CHEAP BUT CREDULOUS:
THE PRICE OF AUDIT RELIABILITY IN THE MARKET FOR AI REASONING

Piyush Akimitsu

STATE UNIVERSITY OF NEW YORK AT ALBANY
Department of Economics
November 2026

No outside funding supported this research. The author has no financial relationships or other potential conflicts of interest to disclose. The views expressed herein are those of the author.

© 2026 by Piyush Akimitsu. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Cheap but Credulous: The Price of Audit Reliability in the Market for AI Reasoning

Piyush Akimitsu

November 2026

JEL No. C51, C93, D82, D83, L15, O33

Keywords: artificial intelligence, large language models, verification, audit technology, costly state verification, credence goods, reliability, verification debt

ABSTRACT

Knowledge work is being increasingly delegated to AI. A principal who delegates such work to an AI agent cannot certify the quality of the output without conducting an audit. Such an audit poses a costly-state-verification problem and is often imperfect. In a within-task experimental design, language models complete graduate level economics tasks with and without a red herring, an irrelevant and distracting passage acting as perturbation, and deliver a final answer, with and without explanation. Four AI Auditors audit the responses given with explanation. The red herring and the solving agent's explanation introduce camouflage and doubt, which displace the signal received by the Auditor and distort its discrimination, the ability to tell apart wrong and correct final answers. A structural model separates each Auditor's discrimination from camouflage and doubt, and then prices various audit policies by their reliability, the share of delivered final answers that are correct, and by the outlay, the time and money prices incurred to deliver that reliability. An audit policy delivers a final answer once affirmation is received from one or several Auditors. Otherwise, it draws a fresh final answer. The solving agents' explanations substitute for the Auditor's own ability, so an audit of the explanation takes less time at the same money price and discriminates well until the camouflage binds. Once the camouflage binds, neither a larger budget nor a panel of Auditors removes that camouflage, since pooling several Auditors' verdicts averages away the errors of the Auditors but not the camouflage which all of them face. An agent auditing its own output can reproduce its own wrong final answer due to entrenchment. Once camouflage and doubt together exceed the discrimination of the ablest Auditor a budget can buy, no audit delivers higher reliability than the absence of audit. Cheapness and credulity are the properties of such machine audit. As the price of machine reasoning falls faster than the reliability of auditing it rises, unverified errors accumulate as verification debt.

Piyush Akimitsu
Department of Economics
State University of New York at Albany
econpi@cloud.com

1 Introduction

The deployment of Artificial Intelligence (AI) by a principal for knowledge and reasoning work hinges on the reliability or the quality of its output. Given that AI is generally substantially faster than humans in solving, an AI agent can be substantially faster in auditing AI output. In the market for AI audit of AI output, the money price is the billed expenditure and the time price is the latency. While the price of solving by AI has fallen substantially, the quality is certified using the benchmarks which the buyers know to be imperfect (Demirer et al., 2026). Final answer accuracy is a key and indispensable component of output quality. Presently, the market for AI reasoning does not sell credible output quality certification or accuracy verification. An agent auditing its own output by re-solving the same task in a separate independent call can reproduce the same wrong final answer, an entrenchment of its own errors. The solving agent's explanation for its final answer could lead an AI Auditor to an affirming verdict because of the fluency of that explanation, a credulous affirmation. Another case is where only the final answer is shown, without the explanation given by the solving agent, and the Auditor is instructed to work the problem out on its own. In such a case, if the Auditor's final answer is the same as the final answer of the solving agent, the Auditor could still affirm the solving agent's incorrect final answer, which could lead to a concordant error. Nevertheless, in the absence of ground truth, it may take lesser time if a Principal has an AI Auditor reading the solving agent's explanation explanation to audit the solving agent's final answer instead of the AI Auditor solving the problem on its own, since the former takes lesser effort. However, this may not be optimal for the Principal since such an audit may have a less time price simply because the Auditor is credulous and easily trusts what it reads. This credulity is exploitable since causes the wrong asnwer to pass as the correct one.

In this audit market, a principal who delegates knowledge work to an AI agent, that involves performing a task that involves solving a problem and giving a final answer, often cannot certify its accuracy upon delivery. Thus, accuracy is a credence attribute of a delivered output (Akerlof, 1970; Dulleck and Kerschbamer, 2006), so a hedonic premium for accuracy as part of the price, 'a la Rosen (1974), cannot arise until the Principal conducts an independent audit that produces an observable signal of accuracy. This is audit may be costly, and comes with the costly state verification problem of Townsend (1979), but it is reliable such that once bought, it reveals the state final answer by model and task, whether it is correct or incorrect. Currently, the rationale for deploying AI at scale rests rapid speed of AI output delivery and affordability of compute, while taking the accuracy of the output as trustworthy. Where the output is not taken for granted, the rationale for deploying AI at scale rests on

the presumption that auditing AI output is much cheaper relative to producing it. This presumption is common in cases such as delegation to AI with audit, autonomous agents and recursive self-improvement (Cunningham et al., 2026).

This presumption has precedents in other fields. In computational complexity theory a problem has stated conditions which a solution must satisfy. A proposed solution is checked by testing it against those conditions, and such checking is precise since a correct solution implies meeting those conditions. A problem belongs to class NP if checking it takes a number of steps or time that grows only polynomially with the number of variables in the problem. The P versus NP question is whether every such problem can also be solved in steps or time that grows only polynomially with the size of the problem. The conjecture is that the hardest of such problems cannot be solved in such a manner. Their solutions remain reachable, since the candidate solutions form a finite list that can be checked against those conditions one by one. However, with every variable added to the problem, the list of candidate solutions keeps getting bigger. Thus, the bigger problems are out of reach in practice. Machine learning makes similar presumption for verifier training (Cobbe et al., 2021). The generator produces many candidate solutions to a problem. The verifier is trained to judge whether a solution candidate is correct. The solution candidate rated highest by the verifier is the one that is accepted. This design rests on the presumption that judging a candidate solution is easier than producing a correct one.

When a capable model gives a wrong final answer, it is displayed with fluent and confident explanations. An AI auditor checking such an answer is often unable to tell it apart from a correct final answer. This accuracy to discriminate between correct and wrong answer can fall towards one half. This is the market analogue of the limit in adversarial estimation where the best discriminator's ability to tell a real from a simulated observations is 0.5 (Kaji et al., 2023).

This paper employs a principal-agent framework and a two-phase experiment to study what the audit reveals and how costly and reliable such an audit is to the principal. There is a two-phase experimental design. The first phase is the agent solving phase, a within-subject experiment where thirty-eight AI models (the agents) answer graduate microeconomics questions (the tasks) independently. One of the arms of the tasks is a clean baseline, the control arm. The other arm of tasks is the red-herring arm, the perturbed arm that has a red herring, an irrelevant and distracting passage that could lead the agents to answer the question incorrectly. A frontier model, different from the 38 agents, acts as a grader that grades their answers using a comprehensive answer key, consisting of the ground truth, the final answer and step-by-step explanation. The grader rules whether the final answer is correct or incorrect.

In the next phase which is the audit phase of the experiment, run independently of the agent solving phase, there are four AI Auditors which audit the responses from the agent solving phase and give a verdict on whether the final answer is correct or incorrect. Three of those Auditors are the solving agents from the agent solving phase which are promoted as Auditors, while the fourth one is the original grader (Anchor Auditor) from the agent solving phase. The three solving agents promoted as Auditors are the one with the highest accuracy (Ceiling Auditor), the weakest accuracy (Floor Auditor)[1] and the third one that was the most robust to perturbation since task perturbation had the least effect on its accuracy (Robust Auditor). There are two passes in the audit phase and the Auditors are not provided with the answer key at all. In the first pass, Pass 1, the Auditor solving pass, the Auditors are shown the problem and only the final answer provided by the solving agents, without the explanation given by the solving agents. The Auditors are asked to solve the problem themselves to give a verdict. In the second pass, Pass 2, the Auditor verifying pass, independent of Pass 1, the Auditors are shown the explanation and final answer given by the solving agents and are asked not to solve, such that the Auditors give their verdict without solving.[2]

An Auditor could give an affirming verdict or affirmation where it rules a response as correct or a rejecting verdict or rejection where it rules a response as wrong. An Auditor ruling on a correct response as correct and on a wrong response as wrong is deemed as a sound verdict. The reduced form exercise gives the sound affirming verdict rate (the share of correct final answers the Auditor rules as correct) and unsound affirming verdict rate (the share of wrong final answers the Auditor rules as correct), for each Auditor, pass and arm cell. The two rates, compared on the scale of a standard normal deviate, give the Auditor's realized discrimination on that pool for each such cell and computed as the signal-detection index d (Table 21).[3] The realized discrimination still doesn't address the Principal's concern for reliability and the time and money price of such an audit. This is because the realized

[1]The weakest solver of the agent solving phase, Gemini 2.0 Flash Lite, was deprecated by its provider before the audit phase, so the Floor Auditor is the weakest solver still served, GPT-5.4 Mini, the second weakest of the 38. Appendix A gives the selection in detail.

[2]Each agent-task run in the agent solving phase is independent of other runs. Similarly, each pass-Auditor-agent-task run in the audit phase is independent of other runs. In the audit phase, the three promoted Auditors do not audit their own responses from the agent solving phase. Each of them audits the responses of all other agents from the agent solving phase.

[3]I refine the terminology from signal detection theory (SDT) rather than simply borrowing it. SDTs discriminability index is the difference between the standard-normal quantiles of two rates, $d' = \Phi^{-1}(C) - \Phi^{-1}(F)$ (Green and Swets, 1966; Macmillan and Creelman, 2005). In SDT d' and ψ would carry the same name, which would blur the distinction. Therefore, I keep d' as the realized discrimination, separate from the discrimination ψ. Sensitivity too cannot be carried over from SDT to this paper. The two states in SDT are whether a signal is present or absent. Here the two states are whether a final answer is correct or wrong. The Auditor sorts one from the other rather than detecting a signal.

discrimination doesn't give any idea about the reasons for the unsound verdict. One reason could be that the Auditor is itself weak and cannot distinguish between a correct and a wrong answer. Another reason is camouflage which leads a model to read a wrong answer as correct and which no auditor may be able to see through. Camouflage has three sources in the structural model which estimate how much each source contributes to the camouflage (Table 2). The first source which acts in both Pass 1 and Pass 2 of the Audit phase is the red-herring, referred to as the perturbation of the task in the structural model, which misleads the solving agents and lowers the average accuracy of final answers given by the solving agents from 0.660 to 0.513 (Section 4.2). It also misleads an Auditor in a similar manner and induces it into giving an unsound affirming verdict. The solving error and the auditing error are correlated. Owing to this correlation, delegating a task to one AI and auditing it with another fails. The effect of camouflage in Pass 1 and Pass 2 of the audit phase differs for each auditor. The magnitude of this camouflage is large and statistically significant for the Floor Auditor and not statistically different from zero for the Robust, Anchor and Ceiling Auditors (with one parameter per Auditor estimated jointly from the verdicts of both passes, Table 2, Section 5.2).

In the solving pass, the Auditor has to solve the problem itself. Just like the red herring distracts the solving agent while solving, the red herring also distracts the Auditor while solving. If the Auditor ends up getting the same wrong answer as the solving agent, that leads the Auditor into giving an unsound affirming verdict, a concordant error. If the solving agents answer is correct and the Auditors wrong answer doesnt match it, that leads the Auditor into giving an unsound rejecting verdict. In the verifying pass the Auditor does not solve but it reads the explanation given by the solving agent. Here, the red herring works on the Auditor through how plausible that explanation looks. A solving agents wrong answer whose explanation is affected by the red herring looks more plausible. This added plausibility is the camouflage due to the red herring. This leads to correct-side displacement in the signal leading the Auditor into giving an unsound affirming verdict. A solving agents correct answer whose explanation leaves the red herring out gets doubted by an Auditor drawn to the red herring. This doubt causes the wrong-side displacement in the signal leading the Auditor into giving an unsound rejecting verdict. All of these effects depend on the Auditor, not on which agent wrote the answer. In the solving pass the red herring makes the task harder to solve. Thus, it is a task level property faced by all Auditors. What differs across Auditors is their own solving ability, which determines how often each Auditor is still able to solve the harder task correctly. The effects in the verifying pass are estimated one Auditor at a time since the Auditors differ in how the red herring affects them individually.

Length of the explanation is a feature of the answer-response itself, which depends on

the solving agent, the task and the arm. The model holds the camouflage created by length common across the four Auditors. Reasoning is a trait of the solving agent and hence differs across agents. Unlike length, it does not vary across tasks or arms for a given agent, since a solving AI agent is either a reasoning or a non-reasoning agent.[4] A reasoning agent writes a more convincing or plausible wrong answer. The model holds the camouflage added by reasoning common across all four Auditors as well.[5]

Thus, the red herring is the only source the model lets vary by Auditor. It is also the only one that works on both sides, the camouflage on wrong answers and the doubt on correct answers. Reasoning and length contribute to camouflage commonly across all four Auditors.

The second source which acts in both Pass 1 and Pass 2 of the Audit phase as well is the reasoning ability of the agent. 17 of the 38 agents reason by default, meaning that they spend tokens on hidden reasoning before they answer. Wrong final answers produced by these agents are affirmed more often than non-reasoning agents' wrong final answers on the same task by the same Auditor (5.3 points more in the Auditor solving pass and 9.8 points more in the Auditor verifying pass, Table 20).[6] The third source which acts only in Pass 2 is the length of the explanation. This works against camouflage because the longer wrong explanations are affirmed less often (5.4 points less per log unit of length, Table 20). A reasoning agent therefore produces the wrong final answers that are hardest to catch. The camouflage an Auditor faces rises with the ability of the agent it audits, so abler agents make for a harder audit.

In the Auditor verifying pass, the Auditor reads a solving agents response and forms a signal. The signal depends on the state (whether the response is correct or wrong) and is built from three parts: discrimination, displacement and noise. A signal threshold separates the two verdicts such that the Auditor affirms when the signal is at or above the threshold (the affirmation side) and rejects when the signal is below the threshold (the rejection side).

There are two kinds of displacements act on the signal. The first one is camouflage, which is formed from the features that make a wrong final answer appear as correct. Camouflage has three sources. The perturbation in the task and the reasoning ability of the solving agent, which add to the camouflage, and the length of the explanation, which works against

[4]See DeepSeek-AI et al. (2025).

[5]Holding reasoning and length common is a modelling choice. The Auditor sees the answer-response. A longer explanation in Pass 2 could in principle affect Auditors by different amounts. The paper makes the choice for two reasons. The contrast between the two arms (red-herring and control) identifies a separate red-herring effect for each Auditor cleanly. Reasoning and length are not randomized. They are observed features of the agent and of the response, so a separate effect for each Auditor would rest on observational variation rather than on an experimental contrast. The second reason is parsimony, which keeps the model tractable by keeping the parameters few and their estimates sharp.

