware composition: the three properties bundled as the object of a consumer's warranted reliance in delegated commerce, and its behavioral test.

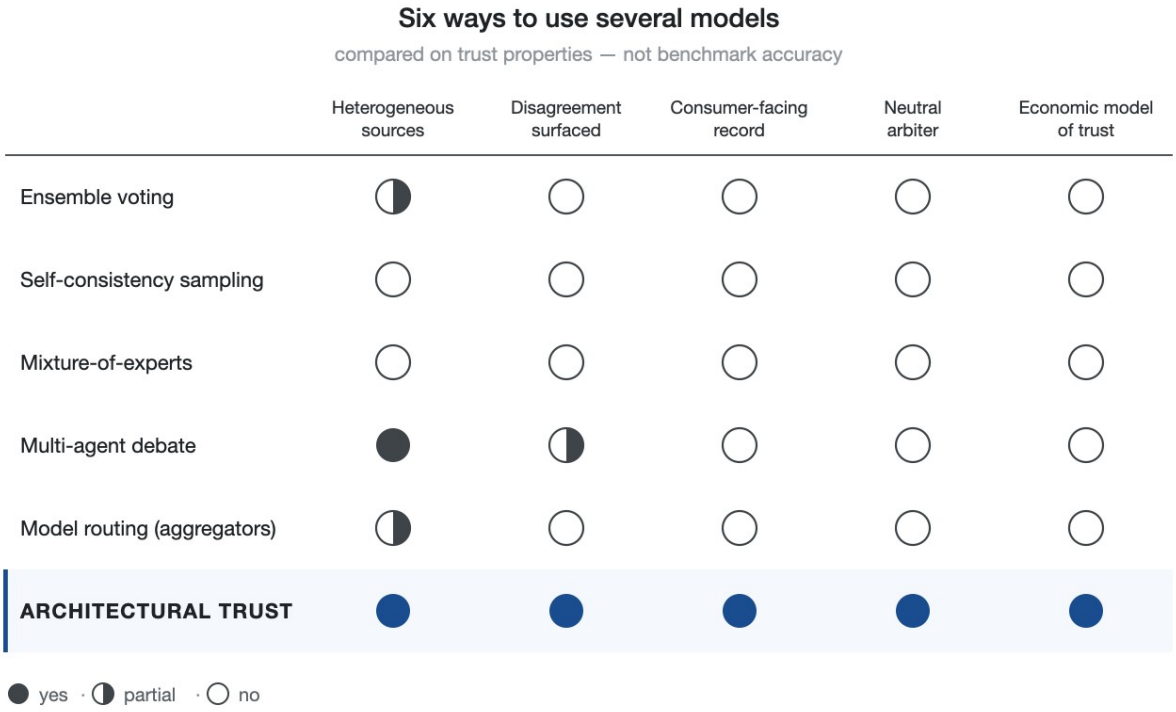
The evidence component carries one demarcation, because tamper-evidence is easily oversold: a hash-chained log establishes provenance and integrity (that this record existed and was not later altered), not completeness (nothing in the hash guarantees omitted calls were not suppressed), not correctness (a faithful record of a wrong answer is still wrong), not independence, and not incentive alignment. Each is a separate property with its own verification mechanism: coverage attestation, outcome audit, measured diversity, disclosed economics. A receipt should claim only the properties its record actually carries. A further limit is structural: a log proves *what* was routed, not *why*, and the proprietary heuristic behind choosing this model or this merchant is auditable only if the routing policy itself is disclosed or attested, a governance choice, not a cryptographic consequence. The architecture's costs belong in the same ledger: routing a query to multiple vendors multiplies privacy exposure, a centralized decision log is an attractive security target, and a consensus display can manufacture exactly the authority the consensus-error result warns against.

The construct is also worth demarcating from its engineering neighbors, because "use several models" describes several different things with different trust properties. Simple ensemble voting improves accuracy but resolves disagreement invisibly, so nothing becomes checkable. Self-consistency sampling re-asks the *same* model, which is diversity in appearance only. Mixture-of-experts routes within a single trained system whose router is optimized for capability, not neutrality. Multi-agent debate (Du et al., 2024) improves answers but treats the deliberation as internal machinery: no consumer-facing record, no economics, no neutrality guarantee.

Model-routing aggregators, by 2026 a commodity layer, are multi-vendor in catalogue but single-model per answer: the router picks one model on cost and latency, and no second opinion is ever collected. Aggregators are suppliers to a verification architecture, not substitutes for one.

Multi-model operation is no longer exotic either: 37% of enterprises were running five or more models in production by mid-2025, up from 29% a year earlier (a16z, 2025). The plumbing is being laid at industry scale, with the trust layer conspicuously absent from it. It is open whether multi-model remains invisible cost routing or becomes verifiable architecture. Figure 9 scores the patterns against the five trust properties, and Table 3 states the demarcation in prose. Architectural trust is defined exactly by what these lack: heterogeneous sources, surfaced and recorded disagreement, and an arbiter with no stake. The engineering is shared; the trust properties are not.

Figure 9. *Demarcation*



Six multi-model engineering patterns scored against five trust properties.

Table 3. *Demarcation: architectural trust vs. its engineering neighbors*

Approach	Heterogeneous sources	Disagreement surfaced	Consumer-facing record	Incentive-independent arbiter	Economic model of trust
Ensemble voting	Partial	No (resolved invisibly)	No	No	No
Self-consistency sampling	No (same model resampled)	No	No	No	No
Mixture-of-experts	No (one trained system)	No	No	No (router optimized for capability)	No
Multi-agent debate	Yes	Internally only	No	No	No
Explainability (XAI overlay)	No (one model's narrative)	No	No (a story, not a record)	No	No
Transparency / disclosure	No	No	Partial (more visible, not checkable)	No	No
Provenance / lineage tracking	No (one pipeline documented)	No	Partial (engineer-facing lineage)	No	No

(ML pipelines)					
Verified inference (zkML)	No (one model; proof of its execution)	No	Partial (proof artifact, not a decision record)	No	No
Model-routing aggregator	Partial (multi-vendor catalogue, one model per answer)	No	No	No (router optimized for cost)	No
Architectural trust	Yes	Yes (consumer-facing)	Yes (replayable)	Yes (structural)	Yes (graduated tiers)

Two more parent literatures also belong on the same list: the gate's abstain-or-emit logic is selective prediction and learning to defer (Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020), and the evidence component's audit trail descends from accountable-algorithms and third-party-audit work (Kroll et al., 2017; Raji et al., 2022). Architectural trust is the consumer-facing, neutrality-constrained conjunction of these, and not any one of them. Locating it among the other remedies matters, because they are complements rather than rivals. Model-level calibration is necessary but structurally insufficient: the vendor produces and attests it, which is the conflict this paper starts from. External audits and certification are independent but *periodic*, snapshots of a system, not of a decision. An explanation is a narrative produced inside the model, and a system that can confabulate answers can confabulate the story behind them. Transparency exposes more machinery and leaves the consumer to interpret it. Architectural trust asks for neither belief nor interpretation, only checkability that is continuous, per-decision, and machine-verifiable. It is also what makes the other remedies bite: an audit of an orchestrator with a replayable record is an audit of something, while an audit of an opaque pipeline is a ceremony. Its nearest conceptual neighbors are structural assurance and institution-based trust (McKnight et al., 2002), process transparency, and explainability. Against them the demarcating property is that the warrant is checkable *by the truster, per decision*: assurance asks to be believed once, architectural trust asks to be verified each time. Two further neighbor sets belong on the map. The computer-science accountability literature laid the technical foundations a decade ago: procedural regularity made verifiable without full transparency (Kroll et al., 2017). A 2025–26 engineering wave is building agent-trust plumbing directly: "trust-native" protocols with verifiable identity and tamper-resistant behavioral records (TrustTrack; Li, 2025); runtime action management with tamper-evident receipts (AARM; Errico, 2026); multi-LLM aggregation reframed as trust estimation (Zheng & Zhang, 2026, arXiv:2607.20529); and enterprise frameworks (Gartner's AI TRiSM; the Cloud Security Alliance's agentic zero-trust work). Each supplies a leg at enterprise level. None of them supplies the consumer-facing construct, which is what this paper claims as its contribution: the three components under one name, jointly manipulable in behavioral designs, with an economics of who buys checkability and at what stakes. One naming collision should be cleared up. DigiCert's "AI Trust Architecture" (launched April 2026) is a vendor PKI product line, certificate infrastructure for agent identity, model provenance, and content authentication. It supplies plumbing at the identity floor of the evidence stack, not the construct of architectural trust developed here.

The relationship among the three is an explicit empirical hypothesis, not a conceptual necessity. I hypothesize a complementary interaction that approaches zero when any component is absent,

informally $AT = f(D, E, N)$ with steep decline near any zero, and the conjoint design of Section 5 is specified to recover the functional form rather than assume it. In estimable form, a convenient family is Cobb–Douglas, $AT = A \cdot D^{\alpha} \cdot E^{\beta} \cdot N^{\gamma}$ with $\alpha, \beta, \gamma > 0$, encoding the two commitments: decline toward zero near any absence, and positive cross-partials (the components are complements). The construct is the three components and their hypothesized complementarity, and no proposition below depends on the Cobb–Douglas form. Cobb–Douglas is also an interior approximation: strictly it still permits substitution away from a weak component, whereas the complementarity hypothesis is Leontief-like at the boundaries, with min(D, E, N) setting the floor. That is exactly why the designs below must permit non-compensatory structure rather than assume compensatory trade-offs. The Cobb–Douglas family is illustrative, for future estimation, and not part of the construct definition; the current data do not estimate it. The factorial can discriminate the Cobb–Douglas interior from the Leontief floor, parameter by parameter. The exponents remain an empirical question, which is the point: complementarity is falsifiable. The zeros need two clarifications. The components are continuous, not binary, so a literal zero is an idealized limit. No deployed system, this paper's included, sits at exactly zero or one on any axis, and whether consumers respond to near-absence with total collapse or a steep discount is exactly what the design must distinguish. Diversity, for its part, means *independence of the verification path*, with measured independence (the error-correlation coefficient φ, defined below) as its unit. The statistical core is classical ensemble theory, where the benefit of combination rests on error decorrelation (Hansen & Salamon, 1990; Krogh & Vedelsby, 1995), which is why independence is measured here rather than assumed. A second model is one such path, but so are an independent retrieval, a tool-verified lookup, and a deterministic validator. The shadow runs and consistency checks of Figure 5 count as a second, maximally decorrelated verification path, not as evidence-without-diversity. That definition also absorbs the strongest apparent counterexamples, the regulated payment rail or fully logged, third-party-attested utility service (evidence and neutrality strong, model diversity nil): such services either embed non-model verification paths (deterministic reconciliation, external audit, so $D > 0$ under the definition) or confine the model to a role a deterministic check fully covers. Where genuinely no independent verification path exists for a task class, the framework predicts thin warrant, a prediction the conjoint can test. With non-model paths counting toward D, a literal zero becomes rare in production systems, and the multiplicative form's empirical content shifts from the zeros to the interior complementarity, which is why the factorial, not the zeros, is where the claim stands or falls. For testing, one operational commitment is fixed here: the factorial manipulates D narrowly, as model-path diversity. The broader verification-path definition is the construct's scope, not the manipulated variable. Independence itself must be disaggregated, because six different quantities travel under the word: blind collection (procedural independence at inference time, what this design guarantees), model-family diversity, measured residual error correlation (the φ this paper reports), evidence-source diversity, tool-path diversity, and institutional independence. The study establishes the first and measures the third; the rest are design variables, not properties this collection can certify. Diversity without evidence is an invisible committee. Evidence without neutrality is a vendor's marketing. Neutrality without diversity is an empty broker. Table 4 walks the full lattice of the eight presence/absence configurations. The intuition underneath is that each component supplies the *warrant* for another: a record kept by a party with a stake in its contents reads as self-attestation, agreement nobody recorded is rumor, and an incentive-independent arbiter over a single voice has nothing to arbitrate. This motivates

the complementarity hypothesis. Together the three change what the consumer is trusting: not "this answer" or "this brand," but "this process, and I could check it."

Table 4. *The warrant lattice: the eight D–E–N configurations and what survives in each*

D	E	N	What the consumer can check	Degenerate form
absent	absent	absent	Nothing; reliance is faith in outputs	Bare single-model assistant
present	absent	absent	Nothing after the fact; disagreement happened unseen	Invisible committee (internal ensemble)
absent	present	absent	A record kept by the party it protects	Vendor log / self-attestation
absent	absent	present	A no-stake claim with no trace behind it	Empty broker
present	present	absent	Recorded disagreement, curated by an interested party	Vendor-run benchmark theater
present	absent	present	Neutral arbitration of checks nobody can replay	Word-of-mouth panel
absent	present	present	A neutral, replayable record of a single unchallenged voice	Notarized monologue (audit without challenge)
present	present	present	An independent check, replayable, held by a party with no stake	Architectural trust

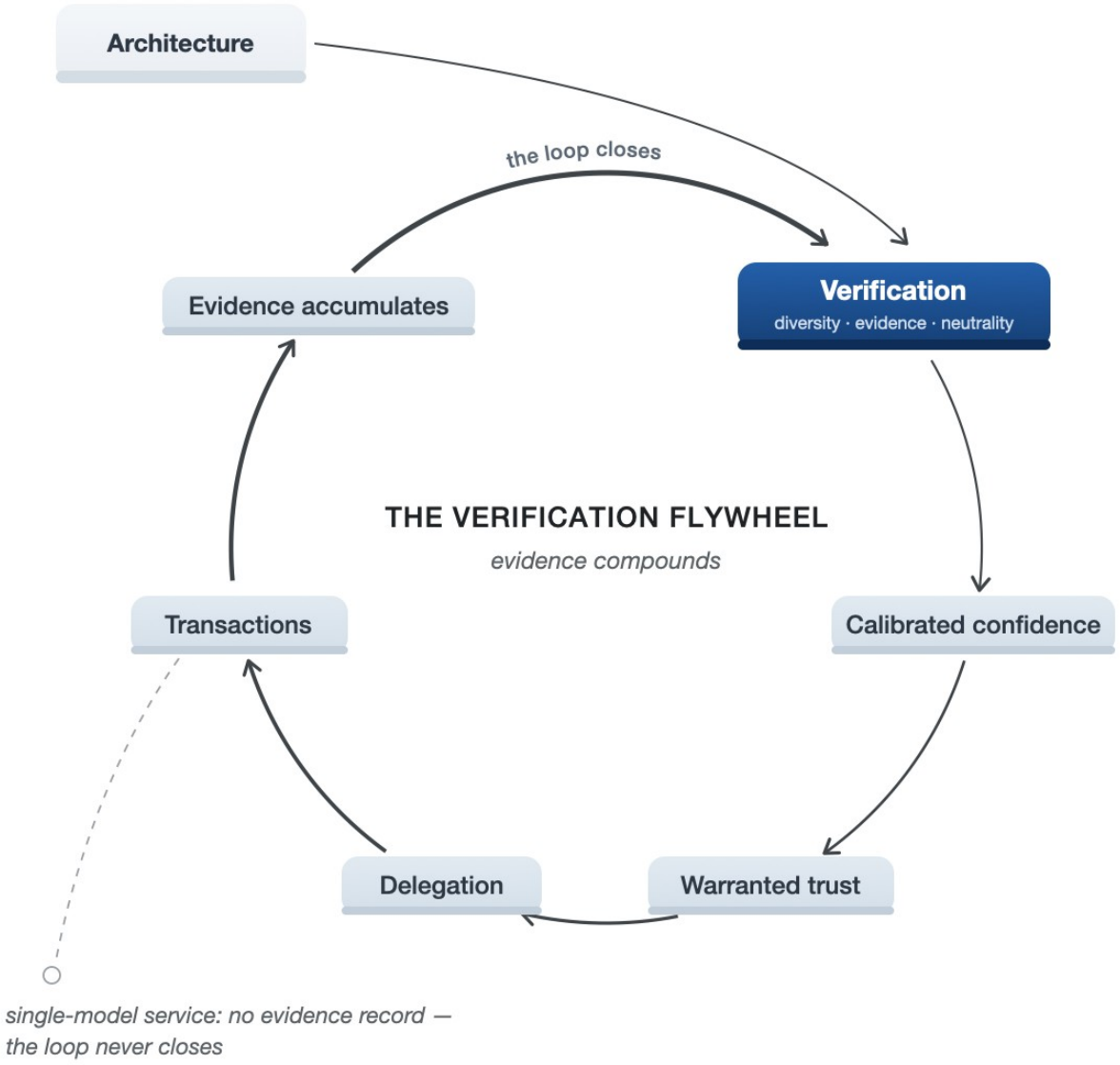
Cells are stated at idealized limits; deployed systems occupy the interior. The E+N-without-D row is the institutionally familiar one — clearing, audit, and ratings operate there — and it is where the framework's demand for an added independent verification path is an empirical prediction, not a definitional exclusion.

The three components map onto the classical anatomy of trustworthiness (Mayer, Davis, & Schoorman, 1995): diversity is an ability mechanism (cross-model agreement calibrates how much competence to infer), evidence operationalizes integrity (claims about what the service did become checkable rather than professed), and neutrality addresses benevolence structurally (a no-stake orchestrator's incentives are inspectable, not believed). Each is re-grounded in verifiable architecture instead of perception management, which is what the human-AI interaction literature asks for when it reaches the same point from its own side: align trust judgments with verifiable trustworthiness cues, not with impressions the interface manufactures (Liao & Sundar, 2022).

The causal chain runs through verification, because "trust" is the consequence rather than the mechanism: *verification → calibrated confidence → warranted trust → delegation → transaction value* (Figure 10). A service earns delegation by being checkable at the moment it

matters, not by feeling trustworthy; Cheng et al. showed how cheaply that feeling is manufactured.

Figure 10. *The verification flywheel*



Persuasion doesn't compound. Verification does

The causal chain, verification (diversity · evidence · neutrality) → calibrated confidence → warranted trust → delegation → transactions, does not terminate: transactions generate evidence, and the loop closes back into architecture. The dashed shunt shows the single-model service, where no evidence record exists and the loop never closes. Persuasion doesn't compound. Verification does.

Table 5. *Single-model service vs. architectural trust*

Dimension	Single-model service	Architecturally trusted service
Object of trust	The model / the brand	The process and its record
Confidence	Asserted by the model	Earned via cross-model agreement
Confident errors	Undetectable from inside	Primary detection target (disagreement signal)
Sycophancy	Rewarded by preference	Separated: empathy and policy-checking on different channels
Error checking	Same system that erred	Independent models + deterministic gates
Evidence of decisions	Logs, if any, vendor-held	Tamper-evident, replayable, consumer-accessible
Quality judge	The vendor itself	Neutral orchestration layer
Dispute position	Consumer's word vs. platform	Replayable record
Insurability	Opaque risk	Priceable from decision histories
Liability allocation (post–AB 316)	"The AI did it" barred; allocation contested in the dark	Computable from the replayable record
Regulatory posture (Art. 50 era)	Retrofit disclosures	Transparency as by-product of operation
Cross-border data promises	Contractual representation	Auditable runtime property (jurisdiction-aware routing)
Cost structure	Flat, cheap	Graduated: expensive only where consequences concentrate

From verifiability to consumer response. The consumer sees the rendering, not the architecture: a receipt showing which independent systems were consulted, where they diverged, and how divergence was resolved or escalated (Figure 8). Three mediating perceptions plausibly carry the effect. The evidence is diagnostic: surfaced, specific disagreement is informative about *this* answer in a way generic disclaimers are not (Detjen et al., 2025; Chen et al., 2026). A replayable record gives perceived process control, since the consumer could check, contest, or escalate after the fact. Procedural assurance is the knowledge that blind collection, independent checks, and a no-stake arbiter were in place (McKnight et al., 2002). The natural dependent variables are reliance calibration, willingness to delegate, and retention after failure. The natural moderators are stakes and reversibility, task objectivity (Castelo et al., 2019), and consumer expertise, since the consumer least able to check an answer personally has the most to gain from an architecture that checks it structurally. The propositions below trace this path.

The framework yields, besides, propositions a consumer study can test, hypotheses rather than findings. Operationalizing them turns on one distinction: *objective* neutrality (the orchestrator is in fact free of undisclosed economic dependence on any vendor, merchant, or outcome), *verifiable* neutrality (the record makes that independence checkable), and *perceived* neutrality (the consumer believes it) are three variables, and an experiment manipulates the verifiable signal while measuring the perception:

P1. Surfacing cross-model disagreement on consequential queries improves consumers' reliance calibration (they rely more when systems agree, less when they diverge) relative to a single confident answer, with perceived risk as the mediating construct (Featherman & Pavlou, 2003). Both failure directions are live: joint presentation reduces over-reliance but can raise under-reliance (Lu, Wang, & Yin, 2024), confidence displays calibrate reliance more readily than accuracy (Zhang, Liao, & Bellamy, 2020), explanations can raise reliance without raising scrutiny (Bansal et al., 2021), cognitive forcing functions cut over-reliance at a cost in satisfaction (Buçinca, Malaya, & Gajos, 2021), and a meta-analysis of 106 experiments finds human–AI combinations on average performing *worse* than the better of human or AI alone (Vaccaro, Almaatouq, & Malone, 2024). P1 tests calibration, not a monotone gain: surfaced disagreement must land in the minority of combination designs that beat their best member. The proposition is split: the main effect is the confirmatory claim, and the moderators are exploratory, derived from the diagnosticity of evidence rather than enumerated after the fact. The calibration gain should increase with stakes and irreversibility, and with task objectivity (divergence on an objective question reads as an error signal; on a subjective one, as taste; Castelo et al., 2019). Consumer expertise is predicted to produce opposite signs: for low-expertise consumers, surfaced divergence without resolution should *decrease* willingness to delegate, and for high-expertise consumers it should *improve* reliance calibration; the crossover is either observed or it is not, which makes expertise a test rather than an escape hatch. This paper's own experiment imposes one design requirement. Disagreement covaries with task difficulty (panel agreement concentrates on easy items), so a P1 design must manipulate or statistically control difficulty, or the calibration effect is confounded with item easiness. The diagnosticity account also yields a boundary prediction that generic transparency theories do not make: divergence disclosure should matter most where the consumer's own priors are weakest, on cross-border or unfamiliar-merchant purchases, because there external diagnostic evidence has no internal competitor.

P2. Access to a replayable decision record increases willingness to delegate consequential tasks to an AI agent, over and above brand trust, with the effect increasing with the stakes and irreversibility of the delegated task. It also increases with the level of delegated agency itself, which is the comparative-statics content of the Trust-Migration Hypothesis: holding actual model performance constant, the marginal effect of verifiable architectural properties on appropriate delegation grows as the consumer surrenders more review and execution control (advisor → copilot → supervised agent → autonomous agent). A design that crosses delegation level with architecture visibility tests the hypothesis directly. The nearest consumer-research evidence sits on this dependent variable: delegating purchasing to AI reduces adoption where it restricts perceived autonomy (Ahmad Husairi & Rossi, 2024), and acceptance of high-autonomy assistants is a directly studied outcome (Frank et al., 2026). P2 predicts the record's effect over and above those established antecedents.

P3. Perceived neutrality of the orchestration layer independently predicts delegation willingness, consistent with procedural-justice effects in service evaluation (Tax, Brown, & Chandrashekaran, 1998). The manipulated variable is the verifiable neutrality signal; the measured variable is the perception it produces. To my knowledge, orchestrator neutrality has never been operationalized in a behavioral experiment, and certification seals are the nearest tested analogue (Cetinkaya & Krämer, 2024), so P3 introduces a new experimental variable. Directionally, the neutrality effect should strengthen as the consumer's own ability to check substance falls, since when outcome quality is hard to judge, process cues carry the evaluative

weight. The manipulation has to separate verifiable neutrality from a neutrality *claim*: one condition carries a receipt whose no-stake attestation can actually be opened and checked, the contrast condition an identical interface bearing only the label, with a manipulation check on perceived neutrality in both. Without that contrast P3 collapses into generic process-fairness and says nothing about verification.

P4. After an agent error, consumers whose service produces an inspectable record show smaller trust collapse and higher retention than consumers of an opaque service: the one-strike dynamic is moderated by evidence. This is the mechanism service-recovery research identifies, where procedurally fair, evidence-based complaint handling restores trust after failure (Tax et al., 1998). Trust-repair research adds that what restores trust depends on the nature of the violation and the account offered (Kim, Ferrin, Cooper, & Dirks, 2004), a requirement human–machine research now treats as a design problem for autonomous systems (de Visser, Pak, & Shaw, 2018), and the buffering should grow with the severity and irreversibility of the failure. P4 predicts: record → smaller trust collapse and higher retention, moderated by failure severity. Cases where the record shows the process itself was not followed are handled by a manipulation check and excluded from the test, since the proposition concerns failures that occur *despite* a followed process. Durability comparisons must control for switching costs and lock-in, which confound retention, and a companion design tests whether a followed-process record shifts blame attribution from the service to the environment. The proposition carries no reversal clause, so no outcome pattern is unfalsifiable by construction.

Empirical anchor: the confidence layer as a tested component. If a service claims its confidence signals mean something, that claim can be put on trial. The confidence layer in the deployment behind this paper's case study went through a pre-registered, placebo-controlled, six-arm evaluation across two independent model families. A measured profile of the system's own weaknesses improved its error prediction, an AUROC gain of roughly +0.07 over placebo, with non-overlapping confidence intervals on both families. A false self-profile collapsed error prediction to a coin flip, and a generic "be aware of your limits" instruction did nothing.

Two earlier versions failed, one falsified by its own authors over a confound, one negative, before the pre-registered design confirmed. A stricter re-test was inconclusive at reduced scope, and a pre-registered four-rung ladder on a frontier family passed no confirmatory endpoint, the effect's direction surviving only exploratorily on three of four rungs. The claim is therefore per-family and bounded: measured self-knowledge improved error prediction where confirmed, and it is not an established property of every model class. The weakness profile runs in production as well, where it adjusts live model routing, bounded and fail-open. The false-profile collapse doubles as a security finding: a fabricated self-model is an attack vector, one more reason self-knowledge needs architectural custody rather than model self-report. The design logic has a consumer-research pedigree older than the technology. Precommitment, binding future behavior through external structure because internal resolve is not expected to hold, is exactly how consumers ration their own vices (Wertenbroch, 1998), and architectural trust extends the same move to machines.