[6]The agent's metacognition traits do not predict which wrong final answers get an unsound affirming verdict by the Auditor in the Auditor verifying pass (Table 17)

the camouflage because longer wrong explanations are affirmed less often. Camouflage moves the mean signal of a wrong answer toward the correct side. It raises the probability that an Auditor affirms that answer, the probability of unsound verdict. Because camouflage belongs to the answer, not to the Auditor, it raises probability of unsound verdict. The second is displacement towards the rejection side. It is a result of the Auditor doubting a solving agent's correct answer where the given explanation rightfully ignores the perturbation. This moves the mean signal of a correct answer toward the rejection side, so fewer correct final answers are affirmed.[7] The control and red-herring arms separate the two and identify each rate one at a time, since the control arm carries no red herring.

Discrimination is the Auditors ability to tell a correct answer from a wrong one. The two kinds of displacements cloud it. The data doesn't reveal the discrimination but the realized discrimination, which is what is left once the displacements have taken place. In the absence of displacement, the discrimination and realized discrimination coincide. In the presence of displacement, a strong Auditor facing heavy camouflage and a weak Auditor facing less or none of the camouflage can show the same realized discrimination. Thus, it gives no idea about the Auditor ability.

Whether a panel can help depends on what the problem is, a weak Auditor or camouflage. A panel or a larger budget raises reliability against a weak Auditor but not against a well-camouflaged answer (Table 29, Section 6.2). A weak Auditors unsound verdicts come from its own signal noise, which it draws from a normal distribution with mean zero, independently of the other Auditors reading the response. Pooling the Auditors verdicts averages the noise toward zero but not the camouflage. Camouflage sets the affirm share the panel converges to, and no larger panel lowers it.

The realized discrimination is discrimination clouded with displacement. A structural model tells them apart. For every policy the principal has the menu of audit policies such as one Auditor's verdict, an Auditor's two passes, a pair of Auditors or a panel of all four Auditors. The audit delivers the answer response to the task once enough of the verdicts affirm it, otherwise a fresh answer response is drawn. The structural model estimates the reliability delivered and the cost in money (money price paid) and in time (time price paid). The framework is similar to costly state verification (Townsend, 1979) but in this paper, the audit technology is not simply assumed but estimated. The audit reveals the state but only with error. The model quantifies this error, for each Auditor-pass-arm. There is no strategic auditee here, since the solving agent does not respond to the audit. This is more of a decision problem rather than a contract to design (Lacker and Weinberg, 1989). The principal takes the estimated audit technology as given and chooses what audit technology

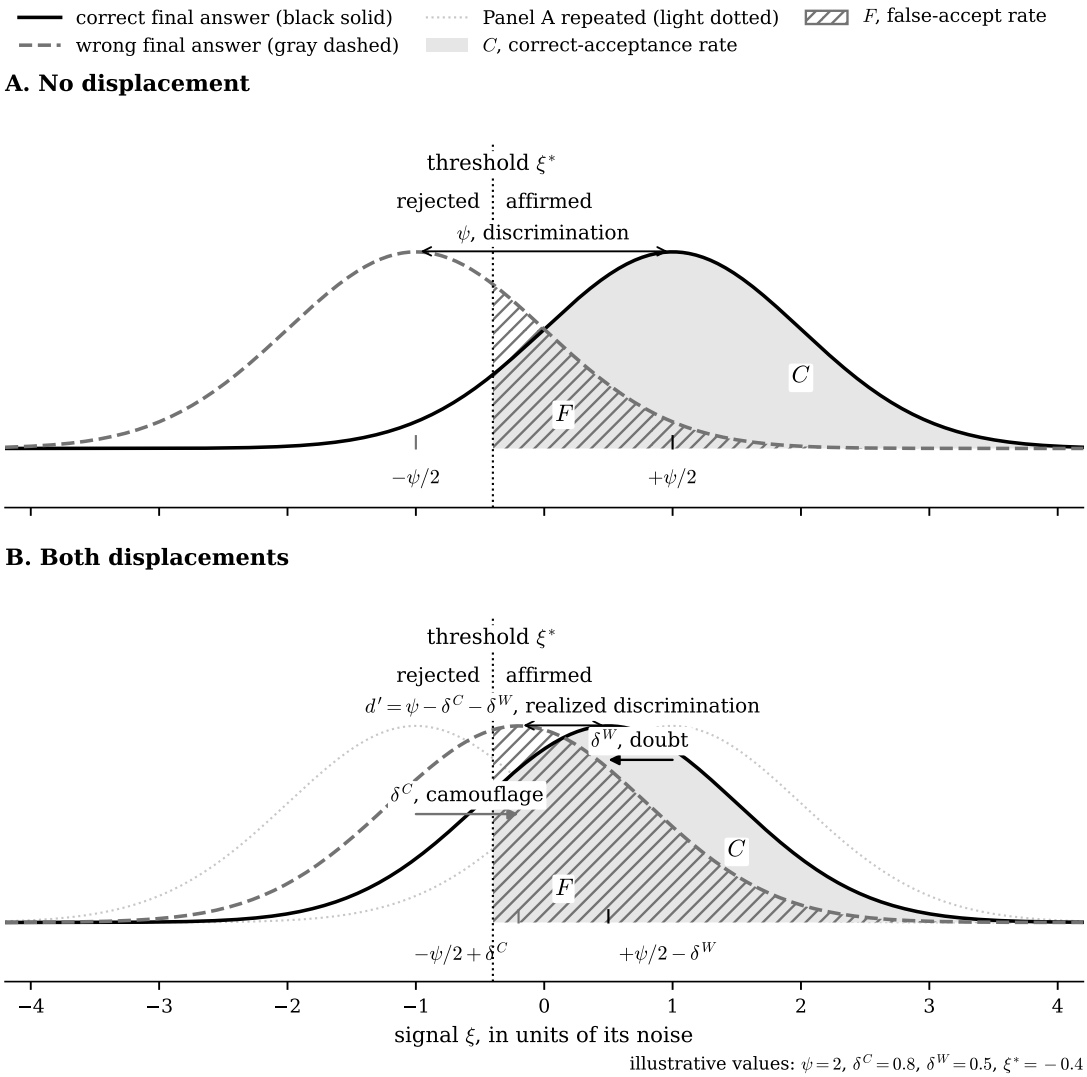
[7] The wrong-side displacement is zero in the control arm, which carries no red herring.

to buy.

In the Auditor verifying pass (Pass 2) the Auditor is not able to tell whether the answer response in front of it is correct or not. It forms a noisy signal about degree of correctness or plausibility of that answer response. In the model, the signal is mean in addition to noise, where the mean is the expected position of the signal on the axis and the noise is a draw from a standard normal distribution (all the distances measured will be in units of standard deviation). The more correct or plausible an answer appears to the Auditor, the higher is the value of the signal. The Auditor gives its verdict by comparing the signal to a threshold it chooses (an auditor-pass-task-arm object). The estimates show how far each Auditor moves that threshold. The responses under audit on one task and arm form that task's pool. A Bayesian Auditor would move its threshold in full with the share of correct final answers in the pool, tightening it when correct final answers are scarce. No Auditor comes close. When the red herring lowers that share, every Auditor barely tightens its threshold. Each moves it by only a small fraction of the Bayesian amount, the strongest included, so base-rate neglect is universal (Table 2, Section 5.2). The threshold therefore gives the principal no protection against a worse pool. The reliability the audit delivers falls with the share of correct final answers in the pool, since the Auditor does not tighten in step with it. The Auditor affirms the answer when the signal clears the threshold and rejects it otherwise. A correct final answer tends to look more plausible to the Auditor than a wrong because its signal is higher on average than the wrong final answer. In the absence of camouflage, the mean signal for a correct answer is at half the value of discrimination above the zero midpoint, while the mean signal for a wrong answer is at half the value discrimination below the zero midpoint (Figure 1). The gap between the two means is the discrimination which rises with the Auditors ability and falls with the tasks difficulty (Section 4.5). On the red-herring arm, there could be enough camouflage coming from length of explanation and red herring itself which could narrow this gap, close it or even carry a wrong answers center above a correct ones. Because of the noise, a correct answers single signal value can lie leftward on the axis and a wrong answers signal can lie in the rightward on the axis. The range for the two signal values is the same, so that the same value of signal can come from either correct or wrong final answer. A single signal value is consistent with both states. Thus, it hides from the Auditor what state produced that response. Camouflage moves the typical value of a wrong final answers signal rightward on the scale, toward the typical value of a correct answers signal. This is the displacement which makes a wrong final answer looks more plausible. The realized discrimination is the reduced gap between the two typical values after the displacement (Section 4.3). The difference between larger discrimination and a larger displacement and the difference between smaller discrimination and a smaller

displacement can give the same value of realize discrimination. Thus, realized discrimination conceals how large the discrimination and the displacement each are.

Figure 1: The audit signal, the criterion and the two displacements



illustrative values: $\psi = 2$, $\delta^C = 0.8$, $\delta^W = 0.5$, $\xi^* = -0.4$

Both panels draw the audit signal ξ of Assumption 1, equation (3), for one Auditor on one task and arm in the Auditor verifying pass. The horizontal axis is the signal in units of its noise. The height of a curve is how densely the signals of that kind of final answer fall at each value. The black solid curve is a correct final answer, whose signal is normal with mean $\psi/2 - \delta^W$. The gray dashed curve is a wrong final answer, whose signal is normal with mean $-\psi/2 + \delta^C$. ψ is the discrimination, the distance between the two means when both displacements are zero. δ^C is the correct-side displacement, the camouflage of a wrong final answer, equation (4). δ^W is the wrong-side displacement, the doubt cast on a correct final answer, equation (5). The dotted vertical line is the threshold ξ^* the Auditor chooses, the criterion of Lemma 1. The Auditor affirms a final answer whose signal lies at or above ξ^* and rejects it otherwise. The hatched area is the false-accept rate F, the rate of unsound affirming verdicts on wrong final answers. The light gray area is the correct-acceptance rate C, the rate of sound affirming verdicts on correct final answers, both from equation (8). Panel A sets both displacements to zero, which is the control arm for a non-reasoning agent at the mean explanation length. Panel B adds the two displacements, which move the two means toward each other, so F rises while C falls at the same criterion. The light dotted curves repeat Panel A for reference. The distance between the displaced means is the realized discrimination $d' = \psi - \delta^C - \delta^W$, which is what the verdicts reveal. The values are illustrative, $\psi = 2$, $\delta^C = 0.8$, $\delta^W = 0.5$, $\xi^* = -0.4$, chosen for legibility. The estimates are in Table 2.

Table 2 shows how the three sources of camouflage lead to the correct side displacement.

The three sources of camouflage enter the correct-side displacement with estimated weights (Table 2). This displacement sits on a wrong final answer and pushes its signal toward the correct side, so a larger displacement is heavier camouflage. The red herring deepens the camouflage of a wrong final answer by 0.283 for the Floor Auditor, an estimate statistically significant. For the Robust, Anchor and Ceiling Auditors the red herrings effect on the camouflage is not statistically significant. Only the red herring differs by Auditor, which is why it is reported one Auditor at a time. Reasoning and length act the same way for every Auditor, so each is one effect shared by all four, not broken out Auditor by Auditor. The reasoning ability of the model deepens the camouflage by 0.414, since a reasoning model writes a more convincing wrong answer. A longer explanation works the other way and thins the camouflage by 0.139 per log unit, since a wrong answer argued at length is easier to catch, an estimate not statistically different from zero. The red herring also reaches a correct final answer. There it works in the opposite direction. An Auditor tempted by the planted passage doubts a correct explanation that ignores it, so the correct answers signal is pushed toward the wrong side. This is a second and separate displacement, the wrong-side displacement, which sits on a correct answer rather than a wrong one (Section 4.3). On the red-herring arm the realized discrimination is therefore the discrimination less both displacements. Camouflage raises the unsound affirming verdict rate. The wrong-side displacement lowers the sound affirming verdict rate. The Floor Auditor suffers both displacements. Its wrong-side displacement is also statistically significant. On red-herring tasks the Floor Auditor affirms more wrong final answers while it also rejects more correct ones, so its collapse is two-sided. The Robust Auditor is immune on both sides (Table 2, Section 5.2). For the Floor Auditor the red herring therefore costs the principal on both sides. Under the Floor Auditor the audit delivers more wrong final answers while it sends more correct ones back for a fresh draw, so the error in what is delivered rises together with the cost of delivering it. The two arms together identify the two rates one at a time. For the Auditor verifying pass the structural model also splits the red herrings effect on each Auditors two rates into three channels. The first is the displacement. The second is the lower share of correct final answers in the red-herring arm, which moves the Auditors threshold. The third is the harder checking the red herring causes. A residual holds their interaction (Table 32, Section 5.2). In the Auditor solving pass the Auditor is instructed to solve. The structural model lets it do one of two things. With a probability the verdicts estimate, 0.61 to 0.79 at the mean task depending on the Auditor, it solves the task and affirms when its own final answer matches the one shown. That match is where the concordant error arises. Otherwise it rules on the bare final answer without solving, which at the estimates means affirming almost all final answers (Table 2, Section 5.2). The same passage that misleads the solving agent misleads

the Auditor that sees it. The errors of solving and of auditing are therefore correlated. That correlation is the configuration in which delegating a task to one AI and auditing it with another fails.

The estimated structural model settles three things the cells alone cannot. First, it prices reliability and finds its cap. The share of correct final answers in the pool is the reliability of delivering an answer without any audit. An audit raises the reliability of the answers it delivers above that share only while the Auditor's discrimination exceeds the two displacements together. Once the displacements meet the discrimination, the Auditor affirms wrong final answers as readily as correct ones. No budget and no panel of any size then raises reliability above the pool's share. This is the camouflage cap (Section 6.2). Beyond it, unchecked errors accumulate as fast as the pool produces them, which is verification debt, the accumulation of unchecked error when the price of solving falls faster than the reliability of checking improves (Goldsmith-Pinkham, 2026). The Floor Auditor on the red-herring arm already sits past the cap in the Auditor solving pass, with a realized discrimination of -0.41, partly through concordant errors, while its Auditor verifying pass clears the cap by only $+0.29$ (Table 21). Second, it places a model's ability as a solving agent and its ability as an Auditor on one scale, where the Robust Auditor audits 0.344 above where it solves, the one such shift that is statistically different from zero (Table 2, Section 5.3). A model can therefore check another model's final answer at a higher ability than it solves the same task. Checking is a capability of its own, which the price of a verdict does not reward (Section 6.3). Third, it extends beyond the Auditors and pools the experiment ran, an extension that rests on the structural model rather than on the data. The wrong final answers that survive an abler solving agent's own checks are the ones that read as correct, so camouflage grows toward the discrimination as solving agents improve. A model auditing its own output inherits its own entrenched errors. On the tasks it reliably fails, it catches its own wrong final answer in the Auditor solving pass with probability between 0.188 and 0.567, so self-audit does not lift the cap, which bounds the self-improvement loop (Table 5, Section 6.4).

Whether machine work can be checked at all is not a question mathematics can settle, even though the celebrated demonstrations of machine proficiency are mathematical. An open problem of the Erdős kind has a structure that dissolves the question. Either a proof is reached or it is not. The verifier is the set of rules of inference, so a search that succeeds certifies itself, while a search that fails has produced no wrong answer, only a not-yet-reached one. The binary is reachable against not-yet-reachable rather than right against wrong. A benchmark task that reaches the model in nearly the form it held in training carries the same defect, with recall standing in for reasoning. Graduate economic reasoning has the

structure the question requires. A definite right answer exists and can be reached, while many plausible wrong answers sit beside it. The reasoner has to identify the operative mechanism, apply it to the case at hand, then land on the one right result rather than on a wrong one that pattern matching alone would suggest. Here a wrong answer is informative. Checking a delivered answer is a genuine problem rather than a second run of the proof checker, so solving and auditing separate in a way that a price can measure. This is why credible accuracy verification is priced on economic reasoning, where the buyer of a machine's answer faces a definite fact of the matter it cannot cheaply certify.

A principal who delegates reasoning has two ways to learn about the output. She can ask the agent about its own output, or she can inspect the output. The companion study measures the first (Akimitsu, 2026b). Asked in advance to predict their own correctness, the models' self-predictions are only weakly informative, with ten of the thirty-six models less accurate on the answers they predict correct than on those they predict wrong. In this paper the agent's metacognition traits do not predict which wrong explanations lead the Auditor to an unsound affirming verdict, Table 17. Because the model has no payoff, the classical incentive toolkit collapses into what the principal asks and how she challenges. Neither instrument prices the failure. What remains is monitoring, so the economics of delegating reasoning to a machine is the economics of verification (Goldsmith-Pinkham, 2026; Dogonowski, 2026). Frontier capability has converged across providers while it keeps rising fast (Egan et al., 2026). That rise is the backdrop for verification debt. Monitoring itself has two technologies. Earlier work detected errors against a verified answer key. This paper removes the key. An Auditor facing a delegated answer with no answer key must either solve the task independently or check the delivered explanation. The Auditor solving pass instructs the first, the Auditor verifying pass the second. The empirical question is how much quicker the Auditor verifying pass is than the Auditor solving pass, then how much of that speed is bought with credulity.

Four reduced-form findings organize the paper. The structural model turns them into prices. First, showing the Auditor the explanation barely moves its accuracy on average. Define verdict accuracy as the share of sound verdicts, verdicts that match the answer key as corrected by the grade ruling of Appendix A.1. A candidate is a response under audit. The within-response gap in verdict accuracy, the Auditor verifying pass (Pass 2) minus the Auditor solving pass (Pass 1) on the same candidate, is +2.3 percentage points (standard error 1.3, Table 9). The average hides a composition. On candidates whose final answer is correct, the explanation helps, raising verdict accuracy from 82.5 to 92.2 percent. On candidates whose final answer is wrong, the explanation hurts, lowering verdict accuracy from 46.8 to 39.6 percent (Table 8). Within task and Auditor, with the control arm as base,

that gap is 24.2 percentage points lower on wrong candidates than on correct ones (standard error 3.8, Table 9), against a raw difference of 16.9 points, the 9.7-point gain on correct candidates plus the 7.2-point loss on wrong ones. The Auditor verifying pass, as deployed, is good at confirming while bad at catching.

Second, the mechanism is credulity rather than information. An explanation raises the sound affirming verdict rate by 9.7 percentage points (standard error 1.9). It raises the unsound affirming verdict rate by 7.5 percentage points (standard error 2.2, Table 9). A fluent explanation makes an affirming verdict more likely regardless of whether the final answer is correct. With the explanation shown, 60.4 percent of wrong final answers pass audit. In the red-herring arm the share is 63.5 percent, while for the Anchor Auditor and the Robust Auditor it exceeds 71 percent (Table 8). The red herring compounds the blindness, with a red-herring-corrupted wrong explanation costing the Auditor a further 7.1 percentage points of the within-response gain in verdict accuracy from seeing the explanation (standard error 2.0, Table 9) beyond an ordinary wrong one. The passage that misled the solving agent also misleads the Auditor. The errors of solving and of auditing are therefore correlated by construction, which is precisely the configuration under which delegating a task to one AI and auditing it with another fails.

Third, the value of seeing an explanation falls with the Auditor's own solving ability. The Floor Auditor gains 13.0 percentage points of verdict accuracy from seeing the explanation (standard error 2.2, Table 10). That gain is not uniform across the arm-task-model responses it rules on. I split the Floor Auditor's verdicts in the Auditor verifying pass by the completion tokens behind each verdict, the tokens the Auditor itself produces in reaching it. On the responses where it produces very few completion tokens, the Floor Auditor affirms more often whether the final answer is correct or wrong. This lifts its verdict accuracy on the majority of responses, whose final answers are correct, while it lowers its verdict accuracy on the responses whose final answers are wrong. The Floor Auditor catches wrong final answers more often only on the responses where it produces more completion tokens (Table 25, Section 3). Part of the Floor Auditor's gain is therefore credulity rather than checking. On the responses where it affirms regardless of the final answer, its verdict accuracy rises only because most final answers are correct, so the Floor Auditor's gain overstates what it checks. The Robust Auditor is unchanged. The Anchor Auditor loses 1.6 points, a change not distinguishable from zero (standard error 2.2). The Ceiling Auditor loses 2.7 points (standard error 1.6), significant at the 10 percent level. The three abler Auditors' changes sit within two standard errors of zero and of one another, so the gradient is a step at the Floor Auditor rather than a smooth decline across all four. An explanation transfers capability to a weak Auditor while transferring contamination to a strong one. This gradient governs

whether the audit can be delegated to cheaper Auditors. It is also the gradient that the recursive-self-improvement literature needs but does not have (Cunningham et al., 2026).

Fourth, the Auditor verifying pass is cheaper than the Auditor solving pass in its time price but not in its money price. The money price of a verdict is the billed expenditure of the Auditor call. Its time price is the latency, the seconds the call takes, which the principal pays in waiting (Acton, 1975; Becker, 1965). Relative to the Auditor solving pass, the Auditor verifying pass writes 0.580 log points fewer completion tokens, the tokens the Auditor itself produces (44.0 percent, standard error 0.035) with a time price 0.420 log points lower (34.3 percent, standard error 0.023), holding task and Auditor fixed (Table 11). Its money price is statistically unchanged on average (log gap +0.041, standard error 0.033), with Auditor-level gaps from +0.35 for the Floor Auditor to −0.38 for the Ceiling Auditor, because the explanation enters the call as tokens the Auditor must take in, which are billed at the posted rate for such tokens. At posted prices the market charges on average the same money price for a credulous fast check as for the slower Auditor solving pass, in which the Auditor is instructed to work the problem out. The Auditor verifying pass buys down the time price rather than the money price. The speed arrives bundled with credulity.

The price of a verdict also depends on which verdict the Auditor gives, holding the task fixed. It costs the Ceiling Auditor more to catch a wrong final answer by rejecting it than to affirm a correct one, in both the money price and the time price. This extra cost is the wrong-answer premium. The premium holds in both passes, so it is present before the Auditor sees the explanation. It costs the Ceiling Auditor less to affirm a wrong final answer than to affirm a correct one. Credulity is the cheap error. The expensive verdicts are the rejections, whether the final answer is correct or wrong. The premium does not arise because the signal sits close to the threshold. Under that account, which I call chronometry, a verdict costs more the closer the signal sits to the threshold. The premium arises instead from what the Auditor does rather than from where the signal falls. I call this composition (Table 1, Section 5.1). The premium is largest for the Ceiling Auditor. For the Floor Auditor, catching a wrong final answer costs less than affirming a correct one. The price of verification therefore rises exactly where verification works, because the Ceiling Auditor is the Auditor that does the catching. Where wrong final answers are well camouflaged, scarce ability rather than cheap capacity must do the checking. This creates a hierarchy of trust that the economics of machine reasoning will have to price. It is the guardian problem of Hurwicz (2008), with the guardians metered by the token.

The estimands are the primitives of that audit market. The structural sections estimate a solving model in the Rasch tradition on the 44,652 scored responses of the agent solving phase, in which the probability that agent m solves task t is a logistic function of the agent's

ability net of the task's difficulty, shifted by the task's red-herring load and by the explanation requirement, the instruction to show a full worked explanation. They then estimate the audit technology of Section 4, in which the Auditor's discrimination in the Auditor verifying pass falls with task difficulty at an elasticity κ, while in the solving model a unit of difficulty lowers the log odds of a correct final answer one for one. The audit market then converts both into the smallest cost per delivered answer, in money and in the principal's time, at which some audit policy reaches a given reliability. The estimated elasticity is $\kappa = 0.218$ (Table 2). On the scale on which solving pays difficulty in full, the explanation shields the Auditor from between three-fifths and four-fifths of that difficulty at the mean task, less for the abler Auditor. The Auditor verifying pass escapes part of the difficulty that solving pays in full by trusting the explanation. The trust is exploitable, since it lets a wrong final answer pass as correct.

The market prices capability in two layers. On the demand side, consumers pay a steep premium for a more capable model and treat providers as closer substitutes as their capabilities converge (Egan et al., 2026). On the supply side, firms buy intelligence through an API priced by the token (Demirer et al., 2026). I price a third layer, the reliability of what a model delivers once an Auditor checks it, together with the audit technology that caps it. No budget raises reliability above that cap.

The paper joins three literatures. In the economics of digitization, Goldfarb and Tucker (2019) made verification cost a first-class cost of digital transactions, where the object verified is the counterparty. Here the object verified is the content of the work itself, which is the axis generative artificial intelligence moves. The frame of the language model as an economic agent comes from Horton et al. (2023) with Brynjolfsson et al. (2025). In machine learning, judging by a language model is standard practice (Zheng et al., 2023), while verifier training is a standard objective (Cobbe et al., 2021). That literature reports agreement rates. This paper reports a technology, prices it, then embeds it in a market. The red-herring lineage (Shi et al., 2023; Mirzadeh et al., 2024; Naeini et al., 2023) supplies the perturbation. The within-response two-pass design follows the within-subject logic of List (2025). Reduced-form standard errors are two-way cluster-robust by task and by agent, 59 tasks and 38 agents in the main sample, with exceptions stated in the exhibit notes (Cameron et al., 2011).

Section 2 details the experiment with the data. Section 3 presents the reduced-form results, which cover verdict accuracy by pass and by whether the candidate's final answer is correct, credulity against information, the capability gradient, the price of checking. Section 4 develops the structural model of solving and auditing. Section 5 reports its estimates. Section 6 takes them to the audit market. Section 7 concludes.

2 Experiment and Data

In the agent solving phase, 38 deployed models, the agents, solve 60 graduate microeconomics tasks under a 2×2 design. The first factor is the red herring, an irrelevant but tempting passage added to the problem text, against a clean control. The second factor is the explanation requirement, the verbatim sentence "Provide a complete explanation." appended to the prompt, against the same prompt with nothing appended. Each agent answers every task version across five waves, which gives five independent draws per cell over 44,652 scored responses on the 59 tasks that the grade ruling of Appendix A.1 keeps. Under the ruled grade, the red herring lowered final-answer accuracy by 14.1 percentage points against a control mean of 61.4 percent, damaging 37 of the 38 agents, while the share giving a complete explanation moved only from 93.1 to 92.3 percent. The corrupted answers did not look corrupted, which is the property this paper needs.[8]

The audit-phase sample is the 4,190 wave-1 responses generated under the explanation requirement, 2,104 from the control arm with 2,086 from the red-herring arm, each carrying a genuine worked derivation. These are real machine answers rather than constructions, so a wrong answer among them is wrong in the way a capable solver is wrong, fluent and confident.

The four Auditors are chosen for their economic roles rather than at random. The Floor Auditor (GPT-5.4 Mini) is the weakest solver of the agent solving phase still served at the audit phase, the Ceiling Auditor (DeepSeek V4 Pro) is the strongest solver in the agent solving phase, the Robust Auditor (Claude Sonnet 4) is the one model the red herring does not move, the Anchor Auditor (Grok 4.3) is the agent solving phase's own grader, which agreed with a second grader on 97.6 percent of regraded cells while it held the answer key. It now audits without the key, so its fall from that agreement prices the key itself. No Auditor grades its own responses, with every candidate response graded by a model that did not produce it.

Each Auditor rules on each response in two independent stateless passes. In Pass 1, the Auditor solving pass, the Auditor sees the task with the candidate's final answer alone, under the instruction to work the problem out and rule. In Pass 2, the Auditor verifying pass, the Auditor sees the task with the candidate's full worked response, under the instruction to work through the steps. In both passes the verdict is about the final answer, an affirming verdict that accepts it or a rejecting verdict that rejects it. The two prompts are reproduced verbatim in Appendix A. Each verdict is scored against a verified answer key the Auditor

[8] That experiment is reported in full in Akimitsu (2026b), with its graded responses in the deposited data (Akimitsu, 2026a), so the new data of this paper are those of the audit phase.

never sees, the corrected key of the grade ruling in Appendix A.1 that retires the automated matcher's errors before any estimate, so a verdict is sound or unsound by construction rather than by a second Auditor's opinion.

The panel of four Auditors over two passes on 4,190 responses is 33,520 calls. I drop truncated calls, calls that ended in a provider error, calls whose verdict does not parse, together with the diagonal pairs where an Auditor meets its own response, which leaves 31,979 usable calls. Correctness comes from the grade ruling of Appendix A.1, which retires the original answer-checker's errors before any estimate and excludes one task whose key was wrong. The ruled analysis sample is 4,116 graded responses on 59 tasks.

3 Reduced-Form Results

The reduced form establishes four facts about verification as it is deployed. Each is computed on the usable sample of Section 2, with its exhibit in Appendix B.

3.1 The derivation confirms, it does not catch

The first question is whether showing the Auditor the worked derivation makes its verdict more accurate. Define verdict accuracy as the share of verdicts matching the corrected answer key. Table 9 reports the within-response contrast between the two passes, the Pass-2 verdict accuracy minus the Pass-1 verdict accuracy on the same candidate. The average gap is +2.3 percentage points, standard error 1.3, so on average the derivation barely moves the verdict. Figure 2 plots verdict accuracy for every Auditor, arm and pass.

The average hides a composition that reverses across the two states of the candidate. Table 8 splits verdict accuracy by whether the candidate's final answer is right or wrong. On candidates whose answer is correct, seeing the derivation raises verdict accuracy from 82.5 to 92.2 percent. On candidates whose answer is wrong, it lowers verdict accuracy from 46.8 to 39.6 percent, a raw gap of 16.9 points. Within task and Auditor the relative reversal is 24.2 percentage points, standard error 3.8, the Wrong_i coefficient of column (3) of Table 9, which takes the control arm as its base. The derivation is a confirming instrument rather than a catching one. It helps the Auditor ratify an answer that is already right while it leads the Auditor to ratify an answer that is wrong, which is the opposite of what an audit is for.

3.2 Credulity, not information

The composition of the first finding invites a mechanism. A derivation could help by adding information, raising the Auditor's acceptance of correct answers while lowering its acceptance

of wrong ones. It could instead work by credulity, raising acceptance regardless of the truth. Columns (4) and (5) of Table 9 regress the verdict on the pass. Seeing the derivation raises acceptance of correct final answers by 9.7 percentage points, standard error 1.9. It raises acceptance of wrong final answers by 7.5 percentage points, standard error 2.2. Both move up. Fluent steps make an affirming verdict more likely whether or not the answer is correct, which is credulity rather than information.