The full arc (protocols, pre-registrations, raw data, the falsified pilot, the negative iteration, the non-confirming extensions) is preserved as a standalone archive, publicly available with the paper's OSF materials at https://osf.io/jtvu9/ (file `self-calibration-archive.zip`). Calibrated confidence is an architectural component with measurable content, and it can fail. The failures

are archived with the successes because a trust layer whose confidence claims have never been experimentally challenged is asking for exactly the blind trust it was supposed to replace. A thirty-author position paper gives this division of labor a formal statement: agentic orchestration should be *Bayes-consistent* (Papamarkou et al., 2026), and calibrated beliefs belong in the control layer, not in the models.

Testing the mechanism: a pre-registered experiment. The claims above are testable, and this section reports the tests. Every experimental result in this section, in both waves, speaks to a *technical antecedent* of the framework, whether independent agreement carries a correctness signal that a model's self-report does not, on factual questions. These experiments do not demonstrate consumer trust, delegation, or commercial behavior, which remain the province of P1–P5 and the behavioral study of Section 6. The protocol was frozen publicly before any confirmatory data existed (OSF: https://osf.io/4y9dv, registered 23 July 2026; hypotheses H1–H6, benchmark subsample fixed by seed and SHA-256 hash, outcomes unfavorable to the hypothesis were pre-declared). The design is deliberately simple. Sixteen models, nine frontier vendors (Anthropic, OpenAI, Google, xAI, DeepSeek, Alibaba, Moonshot, Zhipu, Mistral) and seven local open-weights models, each answered the same 1,500 SimpleQA questions once, blind, without retrieval or tools. That is the regime that maximizes exposure to correlated parametric error, a deliberately hard case for the signal, at the ecological cost stated in the limitations. Answers were given at temperature zero, stating a 0–100 confidence: verbalized uncertainty, elicited in words within the same response rather than read from logits, the format Lin, Hilton, and Evans (2022) showed models can learn to calibrate. Panels are analytic combinations of that single collection of 24,000 nominal answer slots, of which 22,598 were delivered, one local model (deepseek-r1:8b) having stopped at 98 items under a pre-committed stopping rule. The six registered hypotheses map onto the results as follows. H1, that cross-model divergence flags a majority of confident errors, is supported: agreement lifts correctness 1.9–3.2× in the registered wave (Table 6), and in the flagship wave the frozen gate removes 82.1% of confident failures (Table 10). H2, that the panel signal beats the answering model's own confidence, holds on the CN-frontier and cross-bloc panels (ΔAUROC +0.287 and +0.063) and reverses on the US-frontier panel. H3, that a non-trivial share of confident errors survives as unanimous agreement on one wrong answer, is supported: 0.5–2.2% of open-form questions and 4.0–7.9% in the multiple-choice arm. H4(a), that intra-bloc error correlation exceeds cross-bloc, is not supported (φ difference +0.011, 95% CI −0.005 to +0.028), and H4(b), that a cross-bloc panel catches more, is not supported either, the cross-bloc lead being the smaller one. H5, that a $0 local panel flags a majority of frontier confident errors, is not supported: the local tier's agreement signal sits at chance. H6, the lineage-distance ladder, is not tested as a ladder here; its bloc-level step is the H4(a) null.

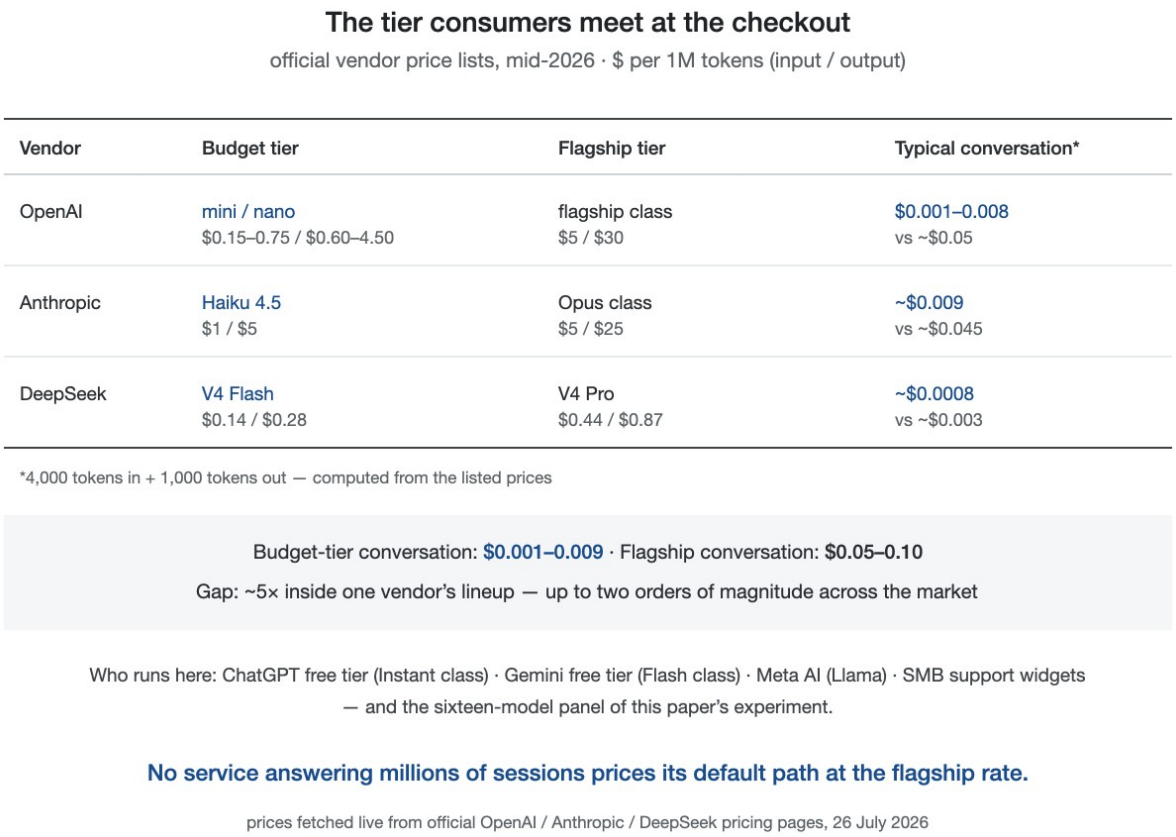
Factual verification is a relevant proxy for agentic commerce because a delegated purchase decomposes, at the moment of execution, into factual sub-claims: the item's current price and availability, the terms of the return policy, whether the product's specifications match the consumer's stated requirements, the identity and rating of the seller. Each of these has a verifiable ground truth of the same kind a SimpleQA item has. The disagreement gate operates on that factual substrate, while the preference weighting (which trade-offs matter, and how much) remains with the consumer, who authored the requirements the facts are checked against. That division of labor is also why the Section 6 design manipulates the display of verification rather than participants' preferences: the architecture's claim concerns the factual layer of a

delegated decision, not its taste layer. Where a judgment has no factual substrate at all, in genuinely subjective choices with no ground truth to verify, the antecedent stops carrying, and the Limitations state that boundary as an open question.

Benchmark choice was measured. On the 2021-vintage TruthfulQA (present in pretraining corpora), the same vendors scored 47–61 percentage points higher than on SimpleQA (e.g., 89% vs 28%), a gap consistent with benchmark-age and exposure effects, though this design cannot separate contamination from differences in benchmark difficulty and construction. No headline number therefore rests on the older benchmark. Grading is objective string matching, with unresolved cases judged for equivalence-to-gold only by a local open-weights model (Qwen3-8B, run locally; equivalence-only prompt in the replication package; never by any panel member's family; whether SimpleQA appears in the judge's training corpus is unknown, but the judge adjudicates equivalence between two supplied strings, not truth, which bounds the exposure) and a 50-item human spot-check (98% agreement), a limitation of scale. That is now extended by a stratified 300-item cross-family audit of judge-resolved cases (a different model family as auditor, seed-fixed sample): 96% agreement (288/300), all twelve disagreements lenient false positives (nine clear, three borderline), a one-directional bias touching roughly 3% of all graded answers (4% of the judge-resolved share). Two-thirds of the verdicts (66.5%) came from the local model grader rather than exact string match, and the headline is robust to grader error: flipping a random 10% of model-graded verdicts shifts the certification ratios by at most ±0.5× and never changes their direction.

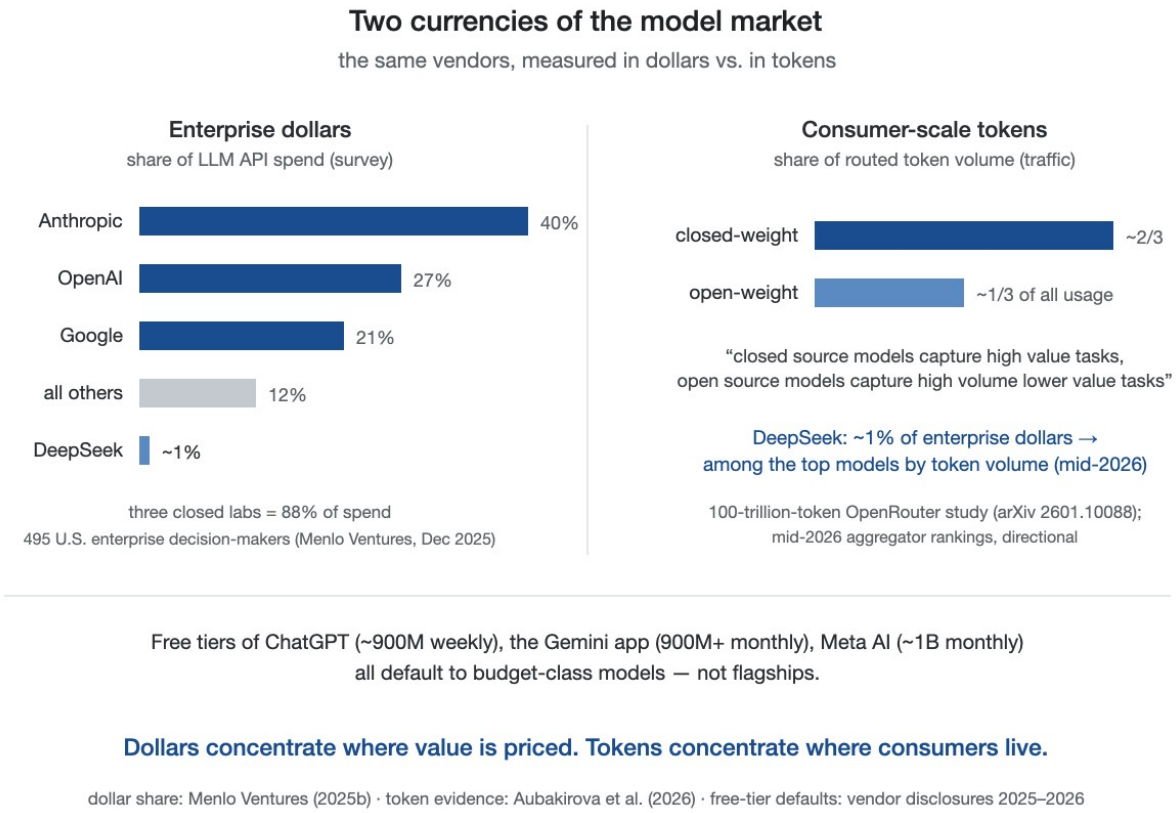
Two design choices are deliberate rather than budgetary. Vendors' mid tiers were used, first, for ecological validity: the agents now transacting on consumers' behalf overwhelmingly run on these cost-optimized tiers, checkable from public price lists (mid-2026; Figure 11). A typical service exchange costs a fraction of a cent to about a cent on the budget tiers, against five to ten cents on flagship weights, and no service answering millions of sessions prices its default path at the flagship rate. The roster mirrors the market: the nine-vendor panel contains all three suppliers that together carry 88% of enterprise LLM API spend, and the US-frontier panel of Table 6 is exactly that trio. The objection that this is enterprise evidence, with the consumer nowhere in it, gets the supply chain backwards. The consumer no longer chooses a model at all: they choose a bank, a retailer, an airline, and the service's routing policy chooses the model. In the currency consumers are actually served in, tokens rather than dollars, the picture confirms the choice (Figure 12): in the OpenRouter–a16z study of 100 trillion tokens of routed traffic, open-weight models reached roughly a third of all usage (Aubakirova et al., 2026), the default tiers of the three largest consumer AI products all sit below the flagship tier, and only about 3% of consumer AI users pay for premium service (Menlo Ventures, 2025c), so the marginal consumer must be served at near-zero marginal cost. Below the API market sits a floor of on-device and self-hosted open weights, invisible to spend-based surveys, and the experiment's local tier of seven open-weight models samples it deliberately. In every currency the market can be counted in, the tiers this study sampled are the ones that serve the consumer. (After a first-party outage, part of the collection was completed through an aggregator, with the actual serving infrastructure recorded per call in the run manifest.)

Figure 11. *The tier consumers meet at the checkout*



Official vendor price lists, mid-2026.

Figure 12. *Two currencies of the model market*



Enterprise dollars vs. consumer-scale tokens.

The mid tiers also suited the paper's claims, which concern the architecture, not flagship capability: an effect that only appeared with premium weights would be a model result. The pipeline replicates on commodity infrastructure for under $15 of API spend. For a paper arguing that verification must be cheap enough to run on every consequential answer, the cost of verifying the paper itself is part of the evidence. The pipeline is a standalone replication package (runners, grader, analysis scripts, raw graded data, and run manifest) requiring nothing of the author's research system. The collection totals 24,000 answers, roughly 30,000 including the benchmark-contamination and conformity arms. On availability: the public pre-registration (osf.io/4y9dv) archives the seed-fixed 1,500-item subsample. The full package is publicly available at https://osf.io/jtvu9/, a public component of the registration's source project (folder `replication-package-v2`).

Three results carry the section. *First, the raw material is abundant:* median stated confidence ran 85–95 at seven of nine frontier vendors, with the two exceptions (medians 42 and 60) exactly the vendors whose confidence discriminates, while actual accuracy ranged from 10% to 66%, producing 289–1,212 confident errors (wrong at confidence ≥ 75) per frontier vendor on 1,500 questions. As a correctness signal, a model's own confidence scored AUROC 0.53–0.79 depending on vendor, and a consumer has no way to know, from the outside, which vendor they got. That vendor spread qualifies, at the tier consumers are actually served from, the frontier-lab finding that large models are often well calibrated at self-evaluation in controlled formats

(Kadavath et al., 2022): some deployed vendors ship that self-knowledge, others ship confident noise, and nothing on the surface distinguishes them.

Second, the headline: verification works as certification, not as alarm. Certification in a probabilistic sense, a posterior of roughly 49–77% depending on panel, not a guarantee. On questions this hard, the naive signal "some panelist disagrees" saturates: it fires on 85–94% of confident errors but also on 91–95% of confident correct answers (the registration anticipated this reframing, moving to the pre-specified AUROC endpoint). The informative direction is the opposite one: correct answers get independently replicated, wrong answers scatter. An answer that at least one independent panelist reproduced was correct 72–77% of the time on the intra-bloc panels and 49% on the cross-bloc panel, against 18–37% without replication. *An answer with independent agreement was roughly two to three times as likely to be correct (1.9–3.2× across the three panels of Table 6).* (Here and throughout, "independent" means blindly and separately produced, with residual error correlation measured rather than assumed, and correctness is always scored against the benchmark's gold labels, never by another model's opinion.) The estimand is associational, not causal: replication marks answers more likely to be correct, and part of what it marks is that the question was easy, a selection effect quantified below. As a correctness signal, panel agreement reached AUROC 0.832 on the panel whose members' own confidence carried the least signal (per-vendor 0.53–0.62; pooled 0.545). The claim that holds up is distributional rather than featured: across all 84 three-vendor panels the agreement AUROC has median 0.79 and exceeds pooled self-report in 89%. Agreement reached 0.722 on the cross-bloc panel against pooled own confidence of 0.659 (pooling unstandardized verbal scales across vendors distorts self-report, because vendors use the 0–100 scale differently and the artifact's direction is panel-dependent, which is why the per-vendor values are printed alongside). That lead's cross-bloc instance is coding-dependent (alternative pre-frozen agreement codings place the cross-bloc panel's agreement signal at 0.544–0.722, the generous unresolved-as-agreement variant falling below that panel's pooled own-confidence of 0.659, while the CN panel's signal spans 0.719–0.894 across the same codings, always above 0.545), so the coding-robust panel-over-self result is the intra-bloc CN case. The one panel where self-report clearly beat the panel signal (0.842 vs 0.697) was the one containing the study's best-discriminating vendor (Anthropic's mid tier, own-confidence AUROC 0.788 on the full 1,500-item collection), the well-ranked vendor's own signal alone outperforming its own panel's agreement signal. Table 7 prints per-panel values, computed within each panel's item population. A paired cluster bootstrap (questions resampled with replacement, all three (item, primary) units per question moving together; B = 2,000) puts confidence intervals on the Table 6/7 contrasts. On the CN-frontier panel, panel agreement beats pooled own confidence by ΔAUROC = +0.287 (95% CI [+0.265; +0.308], $p < .001$). On the cross-bloc panel the lead is smaller but reliable, +0.063 ([+0.037; +0.087], $p < .001$). On the US-frontier panel the direction reverses and the reversal is itself statistically unambiguous: own confidence 0.842 versus panel agreement 0.697. Panel composition is therefore a first-order moderator of the signal's value, and the deployment rule follows: measure the panel members' own calibration before trusting agreement over it, ΔAUROC = −0.145 ([−0.170; −0.120], $p < .001$). All three contrasts are significant, and the heterogeneity across panels is a finding, not noise. The US case is therefore reported as a confirmed reversal, not a null zone, and it bounds the construct: on a panel drawn from the discriminating end of the market, own confidence outperforms agreement as a ranking signal, so the Diversity component's discrimination value depends on panel composition rather than holding universally. Architecture can provide an external correctness signal where exposed

model confidence discriminates weakly, and it audits the calibration claim where one is made. One boundary stays: even a perfectly calibrated model cannot notarize its own history or arbitrate its own conflict of interest, so as vendor discrimination improves, diversity's role shifts from correctness discovery to calibration auditing, while evidence and neutrality do work no amount of calibration can.

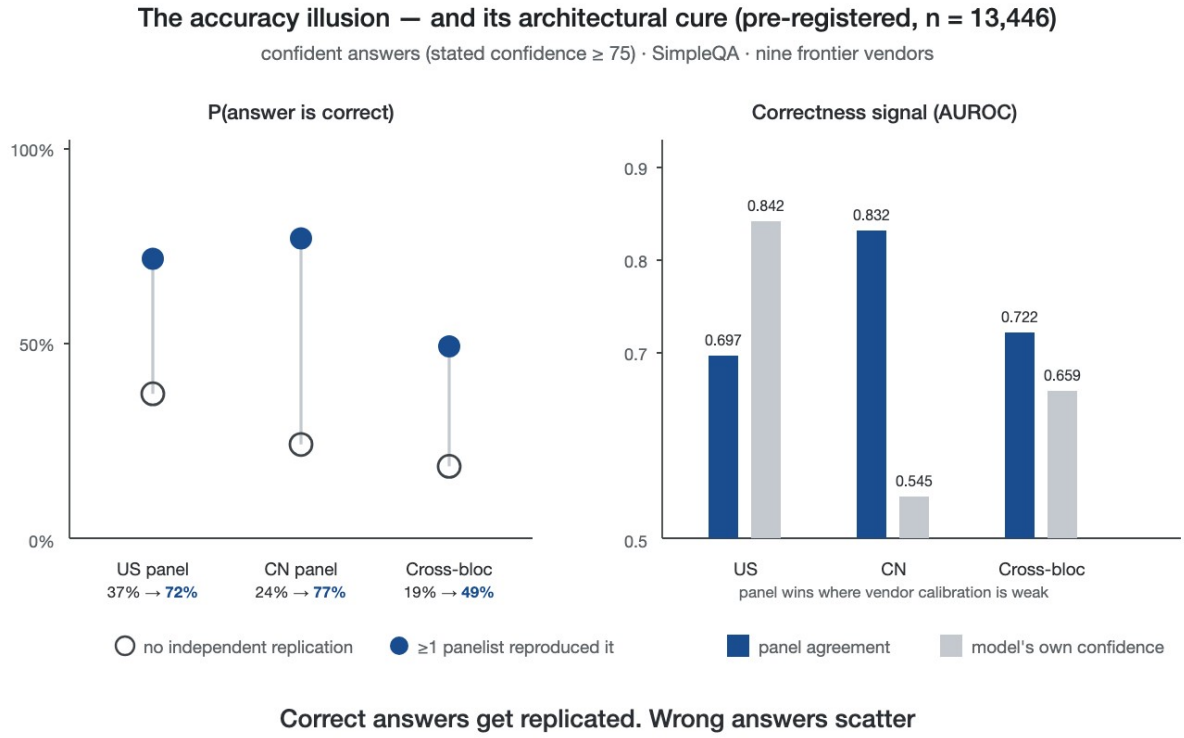
Architectural verification supplies an external correctness signal where a vendor failed to ship a usable confidence signal, and identifying which vendor shipped one requires the architecture anyway. The substitution claim, exactly scoped, rests on the calibration metrics rather than on AUROC. No vendor in the July wave is simultaneously discriminating and well calibrated; xAI comes closest (Table 7a), and the flagship wave reproduces the dissociation's calibration side, so even the consumer who happens to receive the best-ranked confidence cannot cash its level out as a probability without an external reference. That is what the architectural layer supplies. Should vendors eventually ship good calibration everywhere, evidence and neutrality are untouched. The argument survives the world in which models improve.

Third, the failure mode has a shape: certified consensus errors, every panelist confidently converging on the same wrong answer, occurred on 0.5–2.2% of open-form questions but on 4.0 –7.9% in a companion arm that re-posed items in multiple-choice format on the local tier (item construction and outputs in the replication package). Consensus certifies error chiefly where the answer space is constrained, a design warning for any interface that offers agents a menu of options.

The nulls are as instructive as the positives, and I publish them as pre-registered. Geopolitical bloc (including the possibility that same-bloc panels share training corpora and distillation lineages) does not predict error correlation on this benchmark (intra- vs cross-bloc φ difference +0.011 on the frontier tier, 95% CI −0.005 to +0.028; slightly *reversed* on the local tier). Capability drives the correlation, not the flag, echoing Kim et al. (2025). Lineage does not predict catch rates: a local model verifying its own frontier sibling performs like any stranger, and a reasoning-distilled model proved the *least* error-correlated with its own base weights (φ = 0.135). Both are bounded nulls rather than demonstrated equivalence. The bloc interval excludes φ differences larger than +0.028, and that bound, not the failure to reject, does the inferential work. The natural continuous extension, which the replication data support, is to estimate certification lift as a function of measured panel diversity (the surprise-score φ) rather than of panel identity. That would rerun, for model panels, the program the classifier-ensemble literature ran for its own diversity statistics, which found that no single measure cleanly predicts ensemble gains (Kuncheva & Whitaker, 2003).

And cheap local models could not certify frontier answers on open-form questions at all (46% precision with or without their agreement): a verifier must be competent on the task class before its independence is worth anything. The boundary conditions are these: consensus certifies error where answer spaces are constrained, independence must be measured rather than assumed, verifier competence gates certification value. In practice, *diversity is a property to be measured, question by question, not assumed from brands, blocs, or family trees.* Table 6 summarizes the panel metrics. Figure 13 shows the certification lift and the signal comparison side by side.

Figure 13. *The accuracy illusion and its architectural cure*



Certification by independent replication (pre-registered; n = 13,446 graded answers — nine frontier vendors, each answering the same 1,500 SimpleQA questions once, less 54 failed or ungradeable calls).

Concurrent work reinforces the caution: a large-scale audit found high self-consistency co-occurring with wrong answers on reasoning benchmarks, where agreement alone, without independence and blind collection, is a weak and regime-dependent signal (Ding, 2026). I therefore treat *independent, blind* replication as the unit of certification.

Table 6. *Confident-error interception and certification by panel (SimpleQA, confirmatory; conf ≥ 75)*

Panel	Confident errors (n)	Flag rate on errors [95% CI]	False alarms on correct	P(correct | ≥1 agrees)	P(correct | none)	AUROC: agreement vs own conf
US-frontier (Anthropic·OpenAI·Google)	1,630	94.3% [93.1–95.3]	91.0%	71.8%	37.1%	0.697 vs 0.842
CN-frontier (DeepSeek·Alibaba·Moonshot)	2,128	85.5% [83.9–86.9]	94.8%	77.0%	24.1%	0.832 vs 0.545
Cross-bloc (Anthropic·DeepSeek·Mistral)	2,432	89.4% [88.1–90.6]	93.4%	49.3%	18.5%	0.722 vs 0.659

Flag = ≥1 panelist's independent answer non-equivalent to the primary's (refusals and zero-overlap answers count as non-concurrence; a stricter coding is in the replication package). Threshold sensitivity at 60/90 changes no conclusion. Own-confidence AUROC is stated confidence pooled over the panel's full answer population; pooling absorbs between-vendor differences in accuracy and confidence scale, so it can exceed every member's individual value (per-vendor values in Table 7). Answers cluster by question; a question-level cluster bootstrap widens the intervals by roughly a factor of 1.6 without changing any conclusion. Every AUROC in this table and Table 7 reproduces deterministically, with one command and no randomness, from the replication package script compute_panel_vs_own_auroc.py (Mann–Whitney with tie correction).