The credulity is large where it is most costly. With the steps shown, 60.4 percent of wrong final answers pass audit. The share rises to 63.5 percent in the red-herring arm and exceeds 71 percent for the Anchor Auditor and the Robust Auditor. The red herring compounds the blindness. On verdict accuracy the interaction of the wrong state with the red herring is -0.071, standard error 0.020, so a red-herring-corrupted wrong derivation costs the Auditor a further 7.1 percentage points beyond an ordinary wrong one, Table 9 column (3). The passage that fooled the solver fools the checker that sees it. Solver error and verifier error are correlated by construction, which is the configuration under which a delegate-and-audit chain fails.

3.3 The value of the derivation falls with the Auditor's ability

The first two findings pooled the Auditors. The third asks for whom the derivation helps. Table 10 reports the within-response gain in verdict accuracy from seeing the steps, one number per Auditor, ordered by solving ability. The gain belongs to the Floor Auditor. The three abler Auditors show changes within two standard errors of zero and of one another, so the gradient in ability is a step at the Floor Auditor rather than a smooth decline across all four. The Anchor Auditor's place in the ordering rests on the structural ability a_A of Section 4.2, since the Anchor Auditor never solved as an agent. The Floor Auditor, the weakest solver still served at the audit phase, gains 13.0 percentage points of verdict accuracy, standard error 2.2. The Robust Auditor is unchanged, a gain of 0.1 points that does not differ from zero. The Anchor Auditor loses 1.6 points, again not distinguishable from zero, standard error 2.2. The Ceiling Auditor, the strongest solver in the agent solving phase, loses 2.7 points, standard error 1.6, significant at the 10 percent level.

A derivation transfers capability to a weak verifier while it transfers contamination to a strong one. The transfer is confined to the calls in which the weak verifier uses the shown steps. In the 39 percent of the Floor Auditor's Pass-2 calls under fifty completion tokens, acceptance rises by 17.7 points whether or not the answer is correct, which lifts verdict accuracy on the correct majority while lowering it on wrong candidates to 0.299. Wrong answers are caught more often only in the longer calls, Table 25. The Floor Auditor cannot

reliably solve the task on its own, so the delivered steps supply what its own solving would not, which lifts its verdict. The Ceiling Auditor can already solve the task, so the steps add no capability it lacks while a fluent wrong derivation can only mislead it. This gradient governs whether audit can be delegated down the price ladder. It is also the elasticity the recursive-self-improvement argument needs but has not had (Cunningham et al., 2026).

3.4 Cheaper in time, not in money

The first three findings concern accuracy. The fourth concerns price. The money price of a call is its billed expenditure in dollars. Its time price is the production time the principal pays in waiting, valued in its own time units (Acton, 1975; Becker, 1965). Table 11 reports the log change in each resource from the Auditor solving pass to the Auditor verifying pass, holding task and Auditor fixed. The Auditor verifying pass writes 0.580 log points fewer completion tokens, 44.0 percent, standard error 0.035. Its time price is 0.420 log points lower, 34.3 percent, standard error 0.023. Its money price is statistically unchanged on average, a log gap of +0.041 with standard error 0.033, an average that hides opposite signs across Auditors, +0.35 for the Floor Auditor, +0.10 for the Robust Auditor, +0.04 for the Anchor Auditor and −0.38 for the Ceiling Auditor, since a terse Auditor pays more for the derivation's input tokens than it saves in output tokens while an Auditor that reasons at length saves more than it pays.

Verification buys down the time price rather than the money price. Checking a supplied derivation is faster than solving the task afresh, which lowers the time price. The money price does not fall, because the derivation must enter the call as input tokens billed at input rates, so the input tokens rise as the completion tokens fall. At posted prices the market charges, on average across the four Auditors, the same dollar for a credulous fast check as for the slower Auditor solving pass, in which the Auditor is instructed to work the problem out. The pass that is cheaper in time is the same pass that accepts wrong answers more readily, so the speed arrives bundled with the credulity of the second finding. The equality is a fact about the posted schedule, on which the derivation enters the call as input tokens billed at input rates rather than at output rates. A different ratio of the two rates would move the money gap in either direction, so the null on money is conditional on the schedule at which the run was priced, while the time saving is a property of the technology.

4 A Structural Model of Solving and Auditing

The reduced-form sections measure the audit technology point by point. This section gives the points a common origin. I specify the solving technology that generates the candidates, the audit technology of each pass that generates the verdicts, the criterion that each Auditor chooses rather than inherits, then the audit cost block that prices every call. The model reduces the reduced-form surface to a small vector of economic primitives, which Section 6 takes to the delegation market. Every estimand in this section is an average-case description of this experiment's 59 tasks. Nothing here is a claim about worst-case complexity classes.

4.1 Environment

The setting is that of Section 2. Agents produce answers, Auditors audit them, a principal pays for both. Agent m works task t in arm r, where $r = 1$ marks the red-herring arm while $r = 0$ marks the control arm, under the explanation requirement e, where $e = 1$ marks the task version that requires a worked derivation. A produced response i carries a final answer, a worked derivation, a true state $\omega_i \in \{0, 1\}$ equal to one when the final answer is correct under the corrected answer key, an agent $m(i)$, a task $t(i)$, an arm $r(i)$. Auditor j rules on response i in pass p. In Pass 1, the Auditor solving pass, the Auditor sees the task with the candidate's final answer only. In Pass 2, the Auditor verifying pass, the Auditor also sees the worked derivation. The verdict either accepts or rejects the final answer. The Auditor never sees the agent's identity, never sees an arm label, never sees the key.

Two assumptions about knowledge run through the model. The Auditor knows the environment it works in, meaning the composition of the candidate pool for the task version it faces together with the informativeness of its own signal. The model then allows exactly two departures, each a parameter disciplined by the data rather than by assumption. The first is camouflage, which an Auditor cannot see through by definition. The second is a form premium, a belief distortion attached to the mere presence of a worked derivation, whose sign the data leave open. Both are defined below.

4.2 The solving technology

Solving is a latent-ability response technology in the Rasch tradition. A solving draw by agent m on task t in arm r under explanation requirement e is correct with probability

$$q_m(t, r, e) \;=\; \Lambda(a_m - d_t + g\, e - h_t\, r)\,, \tag{1}$$

where $\Lambda(x) = e^x/(1+e^x)$ is the logistic function, a_m is the ability of agent m, d_t is the difficulty of task t normalized to mean zero over the 59 tasks, g is the shift of the explanation requirement, h_t is the red-herring load of task t, while the loading of the index on ability is normalized to one. The five waves of the agent solving phase give five independent draws per cell, so the 44,652 scored responses identify the technology by maximum likelihood with no new data. The clean no-explanation arm anchors a_m with d_t, while the other three task-version families identify g with h_t against that anchor. The index carries no $e \times r$ term because the agent solving phase finds no detectable interaction between the red herring and the explanation requirement.[9] This structure follows the two-equation architecture of detection-controlled estimation, in which a production equation generates the base rate that a detection equation then confronts (Feinstein, 1990).

Three of the four Auditors served as agents, so their abilities sit on this scale by construction. The Anchor Auditor never solved the tasks as an agent. Its ability a_A is a free parameter, identified by its own Pass-1 verdicts through the technology of Section 4.4.

The solving technology fixes what the audit market takes as given. The graded candidates come from the explanation arm of wave 1, so the model share of correct candidates on task t in arm r is

$$\pi_{tr} \;=\; \frac{1}{38}\sum_{m} \Lambda(a_m - d_t + g - h_t\, r)\,, \tag{2}$$

the base rate of the verdict problem.[10] In the graded sample the empirical counterpart is 0.660 in the control arm with 0.513 in the red-herring arm.

4.3 The verdict as a chosen threshold

Every signal-based verdict in the model has one architecture. The Auditor draws a scalar signal about the candidate's state, forms a posterior, accepts when the posterior clears a cost-benefit bar. Three assumptions fix the architecture.

Assumption 1 (Signals)**.** When Auditor j audits response i in pass p on task t in arm r, it draws

$$\xi \;=\; \left(\omega_i - \tfrac{1}{2}\right)\psi \;+\; \delta^{C}_{ip}\,(1-\omega_i) \;-\; \delta^{W}_{ip}\,\omega_i \;+\; \varepsilon, \qquad \varepsilon \sim N(0,1), \tag{3}$$

where $\psi \geq 0$ is the discrimination of that signal, δ^{C}_{ip} is the correct-side displacement of a wrong candidate, δ^{W}_{ip} is the wrong-side displacement of a correct candidate, while ε is

[9] In the agent solving phase the red herring's loss does not differ detectably by the explanation requirement. The interaction is -0.014 ($p = 0.29$) under the two-way clustering convention (Akimitsu, 2026b).

[10] The graded sample keeps responses carrying a genuine worked derivation, which drops 370 of the 4,560 wave-1 explanation-arm responses. Equation (2) abstracts from that filter. A robustness check re-estimates the audit block with the empirical pool shares in place of (2).

independent across calls conditional on the response.

A correct candidate reads as $N(\psi/2-\delta^W_{ip},\,1)$ while a wrong one reads as $N(-\psi/2+\delta^C_{ip},\,1)$. The displacement carries the artifact's camouflage,

$$\delta^C_{ip} \;=\; \rho_j\, r_i \;+\; \rho_R\,\text{Reasoning}_{m(i)} \;+\; \rho_L \log \tilde{L}_i\, \mathbf{1}\{p=2\}, \tag{4}$$

where ρ_j is the red-herring displacement of Auditor j, with the Floor Auditor's value ρ as the base from which the other three deviate, Reasoning_m is the model-level indicator equal to one for a reasoning agent, 17 of the 38, while $\log \tilde{L}_i$ is the logged length of the worked derivation in thousands of characters, demeaned over the graded pool, active only when the derivation is shown. Equation (4) is the decomposition of Table 20 written into the signal. The red herring displaces through the planted final answer with the task text, visible in both passes, rather than through the steps. Reasoning agents write the most credible counterfeits on either margin, the +0.0987 falsification premium of Table 18, with its correct-candidate counterpart in Table 19. Longer wrong derivations are caught more, so the expected sign of ρ_L is negative. Camouflage operates on wrong candidates through δ^C_{ip}. Its red-herring component moves them toward acceptance, while the artifact loadings let the displacement vary in either direction across candidates. It is the costly-state-falsification margin of this market, a state made to look like another at a cost paid upstream by the red herring together with the artifact's own style (Lacker and Weinberg, 1989).

The red herring also touches correct candidates. An Auditor tempted by the planted passage doubts a derivation that ignores it, so a correct derivation can be pushed toward rejection. The wrong-side displacement carries that push,

$$\delta^W_{ip} \;=\; \rho_{Wj}\, r_i\, \mathbf{1}\{p=2\}, \tag{5}$$

where ρ_{Wj} is the wrong-side loading of Auditor j, active only when the derivation is shown. It is zero on the control arm and in the Auditor solving pass, so every control-arm statement below is unchanged by it.

Assumption 2 (Losses)**.** Auditor j bears a loss of one when it accepts a wrong final answer, a loss of $\ell_j > 0$ when it rejects a correct one, no loss otherwise.

Assumption 3 (Beliefs)**.** The Auditor knows its discrimination ψ with the pool share π_{tr}. It does not perceive the displacements δ^C_{ip} with δ^W_{ip}. In Pass 2 it holds the prior log odds

$$\log \frac{\tilde{\pi}_{tr}}{1-\tilde{\pi}_{tr}} \;=\; \log \frac{\pi_{tr}}{1-\pi_{tr}} \;+\; \mu_j, \tag{6}$$

where μ_j is the form premium of Auditor j. In Pass 1 the form premium is absent.

Assumption 2 is a revealed-preference object in the tradition of Chan et al. (2022), whose diagnosticians reveal a skill with a preference jointly through their operating point. A deployed language model optimizes nothing consciously. Its verdicts still trace an operating point, so ℓ_j is the loss ratio that rationalizes the point, exactly as a criterion rationalizes a radiologist's threshold. Assumption 3 carries the model's two departures from rational expectations. Naivety about δ^C_{ip} is what makes camouflage camouflage. An Auditor that priced the displacement would tighten its bar on red-herring tasks, while Table 21 shows no such tightening beyond what the fallen base rate commands. The form premium μ_j says the Auditor treats the presence of a complete worked derivation as evidence about the answer. In this pool that treatment is pure distortion, because every graded candidate carries a complete derivation by sample construction, so form separates nothing. Section 2 built exactly the pool in which fluent form is uninformative, which is what identifies μ_j as a belief distortion rather than as a taste for fluency, if it is identified at all. Separating μ_j from the loss ratio ℓ_j requires the Auditor solving pass to rule on the bare answer with some discrimination. At the estimates it does not, Section 4.8, so μ_j enters the estimates only through its sum with $\log \ell_j$.

Lemma 1 (The chosen criterion)**.** *Under Assumptions 1 to 3 the optimal verdict rule of Auditor j in pass p on task t in arm r is a threshold rule. The Auditor accepts exactly when $\xi \geq \xi^*_{jp}(t,r)$, where*

$$\xi^*_{jp}(t,r) \;=\; -\,\frac{\log \ell_j \;+\; \log \dfrac{\pi_{tr}}{1-\pi_{tr}} \;+\; \mu_j\,\mathbf{1}\{p=2\}}{\psi}. \tag{7}$$

The implied operating point on the true signal distributions is

$$C \;=\; \Phi\!\left(\tfrac{\psi}{2} - \delta^W_{ip} - \xi^*_{jp}\right), \qquad F \;=\; \Phi\!\left(-\tfrac{\psi}{2} + \delta^C_{ip} - \xi^*_{jp}\right), \tag{8}$$

where C is the correct-acceptance rate, F is the false-accept rate, Φ is the standard normal distribution function. For any threshold rule over candidates sharing the displacements δ^C with δ^W, the measured coordinates of Table 21 follow as $d' = \psi - \delta^C - \delta^W$ with $c = \xi^*_{jp} - (\delta^C - \delta^W)/2$. On the control arm and in the Auditor solving pass $\delta^W = 0$, so there $d' = \psi - \delta^C$ with $c = \xi^*_{jp} - \delta^C/2$.