Table 7. *Own-confidence AUROC by panel: pooled panel population and per vendor (same populations as Table 6)*

Panel	Pooled own-confidence AUROC	Per-vendor own-confidence AUROC
US-frontier	0.842	Anthropic 0.784 · OpenAI 0.596 · Google 0.746
CN-frontier	0.545	DeepSeek 0.569 · Alibaba 0.532 · Moonshot 0.620
Cross-bloc	0.659	Anthropic 0.787 · DeepSeek 0.570 · Mistral 0.584

Computed by the deterministic pipeline in the replication package. Two population conventions coexist and are printed explicitly to avoid apparent inconsistency: the table's per-vendor values are computed within each panel's item population, while single-number citations in the text use the full-collection value — for the best-discriminating vendor 0.784 within the US-frontier panel versus 0.788 on the full 1,500-item collection, and for OpenAI 0.596 versus 0.598 (both pairs reproduce from one command of the same pipeline). The pooled column, by contrast, conflates vendor base rates — between-vendor differences in accuracy and confidence scale do part of the work, so the pooled value need not lie inside the per-vendor range (recomputed live: the US-frontier pooled 0.842 exceeds every member's individual value, and the CN-frontier pooled 0.545 sits below that panel's per-vendor macro-average of 0.574); the per-vendor values are the interpretable quantities. The per-vendor spread is the consumer's problem stated numerically: usable discrimination exists at one vendor and is near chance at others, and nothing on the outside of an answer says which vendor produced it. Source values behind the Section 3 vendor ranges, printed per model from the same graded collection (median stated confidence · accuracy · full-collection own-confidence AUROC): frontier — Alibaba 95 · 47.4% · 0.532; Anthropic 42 · 10.3% · 0.788; DeepSeek 90 · 35.7% · 0.570; Google 95 · 66.5% · 0.746; Mistral 90 · 16.5% · 0.584; Moonshot 85 · 51.6% · 0.619; OpenAI 85 · 16.7% · 0.598; xAI 60 · 37.9% · 0.777; Zhipu 95 · 28.4% · 0.611 — the two medians below the 85–95 band belong to the two vendors whose confidence carries usable signal; local tier (enters only the difficulty stratification and the local-tier nulls) — gemma4:e4b 90 · 5.2% · 0.541; glm4:9b 85 · 5.0% · 0.589; llama3.2:3b 100 · 5.3% · 0.494; mistral:7b 95 · 6.3% · 0.506; phi4-mini 95 · 5.5% · 0.505; qwen3:8b 95 · 5.9% · 0.504; deepseek-r1:8b 90 · 9.2% · 0.696 (n = 98, pre-committed stopping rule).

AUROC measures discrimination, whether a vendor's stated confidence ranks its correct answers above its incorrect ones, while ECE, the Brier score, and the confidence–accuracy gap measure calibration, whether the stated number can be read as a probability, and in these data the two dissociate sharply. The study's best-discriminating July vendor, Anthropic's mid tier (own-confidence AUROC 0.788), is still badly miscalibrated in level: mean stated confidence 41.3 (median 42) against 10.3% accuracy, a +31.0-point overconfidence gap, ECE 0.311 (15 bins), Brier 0.206 with a reliability term of 0.126 against an irreducible-uncertainty floor of 0.093. Conversely, the worst calibration in the panel co-occurs with near-chance discrimination (Mistral: gap +74.2 points, ECE 0.742, AUROC 0.584), and only xAI approaches both discrimination and calibration (ECE 0.111, AUROC 0.777). Even there the discrimination is helped by uniformly low stated confidence, while for the other eight July vendors the two properties dissociate sharply. The same dissociation persists at the flagship tier: nine of eleven August flagships carry overconfidence gaps of +12.5 to +71.4 points (ECE 0.125–0.714), while the two exceptions (gpt-5.6-sol, ECE 0.031; claude-fable-5, ECE 0.043) achieve calibration at opposite ends of the accuracy scale. A consumer holding a well-ranked but numerically inflated confidence score cannot cash it out without an external reference. Table 7a prints the calibration metrics next to the discrimination ones.

Table 7a. Discrimination vs calibration, July wave (full collection, resolved; ECE with 15 bins; gap = mean stated confidence − accuracy, percentage points)

Vendor	n	AUROC (own conf)	ECE	Brier	Conf − acc gap	Mean conf → accuracy
Alibaba	1,497	0.532	0.448	0.446	+44.8 pp	92.3 → 47.4%
Anthropic	1,500	0.788	0.311	0.206	+31.0 pp	41.3 → 10.3%
DeepSeek	1,500	0.570	0.536	0.513	+53.4 pp	89.1 → 35.7%
Google	1,461	0.746	0.263	0.271	+26.3 pp	92.8 → 66.5%
Mistral	1,495	0.584	0.742	0.686	+74.2 pp	90.6 → 16.5%
Moonshot	1,500	0.619	0.242	0.292	+24.2 pp	75.8 → 51.6%
OpenAI	1,500	0.598	0.610	0.531	+60.9 pp	77.6 → 16.7%
xAI	1,500	0.777	0.111	0.197	+11.1 pp	48.9 → 37.9%
Zhipu	1,493	0.611	0.642	0.612	+64.2 pp	92.6 → 28.4%
US-frontier panel, pooled units	4,383	0.842	0.393	0.336	+39.3 pp	—
CN-frontier panel, pooled units	4,491	0.545	0.408	0.417	+40.8 pp	—
Cross-bloc panel, pooled units	4,485	0.659	0.529	0.469	+52.9 pp	—

Same populations as the full-collection AUROC and Tables 6–7 (panel rows: pooled (item, primary) units); refusals scored as wrong under the frozen protocol. Brier decomposes by the Murphy identity (reliability − resolution + uncertainty) on the same 15 bins, residual < 0.001 in every row. One flagship-tier note travels with the two calibrated exceptions: gpt-5.6-sol's ECE of 0.031 is earned at 95.4% accuracy, where the Murphy uncertainty term, the floor the base rate sets, is 0.044 — calibration partly purchased by a base rate that carries the Table 9 contamination caveat (61.6% verbatim gold-string matches).

One selection effect must travel with the table: replication is partly a marker of question easiness. Stratifying by question difficulty (mean accuracy across the other fifteen models), the raw certification lift is carried by the easier strata (2.7–4.6×), falls below 1× on the second-hardest quintile, and vanishes on the hardest under this within-collection measure. That stratification is partly endogenous, and an exogenous re-estimate below softens the inversion into monotone attenuation. A robustness analysis addresses the selection concern directly (exploratory, not pre-registered): in a question-clustered logistic model with primary-model fixed effects, controlling for the primary's own confidence and for item difficulty (leave-primary-out mean accuracy across all sixteen models), each additional agreeing panelist still multiplies the odds of correctness by roughly five (pooled b = +1.65 per panelist, 95% CI [+1.49, +1.81], $p < 10^{-88}$; positive in every panel), and adding agreement to a difficulty-plus-confidence baseline raises held-out AUROC from 0.859 to 0.898 and lowers the Brier score from 0.139 to 0.116 (two-fold cross-validation split by question). The same model probes the boundary, and the stratification itself needs care: under the within-collection difficulty measure the conditional coefficient turns negative on the hardest quintile (b = −0.41, p = .011), but that measure includes the panel's own members. The conditional model carries a deployment caveat: its difficulty feature requires labels, so it is diagnostic, not deployable, and the deployable filter is the agreement gate itself. Three further checks address estimation and selection concerns, all under the frozen deterministic protocol coding (the stricter branch the replication package reports alongside the judge-adjudicated headline numbers). Re-estimating the certification lift with item-clustered bootstrap resampling, the clustering the difficulty structure demands, leaves every panel's certification-lift interval far above 1 (point lifts 2.4–4.5 under this deterministic coding, which certifies fewer agreements than the judge-adjudicated pair coding behind the published 1.9–3.2× headline; bootstrap lower bounds 2.2–4.1). The 84-panel distribution reported above makes the US-frontier panel, where self-report wins, the exception rather than the rule (median advantage +0.14, IQR of agreement AUROC 0.76–0.82). And difficulty-stratified AUROC makes the boundary explicit under an exogenous difficulty measure: 0.87 on the easiest quintile falling to 0.61 on the hardest, attenuation without inversion. That measure is leave-panel-out, the thirteen models outside each panel, and re-estimating with it resolves the inversion: the conditional agreement coefficient stays positive on every quintile including the hardest (hardest-quintile b = +0.97, $p < 10^{-7}$; 28% of independently replicated answers on the hardest stratum correct, not 0%). So the within-collection inversion is partly an artifact of the difficulty measure. What survives under either measure is monotone attenuation, with agreement weakest where questions are hardest, and the reality of certified consensus errors; scripts and outputs are in the replication package. The headline ratio is therefore an average over a heterogeneous population. The signal claim rests on the AUROC endpoint, not on the ratio.

The experiment's hypotheses operationalize the research deployment I operate: blind collection mirrors its council mechanism, and the panel compositions existed as its production

configuration before the study was designed. This is the second pre-registered study in this research program. The third, the Section 6 consumer experiment, is complete and reported, and a fourth (field telemetry over production traffic) is drafted. The replication package runs for anyone, on commodity hardware, against public APIs.

A second wave at the flagship tier: eleven laboratories — ten complete arms and one partial — one calibration moat. The experiment above sampled the cost-optimized tiers that actually serve consumers. The objection to test: perhaps the accuracy illusion is a budget-tier artifact, and the flagships, the models each vendor puts forward as its best, are calibrated enough that architecture above them is redundant. A second collection wave (August 2026) tests that objection at the top of the market.

I evaluate one flagship model from each of ten major AI laboratories spanning the United States, China, and Europe. Laboratories were chosen to cover the leading commercial and open-weight ecosystems across these three regions. For each laboratory I selected its most capable generally available model at the time of the evaluation wave. I then verified this coverage against public leaderboards as of August 5, 2026 (LMArena Elo, the Artificial Analysis Intelligence Index, Humanity's Last Exam, and SWE-bench Verified): eight of the ten flagships place in or near the top tier of at least one general-capability leaderboard, while two (Mistral, Baidu) are included to preserve laboratory and regional coverage rather than on ranking grounds (see the coverage limitations below). An eleventh laboratory, Meta, whose flagship became reachable only after the wave was scoped, was added mid-wave; its collection was interrupted by an access revocation and is reported as partial (see below).

Table 8. *Evaluated flagship models and leaderboard positions (as of August 5, 2026)*

Lab	Flagship model (API id)	Country	Leaderboard positions (Aug 5, 2026)
OpenAI	gpt-5.6-sol	USA	AA Index #3 (58.9); LiveBench #1 (82.4); Arena top-6 cluster
Anthropic	claude-fable-5	USA	Arena #1 (~1525 Elo); AA Index #2 (59.9); SWE-bench Verified #1 (95.0)
Google	gemini-3.1-pro-preview	USA	SWE-bench Verified #8 (80.6); HLE #17 (51.4); Arena upper cluster
xAI	grok-4.5	USA	AA Index #8 (53.8); Arena Elo 1499
DeepSeek	deepseek-v4-pro	China	SWE-bench Verified #8 (80.6, Pro-Max); HLE #23 (48.2)
Alibaba	qwen-max (Qwen 3.7 Max)	China	Arena #21 (1488–1496 Elo); SWE-bench Verified #11 (80.4)
Moonshot AI	kimi-k3	China	AA Index #4 (57.1); HLE #9 (56.0); Arena #10 (1500)

Zhipu (Z.ai)	glm-5.2	China	AA Index #13 (51.1); HLE #11 (54.7)
Mistral	mistral-medium-3.5	France	Below top-50 on general leaderboards; included for regional coverage
Baidu	ernie-4.5-vl-424b	China	Outside general top-30 (ERNIE 5.0: HLE #36); ERNIE 5.1: LMArena Search Arena #4
Meta (added mid-wave, partial)	muse-spark-1.1	USA	HLE #4

Scores in parentheses are the aggregator-reported values on the snapshot date; leaderboard positions vary slightly across aggregators and snapshot dates. Snapshot sources are archived with the run records that accompany the replication package.

The protocol is the first experiment's, frozen and reused: the same 1,500-item SimpleQA subsample (fixed by seed and hash), one blind pass per model, temperature zero, identical system prompt and answer parser, a stated 0–100 confidence per answer. Grading is deterministic (exact and normalized string match against gold; refusals scored as errors; unresolved items excluded from the headline and re-scored as errors in sensitivity analysis), and no LLM judge touches the headline numbers. That is a deliberate tightening relative to the first wave, whose grader was a hybrid: objective string match with a local open-weights judge resolving the remainder (66.5% of verdicts model-issued; 50-item human spot-check, 98% agreement). This wave uses only the deterministic equivalence engine (manually validated: 20 of 20 flagged disagreements and 15 of 15 agreements confirmed by hand), so every headline number of this wave reproduces bit-for-bit from the released script. Cross-wave comparisons of absolute accuracy therefore also cross grader regimes, and every July-versus-August comparison carries a grader-regime qualifier. Within-wave audits support each wave's numbers, but cross-wave deltas conflate model change with grader change unless a single grader is re-applied to both frozen response sets, which has not been done here. One effort deviation: OpenAI's arm was collected at the interface's default reasoning effort (*medium*) where the July protocol used minimal, with no web or tool access at inference (verified by enumerating the tools exposed to the model, of which there were none). This shifts that vendor's absolute accuracy, not the agreement mechanism (Section 5). A *confident failure* is a wrong answer offered at stated confidence ≥ 80; the threshold was fixed before the run, with 70 and 90 reported as sensitivity. Union coverage is 1,500/1,500 questions for ten of the eleven laboratories. Meta's collection reached 272 valid answers before access was revoked.

I also tested the cross-regime concern directly. Re-grading the July sixteen-model wave with the second wave's frozen deterministic engine (which replaces the hybrid pipeline that adjudicated 66.5% of verdicts with a local open-weights grader) changes 2.65% of resolved verdicts (566 of 21,338; the engine abstains on a further 4.1%) and moves no headline: under a single deterministic coding the certification lifts read 3.6×/4.6×/4.4× before re-grading and 3.0–3.1×/4.5–4.6×/4.3× after. Both are above the conservative published 1.9–3.2×, which derive from the LLM-adjudicated pair coding, since more agreements certified means more

conservative lifts. Every AUROC in Tables 6–7 shifts by at most 0.016 on the CN and cross-bloc panels and 0.046 on the US panel, preserving each panel's ordering of agreement versus self-reported confidence.

The flagship label guarantees nothing. Resolved accuracy on identical questions spans 11.4% to 95.4% across the eleven flagships (Table 9), an 84-point spread at the top of the market, wider than the mid-tier spread of the first experiment. And the accuracy illusion survives intact at this tier: pooled across the ten complete laboratory arms, **46.6% of all resolved answers offered at stated confidence ≥ 80 were wrong** (4,910 of 10,540; including the partial Meta arm, 46.3%, or 4,935 of 10,665; unresolved items excluded here and re-scored as errors in sensitivity, where the pooled picture is unchanged). The first wave's consumer-facing implication does not soften with price: a flagship badge does not tell the consumer whether the confident answer in front of them came from the 95%-accurate flagship or the 11%-accurate one.

Table 9. *Flagship wave: coverage and accuracy (SimpleQA subsample, n = 1,500 per laboratory, temperature 0)*

Laboratory (flagship)	Coverage	Accuracy (resolved)	Unresolved
OpenAI gpt-5.6-sol	1,500/1,500	95.4% (1,403/1,471)	29
Google gemini-3.1-pro	1,500/1,500	71.1% (837/1,178)[1]	46
Anthropic claude-fable-5	1,500/1,500	70.3% (1,019/1,450)	50
Meta muse-spark-1.1 (partial)	272/1,500 valid	61.4% (148/241)[2]	10
xAI grok-4.5	1,500/1,500	54.8% (794/1,448)	52
DeepSeek deepseek-v4-pro	1,500/1,500	46.6% (666/1,428)	72
Moonshot kimi-k3	1,500/1,500	45.5% (645/1,418)	82
Zhipu glm-5.2	1,500/1,500	30.2% (430/1,425)	69
Baidu ernie-4.5-vl-424b	1,500/1,500	28.9% (411/1,421)	71
Mistral mistral-medium-3.5	1,500/1,500	21.4% (306/1,429)	71
Alibaba qwen-max	1,500/1,500	11.4% (164/1,435)	65

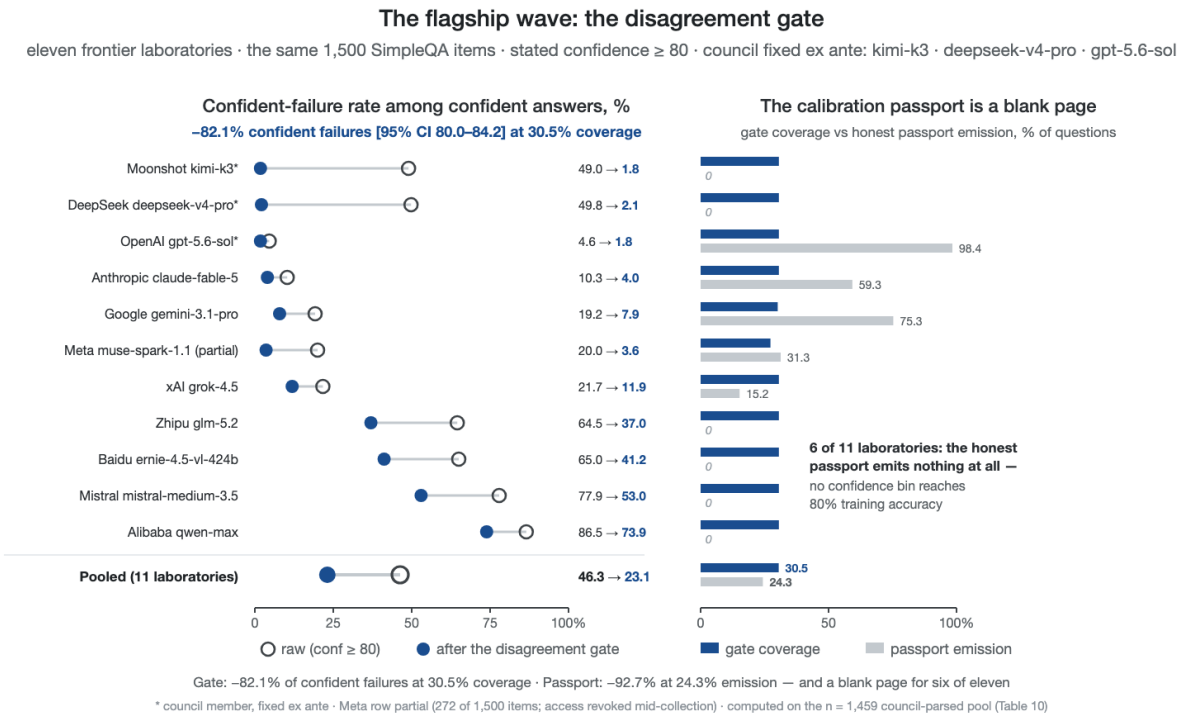
[1] Google's channel returned empty or filtered responses on 18.4% of items (276 of 1,500 — a delivery property, not a coverage gap: all 1,500 were attempted); accuracy is computed on parsed, resolved answers. Both definitions are reported: answered-only, Google's accuracy is 71.1% (837/1,178); counting the 276 undelivered responses as errors it is 57.6% (837/1,454), and 55.8% (837/1,500) with unresolved items also scored as errors. [2] Meta accuracy computed on the answers collected before access revocation; see coverage limitations. Accuracy throughout the table is resolved-only; under the unresolved-as-error sensitivity the top figure, OpenAI's 95.4%, becomes 93.5% (1,403/1,500). Per-call channel provenance for every laboratory is recorded in the run manifest. Unresolved counts delivered answers that did not resolve deterministically; for Zhipu (1,494 answers delivered), Baidu (1,492) and the partial

Meta arm (251) the two columns therefore sum to fewer than the attempted questions, the remainder being undelivered responses.

A contamination caveat scopes the absolute accuracies. This wave reuses the original, publicly released SimpleQA benchmark (October 2024) against models whose training corpora postdate its publication by up to two years, so training-time exposure to the benchmark itself is plausible at the top of the table, and the data carry a concrete signature of it: the top-scoring flagship reproduces the gold answer string verbatim (normalized exact match, down to decimal formatting) on 61.6% of items, against 27.9–45.3% for the other laboratories, at a median stated confidence of 99. The design cannot separate "the model is stronger" from "the model has seen the benchmark and its answer key." The wave's absolute accuracy figures are therefore not comparable to historical closed-book ceilings reported for earlier model generations, and they should not be read as measures of parametric memory. The first wave's benchmark-age check (the TruthfulQA–SimpleQA gap reported in Section 3) covered the July collection; by August 2026, SimpleQA itself is a two-year-old public benchmark, and this caveat covers the flagship wave. The wave's conclusions do not rest on absolute accuracy. They rest on the structure of confident failures, their prevalence at stated confidence ≥ 80 and their interception by the disagreement gate, measured within the wave under one frozen protocol, where any exposure advantage is shared across the laboratories evaluated on the same items. A fresh two-hop probe constructed after the study window confirms this signature directly. See "Why not simply use the best model?" later in this section.

The calibration moat. The first experiment showed that independent agreement carries a correctness signal that self-reported confidence does not. The flagship wave turns that signal into a measured filter. The mechanism is a disagreement gate: a council of three laboratories, fixed in advance from a prior disagreement analysis and not fitted to these outcomes (Moonshot kimi-k3, DeepSeek deepseek-v4-pro, OpenAI gpt-5.6-sol), and a rule with no free parameters. An answer passes the gate only when the three council answers agree pairwise. Because nothing is estimated, no train/test split is needed; the gate is computed once over the frozen collection.

Figure 17. *The flagship gate*



Confident failures at stated confidence ≥ 80 and their interception by the fixed three-model disagreement gate, by laboratory (August 2026 wave).

Pooled across the ten complete laboratory arms (the primary presentation, with the partial Meta arm reported separately and its inclusion treated as a sensitivity), the raw confident-failure rate is 46.6% (4,910/10,540). Pooled across the eleven-laboratory roster, per-laboratory rates vary (Table 9), so this figure characterizes the frontier roster rather than a universal constant. The gate removes **82.1% of confident failures** (bootstrap 95% CI [79.9; 84.1]) **while keeping 30.5% of all pooled answers** (the gate itself opens on 30.6% of questions, 446 of 1,459). Restricted to the complete laboratories outside the council, with no self-reference of any kind, the gate still removes 76.9% of confident failures (CI [74.3; 79.2]). Including the partial Meta arm leaves every estimate effectively unchanged: pooled across all eleven, the raw rate is 46.3% (4,935/10,665; 35.8% of confident answers pass the gate, and correctness among emitted confident answers rises from 53.7% raw to 76.9% post-gate), the gate removes 82.1% (CI [80.0; 84.2]) at the same 30.5% coverage, and the eight-laboratory non-council figure is 77.0% (CI [74.4; 79.3]). Table 10 gives the per-laboratory figures.

Table 10. *Confident-failure interception by the disagreement gate (conf ≥ 80; deterministic grader; council = kimi-k3 · deepseek-v4-pro · gpt-5.6-sol, fixed ex ante)*

Laboratory	Confident answers	Confident failures, raw	After gate (rate on emitted)	dCF
Moonshot kimi-k3 [council]	1,056	517 (49.0%)	7 (1.8%)	−98.6%
DeepSeek deepseek-	1,274	634 (49.8%)	9 (2.1%)	−98.6%

v4-pro [council]				
OpenAI gpt-5.6-sol [council]	1,426	65 (4.6%)	8 (1.8%)	−87.7%
Anthropic claude-fable-5	825	85 (10.3%)	15 (4.0%)	−82.4%
Google gemini-3.1-pro	966	185 (19.2%)	27 (7.9%)	−85.4%
Meta muse-spark-1.1 (partial)	125	25 (20.0%)	2 (3.6%)	−92.0%
xAI grok-4.5	577	125 (21.7%)	36 (11.9%)	−71.2%
Zhipu glm-5.2	1,128	727 (64.5%)	146 (37.0%)	−79.9%
Baidu ernie-4.5-vl-424b	725	471 (65.0%)	107 (41.2%)	−77.3%
Mistral mistral-medium-3.5	1,354	1,055 (77.9%)	224 (53.0%)	−78.8%
Alibaba qwen-max	1,209	1,046 (86.5%)	300 (73.9%)	−71.3%
Pooled (11 laboratories)	10,665	4,935 (46.3%)	881 (23.1%)	−82.1% [80.0; 84.2]
Pooled (8 non-council)	6,909	3,719 (53.8%)	857 (33.5%)	−77.0% [74.4; 79.3]

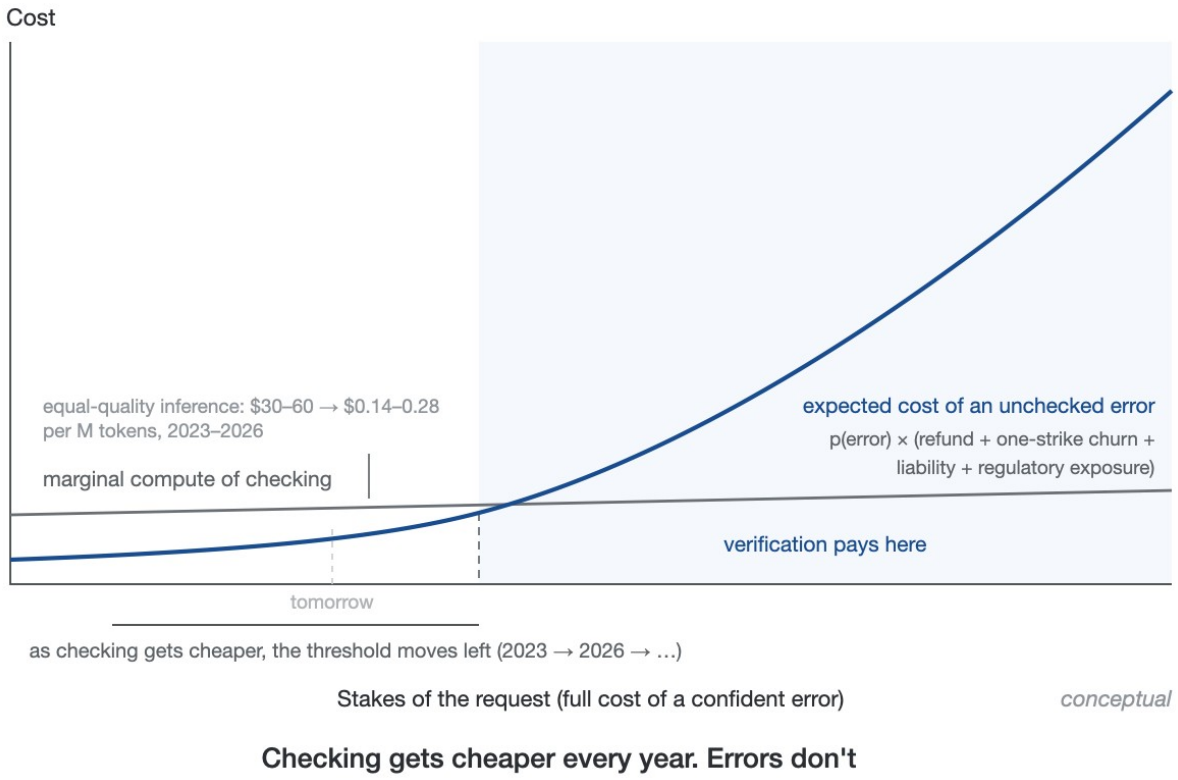
Computed on the pool of questions where all three council members produced parsed, non-refused answers (n = 1,459 of 1,500); the gate opens on 446 of these (30.6%). Meta row is partial (access revoked mid-collection). Bootstrap CIs: 5,000 question-level resamples.