The coordinate $d' = \psi - \delta^C - \delta^W$ is $\Phi^{-1}(C) - \Phi^{-1}(F)$, the two rates of (8) differenced on the standard-normal scale. To see what that difference measures, fix Auditor j and the verifying pass, which is pass 2. Take two of the solving agents responses on the same

task t and arm r, one correct and one wrong. Both responses share the discrimination $\psi_{j2}(t,r)$ (an auditor-task-arm object) and the cutoff $\xi^*_{j2}(t,r)$ (an auditor-task-arm object). The task and the arm set both. The Auditors signal on a response is $\xi = m + \varepsilon$, with noise $\varepsilon \sim N(0,1)$ independent across audits given the response. The mean m depends on the responses state. On the correct response it is $m_C = \frac{1}{2}\psi_{j2}(t,r) - \delta^W$. On the wrong response it is $m_F = -\frac{1}{2}\psi_{j2}(t,r) + \delta^C$. The correct-side displacement δ^C (a response-pass object) raises the wrong-response mean m_F toward the correct side. The wrong-side displacement δ^W (a response-pass object) lowers the correct-response mean m_C toward the wrong side. In the model section these displacements carry a response and pass index, δ^C_{ip} and δ^W_{ip}. I suppress it here. The two responses and the pass are fixed. The two chosen responses are audited once each and draw independent noise, $\varepsilon_C, \varepsilon_F \overset{\text{iid}}{\sim} N(0,1)$. I write the correct responses signal as $\xi_C = m_C + \varepsilon_C$ and the wrong responses signal as $\xi_F = m_F + \varepsilon_F$. Each is normal with unit variance,

$$\xi_C \sim N(m_C,\, 1), \qquad \xi_F \sim N(m_F,\, 1).$$

The discrimination $\psi_{j2}(t,r)$ is the gap between m_C and m_F when both displacements are zero.

The rate $C_{j2}(t,r)$ is the sound affirming verdict probability, the probability that Auditor j affirms the correct response. The Auditor affirms when the signal clears the cutoff. This gives

$$C_{j2}(t,r) = P\big(\xi_C \geq \xi^*_{j2}(t,r)\big)\,.$$

Because ξ_C is a continuous random variable, $P(\xi_C = \xi^*_{j2}(t,r)) = 0$. The events $\xi_C \geq \xi^*_{j2}(t,r)$ and $\xi_C > \xi^*_{j2}(t,r)$ therefore carry the same probability. Standardize the signal. Subtract the mean m_C. The standard deviation is one. There is no division. With $Z = \xi_C - m_C \sim N(0,1)$,

$$C_{j2}(t,r) = P\big(\xi_C - m_C \geq \xi^*_{j2}(t,r) - m_C\big) = P\big(Z \geq \xi^*_{j2}(t,r) - m_C\big)\,.$$

This is an upper tail. By the symmetry of the standard normal about zero the mass above a point equals the mass below its negative, $P(Z \geq a) = 1 - \Phi(a) = \Phi(-a)$. With $a = \xi^*_{j2}(t,r) - m_C$, its negative is $m_C - \xi^*_{j2}(t,r)$. This yields

$$C_{j2}(t,r) = \Phi\big(m_C - \xi^*_{j2}(t,r)\big) = \Phi\big(\tfrac{1}{2}\psi_{j2}(t,r) - \delta^W - \xi^*_{j2}(t,r)\big)\,.$$

The rate $F_{j2}(t,r)$ is the unsound affirming verdict probability, the probability that Auditor j affirms the wrong response. The same steps with ξ_F and mean m_F give

$$F_{j2}(t,r) = P\big(\xi_F \geq \xi^*_{j2}(t,r)\big) = \Phi\big(m_F - \xi^*_{j2}(t,r)\big) = \Phi\big(-\tfrac{1}{2}\psi_{j2}(t,r) + \delta^C - \xi^*_{j2}(t,r)\big)\,.$$

These are the two rates of (8), now carrying their Auditor, pass, task and arm. Φ is strictly increasing and maps the line onto $(0,1)$. Its inverse Φ^{-1} is well defined, with $\Phi^{-1}(\Phi(x)) = x$. Applying it returns each mean net of the cutoff,

$$\Phi^{-1}(C_{j2}(t,r)) = m_C - \xi^*_{j2}(t,r) = \tfrac{1}{2}\psi_{j2}(t,r) - \delta^W - \xi^*_{j2}(t,r),$$
$$\Phi^{-1}(F_{j2}(t,r)) = m_F - \xi^*_{j2}(t,r) = -\tfrac{1}{2}\psi_{j2}(t,r) + \delta^C - \xi^*_{j2}(t,r).$$

The cutoff $\xi^*_{j2}(t,r)$ is the same term in both. It cancels in the difference, which leaves the gap between the two means in standard-deviation units,

$$d'_{j2}(t,r) = \Phi^{-1}(C_{j2}(t,r)) - \Phi^{-1}(F_{j2}(t,r)) = m_C - m_F = \psi_{j2}(t,r) - \delta^C - \delta^W.$$

On one task and arm the index is the discrimination net of the two displacements. That is the wedge.

The realized discrimination d'_{jpr} is the pooled counterpart, auditor-pass-arm, $d'_{jpr} = \Phi^{-1}(C_{jpr}) - \Phi^{-1}(F_{jpr})$, where C_{jpr} is the share of the correct final answers that Auditor j affirms and F_{jpr} the share of the wrong final answers it affirms, over all responses on arm r in pass p. Both are observed shares. I obtain d'_{jpr} directly from the data. These shares pool responses across all tasks in the cell, where each task has its own $\psi_{j2}(t,r)$ and its own cutoff and each response its own displacement. The quantile Φ^{-1} of a pooled share is not the pooled value of Φ^{-1} taken response by response. The realized d'_{jpr} is therefore not any single tasks $\psi_{j2}(t,r)$ net of displacement. It is the index read off the two pooled shares, which is why I call it realized.

The model therefore takes the Pass-2 rows of Table 21 directly. Figure 3 places each Auditor and arm in that (F,C) operating-point space. On the control arm the measured d' is the Pass 2 discrimination while the measured c is the Auditor's chosen Pass 2 cutoff. On the red-herring arm the pool-average displacement compresses the one while contaminating the other, since Table 21 pools candidates within an arm, which the estimation separates through the cross-arm structure. The Pass-1 rows are generated by the mixture of Section 4.4, whose measured coordinates are composites of its two branches rather than a threshold pair. The first economics of the model is a statement about where the cutoff sits.

Proposition 1 (The confirmatory audit)**.** *Let $\pi_{tr} \geq \frac{1}{2}$ under Assumptions 1 to 3, with the pass's criterion constant nonnegative, $\log \ell_j$ in Pass 1 and $\log \ell_j + \mu_j$ in Pass 2, with either $\pi_{tr} > \frac{1}{2}$ or that constant strictly positive, on candidates whose correct-side displacement is nonnegative and weakly exceeds the wrong-side one, $\delta^C_{ip} \geq 0$ with $\delta^C_{ip} \geq \delta^W_{ip}$, which every control-arm and final-answer-pass candidate with $\delta^C_{ip} \geq 0$ satisfies. Then for any discrimination $\psi > 0$ the chosen cutoff satisfies $\xi^*_{jp} < 0$, so the criterion is liberal, while the*

false-accept rate strictly exceeds the false-reject rate. The cutoff falls as the pool improves, $\partial\xi^*_{jp}/\partial\pi_{tr} < 0$*, falls as rejection grows dearer,* $\partial\xi^*_{jp}/\partial\ell_j < 0$*, while it approaches the neutral point as the signal sharpens,* $\partial\,|\xi^*_{jp}|/\partial\psi < 0$*.*

Proposition 1 rationalizes the asymmetry that organizes Table 8. With steps shown, 60.4 percent of wrong final answers pass audit while 7.8 percent of correct ones are rejected. An audit that confirms well while catching poorly is not a malfunction. It is the operating point that any loss-minimizing verifier chooses when most candidates are correct while wrongful rejection is at least as costly as wrongful acceptance. The asymmetry is an equilibrium footprint of the pool, which delegation itself shapes. When the base rate deteriorates, as it does mechanically in the red-herring arm where the pool share falls to one half, the same optimizing Auditor tightens. The Pass-1 criteria of Table 21 sit closer to zero in the red-herring arm than in the control arm for all four Auditors, the direction the fallen base rate commands, while in the Auditor verifying pass the measured red-herring criteria move away from zero for the Robust, Anchor and Ceiling Auditors and toward zero for the Floor Auditor. At the estimates the $-(\delta^C - \delta^W)/2$ contamination of Lemma 1 accounts for the Ceiling Auditor's move, whose wrong-side displacement is negative, together with the Floor Auditor's, whose wrong-side displacement exceeds its correct-side one, -0.57 against -0.62 on the control arm in Table 21. For the Robust and Anchor Auditors the contamination pushes toward zero while the Pass 2 signal's difficulty pushes away, since the red herring lowers the Pass 2 discrimination and the cutoff grows in magnitude as that discrimination falls, Section 4.5. The net move is one to two hundredths. The model warns against taking the audit's kindness as a constant of the technology. It is a function of what the market sends to be checked.

4.4 The Auditor solving pass

In the Auditor solving pass the Auditor holds the task with a bare answer, under an instruction to work the task out. The model lets the Auditor do two things.

Assumption 4 (The Auditor solving pass mixture)**.** In Pass 1 Auditor j re-solves the task with probability ϑ_j, otherwise it rules on the bare final answer without solving, a verdict drawn from a signal with discrimination ψ_0 under the threshold rule of Lemma 1. When it re-solves, its attempt is a solving draw from (1) at $e = 0$, then it accepts exactly when its own final answer matches the candidate's, the compare rule. Conditional on the draw being wrong, its final answer lands on canonical wrong answer k with the share s_k of the version's no-explanation wrong-answer distribution over the other agents.

The Auditor is shown the bare no-explanation problem text, so its re-solving draw is the no-explanation technology, the same object as its arm-matched solving record in Table 22, plotted in Figure 4. The compare rule needs the concordance structure of wrong answers. When the Auditor re-solves correctly, matching reveals the truth. When it re-solves wrongly against a wrong candidate, the two wrong answers can match, in which case a false accept follows. The false accept on this branch is a concordant error, sincere rather than credulous. The Auditor confirms the candidate against its own answer, so agreement with itself reads as correctness. Let χ_{tr} denote the probability that a wrong solving draw on task t in arm r lands on the same final answer as a wrong candidate from the pool, the concordance probability that the last clause of Assumption 4 fixes.[11] The model confines this channel to the Auditor solving pass, whose instruction invites the Auditor to work the task out. The Auditor verifying pass instructs the Auditor to work through the shown steps rather than solve afresh, a checking task that Section 4.5 models with no re-solving branch, hence no concordance term. The verdicts support the restriction, Table 23. With task-version and Auditor fixed effects, Pass-2 verdict accuracy on correct candidates has a slope of +0.020 (standard error 0.024) on the Auditor's own solving share, against +0.216 (0.061) in the Auditor solving pass, so the Auditor verifying pass shows no re-solving fingerprint on correct candidates. On wrong candidates the slope is +0.146 (0.051), which the Pass 2 audit technology produces without re-solving, since within a task version the Auditors that solve it better also audit it with more discrimination at loading λ. The price data carry the asymmetry's fingerprint, since in Pass 2 the input meter grows one for one with the shown derivation while the clock falls along it, Section 4.6. The implied Pass-1 rates for Auditor j are

$$C_{j1}(t,r) \;=\; \vartheta_j\, q_j(t,r) \;+\; (1-\vartheta_j)\,\Phi\big(\tfrac{\psi_0}{2} - \xi^*_{j1}\big)\,, \tag{9}$$

$$F_{j1}(t,r) \;=\; \vartheta_j\,\big(1-q_j(t,r)\big)\,\chi_{tr} \;+\; (1-\vartheta_j)\,\Phi\big(-\tfrac{\psi_0}{2} + \delta^C_{i1} - \xi^*_{j1}\big)\,, \tag{10}$$

where $q_j(t,r) = \Lambda(a_j - d_t - h_t r)$ is the Auditor's own solving probability on the bare task version while ξ^*_{j1} is the cutoff of the branch without solving from (7) at discrimination ψ_0 with no form premium. The false-accept rate is evaluated at the candidate's own displacement, so it varies within a cell exactly through (4).

Three observation statuses run through (10), kept distinct throughout. The candidate's wrongness is observed, graded against the corrected answer key, so the falsification margin conditions on it. Whether the Auditor re-solves is latent. The run stored each Auditor's verdict with its meters, never its reply text, so no call reveals the Auditor's own working,

[11] Appendix C constructs χ_{tr} as the concentration of the version's canonicalized wrong-answer distribution over the 38 agents' no-explanation waves, leaving the Auditor's own model out where the Auditor solved as an agent.

which is why re-solving enters as a branch whose weight ϑ_j the data estimate rather than as an assumption. In the limit $\psi_0 \to 0$ the branch without solving nests an Auditor whose signal from the stated answer tells it nothing about whether the answer is right, which is close to where the estimates land. At the estimated ψ_0 of 0.190 the branch without solving accepts correct and wrong candidates alike on every task whose pool share exceeds a flip point, the share at which the pool odds equal $1/\ell_j$, while its acceptance falls to zero as the share drops below that point, Section 5.2. Whether a latent re-solving draw is wrong is likewise unobserved. Its probability is the model's $1 - q_j(t, r)$, disciplined for the three Auditors that solved as agents by their own graded solving records, in which wrongness is observed for every wave, while the Anchor Auditor's ability is a free parameter that its own Pass-1 verdicts identify, Section 4.2. The product $\vartheta_j(1 - q_j)\,\chi_{tr}$ therefore contains one estimated share, one model-implied latent probability, one measured concordance, while no term claims to observe the Auditor's answer.

Lemma 2 (The collapse representation)**.** *Under Assumption 4, Pass-1 verdict accuracy on correct candidates equals* $\vartheta_j\, q_j + (1 - \vartheta_j)\,\Phi(\psi_0/2 - \xi^*_{j1})$*, linear in the Auditor's own solving probability with slope* ϑ_j*. On wrong candidates it equals* $\vartheta_j\big(1 - (1 - q_j)\chi_{tr}\big) + (1 - \vartheta_j)\big(1 - \Phi(-\psi_0/2 + \delta^C_{i1} - \xi^*_{j1})\big)$*, linear in* q_j *with slope* $\vartheta_j\,\chi_{tr}$*. Pure re-solving is the boundary* $\vartheta_j = 1$*, which forces a unit slope with a zero intercept on correct candidates.*

Lemma 2 is the structural interpretation of the collapse test, the regression of Pass-1 verdict accuracy on the Auditor's own five-wave solving share, Table 22, which estimates a slope of 0.442 with an intercept of 0.517 on correct candidates in its first column, rejecting the boundary. The two coefficients sum to 0.960, with 1.005 in the second column, close to the unit sum the lemma implies when the branch without solving accepts every correct candidate, since the intercept is then $1 - \vartheta_j$ while the slope is ϑ_j. The regression slope is only the projection of ϑ_j, attenuated because its regressor is a five-wave share while contaminated because the Auditor's acceptance threshold, which moves with the pool's share of correct answers, covaries with q across tasks. The joint likelihood estimates ϑ_j free of both problems, because $q_j(t, r)$ enters as a parameter function rather than as a measured share. The concordance term ties the Auditor solving pass to the red herring's signature in the solving data. The red herring makes wrong answers repeat, with the share of waves repeating an identical wrong answer rising from 35.6 to 48.0 percent (Akimitsu, 2026b), so $\chi_{t1} > \chi_{t0}$ on the affected tasks. A re-solving Auditor fails on a fooled candidate exactly when it walks into the same trap, which the concordance term makes literal.

One feature of the data deserves disclosure here. The audit phase stored two meters for every call, the completion tokens the Auditor generated with the latency, the seconds the

call took, which Assumption 5 below calls the cost meters. The completion-token meter counts every token the Auditor generated on the call. The API provider that serves an Auditor may also report reasoning tokens, the tokens the Auditor spends on hidden working before its visible reply. The providers of the Anchor Auditor and the Ceiling Auditor report them, which the completion count carries. The providers of the Floor Auditor and the Robust Auditor report none, so for those two Auditors the meter cannot be split into hidden working and the visible reply. In Pass 1, 1,225 of the Floor Auditor's 4,046 usable calls, three in ten, show fewer than fifty completion tokens. No Pass-1 call of the other three Auditors shows fewer than 89. In Pass 2, 1,572 of the Floor Auditor's 4,046 usable calls show fewer than fifty. A short call does not say whether the Floor Auditor solved the task. Table 25 splits the Floor Auditor's Pass-2 gain by call length. The short calls' gain is state-blind acceptance. The Floor Auditor is a terse model. Its own no-explanation answers as an agent in the agent solving phase have a median of 17 completion tokens. Its Pass-1 calls have a median of 140. A call of fewer than fifty tokens is therefore long enough to hold that Auditor's answer to a task. In the two-step program of Section 4.8 the audit technology is estimated from the verdicts alone, so the meters never enter the identification of ϑ_j, while a short call carries its verdict like any other. The disclosure therefore concerns the meter side, where the Auditor-pass levels of (17) absorb the Floor Auditor's terseness. Appendix C measures how the two meters move with the task clean solve rate, then compares a verdict with the same Auditor's own answer as an agent to the same task version.

4.5 The Auditor verifying pass

In the Auditor verifying pass the Auditor uses the worked derivation. I parameterize the Pass 2 discrimination log-linearly in the primitives,

$$\log \psi_{j2}(t, r) \;=\; c_2 \;+\; \lambda\, a_j \;-\; \kappa \left(d_t + h_t\, r\right), \tag{11}$$

where c_2 is the level, λ is the capability loading of the Pass 2 signal, κ is its difficulty loading. Two comparisons give the loadings their content. Solving pays difficulty at a loading of one on the log-odds scale by (1), a price the Auditor solving pass inherits through its re-solving branch. The Auditor verifying pass pays difficulty at κ, the elasticity of the Pass 2 discrimination to difficulty. On the scale on which solving's loading is one, the Pass 2 accuracy at a criterion of zero is $\Phi(\psi/2)$, whose log-odds falls with difficulty at $\kappa G(\psi)$, with $G(\psi) = \psi\phi(\psi/2)/[2\Phi(\psi/2)\Phi(-\psi/2)]$ increasing in the Pass 2 discrimination ψ, where ϕ is the standard normal density. A small κ says the derivation shields the checker from difficulty the solver paid in full, which is the sense in which checking can be cheap. The shield is thinner for

an abler reader, whose accuracy sits where a given loss of discrimination costs more. A small λ says the derivation does the checking for the checker, so weak Auditors gain most. That is the capability gradient of Table 10, where the value of seeing the steps runs from +0.130 for the Floor Auditor to −0.027 for the Ceiling Auditor. The camouflage displacement δ_{i2}^{C} operates on wrong candidates through (3) with (4), the wrong-side displacement δ_{i2}^{W} operates on correct candidates through (5), while the form premium μ_j operates on every candidate through (7). The next result separates the derivation's two services.

Proposition 2 (Two goods in one document)**.** *Under Assumptions 1 to 4, consider Auditor j on pool (t, r).*

(i) The form premium raises acceptance on both truth margins, since $\partial C_{j2}/\partial \mu_j > 0$ while $\partial F_{j2}/\partial \mu_j > 0$. Persuasion is truth-blind.

(ii) Let $\vartheta_j > 0$, vary the Auditor's ability in its checking role while holding the candidate pool fixed. On the control arm the realized discrimination gain of the Auditor verifying pass is $\Delta d'_j = \psi_{j2} - d'_{j1}$ with $d'_{j1} = \Phi^{-1}(C_{j1}) - \Phi^{-1}(F_{j1})$. It has ability derivative $\lambda\,\psi_{j2} - \partial d'_{j1}/\partial a_j$, where $\partial d'_{j1}/\partial a_j > 0$. If $\lambda\,\psi_{j2} < \partial d'_{j1}/\partial a_j$ over the ability range, the gain declines strictly in the Auditor's own ability, so whenever the gain changes sign there is a unique ability a^* above which showing the derivation lowers realized discrimination.

(iii) The pass contrast in verdict accuracy splits into an information effect, the contrast evaluated at $\mu_j = 0$, plus a persuasion effect. For an Auditor with $\mu_j > 0$ the persuasion effect is strictly positive on correct candidates while strictly negative on wrong candidates, by part (i).

A worked derivation is two goods in one document. It is information, whose value falls with what the reader could produce alone. It is also persuasion, whose incidence does not depend on truth. The weak Auditor buys the first good. Whether any Auditor pays for the second is a question the estimates leave open. Part (iii) is the model's decomposition of the pass contrast that Table 9 estimates, where the derivation raises verdict accuracy by +0.035 on correct control candidates while the wrong-candidate reversal is −0.242. Acceptance rises by +0.097 on correct candidates against +0.075 on wrong ones. That truth-blind rise is a reduced-form fact, which the model carries through the confirmatory criterion of the Pass 2 signal replacing the mixture of the Auditor solving pass. The persuasion premium μ_j is not identified separately from the loss ratio ℓ_j, Section 4.8, so I claim the sign of the information good's gradient in part (ii) and claim nothing about the sign of the persuasion term. Part (ii) is the gradient of Table 10, whose gain is largest for the Floor Auditor and within two

standard errors of zero for the three abler Auditors, the sign pattern the condition in (ii) generates. The condition itself is checked at the estimates rather than assumed.

4.6 The price of a verdict

The audit technology has a cost side, measured in Table 11 with Tables 12 through 16. The model generates it from one behavioral object, the call's tokens and seconds, tied to the same draw that generates the verdict.

Assumption 5 (Audit cost)**.** Each call carries an engagement draw $\zeta \sim N(0,1)$, independent across calls, independent of the state. The call's discrimination is $\psi\, e^{\nu\zeta}$, where ψ is the cell technology of Sections 4.4 with 4.5 while $\nu \geq 0$ is the learning dividend. Each cost meter y of the call, the completion tokens with the latency, satisfies

$$\log y \;=\; t^{y}_{jp} \;+\; \alpha_y \log \tilde{L}_i\, \mathbf{1}\{p=2\} \;+\; \eta_y\, \zeta \;-\; \gamma_y \left|\xi - \xi^{*}_{jp}\right| \;+\; u_y, \qquad u_y \sim N(0, \sigma^2_y), \tag{12}$$

where t^{y}_{jp} is an Auditor-pass level, α_y is a direct derivation-length load, $\eta_y \geq 0$ is the meter's engagement load, $\gamma_y \geq 0$ is the chronometric loading on the draw's distance to the criterion, u_y is measurement noise independent of everything else.

The call's cost carries two services. The chronometric service is selection. Calls near the criterion take more tokens and more seconds, the regularity that decision time peaks near indifference (Card et al., 2024). The engagement service is learning, in the sense the same paper weighs against selection. A call the Auditor engages runs longer while its signal is also sharper, which the dividend ν carries. At $\nu = 0$ the call's cost is a pure thermometer of the draw. The results of Sections 4.3 to 4.5 hold engagement draw by engagement draw, with cell-level objects the engagement average, exactly as stated at $\nu = 0$. The money price needs no new parameter. A call bills the metering identity, the Auditor's posted input rate on the prompt, mechanical in the task text with the shown derivation, plus its posted output rate on the completion meter (Demirer et al., 2026). The money price inherits the derivation-length load on its input meter with the chronometric index on its output meter, while the time price is the clock itself.

Three fingerprints of the price data follow. First, the derivation is billed as input rather than generated as output. In Pass 2 the input meter grows one for one with the derivation shown, so the money price rises with the audited object's length, a derivation-length gradient of $+0.226$ log points, while the time price falls along it, -0.092, since the clock records generation rather than intake, Table 14. Second, difficulty charges twice. A harder task lowers ψ through κ, which at a fixed criterion moves the correct majority of the pool toward

the criterion, raising tokens and seconds there, while the same fall in ψ lowers the sound verdicts per call, so the unit price of a sound verdict carries a difficulty gradient several times the per-call one, -2.19 in Table 16 against -0.49 in Table 14. Third, the meter prices latent wrongness the Auditor is never told, with a sign that turns on the Auditor's own discrimination.

Proposition 3 (The cost premium)**.** *Fix Auditor j in pass p on a pool whose wrong candidates share the displacement $\delta^C \geq 0$ and whose correct candidates share the wrong-side displacement δ^W with $\delta^C + \delta^W \geq 0$, at a criterion $\xi^*_{jp} \leq 0$ with $\delta^C - \delta^W - 2\xi^*_{jp} > 0$, comparing candidates at a common derivation length, under Assumptions 1 with 5 at $\gamma_y > 0$. Let $z_1 = \psi/2 - \delta^W - \xi^*_{jp}$ with $z_0 = -\psi/2 + \delta^C - \xi^*_{jp}$ locate the two candidate types relative to the criterion, let $\mathcal{D}(z) = 2\phi(z) + z\big(2\Phi(z) - 1\big)$ be the expected distance of a unit-noise draw at location z from the criterion. The expected gap in $\log y$ between wrong and correct candidates is $\gamma_y\big[\mathcal{D}(z_1) - \mathcal{D}(z_0)\big]$ for each meter.*

(i) Holding the criterion fixed, the gap is strictly increasing in the discrimination ψ, hence in the Auditor's ability at loading λ.

(ii) When $\delta^C + \delta^W > 0$ the gap is negative at small discriminations while positive at large ones, with a single crossing.

(iii) When $\delta^C + \delta^W = 0$ the gap is positive at every positive discrimination.

The Auditor's meter prices latent wrongness it is never told. The per-Auditor schedules find exactly the gradient of part (i). The wrong-answer premium runs from -0.069 for the Floor Auditor to $+0.517$ for the Ceiling Auditor on the money meter, from -0.053 to $+0.597$ on the clock, Table 15, monotone across the four Auditors. The weak Auditor's negative sign is the signature of part (ii). For an Auditor with little discrimination on a camouflaged pool, a displaced wrong answer is its easiest call rather than its closest, so it spends the least on the answers it is wrongest about. The call's cost is informative in both directions, which the next result turns into a free instrument.

Corollary 1 (The clock's two readings)**.** *Condition on acceptance within the Pass-2 calls of one Auditor on one task version, over candidates at a common derivation length whose wrong members share the displacement δ^C and whose correct members share the wrong-side displacement δ^W, under Assumptions 1 with 5, with the criterion ξ^*_{jp} chosen once from the cell discrimination ψ rather than re-optimized at each engagement draw.*

(i) At $\nu = 0$ with $\gamma_y > 0$, for any engagement load, the probability that the verdict is right is strictly decreasing in each cost meter when $\psi > \delta^C + \delta^W$, strictly increasing when $\psi < \delta^C + \delta^W$.

(ii) *At $\gamma_y = 0$ with $\nu > 0$, $\eta_y > 0$, at a fixed criterion, that probability is strictly increasing in each meter.*

(iii) *With both services active the slope's sign is an estimand, selection against learning.*

A principal already pays the time price, so the clock's reading is free. Which way it reads turns on which of the two services dominates. Under pure chronometry high cost marks the marginal call, so reliability falls with the clock wherever the audit clears the two displacements. Under learning high cost marks the engaged call, so reliability rises with it. The race is testable on columns already in the data, which part (i) turns into a sharp null. The data settle it Auditor by Auditor, Table 24. On the accepted Pass-2 verdicts, within task version at a common derivation length, verdict accuracy rises with both meters for the Floor Auditor and the Ceiling Auditor, rises with the completion meter for the Robust Auditor while its latency slope is positive but imprecise, while it falls with both meters for the Anchor Auditor. Pure chronometry is rejected as a maintained restriction, the untargeted prediction doing its work, so the block carries both services with the mix estimated Auditor by Auditor. The market interpretation follows. Where learning dominates, the call's tokens and seconds are an input, reliability can be bought by the second, with the time price partly an input price rather than a pure waiting cost. Where selection dominates, as for the Anchor Auditor, the clock is a thermometer, informative to read while not for sale. Appendix C adds two descriptive facts about the Auditor solving pass. Each meter's gradient in the task clean solve rate s_t is reported Auditor by Auditor, steepest for the Ceiling Auditor, not distinguishable from zero for the Floor Auditor. The Floor Auditor generates more completion tokens on a verdict than on its own answer as an agent to the same task version under no explanation requirement in the agent solving phase, as does the Robust Auditor, while the Ceiling Auditor generates fewer. The Anchor Auditor never solved as an agent, so it has no own answer.

4.7 What the model reconciles

Each primitive is disciplined by an exhibit already in the paper. The solving block (1) is pinned by the 44,652 scored responses of the agent solving phase. The re-solving probabilities ϑ_j with ψ_0, the discrimination of the branch without solving, are pinned by the slope with the intercept of the collapse test, Table 22. The Pass 2 audit technology (c_2, λ, κ) is pinned by the realized discrimination surface of Table 21 across Auditors with tasks, on the control arm. The displacement loadings (ρ, ρ_R, ρ_L) are pinned by the red-herring compression of that surface on wrong candidates together with the falsification margin of Table 18 with the two-channel decomposition of Table 20, the model's rendering of the -0.071 interaction of

Table 9. The Pass-2 verdicts pin the sum $\log \ell_j + \mu_j$ given the pool shares. The Auditor solving pass pins the pool share at which the branch without solving flips from rejecting to accepting, which fixes $\log \ell_j$ on its own only under the convention that the cutoff of the branch without solving weights the pool share at one, so the split of the sum into loss ratio and form premium is a convention rather than an estimate. The audit cost block turns the price side from input into output. The derivation-length gradient's two signs, the wrong-answer premium with its capability gradient, the doubled difficulty charge of the unit price are predictions of Assumption 5 with the audit block rather than free facts. The metacognition null of Table 17 is honored by omission, since no elicitation-side trait enters the technology. The remaining exhibits are untargeted. The estimation protocol below therefore closes with a fit pass in which the estimated model must reproduce the pass contrast of Table 9, the per-Auditor contrasts of Table 10 and the price schedules of Table 14 with Table 15. It must also match the operating points of Table 8 within a stated tolerance, which Table 30 reports as a mean absolute gap of 0.019 and a largest gap of 0.076. None of these enter the likelihood as moments in their reported form.