Three things go with that headline. The gate also discards roughly half of the confident *correct* answers (the gate retains only 51.3% of confident-correct answers; of the answers it gates out, roughly 41% were in fact right). The trade is coverage for certainty, the canonical risk–coverage trade-off of selective classification (Geifman & El-Yaniv, 2017), executed here at the architecture layer, and the two numbers should always be quoted as a pair: −82.1% confident failures at 30.5% coverage. The system has a floor of its own. In the fully systemic mode, where the open gate emits the council's consensus answer rather than the individual model's, the system's own confident-failure rate is 1.8% (8 of 442 resolved consensus answers), and those eight questions are ones where the entire council confidently agrees on the same wrong answer. That is the correlated-error floor the first experiment predicted from its consensus-certifies-error boundary condition: below it, no amount of agreement-checking helps, because the agreement itself is the error. And the result is flat across the pre-declared sensitivity grid: confidence threshold 70/80/90 gives pooled dCF of −83.1%/−82.1%/−80.6%, and re-scoring unresolved items as errors gives −81.8%, so the number was measured, not selected.

What the 30.5% coverage figure means commercially is set by the break-even economics of Section 5 (Figure 14): the roughly seventy percent of answers the gate declines to certify are requests routed to a cheaper fate than a confident error (escalation to a human, a deeper

verification tier, or an explicit uncertainty display) rather than lost conversions, so the coverage cost is an insurable price of error, not foregone demand. No cost model for the escalation channel is estimated here. The operating point is not fixed. The agreement ladder traced in Figure 19 supplies a family of thresholds, and a service should sit where the expected cost of an uncaught confident failure crosses the margin on the traffic the gate withholds, which makes it a cost-based choice.

Figure 14. *Economics of graduated trust*



The break-even logic of verification.

Why not simply use the best model? The strongest baseline the wave leaves standing is commercial rather than statistical: an operator could skip the architecture and route everything to the wave's most accurate flagship. That baseline needs an answer with numbers. At mid-2026 prices the frozen three-lab council costs $0.0039 per question, less than a single answer from the wave's most expensive flagship (claude-fable-5, $0.0198/question at list) and more than one from its most accurate (gpt-5.6-sol, $0.0027/question at published API rates). "Just buy the best flagship" is a genuine Pareto point: gpt-5.6-sol solo answers every question at a 4.6% confident-failure rate, and no gated configuration matches its coverage at any price. Wherever a service can tolerate confident failure on roughly one answer in twenty and needs full coverage, the flagship wins on both axes. The rule generalizes poorly, however, even at the top of the leaderboard: the wave's other consensus front-runner, claude-fable-5, carries a 10.3% confident-failure rate solo. "Just use whichever model tops the benchmarks" fails outright on the second-strongest model in

the sample, which is why the identification problem below is not hypothetical. Four considerations bound the comparison.

First, the identification problem: knowing *which* model is currently best is itself the product of external, cross-model measurement. The wave's own Table 8 selected its flagships from public leaderboards, continuous, third-party, cross-vendor comparisons that no one proposes replacing with vendor self-report. This paper's own data show why no one does: stated self-confidence discriminates at AUROC 0.53–0.79 depending on vendor, and an 84-point accuracy spread separates models all sold as their laboratory's best. "Route to the best model" is therefore an application of architectural trust, not an alternative to it: the leaderboard that crowns the flagship is an independent panel, blind items, and a published record, the three components of Section 3 operated at market scale. And that crown moves: enterprise model leadership changed hands in a single spring (Kharazian, 2026), models are updated silently under stable names, and agent purchasing behavior reshuffles drastically across model updates (Allouah et al., 2026). Without continuous external measurement, "the best model" is not available to the operator as knowledge.

Second, the 95.4% is partly memory rather than capability: the top flagship reproduces the gold answer string verbatim on 61.6% of items (Table 9's contamination caveat), so its measured edge compresses on questions the benchmark cannot have taught it. The gate mechanism does not inherit that fragility: on the stratum least likely to have been memorized it removes 89.4% of confident failures, its best performance rather than its worst (the contamination-stratified re-analysis below), because agreement among independently trained laboratories is not conditioned on any one vendor's training corpus.

To separate transferable knowledge from benchmark exposure, I composed a fresh 100-item two-hop probe that is guaranteed uncontaminated, machine-generated on 6 August 2026 from pairs of independently verified pre-2023 Wikidata facts (e.g., "In what year was the architect of [minor building] born?"), each with a single computed answer, so the composed question strings cannot have appeared in any training corpus. The set's SHA-256 was frozen before the first model call.

Under the wave's frozen protocol (verbatim system prompt, temperature 0, same deterministic grader), gpt-5.6-sol collapsed from 95.4% accuracy with a 4.6% confident-failure rate on SimpleQA to 33.0% [bootstrap 95% CI 24.0–42.0] with confident failures on 61.0% [51.0–70.0] of its items (67.0% of its confident answers), its mean stated confidence barely moving (98.4 → 93.7), while claude-fable-5 held at 76.0% [67.0–84.0] and 6.0% [2.0–11.0]. The wave's 25-point flagship gap inverts into a 43-point deficit on guaranteed-unseen compositions. The council gate, unchanged from the wave, absorbed this failure: it removed all but one of gpt-5.6-sol's 61 confident failures (60/61; at n = 61 the point rate is imprecise, Wilson 95% CI ≈ 91–99.7%; 24.2% coverage; claude-fable-5: 6 → 1), cutting the emitted confident-failure rate among passed high-confidence answers to 4.5% against a raw 67.0%.

Levels are not point-comparable across benchmarks: two-hop composition is intrinsically harder than one-hop retrieval, and at n = 100 the intervals are wide. But the reversal of the model ordering and the gate's persistence on items no model can have memorized are exactly the signatures the contamination account predicts, and they are the failure mode an architectural trust layer, and not a better model choice, is built to absorb.

Third, there is a floor. The gate reaches a confident-failure rate no single model attains at any price: council-gated configurations cut confident failures on emitted answers to 1.8–2.1% at $0.0039–0.0046 per question, 2.5× below the best solo rate, while the most expensive flagship in the wave sits at 10.3% confident failures for $0.0198 per question, dominated on every axis. The gate is not a rescue device for weak models: qwen-max behind the same gate still fails confidently on 73.9% of its emitted confident answers, so the frontier's gated points are councils of strong models, not gates over weak ones.

Fourth, when the agent is spending money autonomously, the operative question shifts from "which model is smarter" to "which configuration can prove what it did." However accurate it is, a solo flagship cannot notarize its own decision or arbitrate its own conflict of interest. The gated configuration produces an inspectable agreement record per answer. That is the artifact Section 5's liability and insurance markets already price, where coverage is being written against evidence and exclusion-shaped against its absence.

The resolution is a frontier, not a winner. The flagship is the council's strongest member (the frozen council includes it), yet it cannot serve as its own judge. The leaderboard that identifies it already rests on the same principle architectural trust operationalizes, independent cross-model verification, but it supplies that verification only in aggregate, about models in general, not about the answer in front of a consumer. The construct fills that gap, and the gap also bounds it: a claim resting on no independent check is outside architectural trust, not an instance of it. The operator's real choice is between risk classes: full coverage at 4.6% confident failures, or 31% coverage at 1.8% with an audit trail, the selective-classification trade, priced here. The full 26-configuration cost–coverage–risk frontier (every solo flagship, every gated configuration, and the pooled agreement ladders, with the billed-versus-list price basis flagged per point) is released as a CSV with the replication package.

The obvious rival to the disagreement gate is single-model calibration repair: learn each model's confidence-to-accuracy mapping and emit only where mapped accuracy clears a bar, the post-hoc recalibration program whose standard remedies are temperature and Platt scaling (Guo, Pleiss, Sun, & Weinberger, 2017). Fit on disjoint halves (histogram-binned confidence, train on even-indexed questions, evaluate on odd, bins under 20 observations not emitted), this calibration passport removes 92.7% of confident failures, but at 24.3% emission, and **six of the eleven flagships emit nothing at all**: no confidence bin of theirs reaches 80% training accuracy, so the passport for those models is a blank page. For half the flagship market, single-model calibration cannot be repaired from the outside. The disagreement gate holds roughly 30% coverage for every model in the table, including the six the passport abandons. So the first experiment's conclusion holds at this tier too: architecture can supply an external correctness signal where a model's exposed confidence fails as one, in ranking or in level, measured again at the tier consumers are told to trust most.

Robustness of the gate effect. Two referee-driven checks matter: circularity, whether the gate rides on its strongest member, and whether something trivial does the same work. Replacing the council's strongest member (gpt-5.6-sol, 95.4% resolved accuracy) with the wave's fourth ex-ante frontier run (grok-4.5) leaves the result essentially unchanged. The sol-free council removes 81.2% of pooled confident failures at 29.7% coverage (non-council targets: 77.6%) against 82.1% at 30.5% for the frozen trio, and across all 165 possible three-model councils the

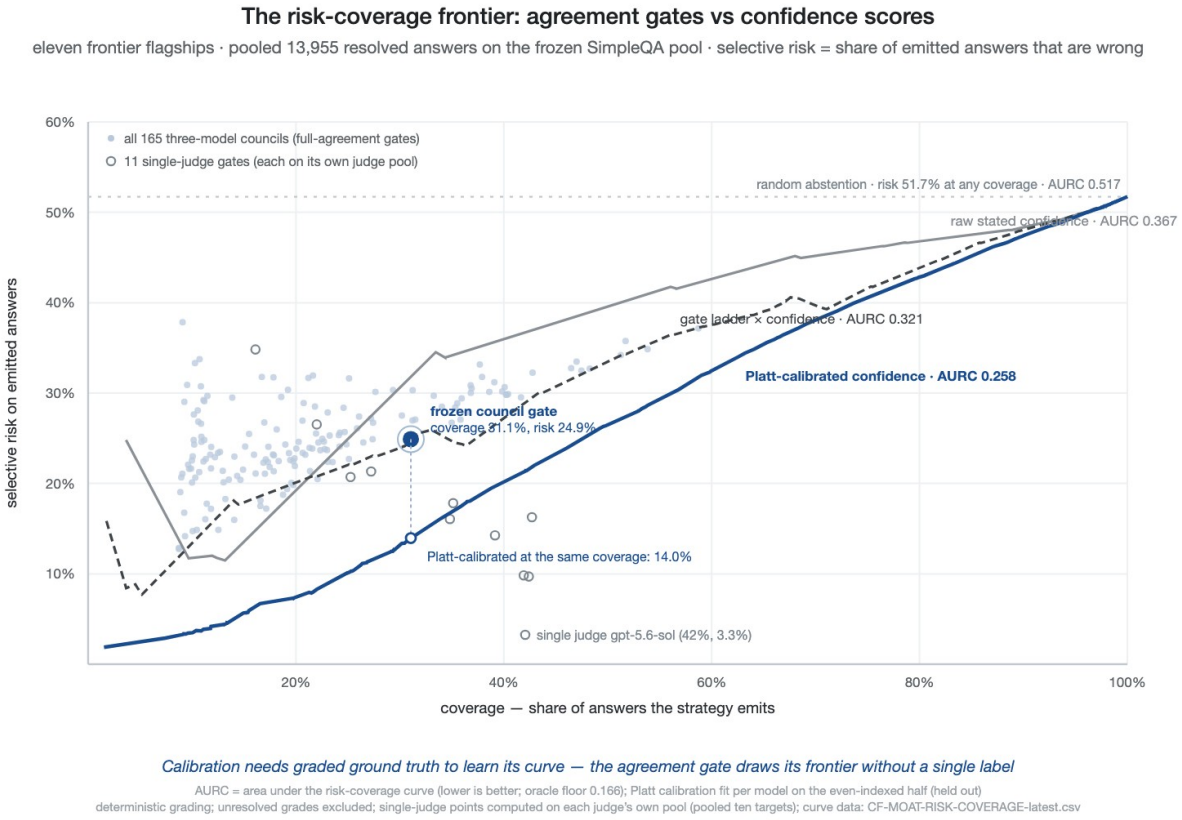
mechanism never fails: every council removes more confident failures than random abstention at its own coverage (median excess +8.5 points), with the ex-ante trio in the top 4% by that margin (+12.6 points). The effect belongs to cross-model agreement, not to any one member's accuracy. Thresholding each model's own stated confidence to the identical per-model coverage, the trivial baseline, removes only 66.7% of pooled confident failures, less than the 69.5% that blind random abstention at the same coverage removes, because confident failures are by construction concentrated at high stated confidence. That threshold also keeps fewer confident-correct answers than the gate (46.4% vs 51.3%). The gate Pareto-dominates both baselines for every one of the eleven laboratories. Using per-answer agreement with the strongest model alone as the filter is an even stronger screen (97.1% of pooled confident failures removed at 40.7% coverage, keeping 97.0% of confident-correct answers), but it is a mechanically different, target-conditioned rule. It passes an answer only after inspecting it, whereas the council gate opens the same fixed question set for every model without looking at the answer under test, so it bounds the council effect without explaining it: without sol the trio still removes 81.2%, and even agreement with a weak single judge (kimi-k3, 45.5% resolved accuracy) removes 86.3% of confident failures. The driver is cross-model agreement as such, while judge accuracy shows up in what is kept (kept-correct 97.0% for sol against 20–63% for the weaker judges).

Calibrated abstention supplies the stronger competing baselines, and they were run too. Two calibrated abstention baselines were fit under the same split as the calibration passport (parameters fit on even-indexed items, evaluated on the held-out odd half, where the frozen gate removes 82.2% of confident failures at 30.8% coverage). The per-model baselines do not reach the gate. Platt-scaled confidence with per-model coverage matched to the gate removes 65.3% of confident failures, and split-conformal abstention ($\alpha = 0.293$, calibrated to guarantee $P(\text{emit} \mid \text{wrong}) \leq \alpha$ and tuned to the same coverage) removes 70.8%, as does a raw stated-confidence top-K control at the same pooled coverage, which removes 69.4%. A pooled variant that ranks all eleven laboratories' answers on one Platt-calibrated probability scale does surpass the gate (89.0% at identical pooled coverage), but it does so by reallocating essentially all coverage to the well-calibrated laboratories: it emits 100% of the best-calibrated vendor's answers and 0% for four of the eleven laboratories, reproducing in softer form the passport's failure mode above. It also requires roughly seven hundred graded answers per model to fit, where the gate requires none. Calibrated confidence and cross-model agreement are therefore complements, not substitutes: the first buys a lower pooled error rate wherever ground-truth labels exist to train it, and the second buys uniform, label-free per-laboratory protection.

Figure 19 places every strategy on the selective-prediction plane (pooled over the 13,955 deterministically resolved answers that the eleven arms returned on the frozen 1,459-question pool, so the total is smaller than Table 9’s, which counts resolved answers over all 1,500 questions; selective risk = share of emitted answers that are wrong). Ranked by area under the risk-coverage curve, lower better: Platt-calibrated confidence 0.258, the gate ladder refined by stated confidence 0.321, the pure agreement ladder 0.356, raw stated confidence 0.367, random abstention 0.517, against an oracle floor of 0.166; held-out-half values differ by at most 0.009. At the frozen gate's own operating point (31.1% coverage in this resolved-answer convention, 4,334 of 13,955 answers emitted; the 30.5% and 30.6% reported elsewhere are the shares of confident answers and of pool questions) the gate's risk is 24.9%, against 33.4% for a raw-confidence threshold and 14.0% for calibrated confidence, subject to the label-cost and coverage-concentration caveats above. The cloud of all 165 three-model councils spans coverages of 9–

59% at risks of 13–38%, so the reported gate is a representative member of its family, not a tuned outlier. Single-judge gates trace the degenerate end of the family, where "agreement" collapses into "ask the most accurate model."

Figure 19. *The risk-coverage frontier*



Every abstention strategy on the selective-prediction plane, coverage against selective risk, pooled over the 13,955 deterministically resolved flagship answers returned by the eleven arms on the frozen 1,459-question pool (Table 9 counts resolved answers over all 1,500 questions): the raw-confidence, Platt-calibrated, and gate-ladder-×-confidence curves with their AURC values, the random-abstention line, the cloud of all 165 three-model councils, the eleven single-judge gates, and the frozen council's operating point (31.1% coverage in the resolved-answer convention, 4,334 of 13,955; 24.9% risk). Lower is better; the oracle floor is 0.166. Every point is recomputable from the replication package's risk-coverage CSV.

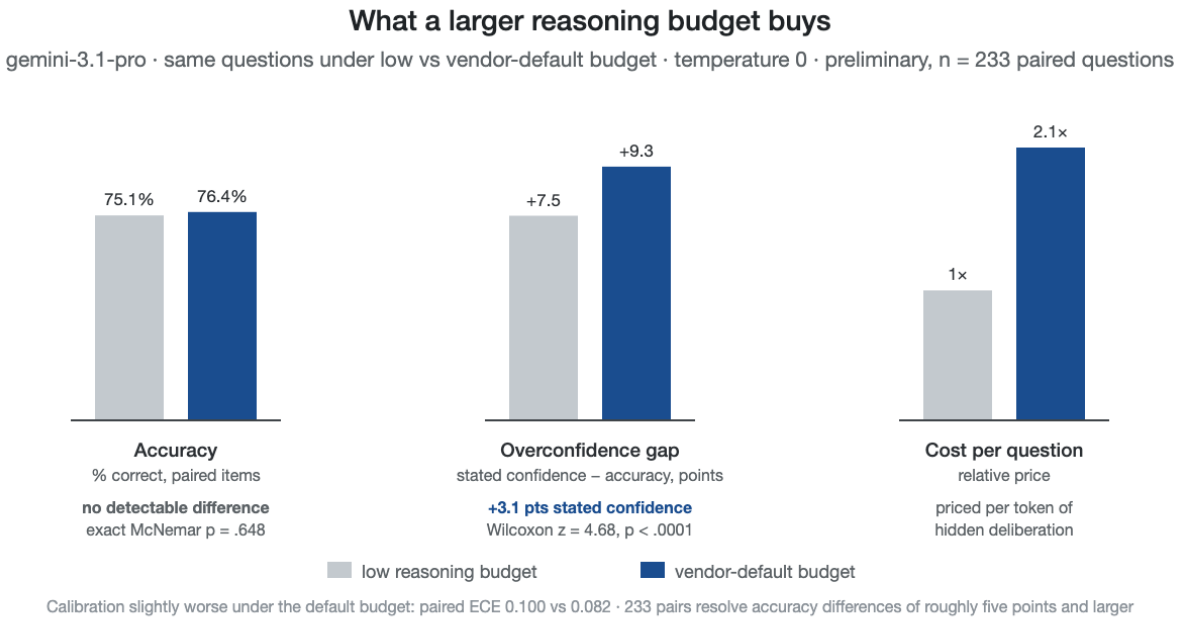
Agreement might carry no information beyond question easiness. To test that, a logistic regression of answer correctness on gate agreement and item difficulty (the share of wrong answers among the other models on that question), with laboratory fixed effects and question-clustered standard errors, was run, and the answer depends on how difficulty is measured. With difficulty measured by the non-council models only (the fair control, since a difficulty measure that includes the council's own answers embeds the gate's signal by construction), agreement retains an independent contribution on the external-laboratory targets: b = +0.76, odds ratio 2.15, z = 17.2. An answer emitted by the gate has roughly twice the odds of being correct as an equally difficult answer the gate would withhold. Fold the council's own correctness into the difficulty regressor instead and the agreement coefficient collapses to zero, which quantifies the

mechanism rather than refuting it: agreement and multi-model difficulty are two measurements of the same latent quantity, and the gate is a label-free instrument for reading it at inference time.

A contamination-stratified re-analysis addresses the exposure concern directly. Flagging a question as likely benchmark-exposed when at least three of the eleven laboratories reproduce the gold string verbatim after normalization (741 of 1,459 pool questions; the same signature that yields gpt-5.6-sol's 61.6% gold-echo rate), the gate removes 89.4% of confident failures on the *non-exposed* stratum (2,793→296; non-council targets: 86.5%) against 72.7% on the exposed one. So the mechanism does not ride on memorized items. It behaves as a selective filter should, opening far less often where the panel is genuinely uncertain (18.7% vs 41.9% coverage) and holding the emitted confident-failure rate nearly constant across strata (25.9% vs 21.8%, from raw rates of 58.5% and 36.4%). Exposure cutoffs of 2–5 echoing models give non-exposed dCF of 87.7–89.7%, so the finding is threshold-robust. The stratification is descriptive: gold-echo is outcome-coupled, and it separates likely-seen from other questions rather than identifying contamination causally.

What a larger reasoning budget buys: confidence, not competence. Flagship vendors now sell a second dial above model choice: a reasoning budget, priced per token of hidden deliberation. If extended reasoning bought calibration, it would be a model-level answer to this paper's problem, purchasable at checkout. A paired experiment on one flagship says it buys something else. Google's gemini-3.1-pro (whose reasoning cannot be disabled) answered the same frozen question set under two budgets, a low reasoning-effort setting and the vendor-default dynamic budget, with temperature zero, identical prompts and parser, and per-question cost recorded. On the 233 questions answered under both budgets at this version's date: accuracy 75.1% (low) versus 76.4% (default), no detectable difference (exact McNemar $p = .648$); stated confidence 3.1 points *higher* under the default budget (Wilcoxon $z = 4.68$, $p < .0001$); calibration slightly *worse* under the default budget (paired ECE 0.100 vs 0.082; confidence-minus-accuracy gap +9.3 vs +7.5 points); price per question 2.1×. The accuracy comparison is a null of a specific kind: these data show *no detectable accuracy difference at n = 233* (exact McNemar $p = .648$), and equivalence is formally established only at a coarse margin. A paired TOST bounds the accuracy difference within ±5 points (difference +1.3 points, 90% CI [−1.8; +4.4]; both one-sided $p \leq .024$) but the study is not powered to establish equivalence at ±2–4 points, while the +3.1-point rise in stated confidence is statistically unambiguous (Wilcoxon $p < .0001$). Within that bound, the doubled spend purchased additional stated confidence and no measurable additional correctness: overconfidence as a paid feature, and a within-model instance of the migration thesis. The extra dollars went to the model layer and produced nothing a consumer could verify, while the architectural layer above it turned confidence into a checkable signal. The result is a pilot, reported as one: the comparison covers 233 paired questions, and the full 1,500-pair design narrows the equivalence bound to about two points; channel composition and token-cap censoring are documented as confounds in the replication package; collection is continuing.

Figure 18. *What a larger reasoning budget buys*



Paired comparison under two reasoning budgets — stated confidence rises, accuracy does not.

Table 11. *Reasoning budget, paired comparison (gemini-3.1-pro, 233 questions answered under both budgets, temperature 0; preliminary)*

	Low reasoning budget	Vendor-default budget
Accuracy (paired items)	75.1%	76.4% (McNemar p = .648)
Mean stated confidence	−3.1 points vs default	+3.1 points (Wilcoxon p < .0001)
Calibration (paired ECE)	0.082	0.100
Confidence − accuracy gap	+7.5	+9.3
Cost per question (relative)	1×	2.1×

Reasoning-effort settings were not identical across arms: each vendor was queried through its standard commercial interface at that interface's default effort (settings logged per-model in the replication package), so a vendor's absolute accuracy can carry an effort component. That does not reach the mechanism. The accuracy ranking is not load-bearing (the flagship label guarantees nothing), the gate survives removing any member (sol-free council removes 81.2%), and a larger effort lifts stated confidence with accuracy flat, as the paired comparison just showed, so any accuracy it inflates surfaces as confident error, the failure the gate intercepts.

Coverage limitations of the flagship wave. The wave is a frozen 5 August 2026 snapshot, so models released afterward, such as Grok 4.6 (12 August) and later Artificial Analysis leaders such as Claude Opus 5, postdate it and cannot enter a pre-committed benchmark. The top position turned over within days. That is the identification problem the gate sidesteps (agreement is roster-robust, surviving removal of any member). Two included laboratories fall below the

current frontier on general-capability leaderboards: Mistral's flagship trails the top tier on every aggregate index checked, and Baidu's general-purpose ERNIE ranks outside the top 30 on Humanity's Last Exam (its strength is search-augmented and agentic tasks, where ERNIE 5.1 ranks #4 on LMArena Search Arena). They are retained to preserve coverage of the European ecosystem and of all major Chinese laboratories. Of the two top-tier laboratories that were instead outside the wave's scope, Meta Superintelligence Labs (Muse Spark 1.1; #4 on Humanity's Last Exam) and ByteDance (Seed 2.1; #10), one was subsequently attempted and interrupted (Meta, below) and one could not be reached at all (ByteDance, below).

ByteDance's Seed 2.1 could not be included: BytePlus ModelArk returned "not available in your country/region" for a freshly registered international account (live attempt, August 5, 2026), no aggregator serves the flagship, and the domestic Volcano Ark channel requires mainland-China verification. Baidu's ERNIE 5.1, the current flagship, is similarly restricted to the domestic Qianfan platform (live attempt, August 5, 2026). Where a laboratory's current flagship is inaccessible, the fallback rule is the strongest openly released model from that laboratory. For Baidu this is ERNIE 4.5 VL 424B (open-weights), which is used instead. That is a finding in itself: the current flagships of two of the six major Chinese laboratories are not accessible to external researchers as of August 2026.

Meta is the wave's third access failure, and the obstacle was my account, not geography. In the first (mid-tier) wave, Meta was absent for a documented infrastructure reason: its first-party Llama API was retired on July 6, 2026, before the protocol freeze (recorded in the run manifest), and the laboratory was represented on the local tier by its open-weight Llama-3.2-3B, which proved one of the panel's most instructive members (the lowest error correlation in the entire matrix). For the flagship wave, Meta's muse-spark-1.1 was added mid-wave when it became reachable, and collection then failed three ways in sequence, each logged: (i) the first-party Meta Model API key died mid-run on August 5 with 251 valid answers collected, and the developer console would not issue a replacement; (ii) an aggregator fallback route yielded 21 further answers before (iii) Meta's provider endpoint began returning HTTP 403 `user_blocked`, "access has been restricted due to repeated policy violations," for closed-book factual-recall questions from a public benchmark, asked at temperature zero. The net result is 272 of 1,500 items (18.1%) collected and reported as partial: the frozen pooled analysis includes the 251 first-party answers (the aggregator top-up landed after the analysis freeze), and Meta's per-laboratory estimate (−92.0% dCF on 125 confident answers) is directionally consistent with the panel but too imprecise for inference. The block is itself a finding: as of August 2026, an independent researcher cannot complete a 1,500-question factual evaluation of Meta's flagship through any authorized channel. Between them the wave's three access failures, a region-locked flagship (ByteDance), a domestic-only flagship (Baidu), and a researcher ban (Meta), are the theoretical argument in empirical form: verifiability at the model layer is enjoyed at vendor grace, and an architectural trust layer cannot assume vendor grace.