4.8 Estimation

I estimate the model in two steps, block by block, with each block identified by its own moments. The solving technology comes first, from the solving data alone. The audit block comes second, from the verdicts alone, at the solving estimates. A dependence layer then prices the correlation of verdicts across Auditors auditing the same response. The audit cost equations come last, conditional on the estimated audit technology, with every criterion frozen. The sequencing is deliberate. The two meters contribute 63,958 observations against 31,979 verdicts, so in a joint likelihood the meter equations would re-identify the audit technology away from what the verdicts support. The meter equation is also the block the reduced-form record disciplines least sharply, with Corollary 1 already rejecting its pure-chronometry form. The limited-information program therefore gives up joint efficiency in exchange for identification of the audit technology that survives misspecification of the cost side. No block inherits another block's sampling error silently, because inference re-estimates the entire chain on every bootstrap draw.

The solving step maximizes the penalized binomial likelihood

$$\mathcal{L}_1 = \sum_{m,t,r,e} \Big[n_m \log q_m + (\bar{n}_m - n_m) \log\big(1 - q_m\big) \Big] \; + \; \tfrac{1}{2} \log \det \mathcal{I}, \tag{13}$$

where $n_m(t, r, e)$ counts agent m's correct waves among its $\bar{n}_m(t, r, e)$ scored waves on the task

version (t, r, e), at most five, q_m is the solving probability (1) with its arguments suppressed, while $\mathcal{I}$ is the Fisher information of the binomial likelihood in the free coordinates. The penalty is the Jeffreys-prior adjustment of Firth (1993). It is needed because two task versions are solved by no agent in any scored wave, a separation under which the unpenalized maximum runs to the boundary. The penalized estimates of every identified contrast are invariant to the normalization, while on the pre-exclusion sample, where the likelihood has an interior maximum, the penalty moves the explanation shift by 0.002 with no ability by more than 0.010, so the penalty is a numerical device rather than a substantive choice. The free coordinates number 156, the 38 agent abilities, 58 free task difficulties under the mean-zero normalization over the 59 tasks, the 59 red-herring loads, the explanation shift, fitted on 44,652 scored responses in 8,962 agent-by-task-version cells.

The verification step maximizes the verdict likelihood

$$\mathcal{L}_2 = \sum_{i,j,p \text{ usable}} \Big[y_{ijp} \log A_{ijp} + (1 - y_{ijp}) \log\big(1 - A_{ijp}\big) \Big] + \sum_{j \in \mathcal{J}_c} \sum_{t,r,e} \Big[n_j \log q_j + (\bar{n}_j - n_j) \log\big(1 - q_j\big) \Big], \tag{14}$$

where $y_{ijp} \in \{0, 1\}$ marks acceptance on a usable call, with calls where an Auditor rules on its own response excluded as everywhere in the paper, while A_{ijp} is the model's acceptance probability, equal to $C_{jp}(t(i), r(i))$ when $\omega_i = 1$ while equal to $F_{jp}(t(i), r(i))$ at candidate i's displacement when $\omega_i = 0$, from (8) in Pass 2 while from (9) with (10) in Pass 1. The second sum runs over the own solving cells of $\mathcal{J}_c$, the three Auditors that solved as agents, so each contestant Auditor's ability is re-estimated from its verdicts jointly with its own solving record. The comparison of this ability with the same model's Stage-1 ability is the paper's estimate of whether a model audits better than it solves. The remaining 35 agent abilities stay at their Stage-1 values, as do d_t, h_t, the explanation shift g. The estimated criterion of the Auditor verifying pass takes the form

$$\xi^*_{j2}(t, r) \;=\; -\,\frac{\log \ell_j \;+\; w_j \log \dfrac{\pi_{tr}}{1 - \pi_{tr}} \;+\; \mu_j}{\psi_{j2}(t, r)}, \tag{15}$$

which generalizes (7) by the base-rate weight $w_j \in (0, 1)$ on the pool-share term. A Bayesian Auditor has $w_j = 1$. An Auditor with $w_j < 1$ underweights the composition of its pool, which is base-rate neglect in the sense of the probability-judgment literature. Under the unit normalization of the signal noise w_j is identified, with one equivalence disclosed, since an Auditor that underweights the base rate is observationally equivalent on the verdict surface to a Bayesian Auditor whose criterion carries noise. The audit block frees 35 parameters. The Pass 2 audit technology (c_2, λ, κ) contributes three. The discrimination ψ_0 of the branch

without solving contributes one. The loss ratios ℓ_j, the form premia μ_j, the base-rate weights w_j contribute four each. The correct-side displacement enters as a Floor-Auditor base ρ with three Auditor deviations, four parameters, active in both passes, while the wrong-side loadings ρ_{Wj}, four more, let the red herring displace correct candidates' signals in the Auditor verifying pass. The artifact loadings ρ_R with ρ_L contribute two. The re-solving probability is $\vartheta_j(t) = \Lambda\big(b_j - \vartheta_1 d_t\big)$, four intercepts with one difficulty slope, so the share of the Auditor solving pass that re-solves falls with intrinsic difficulty. The Anchor Auditor's ability with the three re-estimated contestant abilities completes the count. Three identification facts are used rather than hidden. Only the sum $\log \ell_j + \mu_j$ is identified by the Pass-2 verdicts. I report the sum, with the pool share at which each Auditor's branch without solving flips, in Panel B of Table 3. The split of the sum into loss ratio and form premium rests on the weight the cutoff of the branch without solving puts on the pool share, a choice worth less than one log-likelihood point, so the split is a convention rather than an estimate, Section 5.2. The verdict likelihood's curvature at the optimum has a single near-flat direction, the Robust Auditor's split of that sum, while ψ_0, the discrimination of the branch without solving, sits off that direction, estimated at 0.190, stable across the multistart certification, with a wide bootstrap interval reported in Panel B of Table 3. The pool share π_{tr} for a contestant Auditor leaves that Auditor's own model out of (2), a mean over the other 37, in which the other two contestants enter at their re-estimated abilities while the 35 held agents enter at their first-stage abilities, matching the leave-one-out discipline of the concordance measure, while the concordance probabilities χ_{tr} are measured from the solving data as Appendix C describes, so neither object adds a parameter.

The dependence layer prices what the verdict likelihood treats as independent. The four Auditors audit the same response, so their errors share a response-level component. With $z_{ijp} = \Phi^{-1}(p_{ijp})$ the normal score of the fitted marginal acceptance probability, the layer specifies

$$\Pr\big(y_{ijp} = 1 \mid u_i\big) \;=\; \Phi\big(z_{ijp}\sqrt{1+\tau_{ijp}^2} \;+\; \tau_{ijp}\, u_i\big), \qquad u_i \sim N(0,1), \tag{16}$$

where u_i is a response-level artifact draw shared by all Auditors in both passes, while the loading τ_{ijp} equals τ_w on wrong candidates with τ_c on correct ones in Pass 2, scaled by the Pass-1 multiplier k_1 in Pass 1. The construction preserves every fitted marginal exactly, since $\mathrm{E}_u\,\Phi(z\sqrt{1+\tau^2} + \tau u) = \Phi(z)$, so the layer prices the dependence without moving any acceptance rate. Its three parameters are estimated by maximum likelihood with the margins frozen, a quadrature over u_i per response, with a likelihood-ratio statistic against independence of 1031.5 on three degrees of freedom. This layer is what the panel results of Section 6 price, with the frozen-margin caveat disclosed there, since a loading estimated on frozen margins absorbs any response-level misfit common to the Auditors.

The audit cost equations are estimated conditional on the audit technology, with ψ_{j2}, the criteria, the displacement, the abilities frozen at their Stage-2 values. For Auditor j with meter y, the money meter being the logged dollar cost of the call while the time meter is the logged seconds, the estimating equation is

$$\log y_i \;=\; t^{y}_{jp} \;+\; \alpha^{y}_{jp} \log \tilde{L}_i \;+\; \beta^{y}_{j}\, r_i \;+\; \theta^{y}_{jv_i\omega_i p} \;-\; \gamma^{y}_{j} \left|\hat{z}_{i2}\right| \mathbf{1}\{p = 2\} \;+\; \phi^{y}_{jt(i)} \;+\; u^{y}_{i}, \tag{17}$$

where t^{y}_{jp} is an Auditor-pass level, α^{y}_{jp} is a pass-specific slope in the logged derivation length, β^{y}_{j} is a red-herring shift, ϕ^{y}_{jt} is a task effect, while $\theta^{y}_{jv\omega p}$ is a premium on the verdict-by-state cell of the call, the pair (v_i, ω_i) of the verdict with the true state, measured relative to the accepted-correct cell in each pass. The chronometric regressor $\left|\hat{z}_{i2}\right|$ is the fitted margin of the call, the distance between candidate i's expected Pass-2 signal under the Stage-2 estimates and the frozen criterion, in units of the signal noise, so γ^{y}_{j} is the chronometric loading of Assumption 5 evaluated on the estimated margin. The cell premia are the reduced form of the distance term's conditional mean across verdict-by-state cells, which is what lets the data decide whether the wrong-answer premium is boundary distance or composition, the comparison Section 4.6 set up. The main specification sets the learning dividend ν to zero, so the engagement service lives in the error rather than in the discrimination, with the error pair of the two meters drawn jointly with an Auditor-specific covariance Σ_j. Each equation is estimated by least squares over the Auditor's usable calls. The two meters share one design within Auditor, so per-meter least squares is the efficient system estimator. Standard errors are clustered two ways, by task and by candidate agent (Cameron et al., 2011). One consequence of this design is recorded once here, then respected throughout. The cell structure of (17) nests the reduced-form price schedules of Section 4.6 Auditor by Auditor, so the model reproduces those exhibits by construction, which is a property of the parameterization rather than evidence of fit. The paper therefore keeps a moment ledger. A moment the design nests is targeted, reproduced by algebra, never cited as fit. Only statistics outside the fitted span can test the model.

Inference is by a multiplier bootstrap of the entire chain, the bootstrap analogue of the paper's two-way clustering convention (Cameron et al., 2011). Each draw assigns the 59 tasks multinomial weights with the 38 agents independent multinomial weights, every observation carrying the product of its task weight with its agent weight, while the four Auditors are never resampled, being the technologies under study. Each draw then re-estimates the solving block, re-measures the pool shares with the concordance probabilities, re-estimates the audit block, the dependence layer, the audit cost equations.[12] The second-

[12] A resampled draw can leave the location of the solving block undetermined, since the Rasch likelihood

step intervals therefore carry the first-step estimation error, which is the treatment of the generated-regressor problem in this architecture. I report percentile intervals over 1,000 draws, none of which failed, as the pre-committed inference object. Two standing gates discipline the run. Every tracked interval is compared with its asymptotic counterpart, the two-way clustered standard error of the audit cost equations or the Hessian standard error of the audit block, with any ratio outside its band investigated before the number is quoted, while Table 34 in Appendix C reports the ratios. A frozen-first-stage arm of 192 paired draws decomposes each interval into first-step against second-step contributions, reported with the results it concerns.

The estimation protocol closes with a fit check governed by the moment ledger. The gate simulates data from the estimated model, then confronts only statistics the fitted span does not carry. On the meter side these are the difficulty gradient of the caught-wrong premium, the cell structure of the covariance between the two meters, the Pass-1 gradients of the cell premia in derivation length, the task-specific red-herring cost gradient. On the verdict side the gate confronts the Auditor-pair correlation structure with the exchangeable dependence of (16), six pair correlations per state against one loading. The model bands treat calls as independent draws, which understates the dispersion of statistics computed on dependent calls, so a statistic outside its band is confronted with its own bootstrap interval before it counts as a rejection. The gate's verdicts are reported with the structural results.

5 Structural Estimates

This section reports what the estimated model says. Every interval is the 95 percent percentile interval from the 1,000-draw bootstrap of the full estimation chain, Section 4.8, so every second-step interval carries the first-step estimation error. Point estimates come from the two-step program at the frozen criteria. The moment ledger governs every claim. A moment the design nests is reproduced by construction, so it is never cited as evidence, while the statements below rest on freely estimated parameters, on statistics outside the fitted span, on intervals that could have covered zero yet did not.

Table 1: The wrong-answer premium is composition, not chronometry

	Money meter, log dollar cost		Time meter, log seconds	
	Caught wrong $\theta^{y}_{j,\mathrm{CW},2}$	Chronometric γ^{y}_{j}	Caught wrong $\theta^{y}_{j,\mathrm{CW},2}$	Chronometric γ^{y}_{j}
Panel A. The Pass-2 caught-wrong premium with the chronometric loading, by Auditor				
Floor Auditor	+0.109	+0.110	+0.163	+0.054
	[+0.016, +0.238]	[−0.069, +0.238]	[+0.074, +0.272]	[−0.074, +0.170]
Robust Auditor	+0.019	−0.034	+0.072	−0.032
	[−0.112, +0.120]	[−0.102, +0.050]	[−0.047, +0.163]	[−0.094, +0.051]
Anchor Auditor	+0.125	+0.018	+0.220	+0.073
	[−0.038, +0.322]	[−0.116, +0.128]	[+0.043, +0.450]	[−0.080, +0.187]
Ceiling Auditor	+0.716	−0.017	+0.844	+0.044
	[+0.489, +1.006]	[−0.232, +0.175]	[+0.604, +1.127]	[−0.135, +0.220]
Panel B. The Ceiling Auditor's Pass-2 decomposition				
Caught wrong, $\theta_{\mathrm{CW},2}$	+0.716		+0.844	
	[+0.489, +1.006]		[+0.604, +1.127]	
Accepted wrong, $\theta_{\mathrm{AW},2}$	−0.277		−0.323	
	[−0.565, +0.025]		[−0.634, −0.007]	
Catching minus credulity, $\theta_{\mathrm{CW},2} - \theta_{\mathrm{AW},2}$	+0.992		+1.167	
	[+0.798, +1.194]		[+0.943, +1.385]	
Rejected correct, $\theta_{\mathrm{RC},2}$	+0.655		+0.750	
	[+0.274, +0.958]		[+0.333, +1.101]	
Chronometric share of the premium	+0.017		−0.038	
	[−0.170, +0.256]		[−0.193, +0.138]	

Pass-2 caught-wrong premium with the chronometric loading by Auditor (Panel A), and the Ceiling Auditor's premium decomposition (Panel B), with 95 percent bootstrap percentile intervals in brackets.

5.1 The Ceiling Auditor's wrong-answer premium is composition, not chronometry

Table 1 reports the audit cost equations' verdict on the price of latent wrongness. For the Ceiling Auditor a caught wrong answer, a wrong final answer the Auditor rejects, carries a Pass-2 premium of +0.716 log points on the money meter with +0.844 on the clock, within task relative to an accepted correct answer, intervals [+0.489, +1.006] with [+0.604, +1.127]. The sharper object is the gap between catching a wrong answer and being fooled by one, the catching-minus-credulity contrast $\theta_{\mathrm{CW},2} - \theta_{\mathrm{AW},2}$, which is +0.992 log points of money with +1.167 of time, intervals [+0.798, +1.194] with [+0.943, +1.385]. The contrast survives every convention. The raw within-cell gap is +1.063 with +1.165, the task-effects-only design gives +1.001 with +1.132, the full design gives +0.992 with +1.167.