4. An extreme deployment: delegation with irreversible stakes

The usual way to study consumer trust is a survey or a lab task. I had a harsher instrument available: an extended, high-stakes deployment in which every consequence fell on me. This section is an existence proof and design testimony from a single, conflicted operator, not empirical evidence of generality, and none of P1–P5 rests on it. The two kinds of evidence

interlock. In the laboratory the calibration effect was measured under placebo control; in the deployment, where errors were irreversible, the same three properties were the difference between delegation being possible and being reckless.

The deployment is not a client's and not a simulation: I am simultaneously the operator, the developer, and the analyst. The evidence layer is hash-anchored and every measurement independently recomputable, so the conflict bears on the advocacy and on what I chose to measure, not on whether the numbers reproduce. The setting combines ordinary consumer competence with irreversible stakes, where every architectural weakness gets punished quickly and visibly.

Methodologically this is an extreme, revelatory single case (Yin, 2018) in the interpretive and analytic-autoethnographic tradition (Walsham, 1995; Gould, 1991; Anderson, 2006), with that tradition's evidentiary cautions in view (Wallendorf & Brucks, 1993): records first, interpretation marked as interpretation. The justification is access: a researcher cannot ethically assign participants to bet irreversible real-world outcomes on an AI architecture, so the only available observer is a participant who chose the bet himself.

The analysis relies on contemporaneous, hash-anchored logs. The deployment's evidence layer is itself the record: load-bearing steps (which models were consulted, where they diverged, what was gated, what I confirmed) were logged at the time, so the observations rest on replayable records and not on memory alone. Hindsight bias cannot be removed from the interpretation, so I mark each place where interpretation exceeds the record. The deployment-period records of the system's evidence layer are hash-anchored as of 2 August 2026: a per-file SHA-256 manifest of 129 files (7,414,043 bytes) whose root is 2f88a16611e51c0c12c8015754c89269e782ea167a89b5ed2a881f18eee4defc, and the anchor is published with this paper. The records remain sealed until the deployment closes, because selective quotation from a live deployment would be worse than none. The commitment can be checked: any record released later either matches the anchored manifest or it does not, so retroactive editing, substitution, or back-dating is detectable by any reader of this page. A fuller quantitative treatment of the telemetry is left to future work, for the same reason.

Three observations from living inside this loop are offered as hypotheses about what transfers beyond the original setting, feeding the propositions of Section 3, and the events they rest on were logged by the deployment at the time.

Disagreement is a consumer-usable signal. When independent models diverged on a consequential question, the divergence itself, surfaced instead of suppressed, told me where the genuine uncertainty lay: a free early warning no single-model service can provide.

Evidence changes behavior before it ever changes a dispute. Knowing that every consequential step was recorded made delegation psychologically viable at all; the record was the precondition of reliance, not an audit afterthought.

Confidence gating matters more than average accuracy. The practically dangerous events were the confident errors, never the mediocre outputs. The architecture's real contribution was to make

high confidence expensive: an answer had to survive cross-model agreement to earn it, instead of getting it for free from one model's fluency.

5. Implications: from persuasion to infrastructure

Design and experience

For service design. The unit of trustworthy AI service design is shifting from the model to the pipeline. That means routing consequential queries through independent models, treating disagreement as a first-class UX element, attaching evidence by default, and pricing multi-model processing against the asymmetric cost of confident errors, with the expensive path reserved for requests where stakes concentrate, as fraud screening reserves scrutiny for risky transactions. The infrastructure already exists: cost-driven cascades that escalate hard requests to expensive tiers are standard agent engineering, and for a service that already routes by difficulty, the marginal architecture of routing by *consequence* is close to zero. The failure modes both experiments measured leave two hard rules, and they are design requirements, not preferences. Positions must be collected blind and recorded *before* any cross-model deliberation, since a deliberation-first pipeline manufactures the very consensus the architecture is supposed to police (the conformity pressures of Section 3). The second rule: unanimous confident consensus over a constrained option set is a red flag for review, not a green light for execution: both the model panels and the human participants of Section 6 treated exactly that consensus as certification.

The economics work because escalation is graduated, a *graduated trust architecture*: in the deployment described here, processing tiers span roughly three orders of magnitude in cost per request between a single-model answer and a full multi-model deliberation (as of July 2026), so the expensive tier is bought only where consequences concentrate. The break-even logic is fraud scoring's (Figure 14). Orchestration pays whenever the probability of a confident error times its full cost (refund, churn under one-strike tolerance, liability, regulatory risk) exceeds the marginal compute of checking. For high-stakes transactions that inequality is not close.

Falling inference prices (roughly two orders of magnitude at fixed capability in three years, a conservative reading of tracked declines; Appenzeller, 2024) do not dissolve the argument, because the ratio, not the absolute cost, is the structural quantity: both tiers ride the price curve down together, and cheap inference makes the expensive tier affordable at ever-lower stakes. Checking rides its own second curve: ensembles of weak verifiers, with no labeled data, reach the answer-selection accuracy of frontier reasoning models from open 70-billion-parameter components (Saad-Falcon et al., 2025). Given one-strike consumer tolerance, the comparison is not "cheap answer versus expensive answer" but "expensive answer versus lost customer."

Two consumer-facing primitives complete the design: deliberate friction (a marked pause before irreversible actions) and receipts. Both intervene on how delegation *feels*. Consumers experience AI sociologically, not statistically (Puntoni, Reczek, Giesler, & Botti, 2021), and as delegated execution removes the service encounter from view, the receipt inherits the encounter's old role as the moment of truth. A risk travels with that framing: making disagreement salient can *advertise uncertainty* and depress trust rather than calibrate it (Detjen et al., 2025; Lu, Wang, & Yin, 2024), so how much divergence to surface, and how, is a design variable. It is not a free good.

An *orchestration receipt* can be minimal and still do its work (Figure 8 shows the mock-up). Three fields carry most of the value: (1) *who acted*, the models and versions consulted, under what authority; (2) *how they agreed*, a plain-terms divergence summary; (3) *what was checked*, with a pointer into the full replayable record: small enough to read at a glance and specific enough to anchor a dispute, an insurance claim, or an audit. The same artifact answers the operator's own bookkeeping question, so the receipt sells twice: inward as cost governance, outward as warranted trust. That two-sidedness is a necessity. Measured consumer willingness to pay for transparency alone is thin (in the one priced conjoint, explainability's value fell short of €1.99 a month; König et al., 2022), so a trust layer funded by the consumer side alone would starve, and the operator's cost-governance side carries it. The receipt's nearest engineering neighbor is *tool receipts* logged inside agent frameworks for hallucination detection (Basu, 2026), which share the instinct but not the audience: those are developer-facing records of individual tool calls, while the orchestration receipt is their consumer-facing, per-decision generalization. The receipt invites its own failure mode: an artifact engineered to look verified without exposing anything a consumer could re-run. The design line that matters is whether the record is genuinely replayable and not merely receipt-shaped. A constraint arrives from the privacy side too: an artifact that proves what was checked also records what was preferred, so verifiability and privacy have to be optimized jointly (Alavi & Nozari, 2026).

One last piece of receipt economics goes under a name of its own, *the economics of checking*. A tamper-evident record is worth only what someone's incentive to open it is worth, and the decentralized-systems decade is the caution here: public, tamper-evident blockchain records did not turn ordinary users into auditors, and "trustless" architectures quietly reintroduced trusted intermediaries at every practical seam (Werbach, 2018). The Verifier's Dilemma is the formal version: where verification is costly and correct results are the common case, the rational verifier skips the check. Verification demand must be concentrated where a private payoff to checking already exists, among insurers, auditors, and dispute processes, with the consumer inheriting verification as a by-product of those markets; the receipt's consumer function is availability in dispute, not nightly reading. The funding question follows from the same point, and both naive answers to it fail: a consumer-pays orchestrator prices neutrality as a luxury good, and a merchant-pays orchestrator reimports the conflict it exists to remove. That leaves the model the receipt's two-sided economics already implies. The layer is paid for the way audit is paid for, by parties who need the risk priced rather than the story told (operators buying cost governance and liability reduction, insurers buying underwritable risk profiles, dispute processes buying evidence), with regulatory floors where those markets under-demand it. The consumer buys the service, not the neutrality, and inherits the neutrality as its by-product. The same problem has a market-structure version. A platform that is simultaneously marketplace, model vendor, and advertising network is structurally unable to supply neutrality, which is the conflict-of-interest point in market form, and the historical record of algorithmic prioritization (sponsored placement, preferred partners) says dominant players will monetize routing if they can. Where that pressure operates, the neutrality floor is regulatory or open-protocol, not voluntary. If first-exposure consumer demand for evidence is thin, adoption of architectural trust may be driven less by consumer choice than by the compliance and insurance side, a migration through institutions rather than preferences, with consumers inheriting verifiable architecture as a default. The calibration result keeps that route from being decorative: consumers demonstrably use the architectural signal once it is put in front of them, but they do not yet pay to summon it. There is a rival hypothesis here: trust may migrate to liability and institutions instead of to verifiable

architecture. Consumers rely on chargeback rights, not on ledger audits; card networks scaled remote commerce through dispute rights instead of through consumers inspecting authorization records; on that reading the decision record is plumbing that makes liability computable rather than an object of consumer trust. The two accounts predict different things: on the architectural account consumers reward verifiability when they can see it; on the institutional account they reward it only when a remedy is attached, and the consumer's trust then stays anchored in the institution while the institution's trust migrates to the architecture. The dispute-time design specified below discriminates them. And every record layer inherits the oracle problem: the record certifies the process, and only the process's own checks reach the facts. The asymmetry the receipt inverts is the one the platform-critique literature documents: architecture, not stated policy, is the operative regulator of digital conduct (Lessig, 2006), and to date its instruments have pointed at the consumer rather than being held by the consumer's side (Zuboff, 2019; Möhlmann & Zalmanson, 2017), and the receipt turns those instruments back the other way.

The evidence layer should be visible in the product itself, not filed away in a back-office log. Software engineering has recently converged on a name for the runtime wrapper around agentic models, the *agent harness*, and on the recognition that the harness, and not the model, is what can be certified, regulated, and trusted. Architectural trust specifies what a harness must *produce* to deserve that role.

Both the scale this section designs for and the snap-back it designs against are already on the record. Klarna's assistant handled 2.3 million conversations in its first month, two-thirds of the company's customer-service chats and the estimated work of 700 full-time agents (Klarna, 2024, company-reported figures). Fourteen months later the CEO reversed course in public: output quality had suffered, human recruitment restarted, and the standing commitment became that "there will always be a human if you want" (Daly, 2025). Read architecturally, as my reading rather than the company's stated reason, the pair of announcements brackets exactly the gap this paper names. The volume tier works at machine cost, but with no graduated escalation and no checkable quality signal, delegation snapped back to the human baseline. What would make the position between those announcements stable is a graduated trust architecture: cheap where stakes are low, checked where stakes concentrate, and human where checking fails.

Markets and regulation

For cross-border services. Neutrality extends beyond model vendors to jurisdictions. An orchestration layer can enforce jurisdiction-aware routing and record the enforcement, turning data-sovereignty compliance from a legal representation into an auditable runtime property. The differentiation it responds to has been measured: across 25 countries, a median 53% of respondents trust the EU to regulate AI effectively, against 37% for the United States and 27% for China (Pew data, reported in Stanford HAI, 2026, an institute report, not peer-reviewed). Where trust varies by jurisdiction, an architecture that can prove which jurisdiction's rules applied has real value.

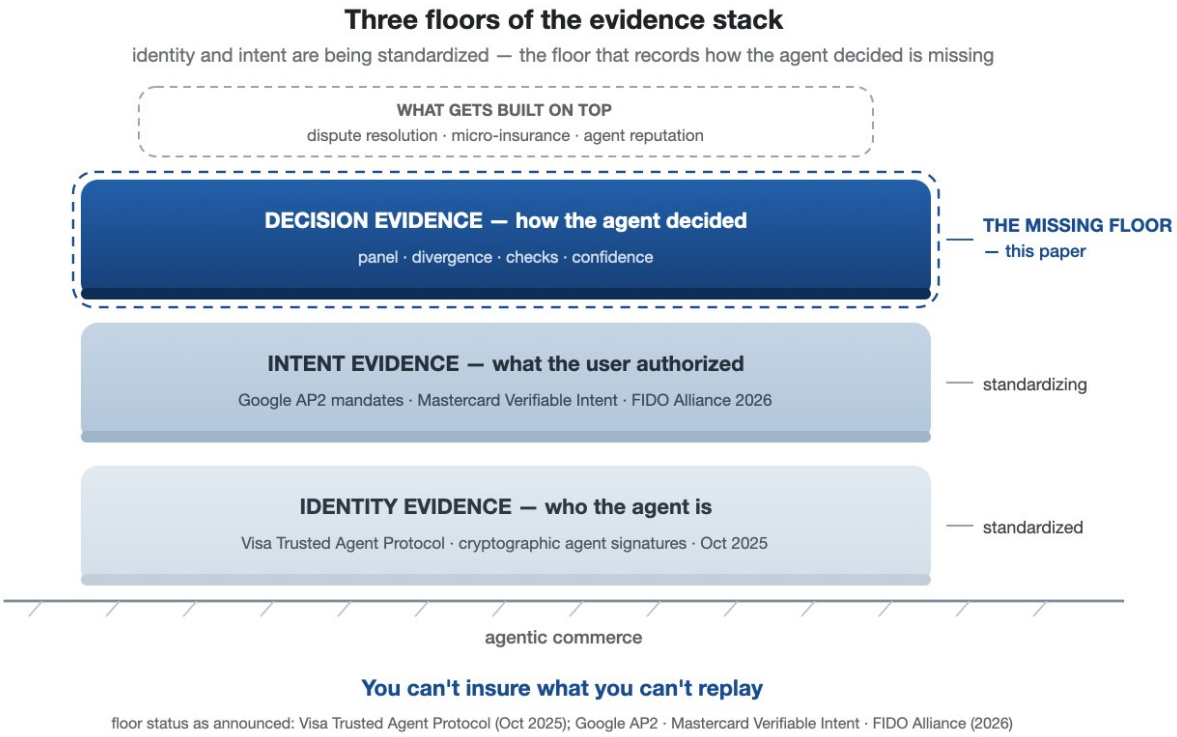
For agentic commerce. The Visa–ChatGPT integration makes the three components concrete. Caps and whitelists are static, consumer-side guardrails. Architectural trust adds the dynamic, service-side layer: purchases cross-checked by independent models (diversity), a replayable

record of why the agent chose this merchant at this price (evidence), and an orchestration layer whose revenue does not depend on which vendor or merchant wins (neutrality).

The payments industry has already begun building the lower floors of this stack. Visa's Trusted Agent Protocol (October 2025, with Cloudflare; Visa, 2025) signs an agent's *identity* cryptographically into its requests: merchant- and purpose-specific, time-bound, replay-proof signatures. Google's Agent Payments Protocol (Google Cloud, 2025), contributed alongside Mastercard's framework to the FIDO Alliance in 2026, encodes the user's *intent* as signed mandates carried as W3C Verifiable Credentials (W3C, 2025). Mastercard's Verifiable Intent (a joint framework with Google; announced March 2026; Mastercard, 2026) states its purpose in words this paper could have written, a tamper-resistant record of what the user authorized when an AI agent acts on their behalf, with early pilots reported in Asian and Latin American markets in 2026. Identity and intent evidence are being standardized now, and so is the record format: the IETF's SCITT architecture (RFC 9943, 2026) specifies the same pattern, signed statement → append-only transparency log → receipt, that this paper's evidence layer presupposes. The interface through which delegation happens is not even morally neutral. Across thirteen experiments, people who delegated to AI behaved less honestly, most sharply under vague goal-setting, and machine agents carried out fully dishonest instructions far more often than human agents did (Köbis et al., 2025). Interfaces that force intent to be stated in specific terms are ethical infrastructure.

No protocol yet covers the third floor, *decision* evidence: how the agent chose this merchant, this product, this price; whether independent checks agreed; what its confidence was worth (Figure 15 draws the three floors). That is the layer this paper names, and "know-your-agent" should mature into exactly this process question, with the unit of oversight migrating down to the *transaction*. What the architecture can guarantee is the record's existence, not its readership. Whether anyone opens such records is an empirical question only deployment can answer.

Figure 15. *The evidence stack*



Identity and intent evidence are being standardized now; decision evidence is the missing floor this paper names.

The first systematic audits of agent purchasing behavior show why decision evidence cannot wait. Allouah et al. (2026), running randomized trials in a controlled e-commerce sandbox, found that shopping agents exhibit strong positional biases, concentrate demand on a few "modal" products, and reshuffle those preferences drastically when the underlying model is updated. Without a record of how the choice was made, none of this is visible to the consumer, the merchant, or the regulator. Diversity is the design answer here, since a panel dilutes any single model's positional quirks, and accountability is what decision evidence supplies: what those auditors had to build a sandbox to observe, an architecturally trusted service would log in production.

The frame also extends past the consumer, to machine-to-machine trust. In B2B agentic workflows there is no human in the loop to *feel* trust at all, so the only possible substrate is verifiable process, with agents exchanging exactly the artifacts described here and computing whether to transact. How trust aggregates and propagates across such delegation chains at scale, and what the evidence layer's own logging and verification overhead costs at production volume, are open problems here. Chains sharpen the requirement, since a chain is only as trustworthy as its least evidenced link: the records must *compose*. Where no consumer is present, architectural trust is, plausibly, the *only* form of trust available.

A structural development the trust literature has not caught up with is the protocol war. Competing agent-commerce stacks (OpenAI with Stripe, the Agentic Commerce Protocol; Google's Universal Commerce Protocol coalition; Perplexity with PayPal; card networks and open standards underneath) are fragmenting commerce along vendor protocol lines exactly as AI

services fragmented along vendor model lines, and the neutrality argument generalizes almost word for word: a consumer's agent that can transact only inside one vendor's protocol is the vendor's agent, not the consumer's. Whatever layer arbitrates across protocols will face the same three requirements, independence, evidence, and no stake, and if such a layer takes hold the battle for the shelf becomes the battle for the checker.

For the evidence market. In 2026 this stopped being a forecast. On the liability side, insurance moved first and in two directions at once. Optional exclusion endorsements for generative-AI claims entered the standard general-liability toolkit in January 2026 (ISO forms CG 40 47 and CG 40 48, endorsements carriers may elect, not automatic exclusions; within months, major carriers had filed to adopt them or proprietary equivalents). In the same window the first AI-agent policies were written the opposite way, *against evidence*: the AIUC-1 agent standard exists in order to generate, through red-team audits, the empirical risk profiles insurers require to underwrite, and February–March 2026 brought its first accredited auditor, first agent policy, and first enterprise certifications. Squeezed between blanket exclusion and evidence-based underwriting, the record has become the unit of liability allocation: an agent with no inspectable history faces exclusion-shaped coverage, and an agent with one is a quotable risk.

On the regulatory side, the EU AI Act's Article 50 transparency obligations, including mandatory disclosure of AI interaction to consumers, become enforceable on 2 August 2026 (a transitional marking sub-obligation runs to 2 December 2026), with fines, on this paper's reading of the Act's penalty provisions, of up to €15 million or 3% of global turnover. The Commission adopted its Article 50 guidance on 20 July 2026 (European Commission, 2026), and the parallel Digital Omnibus simplification (signed 8 July) does not delay the August obligations. Existing systems receive only a limited deferral, to December 2026, for machine-readable content marking. Broader high-risk obligations phase in through 2027–2028.

The evidence layer changes this obligation: a tamper-evident record that includes *proof that the required AI disclosure was shown to this user in this session* converts, on this paper's reading, a legal exposure of up to €15 million into an auditable runtime property, compliance as a by-product of architecture. Insurers underwriting AI-agent liability need exactly these records, and traceability and accountability are core characteristics in the NIST AI Risk Management Framework. The regulatory precedent is older than the technology. Banking's model-risk guidance has required independent validation by parties independent of a model's developers since 2011 (Federal Reserve SR 11-7), a requirement an orchestrated panel automates, makes continuous, and extends to every consequential answer. Consumer law already reserves the human exit: GDPR Article 22 grants the right to human intervention in solely automated decisions with significant effects, and escalation-on-divergence is, on this paper's reading, the runtime form of an existing legal right. The AI Act reaches the same design from the oversight side: Article 14 of Regulation (EU) 2024/1689 requires high-risk AI systems to be built for effective human oversight, and an architecture that surfaces divergence and routes unresolved cases to a human implements that oversight duty as a runtime property rather than a compliance document.

Once the record exists, two derivative products follow, both near-term conjecture. *Transaction-level micro-insurance* would insure a specific agent purchase for cents, priced from the evidence record, unwritable against an opaque agent and trivial against a replayable one. The other is

agent reputation: longitudinal evidence histories aggregate naturally into reputation scores for agents and orchestrators. Services that produce audit-grade histories as a by-product of operation would hold an asset their competitors would have to retrofit.

How infrastructure of this kind becomes institutional is itself a hypothesis. The trajectory this paper bets on, products first, standards after, mandates last, is one of several the history of trust infrastructure exhibits (shared ATM networks, card-network rulebooks later ratified by regulation, blockchain's verification supply still waiting for demand). It also has a named research program behind it. Hadfield's regulatory-markets line, private regulators licensed by public authorities, with independent verification organizations as the load-bearing institution (Hadfield, 2017; Clark & Hadfield, 2020; Hadfield & Clark, 2026), reaches it from the supply side of governance. This paper reaches it from the demand side of consumer delegation, and the two programs meet at the receipt. Underneath sits the harder question of whether a receipt sold as a product feature ever hardens into an institution. If it does not, it remains a governance ornament rather than a governance layer, which would refute this paper's central claim. The refutation point is pre-registered: call the receipt a governance *layer* when at least two industry standards require traceable decision records for agentic transactions, services carrying all three components handle a majority of high-stakes agentic transaction volume, and at least one national regulator recognizes such records in its allocation of liability. Persistent failure of all three is what "ornament" means operationally. If verification succeeds as infrastructure, it commoditizes, and trust migrates into the shared utility itself, the present architecture included, which would end any moat even as the infrastructure succeeded.

Who is liable when an *orchestrated* decision errs remains an open question: the orchestrator, the model vendors, or the consumer who authorized the agent.

The law has begun closing off the easy exit. California's AB 316, effective 1 January 2026, bars any defendant who developed, modified, or used an AI system from arguing that the AI *autonomously* caused the harm, while leaving ordinary defenses and comparative allocation of fault intact. It does not impose strict liability. That makes the allocation question unavoidable, and whichever way doctrine settles it, the replayable record is what any allocation of responsibility would be computed from. Legal scholarship has been arriving at the same destination, holding that the classical agency-law toolkit does not transfer to AI agents and that what is required is new infrastructure of visibility and liability (Kolt, 2025), the nearest predecessor of this paper's argument in the governance literature, approached from the supply of oversight where this paper approaches from the demand for it. Architectural trust is the evidentiary substrate that an emerging liability doctrine could plausibly run on.

Measurement and the research agenda

For trust measurement. Standard instruments measure attitudes toward a brand or an interface. The attitude worth measuring is *willingness to delegate*, not the answer to "do you trust this AI": what class of decisions, at what stakes, will the consumer hand to the agent without reviewing each step? Delegation is behavioral, monotone in warranted trust, and directly tied to revenue. Beside it sit the questions of whether the consumer knows how the answer was produced, whether they can reach the record, and whether surfaced disagreement improves or degrades their reliance calibration.

The survey was the receipt of the human era: while the consumer performed the search, the comparison, and the choice, asking the consumer was a defensible way to learn what happened. Delegation removes the actor the questionnaire was built to interrogate, so the primary trace of the episode becomes the artifact the architecture produces, and the receipt is in that precise sense the primary *transaction trace* of the delegated era. The two instruments do not compete: the receipt logs what the machine did and says nothing about subjective appropriation, felt autonomy, or cognitive load, which remain the survey's territory. (This dissolves an apparent tension with Section 1, whose motivating evidence is itself stated-preference surveys: surveys measure what consumers currently feel about a market that barely exists, while the receipt measures what delegated services actually did once they exist.) That bridge can be tested by pairing attitude measures with the artifacts of the same episodes, stratified by outcome, and estimating where questionnaire and record agree and which better predicts subsequent delegation behavior. The proposed conjoint carries part of this validation by design, and until such studies exist, artifact-based metrics complement the survey rather than retire it.

Industry surveys already show the behavioral baseline: in a Gartner consumer survey released May 2026 (846 U.S. consumers, fielded November–December 2025), 54% of shoppers who used generative AI while shopping report double-checking all of the information it gave them. Architectural trust is the productization of what consumers already do for themselves.

The measurement program has natural allies in the holistic-evaluation movement (Liang et al., 2023) and the emerging science of agent reliability, which decomposes reliability into twelve metrics because a single success rate hides the failures that matter (Rabanser et al., 2026). Architectural trust operationalizes the same way, behaviorally: confident-error rate, appropriate reliance, weight-of-advice shifts, override rates, decision time, and one especially sharp statistic, the gap in approval rate between correct and incorrect outputs, which is reliance calibration read directly off behavior. On the record side there is evidence coverage, measured panel divergence, and receipt-access rates, triangulating survey, behavior, and logs on the same episodes. None of this requires new science. The requirement is a decision that these, and not engagement, are the numbers a trust layer is accountable to. My own observation from the deployment points the counterintuitive way: well-presented uncertainty increased my appropriate reliance, and I trusted the system more when it visibly knew what it did not know.

One further dependent variable belongs on the agenda because trust is not the only psychology delegation strains. Consumers respond less warmly to a favorable outcome when an algorithm produced it, because an outcome one did not produce is harder to claim as one's own (Yalcin, Lim, Puntoni, & van Osselaer, 2022). Delegation can raise a decision's objective quality while draining its subjective ownership. The receipt sits on that fork: read one way, it restores authorship; read the other, it certifies that the consumer did nothing. Which reading prevails is a design question the framework has already instrumented: the Evidence component records what was checked and against what standard (Section 3), and the receipt is built to carry the consumer's own criteria into that record, so authorship of the evaluation criteria is a variable the architecture exposes rather than one an experimenter invents. Yalcin et al.'s mechanism, that an outcome one did not produce is harder to claim as one's own, then makes criteria authorship the natural moderator, and P5 is the Evidence component's consequence for the psychology of appropriation rather than a side branch of the agenda. **P5.** Receipt salience interacts with outcome favorability in determining subjective appropriation of a delegated outcome: for

favorable outcomes, a salient receipt restores appropriation when it encodes consumer-authored criteria (*the standard the agent had to meet was mine, and here is the proof it met it*) and depresses appropriation when the criteria read as the platform's; consumer authorship of the evaluation criteria is the design moderator. Yalcin et al.'s effect was observed where consumers expected to make the decision themselves, so P5 is scoped to that boundary and may reverse under habitual delegation, and what distinguishes the receipt from existing process-transparency manipulations must be implemented, not assumed. The receipt is checkable and can carry the consumer's own criteria; a design that omits both is testing transparency, not architectural trust.

The transparency literature supplies two cautions here: *too much* procedural transparency can depress trust (Kizilcec, 2016), and mere *disclosure* of AI use erodes it (Schilke & Reimann, 2025). The second is a boundary condition on the Evidence component: across thirteen experiments, actors who disclosed AI use were trusted less than those who did not, a *transparency dilemma* in which making AI involvement more visible, by itself, lowers trust. If that penalty attaches to any signal of AI involvement, then an architecture that foregrounds its own machinery could depress trust even while making it more checkable, and the Evidence component would backfire exactly where it is most salient. The framework's answer rests on a distinction that has to be demonstrated: verifiability is not mere disclosure. The receipt is a checkable record of *how* the answer was produced, not a declaration *that* AI was used, and P2 and P3 (with H2's pre-registered two-sided test, disclosure backfire named in advance as a reportable outcome) are exactly where the distinction either earns its keep behaviorally or fails. Checkability is what builds trust, delivered on demand rather than as an ever-present wall of provenance, and neither hiding the machinery nor announcing it does that work. How much of the record to surface, to whom, and when is itself a design variable.

The proposed mechanism for Evidence overcoming the disclosure penalty rather than compounding it is that verifiability and disclosure travel through different psychological channels. Disclosure delivers a category label, that an AI was involved, which Schilke and Reimann's evidence ties to depressed legitimacy. A receipt delivers a procedure, a record that can be opened, contested, and re-run, and the procedural-justice literature holds that a process one can challenge and check legitimates the actor running it (Tax, Brown, & Chandrashekaran, 1998). The channel does not require the checking to happen: what carries the trust is the consumer's awareness that the check exists and is openable in a dispute, the availability of verification rather than its exercise, which is exactly the consumer function the economics of checking assigns to the receipt above. Checking concentrates where a private payoff to checking exists, and the consumer inherits verification as a by-product, holding the record as an option rather than reading it as a document. The same distinction bears on subjective appropriation (P5): a disclosure label attributes the decision to the machine and drains ownership, whereas a receipt that encodes consumer-authored criteria documents that the standard the agent had to meet was the consumer's own. Whether the two channels in fact dissociate, with legitimacy rising with checkable procedure while falling with bare disclosure, is exactly what H2's pre-registered two-sided test is built to detect.

Limitations. The model is derived from architectural analysis plus one extended, high-stakes, single-operator deployment. Every empirical anchor was obtained on that one stack, which I built and operate. Applied to itself, the paper's warrant is mixed by its own lattice: one author and one stack (low diversity), a public replication package beside sealed deployment records and

an embargoed behavioral registration until publication (partial evidence), and a declared commercial stake (compromised neutrality). Independent replication on other stacks is required before any claim generalizes. The deployment records' hash anchor makes any later editing, substitution, or back-dating detectable, so on unsealing the characterizations here can be checked against an artifact that provably has not changed. Readers can verify integrity, not contents, until unsealing. The mechanism numbers come from public model APIs, and the pipeline replicates the collection end-to-end for under $15. Stimulus rendering remains an interface confound: the consumer experiment's stimuli, the receipt's styling included, are my own UX rendering, and interface-level confounds untangle only through replication with independent renderings. Blind-first collection, tamper-evident records, and no-stake routing are the construct; the aggregation rules and escalation thresholds are engineering choices. The experiment's panels are analytic combinations of one blind collection rather than a deployed orchestration loop, and the drafted field-telemetry protocol closes that gap. The non-model verification paths the framework names (shadow runs, temporal-consistency checks) are described here, not tested. The collection is a single run at temperature zero, which suppresses within-model variability and leaves shared-lineage correlation in place, and hosted APIs are not strictly deterministic even there. SimpleQA's short factual answers sit near the constrained-answer-space regime in which consensus certifies error, and the difficulty stratification shows exactly that pattern on the hardest quintiles. A mid-tier panel may overstate diversity's benefit, since error correlation rises with capability (Kim et al., 2025). The mechanism experiment tests open-response factual questions with single verifiable answers, and on open-ended generative tasks, where semantic equivalence between answers is itself a judgment call, the certification signal must be re-established rather than assumed. A related boundary is the answer space's dimensionality: SimpleQA items have a single verifiable answer, whereas the consumer choices agents execute are subjective, multi-attribute utility judgments with many defensible answers, so how consensus behaves in preference space, and whether agreement among independent agents even identifies a better choice there, is an open empirical question (semantic equivalence becomes preference equivalence, and the grading itself a modeling choice). One label-noise check is run: on the 351-item intersection of this study's subsample with SimpleQA Verified (Google DeepMind, 2025), regrading every panel answer against the corrected gold labels leaves the certification lift unchanged or slightly stronger (2.7 → 2.8, 4.4 → 4.4, 4.0 → 4.3 across the three panels, deterministic grader), so label noise is not the lift's driver. All confidence figures are verbalized 0–100 self-reports elicited by a single prompt format, and sensitivity to the elicitation protocol may itself vary by vendor, so per-vendor discrimination and calibration conclusions should be read as properties of this configuration under this elicitation protocol, not as fixed vendor traits. The self-confidence baseline in the sixteen-model wave's Tables 6–7 is raw stated confidence. The stronger calibrated baselines in the tradition of Guo et al. (2017) were run on the flagship wave (Platt scaling and split-conformal abstention, Section 3), where they emerge as complements to the gate rather than substitutes, and the first wave's panel-versus-self comparison should be read against the baseline consumer products actually deploy. The headline gate estimate (−82.1% of confident failures, bootstrap 95% CI [79.9; 84.1] for the ten complete arms, [80.0; 84.2] with the Meta arm included) is a single pre-specified quantity, and the surrounding analyses (the threshold sensitivity grid, the difficulty and contamination strata, the council ablations and trivial-baseline comparisons) are descriptive robustness checks reported without any correction for multiple comparisons, and they claim no confirmatory status of their own. The only family-wise error control in the paper is the Holm correction across the consumer

experiment's pre-registered {H1, H2}. If the evaluation regime that rewards confident guessing were repaired so that truthfulness could be *trained* into models (Kalai et al., 2025), architecture would insure against a problem with a cheaper solution, though evidence and neutrality would remain irreducible to any model-level fix. Further boundaries are stated where they are reported: the flagship wave's coverage limitations (Section 3: two inaccessible Chinese flagships and one interrupted Meta collection), and the preliminary 233-pair reasoning-budget comparison, whose TOST bounds equivalence at a ±5-point margin without power for narrower bounds. Delegation measures come from a hypothetical vignette, where stated measures overstate real willingness by roughly 21% on meta-analysis (Schmidt & Bijmolt, 2020); the consumer experiment realized 119 primary and 133 pooled of 200 registered participants over six fixed items. A commerce-substrate replication extends the mechanism beyond the benchmark: a fresh, pre-frozen set of 267 Wikidata-verified product facts (which company manufactures a given car or aircraft, what operating system a phone runs, which company published a game, in what year it shipped), run blind through an independent six-model panel (Claude Opus, GPT-5.4, Grok, Qwen, GLM, and Mistral), the cost-optimized tier on which consumer-facing agents actually operate, not the flagship, reproduces the dissociation, cross-model agreement predicting correctness at AUROC 0.77–0.81 against 0.63–0.70 for stated confidence and a 4–6× certification lift, robust across independently-composed panels of three to six models spanning US, Chinese, and European laboratories, and to the strictness of grading (the deterministic string-match here is the primary, conservative rule). The subjective, multi-attribute end of the commerce spectrum (preference tradeoffs, adversarial merchant incentives, and tool-grounded retrieval over changing inventory) remains untested.

Run-to-run variance under sampling. A variance arm probed the single-deterministic-run boundary directly: a difficulty-stratified subsample of 100 questions (5 difficulty quintiles × 20, seed 20260806) was re-run at temperature 0.7 with 3 independent samples per question for 9 of the 11 flagship-wave panel models; Meta's endpoints were inaccessible, and Claude Fable 5's API does not expose a temperature parameter. Across models, all three samples produced mutually equivalent answers on 22–77% of questions (median 54%), per-question correctness was identical across the three samples on 63–97% of questions, and the temperature-0 answer matched the majority temperature-0.7 answer on 40–85% of questions. gpt-5.6-sol's leg carries a confound: it ran at reasoning effort *none*, whereas 1,486 of its 1,500 wave records used the default effort (medium, as noted above). Its 92.0% → 46.3% accuracy delta on this subsample therefore confounds temperature with reasoning budget, though its inter-sample consistency figures are unaffected. Excluding that confounded leg, mean per-sample accuracy at temperature 0.7 differed from the deterministic temperature-0 accuracy on the same questions by at most 14.0 points (gemini-3.1-pro, 49.0% → 63.0%), with the remaining seven models within 4.4 points. At the panel level, the council gate (three-way answer agreement among kimi-k3, deepseek-v4-pro and gpt-5.6-sol) opened on 31.0% of the subsample at temperature 0, against 25.0%, 23.7% and 29.8% across the three temperature-0.7 runs. The agreement signal underlying the headline result is stable under sampling noise.

Evidential status of claims. The claims fall at three evidence levels. *Established by these data:* that independent agreement carries a correctness signal a model's self-report does not, under the stated boundary conditions (blind collection, competent verifiers, open-response factual questions with single verifiable answers); that a fixed disagreement gate cuts the large majority of confident failures at a substantial, stated coverage cost (−82.1% at 30.5% coverage); and the

strong boundary conditions themselves: consensus certifies error on constrained answer spaces, the raw certification lift attenuates monotonically with difficulty (vanishing on the hardest quintile under the within-collection stratification; positive but weakest there under the exogenous one), and verifier competence gates certification value. *Supported by the behavioral experiment (one design, modest single-study scale, n = 119):* that surfacing real cross-model disagreement improves reliance calibration in both directions ($d = 0.67$, Holm $p = .0005$; carried by one of six items; main-wave manipulation checks 88–98%; scale reliabilities $\alpha = .83$–$.93$); and, as a registered inconclusive null, that a static first-exposure receipt does not detectably move delegation. *Theoretically proposed, not yet tested:* the D×E×N complementarity; the neutrality effect (P3); trust migration under increasing delegation; post-failure durability of architecture-based trust (P4); willingness to pay for verifiability; and the market-category claim of Section 7.

The natural next step is experimental: measure trust, reliance calibration, and willingness-to-pay across single-model and orchestrated conditions, with disagreement alternately surfaced and suppressed. Section 6 reports the realized experiment, pre-registered and incentive-compatible. A standard choice-based conjoint estimates part-worths additively and assumes compensatory trade-offs, which is what the complementarity hypothesis denies, so the design class has to permit non-compensatory structure, whether elimination-by-aspects or forced-choice designs with deal-breaker attributes, estimated with explicit interaction terms. Two design tiers are specified. The minimal tier is a targeted-profile study, four to five profiles spanning none, single-component, two-component, and full configurations, with the theoretical work done by one planned contrast, the full configuration's effect exceeding the sum of the component effects (super-additivity), against the additive null. The full tier crosses three levels per component (absent, present-but-uninspectable, present-and-inspectable) in a 3×3×3 between-subjects factorial (27 cells; stakes varied within subject across four levels), with at least 80 respondents per cell ($N \geq 2{,}160$) for power ≥ 0.85 on the interaction. A binary 2^3 design cannot recover the Cobb–Douglas exponents and is underpowered at plausible effect sizes ($d \approx 0.3$–0.4). Three stimulus disciplines apply: vary the architectural magnitudes themselves and not only their labeling (a missing receipt label is consumer ignorance of evidence, not its absence), give each component a verifiable-versus-merely-asserted contrast, and pair choice data with process-tracing on the receipt's fields, since the theory predicts *use* of evidence. To my knowledge this would be the first design to vary the three components jointly. The single-component building blocks are already priced (König et al., 2022; Ioku et al., 2024), and adjacent bundles have been run (Cetinkaya & Krämer, 2024, the nearest instrument and a power-calibration source alongside Köbis et al., 2025), so the study is specified, powered, and ready to run. The construct's boundary conditions are these: architectural trust does not pay where stakes are low and volume is high; it adds little where the verifier pool is not competent on the task class; it can certify error where answer spaces are tightly constrained; and it degrades toward theater wherever panel independence is asserted rather than measured, sharpest as the open red-team question *what does a gamed receipt look like?* Once hyperlinks became authority signals, link farms arose to manufacture them, and agreement records should be expected to invite their own farming. Two temporal boundaries apply as well: cues win attention while evidence wins persistence, so the claim concerns which trust survives failure rather than which onboards fastest, and even a working check may only slow a spiral of the kind Chandra et al. (2026) model. Where consequences concentrate and independence is real, architectural trust earns its cost and its name (Table 12 collects the four).

Table 12. *Boundary conditions: where architectural trust does not pay*

Condition	Why it fails there	Evidence
Low stakes, high volume	Verification cost exceeds the cost of errors; single-model serving is the right answer	Break-even logic (Figure 14)
Verifier pool below task competence	Independence without competence carries no signal	Local-tier null: 46% precision with or without agreement
Tightly constrained answer spaces	Consensus certifies error rather than catching it	Consensus errors 4.0–7.9% under multiple choice vs. 0.5–2.2% open-form
Independence asserted, not measured	A panel of near-clones manufactures confidence	Correlated errors; bloc and lineage nulls

A last boundary concerns what any of the proposed designs can identify. Trust in an agentic market is, in the end, an equilibrium object. Once sellers, platforms, and competing agents have adapted to a verification architecture, its effects are properties of the adapted market, not of the treatment cell. Observational data inherit selection, because verification will be adopted first where trust is already breaking. Unit-level randomization repairs the selection but stops at interference, since consumers randomized to receipts still transact in a market whose sellers respond to everyone's receipts at once. The equilibrium quantity needs clustered, switchback, bipartite, or market-level designs; randomizing architecture, rather than measuring trust, is the next hard problem. And early agentic-commerce cohorts self-select on trust propensity, so effects estimated on early adopters bound, rather than estimate, the population effect. Neutrality is valuable exactly because nobody with a stake in a model can supply it, including anyone with a stake in orchestration. The argument also describes the kind of system I build (Author's note). That interest touches the advocacy and not the reproducibility of the evidence: the models evaluated are independent, publicly accessible frontier systems, and every reported number is recomputable by third parties from the frozen replication package. A reader who discounts the advocacy entirely can still verify the measurements.

6. Can consumers use the signal? A preliminary behavioral test

Whether *consumers* can use the correctness signal the mechanism establishes is a separate, behavioral question, and I report a pre-registered, incentivized experiment on it. The fielded study is an existence proof rather than a test of a general effect; it asks whether consumers can act on the signal at all. A powered, confirmatory, many-item, stimulus-sampled replication with item random slopes is planned. A hosting outage in the final recruitment window ended collection at 119 valid participants against 200 planned (133 pooling a soft-launch cohort; US and UK; three between-subjects conditions around an agentic shopping assistant; calibration measured over six incentivized factual trials as $\Delta = P(\text{accept} \mid \text{correct}) - P(\text{accept} \mid \text{incorrect})$, the appropriate-reliance discrimination index of Schemmer et al., 2023). Full method, the pilot, the five deviations from the registration, the ethics and exemption disclosure, and all ancillary statistics are in the detailed report below (§6.1–6.2) and the replication package.

Displayed cross-model disagreement improved calibration overall (Control $\Delta = 0.24 \rightarrow 0.45$; $d = 0.67$, 0.68 pooled, $p = .0001$), but the pooled estimate is fragile: it is carried by one of six stimuli, the single high-confidence dissenter item, where the display cut acceptance of a confident wrong answer from 74% to 24% (pooled cohort, Table 18). Removing that item drops the effect to $d = 0.09$ ($p = .54$), while dropping any other single item leaves it intact. The benefit concentrated in the UK subsample. The US control cell ($n = 16$) already discriminated at baseline (interaction $p = .002$). On the one item where the panel itself agreed on a wrong answer, the display *increased* acceptance in both countries, which is Section 3's certified-consensus-error mode reproduced in human behavior and a finding in its own right: an architecture that surfaces agreement transfers its one systematic failure to its users. The first-exposure test of P2's evidence component, a static orchestration receipt, produced no detectable effect on willingness to delegate ($d = 0.21$, inconclusive at this power; a live drill-down receipt at dispute time is the registered follow-up).

A dissenting panel display can sharply cut acceptance of a confident wrong recommendation where a consumer would otherwise accept it, and it fails where the panel certifies a shared error. The calibration contrast therefore has behavioral support under these six stimuli; P1 as stated, with perceived risk as mediator, is not supported by this test. P2–P4 remain the open program.

6.1 Behavioral test: method and pilot

The propositions of Sections 3 and 5 are written to be tested, and the measurement agenda above specifies how. The experiment is a pre-registered, incentive-compatible online test of P1 (reliance calibration under surfaced disagreement) and P2 (willingness to delegate under a replayable record), with stimuli drawn from the real multi-model outputs of the flagship wave, extending the second-opinion paradigm (Detjen et al., 2025; Chen et al., 2026) to real, incentivized frontier-model disagreements under an architectural framing. The instrument, the protocol freeze, the pilot, the main wave, and the frozen-code analysis are complete and reported below. The confirmatory tables were carried in earlier versions of this paper as a pre-registered skeleton with cells marked [MAIN DATA PENDING] and are now filled by the frozen analysis code. Two outcomes unfavorable to the hypothesis were pre-registered as reportable primary findings: surfaced disagreement worsening calibration (net under-reliance), and the receipt lowering willingness to delegate (disclosure backfire); neither materialized, but one pre-registered null did.

Method.

Design and hypotheses. Reliance calibration requires crossing displayed agreement with answer correctness, which a single vignette cannot do. The design is therefore hybrid. Participants first read a shopping-delegation vignette (an AI shopping agent, "ShopAssist," recommending a laptop purchase within a $1,200 budget), the primary test of P2, and then complete a short incentivized judgment block on real multi-model outputs, the primary test of P1. H1 (P1): displaying cross-model agreement and disagreement from several independent AI systems improves consumers' reliance calibration (they accept the agent's proposed answer more when it is correct and less when it is incorrect) relative to a single confident answer; two-sided, because the literature documents both reduced over-reliance and increased under-reliance under joint presentation. H2 (P2): access to an orchestration receipt (who acted, how they agreed, what was checked, and a full re-verifiable record) increases willingness to delegate a consequential

purchase over and above the effect of displayed disagreement alone; two-sided, because disclosure can also depress trust.

Conditions. Three between-subjects conditions, randomized 1:1:1 and held constant across the vignette and the judgment block: (1) *Control*, a single confident recommendation, no architectural signals; (2) *Disagreement*, the same recommendation preceded by a panel of three independent AI systems (anonymized as System A, B, C) with explicit agreement/disagreement marking, the systems' alternative answers and stated confidences, and a plain-language resolution summary; (3) *Disagreement + Evidence*, the Disagreement screen plus an orchestration receipt card structured as who acted / how they agreed / what was checked / full record. One pre-registered design constraint controls the obvious confound: the final recommendation block (product, price, and the agent's stated final confidence) is verbatim identical across all three conditions, in the vignette and on every judgment trial, and only the architectural signals around it differ.

Stimuli and receipts. The judgment-block stimuli are real outputs of three frontier models from the flagship wave on real factual questions on which the models genuinely agreed or disagreed at high stated confidence (95–100); no answers or confidences are fabricated. Model names are replaced by neutral labels to avoid brand-prior confounds, and the mapping is disclosed in the replication package. The receipts shown in condition 3 (one vignette receipt and six trial receipts) are minted by the research deployment's production receipt minter (schema `decision-receipt@v1`) over the actual provenance logs of the stimulus panel, and their verification hashes are genuinely recomputable from the record (verification log: 7/7 valid). The experiment's evidence manipulation is therefore an instance of the artifact Section 3 defines. Displayed focal confidence was closely matched across correctness cells (incorrect items 95/98/99; correct items 99/99/100) and, decisively for the treatment contrast, identical across conditions. Control saw the same focal confidences and discriminated far less, which a confidence-cue account cannot explain.

Measures. DV1 (H2): a 4-item willingness-to-delegate scale (7-point; e.g., "I would let ShopAssist complete this purchase without reviewing the recommendation myself"), averaged; a single-item delegation ladder (0–4, from "recommendations only" to "acts autonomously including non-refundable purchases") is secondary. DV2 (H1): six judgment trials; on each, the agent proposes one answer and the participant decides whether to let the assistant proceed (yes/no) and rates trust in the answer (0–100 slider). Reliance decisions are accuracy-incentivized (a bonus for each correct reliance decision, accepting a correct answer or rejecting an incorrect one). The proposed answer is correct on exactly three of six trials. The per-participant calibration score is $\Delta = P(\text{accept} \mid \text{correct}) - P(\text{accept} \mid \text{incorrect})$, the discrimination index of the appropriate-reliance literature (Schemmer et al., 2023, whose RAIR/RSR measures decompose the same object); a trial-level mixed-effects model recovers item-level resolution. Perceived risk and perceived process control (4 items each) are measured as mediators in pre-registered secondary analyses for the P1 and P2 pathways respectively; perceived orchestrator neutrality (4 items) is measured as an exploratory correlate of P3, which this study does not manipulate.

Frozen protocol. The pre-registration was frozen before main data collection (analysis code and protocol frozen by SHA-256 on August 5, 2026; the OSF registration was submitted on 13

August, before the main wave entered the field on 13–14 August; registered with embargo (pre-registration, view-only: https://osf.io/pqeg6/?view_only=87eb0a88579647b48e93d4a475795d5f), to be released into the paper's public OSF component, https://osf.io/jtvu9/). The six-trial item set is frozen by seed: three disagreement items fixed by design and hand-verified, plus consensus items drawn by a seeded shuffle from the pools where all three systems answered identically at stated confidence ≥ 95, including one genuine consensus error, so that displayed agreement is diagnostic but imperfect, exactly as in the source data. The six items were identical for every participant; presentation order was randomized in the instrument, though the fielded export did not retain per-trial display order. The fielded display did not surface an estimated independence score for the panel, which is the measured-diversity disclosure the framework itself calls for, so participants saw disagreement without seeing how independent the disagreeing systems were. The panel's independence was measured but not shown, and its value on the source collection is reported here post hoc: mean pairwise error correlation $\varphi = 0.22$ (0.10–0.44 across the three pairs, the two same-bloc members by far the most correlated at 0.44; graded by the study's frozen deterministic grader, which resolves 95–98% of answers, and insensitive to how the unresolved remainder is coded), which puts the panel far from independent and far from clonal. That is the number the next instrument should put in front of participants. One stimulus boundary: with six fixed items, treatment-by-item variation cannot be estimated, so the effect is established for these stimuli. Sessions began with informed consent and ended with a debrief disclosing the hypothetical nature of the vignette purchase. The study was conducted in accordance with the Declaration of Helsinki: anonymous, minimal-risk survey research with consenting UK and US adults on a GDPR-compliant panel, compensation above platform minimum, and no deception beyond the hypothetical vignette disclosed at debrief. Under these criteria it met the conditions for exemption from full ethics-board review, and the exemption assessment is documented in the pre-registration. The pool is now closed; the verbatim item set, answers, and per-trial receipt identifiers are part of the embargoed pre-registration and will be released with the data.

Sample and analysis plan. Target N = 200 valid participants (100 US, 100 UK; separate Prolific studies; recruitment capped at 210), age 18+, fluent English, ≥ 98% approval rate, ≥ 100 prior submissions, at least one online purchase in the past three months. Pre-registered exclusions (applied before viewing substantive responses): two or more failed attention checks of three; completion under 3 minutes; straight-lining; failed manipulation checks (agreement item in conditions 2–3, receipt item in condition 3, product-comprehension item); twice-failed judgment-block comprehension; duplicate ID or IP; incomplete session. Confirmatory analyses: for H1, OLS of Δ on pooled disagreement conditions (2 and 3) versus control with country covariate; for H2, restricted to conditions 2 and 3, OLS of the willingness-to-delegate composite on condition with country covariate; Holm correction across {H1, H2}, family-wise $\alpha = .05$; Welch t-tests as robustness; manipulation-check gating reported first. Cohort ordering is fixed here: the primary confirmatory analyses are computed on the main-wave cohort *excluding* the pilot's soft-launch participants, and the branch that pools the soft-launch cohort into the confirmatory sample is reported as a robustness analysis alongside. The pilot's exploratory Δ descriptives were seen before the main wave, so excluding that cohort from the primary estimate removes even the soft form of double-dipping while retaining the pre-registered soft-launch inclusion as a reported branch. Primary reporting will include an intention-to-treat analysis retaining all randomized participants alongside the pre-registered per-protocol analysis. Manipulation-check exclusions will be reported with differential-attrition checks by condition, since architectural-signal salience

could itself differ across conditions. Sensitivity: the design detects d ≈ 0.42 for H1 and d ≈ 0.49 for H2 at power .80, adequate for the medium effects in the adjacent reliance and transparency literatures. Smaller effects are left to the full factorial of the research agenda. Three pre-registered elements answer mechanism questions the confirmatory contrasts alone cannot. The instrument carries a four-item perceived-risk scale (pilot α = .881) on the P1 pathway, so risk perception enters the analysis as a measured construct rather than an assumption. Mediation is pre-registered as a secondary analysis, with perceived risk mediating the disagreement effect on reliance and perceived process control mediating the receipt effect on willingness to delegate, estimated by bootstrap path analysis (PROCESS Model 4, 5,000 resamples). Time on the condition-3 stimulus screen is collected as the pre-registered proxy for receipt reading, a direct check on the cognitive-load alternative that participants pass the receipt by without processing it. Analysis code was frozen before data collection and is archived with the replication package.

Pilot (soft launch): criteria and outcomes.

A pilot of n = 20 (US), run on the final instrument with pre-registered pass criteria fixed in advance, was completed on August 6, 2026. Table 13 reports every criterion against its outcome.

Table 13. *Pilot acceptance criteria (pre-registered) and outcomes (n = 20, US)*

Criterion (fixed ex ante)	Outcome	Verdict
Manipulation noticed: > 80% correct on agreement check (conditions 2–3)	81.3% (13/16)	Pass
Manipulation noticed: > 80% correct on receipt check (condition 3)	88.9% (8/9)	Pass
Cronbach's α > .70, willingness to delegate (4 items)	.854	Pass (also clears the .85 target)
Cronbach's α > .70, perceived risk (4 items)	.881	Pass
Cronbach's α > .70, process control (4 items)	.965	Pass
Cronbach's α > .70, neutrality (4 items)	.851	Pass
Median completion time within 10–15 minutes	9.60 minutes	Fail, by 24 seconds, in the fast direction

One criterion failed. The median ran 24 seconds under the pre-registered window, a window built against over-length, not against efficiency. The fast tail shows no evidence of careless responding: no participant fell below the 3-minute exclusion threshold; the one participant under five minutes passed every attention, manipulation, and comprehension check; attention-check pass rates were 95–100%; there were zero missing cells on key variables, zero duplicate IDs or IPs, and zero straight-liners; and the recruitment-platform completion times independently confirm the survey timer (median 9.74 minutes). The instrument is shorter in the field than budgeted. Condition balance (4/7/9 across conditions 1/2/3) is compatible with 1:1:1 randomization at n = 20 ($\chi^2(2) = 1.90$, p = .387). Under the pre-registered analysis exclusions, 6 of 20 pilot participants are excluded from confirmatory analysis (five on manipulation-check

grounds, one on twice-failed comprehension), leaving 14 valid. Per the soft-launch rule, the pilot cohort proceeds into the pooled sample with the timing deviation disclosed here and in the registration itself, which was submitted to OSF on 14 August 2026 at 03:09 UTC, before main-wave collection, and which states that the soft-launch pilot is included in the confirmatory sample with the timing deviation stated; treating the main-wave cohort as the primary confirmatory estimate and the pooled branch as robustness is a post-registration analytic decision, taken because pilot descriptives had been seen before the main wave, and both branches are reported; the product-comprehension margin (one participant above the 80% bar) is flagged for monitoring in the main wave. One exploratory descriptive, reported without inference at pilot n: the calibration score Δ by condition ran +0.33 (control, n = 4), +0.52 (disagreement, n = 7), +0.22 (disagreement + evidence, n = 9), so the disagreement condition moves in H1's direction; the pilot cannot say more.

6.2 Behavioral test: results, exploratory analyses, and remaining ledger

Confirmatory results.

Main-wave collection ran on 13–14 August 2026 (190 slots: 85 US + 105 UK on top of the soft-launch cohort), and it did not complete: the survey host suffered a technical outage during the final recruitment window, and collection ended at 158 completed sessions in total (138 in the main wave and 20 in the soft launch) rather than the 200-valid target. Recruitment ended with the platform outage, under the pre-registered optional-stopping rule and not in response to the observed outcomes, although results had already been seen by then; the consequence is reduced power. At the realized cells the minimum detectable effects rise from the registered $d \approx 0.42$ (H1) and $d \approx 0.49$ (H2) to roughly $d \approx 0.54$ and $d \approx 0.64$ at the same power, a bound the confirmatory outcome renders moot in one direction only, and one that applies to every null and exploratory contrast below. Cohort ordering follows the plan of §6.1. The pilot missed one pass criterion (the completion-time window, by 24 seconds in the fast direction), and its exploratory descriptives had been seen before the main wave, so the primary confirmatory cohort is the main wave alone (N = 119 valid) and the branch pooling the identically-fielded soft-launch cohort back in (N = 133) is reported as robustness. Every number below is the output of the analysis code frozen by SHA-256 before data collection. Deviations from the registration, collected in one place: (i) the N shortfall (133 valid of 200 registered; hosting outage); (ii) the pilot's completion-time criterion missed by 24 seconds in the fast direction; (iii) per-trial display order not retained in the fielded export; (iv) screen-level timing not captured; (v) the trial-level mixed model run on a variational approximation (the package's final engine may differ). Nothing else deviated; every confirmatory test ran as registered, the primary-cohort decision above being the one post-registration analytic choice.

Table 14. *Sample flow and exclusions (main wave; soft-launch cohort in note)*

	US	UK	Total
Completed sessions	50	88	138
Excluded: attention (≥ 2 of 3 checks)	0	0	0
Excluded: speed (< 3 min) /	0	0	0

straight-lining			
Excluded: manipulation / comprehension checks	10	9	19
Excluded: duplicates / incomplete	0	0	0
Valid N (by condition 1/2/3)	40 (16/15/9)	79 (26/28/25)	119 (42/43/34)

The soft-launch cohort adds 20 completed US sessions, of which 14 survive the same exclusions; it enters only the pooled robustness branch (N = 133). Every exclusion in the main wave came from the manipulation- and comprehension-check rules — no participant was excluded for inattention, speeding, straight-lining, duplication, or incompleteness — and manipulation-check gating shows no between-condition asymmetry (MC1 $\chi^2(1) = 0.12$, $p = .73$; MC3 $\chi^2(2) = 3.67$, $p = .16$, computed on the pre-exclusion sample).

Table 15. *H1 (P1): reliance calibration Δ by condition (primary cohort, N = 119)*

	Control	Disagreement	Disagreement + Evidence
Δ, mean (SD)	0.238 (0.332)	0.465 (0.274)	0.422 (0.299)
Over-reliance: P(accept | incorrect)	0.603	0.434	0.510
Under-reliance: 1 − P(accept | correct)	0.159	0.101	0.069
Confirmatory contrast: pooled (2,3) vs control, OLS with country covariate	b = +0.217, 95% CI [+0.104, +0.330], Holm-adjusted p = .0005 (d = 0.67)		

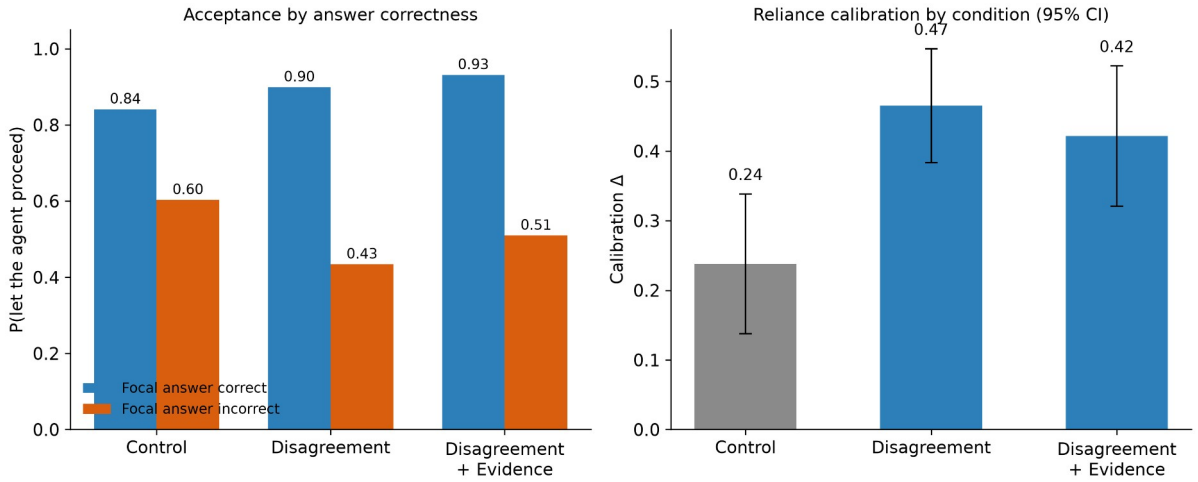
Table 16. *H2 (P2): willingness to delegate, receipt vs disagreement alone (primary cohort)*

	Disagreement (2)	Disagreement + Evidence (3)
Willingness-to-delegate composite, mean (SD)	3.09 (1.61)	3.45 (1.79)
Delegation ladder, median	2	2
Confirmatory contrast: 3 vs 2, OLS with country covariate	b = +0.41, 95% CI [−0.35, +1.18], Holm-adjusted p = .29	

H1 is supported in this design for the calibration contrast; the pre-registered mediation through perceived risk was null, so P1 as stated is not supported by this test. Facing a single confident answer, participants barely discriminated: they let the agent proceed on 84% of correct answers but also on 60% of incorrect ones (Δ = 0.238, 95% CI [0.14, 0.34]; treatment conditions 0.465 [0.38, 0.55] and 0.422 [0.32, 0.52]). The disagreement display nearly doubled calibration (pooled Δ = 0.446; b = +0.217, p = .0002, Holm-adjusted p = .0005; Welch d = 0.67), and it did so on both sides at once: over-reliance fell (60% → 43% and 51%) while under-reliance also fell

(16% → 10% and 7%) (Figure 20). That pooled average is carried by a single item and is best read as an existence proof rather than a robust mean: on the one high-confidence dissenter trial that control participants were accepting at 74%, the display cut acceptance of the wrong answer to 24%, and with that trial removed the confirmatory contrast is no longer significant (d = 0.09, p = .54), while dropping any other single item leaves it intact (d = 0.47–0.89). The exploratory decomposition below (Tables 17–18) locates the effect item by item. Its concentration is the same headroom dependence reported there at the country level: the display closes a calibration gap only where one is open, and one of six items opened it wide.

Figure 20. *What the disagreement display does to reliance (primary cohort)*



Left: acceptance rates by condition, split by whether the focal answer was correct (calibration = the gap between the pair of bars). Right: the per-participant calibration score Δ by condition with 95% CIs; the confirmatory contrast pools the two disagreement conditions against control.

The pre-registered two-sided framing existed because the second-opinion literature documents interventions that buy reduced over-reliance at the price of increased under-reliance; this one did not pay that price. The signal made participants more discriminating, not more skeptical, and a signal-detection decomposition of the aggregate rates puts numbers on that: sensitivity d′ rose from 0.74 in Control to 1.44 and 1.46 in the treatment conditions (computed from unrounded rates) while the response criterion barely moved (−0.63 vs −0.56 and −0.76). The trial-level mixed-effects model recovers the same structure, with large positive correctness × condition interactions for both treatment conditions. Every registered branch agrees: the pooled robustness branch (b = +0.215, p = .0001, d = 0.68), and the intention-to-treat branch retaining all completed main-wave sessions regardless of check failures (b = +0.2174 vs +0.2169 per-protocol, p = .0001, n = 138), with exclusions balanced across conditions (8/6/5; $\chi^2(2) = 0.34$, p = .85), the differential-attrition check §6.1 commits to (manipulation-check exclusions are post-treatment conditioning, per Montgomery, Nyhan, & Torres, 2018, hence the equal standing of the unconditioned branch). One design confound qualifies the reading. The disagreement conditions differ from Control not only in what the signal says but in how much is shown, three positions and confidences instead of one, so the calibration gain could in principle reflect information volume rather than the agreement signal's content. Two observations cut against the volume reading without settling it: the effect tracks the *direction* of the displayed signal item by

item, since on incorrect-focal items where the display showed dissent, acceptance fell, while on the one item where the display showed unanimous confident agreement, acceptance of the wrong answer *rose*; and the trial-level interaction is with answer correctness, which volume alone does not predict. The clean discriminating arm, several systems shown agreeing on every trial with information volume held constant, is specified for the next wave.

H2 is not supported. Among participants already shown the disagreement display, the receipt produced no detectable effect on the delegation composite at realized power (3.45 vs 3.09; b = +0.41, p = .29; pooled branch b = +0.34, p = .33; composite reliability: main-wave Cronbach α = .89). Three pre-registered secondary results shape how the null should be read. The receipt *was* received: median session time in the receipt condition ran about a minute and a half longer than in the disagreement condition (11.0 vs 9.4 minutes; screen-level timing was not captured in the fielded export, so total duration stands in for the registered time-on-stimulus proxy). The extra time does not establish that participants used the receipt. The fielded receipt was a static card, and its "view full record" affordance was not live inside the survey, so the manipulation tested the receipt's *presence*, not its *exercise*. A live-drill-down receipt is the stronger manipulation the next design includes. This is a null for one static receipt display, not for records, checkability, or P2 as written. The receipt also moved its pre-registered mediator. Perceived process control rose in the receipt condition (a = +0.60; scale reliabilities, main wave: process control α = .93, risk α = .83, neutrality α = .90), process control strongly predicts delegation (b = +0.60), and the bootstrap indirect path is positive and excludes zero (indirect = +0.35, 95% CI [+0.02, +0.78], 5,000 resamples), while the direct path is near zero. This is exploratory mediation evidence: the mediator was measured post-treatment in the same instrument, so common-method variance and reverse interpretation remain possible. Sensitivity bounds what this null can settle. The realized cells detect d ≈ 0.49 at power .80 only at the registered N, so a receipt effect of the size the mediation path implies is exactly the size this sample cannot resolve. The result is inconclusive about small and medium effects and decisive only against larger ones. Formally, equivalence testing bounds the effect only coarsely (TOST establishes $|d| < 0.64$ at $\alpha = .05$, not $|d| < 0.5$), and a JZS Bayes factor gives $BF_{01} \approx 2.9$ (pooled branch 3.1), anecdotal evidence favoring the null on the standard bands (at or just past the anecdotal/moderate boundary of 3); it neither demonstrates absence nor supports presence. The measurement section predicted as much: checkability is an on-demand property, and its value should surface at the moment of dispute, claim, or audit, so a static receipt read once in a hypothetical purchase may be the wrong moment to observe it. The registered conclusion: the first-exposure test of P2 failed to move delegation, and the operator-side economics of the receipt (Section 5) do not rest on this consumer-side result. Two readings remain: either the evidence component contributes less at the moment of decision than the framework claims, or this operationalization is the wrong test of it. The dispute-time account is a post-hoc hypothesis, scheduled as the follow-up design named at the end of this section, not the resolution of the null. A third reading comes from the aversion literature. A detailed receipt is a salience manipulation, an exposure of machinery that reminds the consumer how much can go wrong at exactly the moment of delegation, the transparency paradox in which procedural detail depresses rather than builds trust (Kizilcec, 2016; Dietvorst, Simmons, & Massey, 2015). Those competing readings, a weaker-than-claimed component, a wrong-moment test, and salience backfire, make different predictions and are separable by design.

The certified-consensus-error mode reproduced in human behavior. One incentivized trial carried, by design, a genuine consensus error, with all three systems confidently converging on

the same wrong answer, mirroring the 0.5–2.2% open-form consensus-error rate the mechanism study measures. On that trial, a single item and so an observation rather than an estimate, the disagreement display did not protect anyone and, in both countries, directionally *increased* acceptance of the wrong answer (by 4 and 14 percentage points relative to control): participants rationally read unanimous independent agreement as certification. So the architecture's one systematic failure mode transfers to its users with high fidelity. Section 3 gave the design warning, that consensus certifies error exactly where the panel's diversity fails, and the probe shows the same failure in behavior.

Exploratory: where the calibration benefit lands.

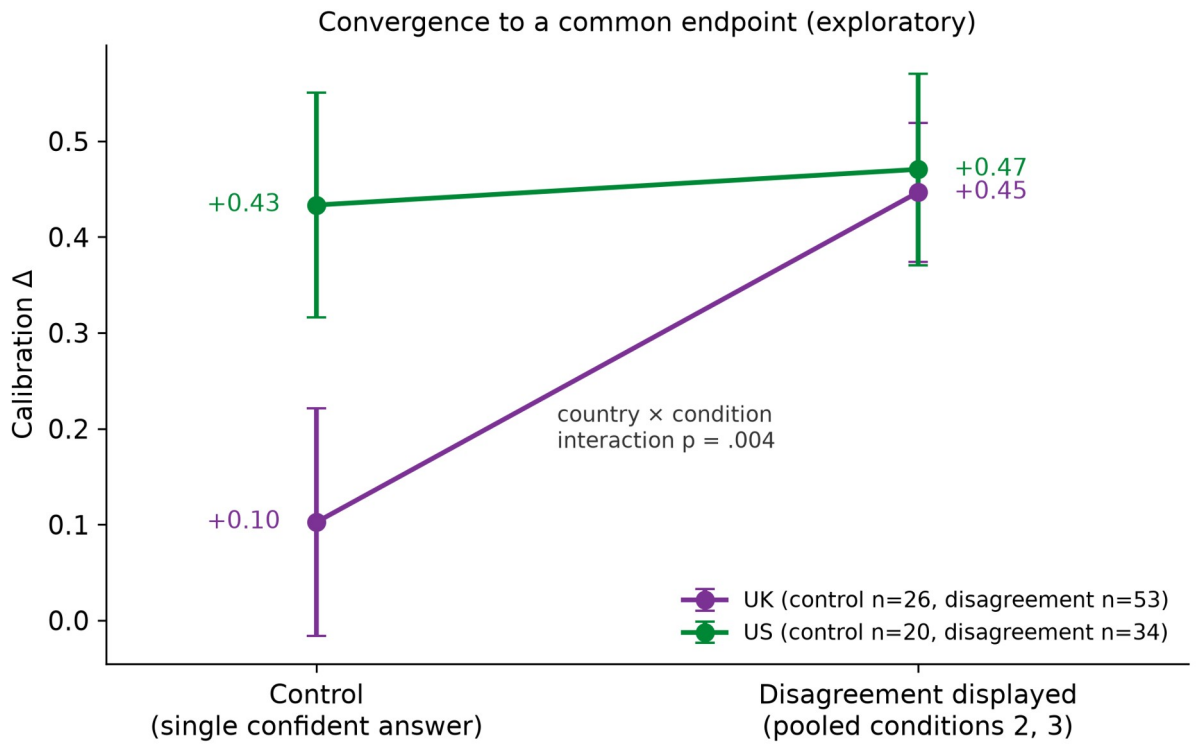
All of what follows is exploratory and hypothesis-generating. The country cells are small, and every within-country p-value below should be read as descriptive. The calibration effect concentrated almost entirely in the UK sample and was absent in the US sample:

Table 17. *Calibration Δ by country and condition (exploratory)*

	Control Δ	Disagreement (pooled 2, 3) Δ	Within-country effect size
UK (n = 79)	+0.103 (n = 26)	+0.447 (n = 53)	d = +1.19
US (n = 40 primary / 54 pooled)	+0.458 / +0.433	+0.444 / +0.471	d = −0.05 / +0.13
Country × condition interaction			b = −0.36, p = .002 (primary); b = −0.31, p = .004 (pooled)

The decomposition is more informative than the interaction. US Control participants already discriminated sharply without any architectural signal, accepting 92% of correct and only 48% of incorrect answers, while UK Control participants barely discriminated at all (78% vs 68%). Under the disagreement display the two countries became nearly indistinguishable, item by item (maximum cell difference 0.07 across all six trials). The intervention converged both samples to the same calibration endpoint ($\Delta \approx +0.45$), equalizing a baseline difference it did not create (Figure 21). Part of this pattern is mechanical. Δ is bounded, so available headroom shrinks as baseline calibration rises, and "benefit concentrates where baseline is weakest" is partly a property of the measure. The ceiling alone, however, does not force identical endpoints or the item-by-item profile match. The most defensible reading is that the six items were easy enough for one panel to discriminate unaided, so the display closes a calibration gap where one exists. That reading is about headroom, not about nations, and equally about stimuli: the item set itself was culturally tilted (two dissenter items carried US-local content), so the convergence profile is a property of these six items as much as of these two panels.

Figure 21. *Convergence, not uniform response (exploratory)*



Calibration Δ by country and condition: the US control group already discriminates (Δ ≈ +0.45); the UK control group barely does (Δ ≈ +0.10); under displayed disagreement both countries land on the same endpoint (Δ ≈ +0.45). Benefit concentrates exactly where baseline calibration is weakest.

The US advantage in Control, meanwhile, is concentrated on the two dissenter items with plausibly US-local content. One is a Museum of Bad Art question on which the wrong answer, "the Mona Lisa," is absurd to anyone who knows the museum's premise (US acceptance 5% vs UK 27%). The concentration on those two items points at item-level familiarity and not at national temperament.

Table 18. *Acceptance of incorrect focal answers by item and cell (exploratory; pooled cohort)*

Cell	Dissenter wrong, conf 95 ("Mona Lisa" item)	Dissenter wrong, conf 98 (ICM-city item)	Consensus wrong, conf 99
UK Control (n = 26)	0.27	0.85	0.92
US Control (n = 20)	0.05	0.60	0.80
UK Disagreement (n = 53)	0.21	0.25	0.96
US Disagreement (n = 34)	0.18	0.24	0.94

The intervention's protective effect sits almost entirely on the high-confidence dissenter item (UK −0.60, US −0.36); on the consensus-error item the display slightly increases acceptance in both countries — unanimous agreement read as certification, the behavioral image of the mechanism study's certified consensus errors. Because the protective effect sits on this one item, the pooled confirmatory contrast does not survive its removal (d = 0.09, p = .54) though it survives removal

of any other single item (d = 0.47–0.89) — the quantitative form of the effect being carried by one item.

Table 19. *Attitude scales by country (exploratory; pooled cohort): the gap is behavioral, not attitudinal*

Scale (1–7 unless noted)	UK mean	US mean	Direction
Pre-exposure trust in AI	4.68	4.70	identical
Willingness to delegate (composite)	2.94	3.43	US higher (p = .08)
Delegation ladder (0–4)	1.71	1.98	US higher (p = .08)
Perceived risk	5.54	5.13	US lower (p = .09)
Perceived process control	4.17	4.56	US higher (p = .14)
Perceived neutrality	4.75	4.62	n.s.

Every attitudinal difference that exists points the "wrong" way for a US-skepticism account: US participants are more delegation-willing and less risk-perceiving, yet discriminate better behaviorally in Control. Measured attitudes explain none of it: adjusting the baseline country gap for demographics and all five scales attenuates it by at most 15%, and bootstrap indirect paths through the scales are null.

Cross-national surveys agree with the attitudinal null. Stated willingness to trust AI is essentially identical in the US and UK (41% vs 42%: Gillespie et al., 2025), and both countries sit in the same high-wariness "Anglosphere" cluster on AI nervousness (Ipsos, 2025). The one population-level asymmetry pointing the observed direction is AI literacy. UK respondents self-report less skill and confidence in using and evaluating AI (36% vs 42% on skills: Gillespie et al., 2025), and US adults are among the most AI-exposed publics surveyed (Pew Research Center, 2025). Two accounts therefore survive the exploratory record. On a familiarity account, the population less equipped to evaluate AI advice gains most from an architecture that evaluates it structurally. The other account is panel composition: the UK arm of the recruitment platform is its older and more experienced pool, and online panels are imperfect instruments for between-country inference (Gvirtz & Sabherwal, 2024), so the "country" label may partly proxy for panel tenure. The design cannot separate the two, and does not need to, because on either account the substantive finding is the compensatory profile itself. Given the cell sizes, the safer default is to read the pattern as a property of these two panels until a powered replication speaks; no theoretical claim in this paper rests on it.

That profile is not the default outcome of overreliance interventions. Cognitive forcing functions reduce over-reliance but concentrate their benefit among users already high in Need for Cognition, so the analytically inclined get more analytical (Buçinca et al., 2021), and adaptive trust-calibration research treats targeting the miscalibrated as a design goal requiring active monitoring (Okamura & Yamada, 2020). Here the targeting came free. Displayed disagreement is self-triggering: it appears exactly when the panel disagrees and demands nothing of the user in advance. Its benefit landed squarely on the group whose baseline discrimination was poorest, and it left the already-calibrated group's accuracy intact rather than degrading it. For a consumer-

protection mechanism that is the distributional profile one would specify ex ante, with help concentrated where miscalibration concentrates and no tax on those who do not need it.

Pre-registered secondary analyses: the remaining ledger.

The remaining pre-registered secondary analyses sit here in one place (primary cohort; pooled-branch values in the replication package). Mediation of the disagreement effect through perceived risk, the P1 pathway the design anticipated, is null. The display, if anything, slightly *reduced* perceived risk ($a = -0.40$), and the bootstrap indirect path is indistinguishable from zero ($+0.006$, 95% CI $[-0.013, +0.029]$). The calibration effect is direct, consistent with participants using the disagreement pattern as information about the answer rather than as a general alarm about the transaction. Perceived orchestrator neutrality, measured rather than manipulated, is a strong correlate of willingness to delegate controlling for condition and country ($b = +0.47$, $p < .0001$), the correlational trace P3 predicts, awaiting its manipulation. The delegation ladder shows the same ordering as the composite without reaching significance (ordered logit, receipt vs disagreement: $b = +0.70$, $p = .11$; medians tie at 2 of 4). The unregistered condition 3-versus-1 contrast on delegation is directionally positive and non-significant ($b = +0.52$, $p = .15$). One deviation: the fielded export did not capture screen-level timing, so the registered time-on-stimulus receipt-reading proxy is approximated by total session duration (reported under H2 above). The intended screen-level check should be built into the next wave's instrument.

This study tests P1 and P2 only. P3 is measured, not manipulated, and P4 and the full joint-variation conjoint of Section 5 remain the specified next study. The behavioral validation of the triad is therefore incomplete: one component supported, one inconclusive, one measured but unmanipulated. One longitudinal question sits above the program. Persistent disagreement display may stabilize long-run adoption, as a dampener on the oscillation between algorithm appreciation and one-strike aversion, or it may exhaust that adoption through continuous exposure to machine fallibility. A single-session design cannot say. Two studies motivated by the results above are added to the program. One is a dispute-time design that re-tests the receipt at the moment its value is predicted to live, using a false-receipt control, a receipt whose hash does not recompute, to separate verifiability from its rendering. The other is a replication powered for the country moderation reported above (Table 17).

7. Conclusion

As AI mediates consequential consumer decisions, and increasingly makes them, I have argued that trust must migrate upward in the stack: from the model, to the architecture, to the record. Diversity, evidence, and neutrality together produce a service whose trustworthiness can be checked rather than taken on faith. What is being built is not better persuasion but economic trust infrastructure, the accountability chain that lets delegation scale. The flagship wave measures that at the top of the market. Pooled across the wave's ten complete flagship arms, models sold as each laboratory's best were confidently wrong on nearly half of their resolved answers offered at stated confidence ≥ 80 (46.6%, and the figure barely moves when unresolved items are re-scored as errors). The partial Meta arm is reported separately, and including it leaves the figure effectively unchanged. A fixed three-model disagreement gate above them, with no training and no tuning, removes four-fifths of those failures at a stated, inspectable cost in coverage. Because the substrate here is purely factual, that rate should be read as an upper bound for mixed

commerce tasks. What would license the claim for commerce proper is a retrieval-grounded, tool-using replication on a constrained product menu, where consensus errors behave differently than in open factual space. A pre-registered experiment (Section 6) put the same architectural signal in front of consumers and returned evidence that they can act on it as a correctness signal: the disagreement display cut acceptance of a confident wrong recommendation from 74% to 24% on the qualifying item. In an exploratory country split it concentrates in the over-reliant UK cell and rests on that single item (the pooled contrast does not survive its removal, d = 0.09). That is an existence proof under these stimuli, not yet a general effect; a stimulus-sampled confirmatory test is planned.

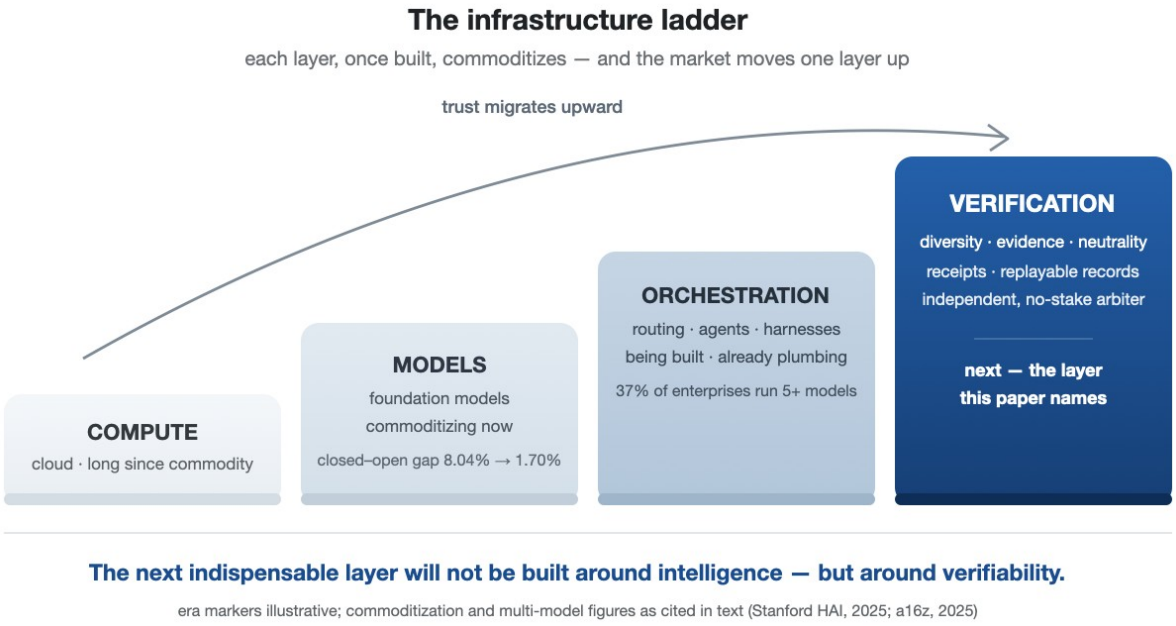
This layer deserves a market name of its own, *evidence infrastructure*, to be sold the way cloud, identity, and payments infrastructure are sold, because it answers the same kind of question: "can this be relied on, and how would we know?" That is a different question from "is this smart?" Security and privacy both went this way: differentiator, then expectation, then entry requirement. Architectural trust may plausibly follow the same path, at a pace set by the regulatory and liability pressure already visible.

My claim on the theory side is that the unit of analysis for consumer trust in AI services should move from the model to the architecture, because the trust-relevant properties actually live there: independence of verification, replayability of decisions, absence of conflicted interest. Models keep changing under every service, faster than any reputation can track. The closed-open performance gap collapsed from 8.04% to 1.70% in a single year (Stanford HAI, 2025), and leadership of the enterprise model market changed hands in a single spring (Kharazian, 2026). A layer that commoditizes this fast is not where trust can anchor. What serves a consumer is increasingly routing across models and price tiers, cost-driven, invisible, and unverified, and a chosen best model serves fewer and fewer. The step that remains, and the one I specify here, is to move from the best-routed architecture to one whose verification a consumer, an insurer, or a regulator can hold still long enough to check. Underneath that sits a transition from AI assistants to AI fiduciaries. Fiduciaries, in every domain that has ever had them, are constituted not by their talents but by their duties: to keep records, to submit to audit, to have no adverse interest. Architectural trust builds those duties into the machinery.

The same requirement applies to how the system itself grows. A system asking to be trusted with consequential action should have to grow the way it answers, verifiably. Configuration changes to an orchestration layer should commit or roll back through audited statistical gates against held-out evidence, with no automated process weakening a safety rule, only adding one. The research deployment in the case study implements those gates.

AI infrastructure now needs a verification layer (Figure 16). The industry has built the compute layer and the model layer, and is now building the orchestration layer. After it comes *verification*, the machinery that makes an AI system's behavior checkable rather than credible. *The* warrant *for AI trust should become machine-verifiable rather than reputation-based; trust itself stays human, and what can be machine-checked is its grounds.* If that is right, the next indispensable layer of AI infrastructure will be built not around intelligence but around verifiability.

Figure 16. *The infrastructure ladder*



Compute, then models, then orchestration, and verification as the next indispensable layer.

Author's note and disclosure

The deployment described in Section 4 is a reference implementation built and operated by the author; the methods it embodies are open, and the author is a named applicant on pending patent applications, not granted patents, covering AI orchestration and multi-model validation systems of the kind this paper examines. It serves here as the study's research instrument, and the author has a developer's interest in the architecture it embodies; Section 4 should be read as an existence proof and design testimony, not as independent evaluation. All models in this study were accessed through the vendors' standard commercial APIs on ordinary terms. No external funding supported this work.

Tools and verification. AI assistants were used as instruments under the author's direction; the ideas, research and conclusions are the author's. Load-bearing factual claims were checked against primary sources where available and are attributed in-text with their evidentiary status (press reports as reports, allegations as allegations, preprints as preprints), and a dated claim-verification log is maintained with the manuscript's working materials. Correctness in the experiment is never judged by any panel model (scoring is objective string matching or a local open-weights grader), the per-vendor discrimination and calibration numbers are printed in the main text (Tables 7 and 7a), and every reported number is recomputable from the deterministic replication package (publicly available at https://osf.io/jtvu9/).

Data and code availability

Every reported number in the mechanism experiments (Section 3) is recomputable from a deterministic replication package — the raw model responses of both waves (including the

sixteen-model collection of 24,000 answers and the flagship wave's graded records), the collection runners, the deterministic grader, the analysis scripts, and the derived CSVs (including the risk-coverage curves and the 26-point cost–accuracy frontier, with each point's billed-versus-list price basis flagged) — publicly available at https://osf.io/jtvu9/, a public component of the pre-registration's source project (folder `replication-package-v2`; the self-calibration program archive of Section 3 is deposited alongside it). The frozen SimpleQA subsample (fixed by seed and SHA-256 hash) is already public at the mechanism study's OSF pre-registration (https://osf.io/4y9dv, registered 23 July 2026). The consumer experiment's pre-registration is registered on OSF under embargo (frozen before main data collection; registration https://osf.io/pqeg6/?view_only=87eb0a88579647b48e93d4a475795d5f). The participant pool is closed; the survey instrument, the vignette and judgment-block stimuli, the receipt-minting records of Section 6, and the consumer-experiment data with the frozen analysis code will be released into the same OSF component upon publication.

References

Entries marked arXiv are preprints; industry announcements and 2026 press items are contemporaneous reports rather than settled literature and are cited as such.

a16z — Andreessen Horowitz (2025). *How 100 enterprise CIOs are building and buying gen AI in 2025.* a16z.com/ai-enterprise-2025.

Accenture (2026). *Talk to my AI agent: The new rules of brand value.* Survey of 25,590 consumers in 16 countries, January 2026.

ACI Worldwide (2026). *Consumer attitudes to AI-driven shopping and payments.* YouGov survey of 2,080 UK adults, fielded 19–22 June 2026. Press release.

Ahmad Husairi, M., & Rossi, P. (2024). Delegation of purchasing tasks to AI: The role of perceived choice and decision autonomy. *Decision Support Systems*, 179, 114166. DOI 10.1016/j.dss.2023.114166.

AIUC (2025). *AIUC-1: The agent standard.* Artificial Intelligence Underwriting Company.

Alavi, S., & Nozari, S. (2026). When agents shop for you: Role coherence in AI-mediated markets. *arXiv:2604.26220*.

Allouah, A., Besbes, O., Figueroa, J. D., Kanoria, Y., & Kumar, A. (2026). What is your AI agent buying? Evaluation, biases, model dependence, and emerging implications of agentic e-commerce. *Proceedings of the ACM Web Conference 2026*, 8697–8700. https://doi.org/10.1145/3774904.3792943

American Express (2026). American Express debuts Agentic Commerce Experiences (ACE) developer kit and announces industry-first protection for registered agent purchases. Press release, 14 April 2026.

Anderson, L. (2006). Analytic autoethnography. *Journal of Contemporary Ethnography*, 35(4), 373–395.

André, Q., Carmon, Z., Wertenbroch, K., Crum, A., Frank, D., Goldstein, W., Huber, J., van Boven, L., Weber, B., & Yang, H. (2018). Consumer choice and autonomy in the age of artificial intelligence and big data. *Customer Needs and Solutions*, 5(1–2), 28–37.

Appenzeller, G. (2024). *Welcome to LLMflation: LLM inference cost is going down fast.* Andreessen Horowitz, November 2024.

Aubakirova, M., Atallah, A., Clark, C., Summerville, J., & Midha, A. (2026). *State of AI: An empirical 100 trillion token study with OpenRouter.* arXiv:2601.10088 (report released December 2025).

Bain & Company (2025). *Agentic commerce market projection: $300–500 billion U.S. e-commerce by 2030.* December 2025.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779.

Balaskas, S. (2026). From recommendations to delegation: A systematic review mapping agentic AI in e-commerce and its consumer effects. *Information*, 17(3), 222.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Basu, A. (2026). Tool receipts, not zero-knowledge proofs: Practical hallucination detection for AI agents. *arXiv:2603.10060*.

Board of Governors of the Federal Reserve System & OCC (2011). *Supervisory guidance on model risk management* (SR Letter 11-7 / OCC 2011-12).

Buçinca, Z., Malaya, L. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–21.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239.

California State Legislature (2025). Assembly Bill 316 (artificial intelligence: defenses), Stats. 2025, ch. 672 (adding Cal. Civ. Code § 1714.46), effective 1 January 2026.

Carlson, S. C. (2026). Trust migration in the digital era: Information regimes and the transformation of buyer–seller relationships. *Atlantic Marketing Journal*, 15(1), Article 6.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825.

Cetinkaya, N. E., & Krämer, N. (2024). Understanding vs. trust: A conjoint analysis. Working paper, OSF Project mw4dq.

Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., & Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv:2602.19141*.

Chen, C., Sun, Y., Liao, M., & Sundar, S. S. (2026). When AI disagrees: The effect of second opinion on patients' trust in doctors. *International Journal of Human-Computer Studies*, 213, 103824.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. DOI: 10.1126/science.aec8352.

Choi, H. K., Zhu, X., & Li, S. (2026). When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. *Proceedings of the 64th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), 14284–14311. https://doi.org/10.18653/v1/2026.acl-long.650

Choi, M., Kim, K., Chae, S., & Baek, S. (2025). An empirical study of group conformity in multi-agent systems. *Findings of the Association for Computational Linguistics: ACL 2025*, 5123–5139. https://doi.org/10.18653/v1/2025.findings-acl.265

Clark, J., & Hadfield, G. K. (2020). Regulatory markets for AI safety. arXiv:2001.00078.

CNBC (2026, March 24). *OpenAI revamps shopping experience in ChatGPT after struggling with Instant Checkout offering* (reporting The Information's March 2026 account of the native-checkout pullback).

Daly, C. (2025). Klarna slows AI-driven job cuts with call for real people. *Bloomberg News*, 8 May 2025.

de Bellis, E., & Johar, G. V. (2020). Autonomous shopping systems: Identifying and overcoming barriers to consumer adoption. *Journal of Retailing*, 96(1), 74–87.

Detjen, H. H. J., Densky, L., von Kalckreuth, N., & Kopka, M. (2025). Who is trusted for a second opinion? Comparing collective advice from a medical AI and physicians in biopsy decisions after mammography screening. *CHI 2025*.

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics, 61*(10), 1409–1427.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.

DigiCert (2026). *AI Trust Architecture.* Product announcement, April 2026.

Ding, K. (2026). When LLMs agree, are they right? Auditing self-consistency and cross-model agreement as confidence signals. *arXiv:2607.08065*.

Dranove, D., & Jin, G. Z. (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature*, 48(4), 935–963.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 11733–11763.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*(2), 381–394.

Errico, H. (2026). Autonomous action runtime management (AARM): A system specification for securing AI-driven actions at runtime. arXiv:2602.09433.

European Commission (2026). Guidelines on the implementation of the transparency obligations for certain AI systems under Article 50 of the AI Act. Adopted 20 July 2026.

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625–630.

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: A perceived risk facets perspective. *International Journal of Human-Computer Studies*, 59(4), 451–474.

Frank, D.-A., Folwarczny, M., & Otterbring, T. (2026). Consumer acceptance of high-autonomy AI assistants is driven by perceived benefits in online shopping settings characterized by scarcity. *Psychology & Marketing*, 43(3), 538–555. https://doi.org/10.1002/mar.70074

Gartner (2025). *Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027.* Press release, 25 June 2025.

Gartner (2026). *Consumer survey on generative AI in shopping* (May 2026 press release).

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 27(1), 51–90.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 4878–4887 (arXiv:1705.08500).

Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). *Trust, attitudes and use of artificial intelligence: A global study 2025.* The University of Melbourne and KPMG. https://doi.org/10.26188/28822919

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.

Google (2026). *Universal Commerce Protocol.* Announcement at NRF, 11 January 2026; cart, catalog, and identity-linking extension, March 2026.

Google Cloud (2025). *Agent Payments Protocol (AP2).* Announcement, 16 September 2025.

Google DeepMind (2025). SimpleQA Verified. *arXiv:2509.07968*.

Gorbett, M., & Jana, S. (2026). Cross-model disagreement as a label-free correctness signal. *arXiv:2603.25450*.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Sirkovic, P., Myaskovsky, A., Glowaty, G., Weissenberger, F., Orlandi, A., Popovici, D., Palepu, A., Rong, K., Tanno, R., Saab, K., Zhang, F., Blum, J., Carroll, A., Kulkarni, K., Tomašev, N., Zverinski, D., Rendulic, I., Vedadi, E., Hasler, F., Rimanic, L., Boia, M., Budiselic, I., Feinstein, B., Bellaiche, M., Sheffer, T., Freyberg, J., Ratcliff, J., Bertolli, O., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Matias, Y., Manyika, J., Hassabis, D., Xu, Y., Kohli, P., Pawlosky, A., Karthikesalingam, A., & Natarajan, V. (2026). Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122), 487–496. https://doi.org/10.1038/s41586-026-10644-y

Gould, S. J. (1991). The self-manipulation of my pervasive, perceived vital energy through product use: An introspective-praxis perspective. *Journal of Consumer Research*, 18(2), 194–207.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *ICML 2017*, PMLR 70, 1321–1330.

Gvirtz, A., & Sabherwal, A. (2024). The limits of doing global, cross-cultural behavioral science research. *Proceedings of the National Academy of Sciences*, 121(36), e2316690121.

Hadfield, G. K. (2017). *Rules for a flat world: Why humans invented law and how to reinvent it for a complex global economy.* Oxford University Press.

Hadfield, G. K., & Clark, J. (2026). Regulatory markets: The future of AI governance. *Jurimetrics*, 65, 195–240 (arXiv:2304.04914).

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), 993–1001.

Hao, X., Wu, Z., Qiu, Y.-X., Xiao, C., Xu, R., Zheng, S., & Qin, J. (2026). Not all flips are conformity: Decomposing stance convergence in multi-agent LLM debate. *arXiv:2606.00820*.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46), 16385–16389.

Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, 49(1), 30–50.

Insurance Services Office (2026). *Endorsement forms CG 40 47 and CG 40 48: generative artificial intelligence exclusions.* Effective 1 January 2026.

Internet Engineering Task Force (2026). *An architecture for trustworthy and transparent digital supply chains* (RFC 9943; SCITT working group). IETF. https://datatracker.ietf.org/doc/rfc9943/

Ioku, T., Song, J., & Watamura, E. (2024). Trade-offs in AI assistant choice: Do consumers prioritize transparency and sustainability over AI assistant performance? *Big Data & Society*, 11(4).

Ipsos (2025). *The Ipsos AI Monitor 2025.* Ipsos Global Advisor, 30-country survey, June 2025.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *FAccT 2021*, 624–635.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv:2207.05221*.

Kalai, A. T., Nachum, O., Vempala, S., & Zhang, E. (2025). Why language models hallucinate. *arXiv:2509.04664*.

Kharazian, A. (2026). Anthropic beats OpenAI on business adoption. *Ramp Economics Lab / Ramp AI Index*, 13 May 2026 (corporate-payment data across 50,000+ U.S. businesses).

Kim, D., & Benbasat, I. (2010). Designs for effective implementation of trust assurances in internet stores. *Communications of the ACM*, 53(2), 121–126.

Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 30038–30066.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104–118.

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *CHI 2016*, 2390–2395.

Klarna (2024). *Klarna AI assistant handles two-thirds of customer service chats in its first month.* Press release, 27 February 2024.

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, 646(8083), 126–134. DOI: 10.1038/s41586-025-09505-x.

Kolt, N. (2025). Governing AI agents. *Notre Dame Law Review*, 101 (SSRN working paper 4772956; arXiv:2501.07913).

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 30(4), 941–960.

König, P. D., Wurster, S., & Siewert, M. B. (2022). Consumers are willing to pay a price for explainable, but not for green AI: Evidence from a choice-based conjoint analysis. *Big Data & Society*, 9(1).

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 7, 231–238.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2), 181–207.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.

Lessig, L. (2006). *Code: Version 2.0.* Basic Books.

Li, M. (2025). From cloud-native to trust-native: A protocol for verifiable multi-agent systems. arXiv:2507.22077.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Liao, Q. V., & Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. *FAccT 2022*, 1257–1268.

Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *arXiv:2205.14334*.

Lindsey, J. (2025). Emergent introspective awareness in large language models. *Anthropic / Transformer Circuits*.

Liu, J., Bursztyn, V. S., Ai, L., Wang, H., Choudhary, S., Mitra, S., & Wu, Q. (2026). TeamFusion: Supporting open-ended teamwork with multi-agent systems. *arXiv:2604.19589* (ACL 2026; Adobe Research).

Lizzeri, A. (1999). Information revelation and certification intermediaries. *RAND Journal of Economics*, 30(2), 214–231.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4), 629–650.

Longoni, C., Cian, L., & Kyung, E. J. (2023). Algorithmic transference: People overgeneralize failures of AI in the government. *Journal of Marketing Research*, 60(1), 170–188.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22), 9020–9025. DOI: 10.1073/pnas.1008636108.

Lu, Z., Wang, D., & Yin, M. (2024). Does more advice help? The effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1–31.

Marchal, N., Chan, S., Franklin, M., Revel, M., Keeling, G., Fischli, R., Chandra, B., & Gabriel, I. (2026). Architecting trust in artificial epistemic agents. *arXiv:2603.02960*.

Mastercard (2026). Verifiable Intent (co-developed with Google). Newsroom announcement, 5 March 2026; early pilots reported in Asian and Latin American markets, 2026.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734.

McKinsey & Company (2025). *The agentic commerce opportunity* (Schumacher, Roberts, & Giebel). 17 October 2025.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce. *Information Systems Research*, 13(3), 334–359.

Menlo Ventures (2025a). *2025 mid-year LLM market update.* 31 July 2025 (survey of 150 technical decision-makers).

Menlo Ventures (2025b). *2025: The state of generative AI in the enterprise.* 9 December 2025 (survey of 495 U.S. enterprise AI decision-makers; Menlo Ventures is an Anthropic investor).

Menlo Ventures (2025c). *2025: The state of consumer AI.* 26 June 2025 (survey of 5,031 U.S. adults, with Morning Consult; the ~3% paying-user figure is a revenue-derived estimate).

Möhlmann, M., & Zalmanson, L. (2017). Hands on the wheel: Navigating algorithmic management and Uber drivers' autonomy. *ICIS 2017*.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3), 760–775.

Morgan Stanley (2025). *Agentic commerce forecast: $190–385 billion U.S. e-commerce spending by 2030.* December 2025.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Lawrence Erlbaum Associates.

Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. *ICML 2020*, PMLR 119, 7076–7087.

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, 341(6146), 647–651. DOI: 10.1126/science.1240466.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905–1922.

NIST (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology.

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, 15(2), e0229132.

Papamarkou, T., Alquier, P., Bauer, M., Buntine, W., Davison, A., Dziugaite, G. K., Filippone, M., Foong, A. Y. K., Fortuin, V., Fouskakis, D., Frellsen, J., Hüllermeier, E., Karaletsos, T., Khan, M. E., Kotelevskii, N., Lahlou, S., Li, Y., Liu, F., Lyle, C., Möllenhoff, T., Palla, K., Panov, M., Sale, Y., Schweighofer, K., Shelmanov, A., Swaroop, S., Trapp, M., Waegeman, W., Wilson, A. G., & Zaytsev, A. (2026). Position: Agentic AI orchestration should be Bayes-consistent. *arXiv:2605.00742*. (Accepted at ICML 2026.)

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. DOI: 10.1177/0018720810376055.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans, 30*(3), 286–297.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, 7(3), 101–134.

Pew Research Center (2025). *How people around the world view AI.* 25-country survey, October 2025.

Piehlmaier, D. M. (2023). The one-man show: The effect of joint decision-making on investor overconfidence. *Journal of Consumer Research*, 50(2), 426–446. https://doi.org/10.1093/jcr/ucac054

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, 85(1), 131–151.

Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S., & Narayanan, A. (2026). Towards a science of AI agent reliability. *arXiv:2602.16666* (ICML 2026).

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. *AIES 2022*, 557–571.

Regulation (EU) 2016/679 (General Data Protection Regulation), Article 22 — automated individual decision-making.

Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14 — human oversight of high-risk AI systems; Article 50 — transparency obligations for providers and deployers of certain AI systems.

Ríos-García, M., Alampara, N., Gupta, C., Mandal, I., Mannan, S., Aghajani, A. A., Krishnan, N. M. A., & Jablonka, K. M. (2026). AI scientists produce results without reasoning scientifically. *arXiv:2604.18805*.

Saad-Falcon, J., Buchanan, E. K., Chen, M. F., Huang, T.-H., McLaughlin, B., Bhathal, T., Zhu, S., Athiwaratkun, B., Sala, F., Linderman, S., Mirhoseini, A., & Ré, C. (2025). Shrinking the generation-verification gap with weak verifiers. *arXiv:2506.18203*.

Sabater, J., & Sierra, C. (2005). Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1), 33–60.

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *IUI '23*, 410–422.

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405.

Schmidt, J., & Bijmolt, T. H. A. (2020). Accurately measuring willingness to pay for consumer goods: A meta-analysis of the hypothetical bias. *Journal of the Academy of Marketing Science*, 48, 499–518.

Shapiro, S. P. (1987). The social control of impersonal trust. *American Journal of Sociology*, 93(3), 623–658.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *ICLR 2024* (arXiv:2310.13548).

Shehata, D., & Li, M. (2026). The inverse-wisdom law: Architectural tribalism and the consensus paradox in agentic swarms. *arXiv:2604.27274*.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). MIT Man-Machine Systems Laboratory.

Smit, A. P., Duckworth, P., Grinsztajn, N., Barrett, T. D., & Pretorius, A. (2024). Should we be going MAD? A look at multi-agent debate strategies for LLMs. *ICML 2024*, PMLR 235, 45883–45905 (arXiv:2311.17371).

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. *European Journal of Information Systems*, 25(3), 274–287.

Spiro, T. (2026). The Oracle's fingerprint: Correlated AI forecasting errors and the limits of bias transmission. *arXiv:2605.00844*.

Stanford HAI (2025). *Artificial Intelligence Index Report 2025.* Stanford Institute for Human-Centered Artificial Intelligence (Technical Performance chapter).

Stanford HAI (2026). *Artificial Intelligence Index Report 2026.* Stanford Institute for Human-Centered Artificial Intelligence (Public Opinion chapter, reporting Pew 25-country survey data).

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing*, 62(2), 60–76.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. DOI: 10.1038/s41562-024-02024-1.

Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., & Lewis, P. (2024). Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv:2404.18796*.

Visa (2025). Visa introduces Trusted Agent Protocol: An ecosystem-led framework for AI commerce. Press release (with Cloudflare), 14 October 2025.

Visa (2026). Visa partners with OpenAI to power the next generation of AI commerce. Press release, Visa Payments Forum, 10 June 2026.

W3C (2025). *Verifiable Credentials Data Model v2.0.* W3C Recommendation.

Wallendorf, M., & Brucks, M. (1993). Introspection in consumer research: Implementation and implications. *Journal of Consumer Research*, 20(3), 339–359.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems*, 4(2), 74–81.

Wang, W., & Benbasat, I. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, 6(3), 72–101.

Wang, Y., Zhang, J., Cai, T., Liu, Z., Sun, Q., Sun, Z., Wu, Z., Dong, M., Zheng, M., Yin, X., & Zhu, Y. (2026). From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents. *arXiv:2606.04990*.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., & Fedus, W. (2024). Measuring short-form factuality in large language models. *arXiv:2411.04368*.

Werbach, K. (2018). *The blockchain and the new architecture of trust.* MIT Press.

Wertenbroch, K. (1998). Consumption self-control by rationing purchase quantities of virtue and vice. *Marketing Science*, 17(4), 317–337.

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, 31(1), 137–209.

Yalcin, G., Lim, S., Puntoni, S., & van Osselaer, S. M. J. (2022). Thumbs up or down: Consumer reactions to decisions by algorithms versus humans. *Journal of Marketing Research*, 59(4), 696–717.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 295–305.

Zheng, J., & Zhang, J. (2026). Uncertainty-aware trust estimation for multi-LLM systems via structured expert judgement. arXiv:2607.20529.

Zuboff, S. (2019). *The age of surveillance capitalism.* PublicAffairs.

Zucker, L. G. (1986). Production of trust: Institutional sources of economic structure, 1840–1920. *Research in Organizational Behavior*, 8, 53–111.