The accepted-wrong premium is negative within task, −0.277 money with −0.323 time. The time interval [−0.634, −0.007] excludes zero while the money interval [−0.565, +0.025] does not. The within-task clause carries real content rather than a caveat. Across tasks an accepted wrong answer costs more than an accepted correct one, +0.258 money with +0.208 time in the raw cell means, because wrong answers live on harder tasks. Within task the Auditor spends less on the wrong answers it accepts than on the correct answers it accepts. Credulity is the cheap error. The rejected-correct premium, +0.655 money with +0.750 time, completes the pattern, since the expensive verdicts are the rejections on either state.

The chronometric channel is estimated freely against this composition. Its loading γ_j^y multiplies the fitted distance of the candidate's expected signal from the criterion, so pure chronometry, the account in which the premium is boundary distance, needs a positive γ carrying the premium. For the Ceiling Auditor γ is −0.017 money with +0.044 time, intervals [−0.232, +0.175] with [−0.135, +0.220]. The chronometric share of the Ceiling Auditor's premium, computed draw by draw, spans zero on both meters, [−0.170, +0.256] money with [−0.193, +0.138] time, so at no point of either interval does the chronometric term carry more than 25.6 percent of the money premium or 19.3 percent of the time premium. The premium is composition.

The exclusion names its Auditor. The wrong-answer premium belongs to the Ceiling Auditor, whose reduced-form premium is +0.517 money with +0.597 time. The Floor Auditor's premium is negative, −0.069 with −0.053, the single-crossing signature of Proposition 3(ii), so there is no premium there for chronometry to explain. The Robust Auditor's premium is indistinguishable from zero, which leaves its share bound uninformative. The Anchor Audi-

identifies only contrasts, so every draw is exported on a fixed location convention that leaves each identified contrast unchanged to machine precision. The replication code verifies the convention with an executable test rather than an argument.

tor carries a modest premium whose chronometric bound reaches 101.6 percent on the clock, so chronometry is not excluded for the Anchor Auditor. Every chronometry statement in this paper is therefore about the Ceiling Auditor, never about the panel.

Two statements survive from outside the fitted span. The composition account predicts a premium flat in difficulty, since a cell premium moves with what the Auditor did rather than with where the signal fell, while chronometry ties the premium to the margin, which moves with difficulty through κ. The fit check estimates the difficulty gradient of the Ceiling Auditor's premium at -0.005 money with $+0.003$ time, inside even the independence bands of the gate. The ordering is claimed exactly where it separates. The Ceiling premium is the panel maximum in 100.0 percent of the 1,000 draws on both meters, ahead of the runner-up by $[+0.297, +0.847]$ money with $[+0.321, +0.882]$ time, while no complete ranking of the four Auditors is claimed, since the middle pair sits within sampling variation of each other.

The premium is present before the derivation is shown. The Pass-1 caught-wrong premium for the Ceiling Auditor is $+1.017$ money with $+1.076$ time, standard errors 0.083 with 0.076 under the two-way clustering convention. One gradient the common-slope specification does not carry is disclosed by the fit check. That premium declines with the length of the candidate's response, -0.159 log points of money per unit of logged length, which is 0.115 log points per standard deviation of length, about a ninth of the premium. The block is also first-stage robust. Re-estimating the chain with the solving block frozen moves the composition cells' interval variance by under one percentage point with the chronometric share by essentially nothing, so nothing in this subsection leans on the estimated solving technology beyond its point estimates. The economic interpretation is the section's headline. The cost of a call prices the act of catching. A counterfeit that is caught is expensive to catch, while a counterfeit that passes is cheap to wave through, so the price of verification rises exactly where verification works. The Ceiling Auditor pays the largest premium because it is the Auditor doing the catching, the capability gradient of Proposition 3(i) recast as composition.

5.2 The audit technology: neglect, displacement, the price of difficulty

Table 2 reports the audit block. Base-rate neglect is universal. The equivalence of Section 4.8 stands, since an Auditor that underweights its pool is observationally identical on the verdict surface to a Bayesian Auditor whose criterion carries noise, so w_j measures how little the bar moves with the pool rather than why. The weight w_j on the pool-share term of the criterion (15) is 0.115 for the Floor Auditor, 0.193 for the Robust Auditor, 0.212 for the Anchor Auditor, 0.185 for the Ceiling Auditor, every interval far below the Bayesian benchmark

Table 2: The audit technology with the ability shifts

	Floor Auditor	Robust Auditor	Anchor Auditor	Ceiling Auditor
Panel A. The audit block, by Auditor				
Base-rate weight w_j	+0.115	+0.193	+0.212	+0.185
	[+0.061, +0.164]	[+0.118, +0.256]	[+0.122, +0.305]	[+0.024, +0.277]
Correct-side displacement ρ_j	+0.283	−0.090	−0.045	+0.079
	[+0.053, +0.562]	[−0.316, +0.207]	[−0.235, +0.213]	[−0.130, +0.342]
Wrong-side displacement ρ_{Wj}	+0.391	+0.076	+0.058	−0.213
	[+0.169, +0.625]	[−0.211, +0.365]	[−0.325, +0.345]	[−0.561, +0.080]
Re-solve probability, mean task $\vartheta_j(0)$	0.732	0.614	0.745	0.789
	[0.638, 0.850]	[0.513, 0.725]	[0.666, 0.832]	[0.676, 0.882]
Panel B. Common parameters				
Pass 2 level c_2		+0.299 [+0.075, +0.476]		
Pass 2 difficulty κ (solving = 1)		+0.218 [+0.134, +0.298]		
Pass 2 capability λ		+0.145 [+0.062, +0.244]		
Re-solve difficulty slope ϑ_1		+0.211 [+0.101, +0.347]		
Reasoning-agent displacement ρ_R		+0.414 [+0.173, +0.727]		
Derivation-length displacement ρ_L		−0.139 [−0.299, +0.015]		

The audit block by Auditor (Panel A) and the common parameters (Panel B), with 95 percent bootstrap percentile intervals in brackets.

Table 3: Solving versus auditing, with the criterion objects the passes identify

Panel A. Solving versus auditing, the three Auditors that solved as agents

	Solving a_m	Auditing a_j	Shift, pooled	Anchored stratum
Floor Auditor	−1.342	−1.064 (0.055)	+0.279	[−0.153, +0.726]
			[−0.175, +0.782]	
accuracy, mean task	20.7%	25.7%		
Robust Auditor	+0.502	+0.846 (0.055)	+0.344	[+0.058, +0.571]
			[+0.034, +1.208]	
accuracy, mean task	62.3%	70.0%		
Ceiling Auditor	+2.489	+2.619 (0.062)	+0.129	[−0.314, +0.402]
			[−0.350, +0.799]	
accuracy, mean task	92.3%	93.2%		

Panel B. The criterion objects the passes identify

	Floor Auditor	Robust Auditor	Anchor Auditor	Ceiling Auditor
Pass-2 criterion constant $\log \ell_j + \mu_j$	0.468	1.280	1.446	0.606
	[0.288, 0.658]	[0.932, 1.733]	[1.098, 1.848]	[0.246, 1.011]
Pass-1 flip point, pool share	0.016	< 0.0018	0.0015	0.062
	[0.001, 0.055]	< 0.0004	[0.0001, 0.0025]	[0.002, 0.123]
ψ_0 (branch without solving)		0.190 [0.006, 1.905]		

Solving ability against auditing ability on one scale for the three Auditors that solved as agents (Panel A), and the criterion objects the passes identify (Panel B).

of one, the highest upper endpoint 0.305. When the red herring cuts the share of correct candidates in the pool, the Auditor barely tightens its bar. The three abler Auditors' weights are statistically indistinguishable, since the Robust Auditor's interval $[+0.118, +0.256]$, the Anchor Auditor's $[+0.122, +0.305]$ and the Ceiling Auditor's $[+0.024, +0.277]$ overlap over most of their length, so the data do not order them by ability. The Floor Auditor's 0.115 is the lowest, with its interval $[+0.061, +0.164]$ reaching into the other three. What the block establishes is that no Auditor, the strongest included, moves its bar by more than 0.212 of the Bayesian amount, about a fifth.

The displacement estimates retire a simpler story this paper could have told, in which the red herring works only through wrong candidates against the Floor Auditor. The collapse of the Floor Auditor is two-sided. Its correct-side displacement ρ_{Floor} is $+0.283$, interval $[+0.053, +0.562]$. Its wrong-side displacement $\rho_{W,\text{Floor}}$ is $+0.391$, interval $[+0.169, +0.625]$. Both exclude zero. On red-herring tasks the Floor Auditor accepts more counterfeits while it also rejects more correct answers, which is what the model-free arm gaps of its confusion table show. The Robust Auditor is immune on both margins, with both displacement intervals spanning zero. The Ceiling Auditor's correct-side displacement is $+0.079$ with its wrong-side -0.213, both spanning zero. The artifact loadings complete the falsification margin. Reasoning agents displace signals by $\rho_R = +0.414$, interval $[+0.173, +0.727]$, the structural rendering of the falsification premium of Table 18, while longer wrong derivations lean toward being caught, $\rho_L = -0.139$, interval $[-0.299, +0.015]$. One limit is carried from the estimation. The Pass-2 verdicts identify only the sum $\log \ell_j + \mu_j$, the Pass 2 criterion constant, which is positive for every Auditor, 0.468, 1.280, 1.446 and 0.606 in the order Floor, Robust, Anchor, Ceiling, with 95 percent intervals from the re-executed draws of $[0.288, 0.658]$, $[0.932, 1.733]$, $[1.098, 1.848]$ and $[0.246, 1.011]$, every one excluding zero, so the confirmatory criterion of Proposition 1 holds in the Auditor verifying pass throughout the intervals. The Auditor solving pass identifies only the pool share at which the branch without solving flips, 0.016, 0.0015 and 0.062 for the Floor, Anchor and Ceiling Auditors, with the Robust Auditor's below 0.0018. For the Floor and Ceiling Auditors two to four tasks per arm sit below that share, which pins it. For the Anchor Auditor no task sits below it and two sit in its transition, where the branch without solving accepts about five in six, which pins it weakly. For the Robust Auditor no task comes near it, so its flip point is bounded above by the lowest pool share in the sample and its loss ratio is bounded below, not estimated. How the flip point is written in terms of ℓ_j depends on the weight the cutoff of the branch without solving puts on the pool share, a choice worth less than one log-likelihood point. Under weight one, the convention of (7), the form premia are -3.63, -5.04 and -2.11 for the Floor, Anchor and Ceiling Auditors and at most -5.0 for the Robust Auditor. Under

the Pass 2 weights w_j they are 0.00, +0.07 and +0.10, with at most +0.06 for the Robust Auditor. The sign of μ_j, on which the persuasion effect of Proposition 2(iii) rests, is therefore not identified. I claim nothing about it.

Difficulty is where checking earns its keep. Solving pays intrinsic difficulty at a loading of one on the log-odds scale by (1). Checking the derivation has difficulty elasticity $\kappa = 0.218$, interval $[+0.134, +0.298]$. On the odds scale on which solving's loading is one, the Pass 2 loading is $\kappa G(\psi_{j2})$, which at the mean task runs from 0.21 for the Floor Auditor through 0.29 for the Robust Auditor and 0.31 for the Anchor Auditor to 0.39 for the Ceiling Auditor, so the worked derivation shields the checker from between three-fifths and four-fifths of the difficulty the solver paid in full, less for the abler Auditor. The capability loading of checking is $\lambda = 0.145$, interval $[+0.062, +0.244]$, small against solving's unit loading, so the derivation does much of the checking for the checker, which is why weak Auditors gain most from seeing the steps. The Auditor solving pass mixes the two technologies. At the mean task the Auditors re-solve with probability 0.61 to 0.79, intervals from 0.51 to 0.88 that all sit above one half, with the share falling in intrinsic difficulty at slope $\vartheta_1 = 0.211$, interval $[+0.101, +0.347]$, so on harder tasks the pass shifts from re-solving toward ruling on the bare final answer without solving.

At the point estimate the branch without solving itself does not discriminate. Its discrimination ψ_0 is 0.190, with a bootstrap interval of $[0.006, 1.905]$ from the re-executed draws that is wide because only the tasks the pool almost never solves inform it, 80 percent of the draws placing it below the point estimate and 92 percent below 0.5. At the estimated criterion the branch without solving accepts a correct candidate and a wrong candidate alike on every task whose pool share exceeds a flip point, the share at which the pool odds equal $1/\ell_j$, which is 0.016 for the Floor Auditor, 0.0015 for the Anchor Auditor and 0.062 for the Ceiling Auditor, with the Robust Auditor's below 0.0018, the lowest pool share in the sample. Its acceptance falls to zero as the share drops below that point. That leaves every task for the Robust and Anchor Auditors and all but two to four tasks for the Floor and Ceiling Auditors, the tasks the pool almost never solves. The Auditor solving pass therefore discriminates only when the Auditor re-solves. With the branch without solving accepting everything, (9) and (10) reduce to $C_{j1} = 1 - \vartheta_j(1 - q_j)$ and $F_{j1} = 1 - \vartheta_j\big[1 - (1 - q_j)\chi_{tr}\big]$. Every rejection then comes from re-solving, of correct and wrong candidates alike, which is why the Floor Auditor, the weakest solver among the three Auditors that solved as agents, re-solving with probability 0.732 at the mean task, delivers 0.632 of the correct control candidates in Pass 1 against 0.882 in Pass 2, Table 30.

One reconciliation is owed, because the model-free arm gaps could otherwise read as contradicting the parameters. The Ceiling Auditor's Pass-2 false-accept rate rises by +0.119

under the red herring while its correct-side displacement is only +0.079 with an interval spanning zero. Table 32 in Appendix C resolves this. It decomposes each Auditor's model arm gap one channel at a time from the full model, displacement, pool and difficulty, with the interaction residual, so the four parts add to the model gap in every row. The Ceiling Auditor's false-accept gap of +0.090 in the model is mostly the difficulty channel, +0.094, since the red herring raises the effective difficulty κh_t of the Pass 2 signal itself, against +0.029 from displacement and −0.044 from the pool channel, the bar tightening that w_j damps. The red herring beats the Ceiling Auditor mainly by making the checking harder, not by camouflage. The Floor Auditor is the mirror case. Displacement contributes +0.103 of its false-accept gap of +0.169 and −0.100 of its wrong-side gap of −0.117, so the Floor Auditor's two-sided collapse is camouflage. The model overshoots the Floor Auditor's measured false-accept gap, +0.123, by 0.045, because its control-pool false-accept rate is 0.497 against a measured 0.540, Table 30. The Robust Auditor's measured gap of −0.001 hides offsetting channels, +0.120 from difficulty against −0.027 from displacement and −0.022 from the pool, with a model gap of +0.040. These are readings of the fitted mechanism rather than tests, since the verdict likelihood targets the underlying cells, which the moment ledger records.

5.3 Auditing against solving on one scale

Panel A of Table 3 puts each contestant Auditor's two abilities on one scale. The Floor Auditor solves at ability −1.342 while it verifies at −1.064, standard error 0.055 on the auditing ability, a shift of +0.279 whose pooled interval is $[-0.175, +0.782]$. The Robust Auditor shifts +0.344 from +0.502, the Ceiling Auditor +0.129 from +2.489. At the mean task the shifts translate to accuracy gains from 20.7 to 25.7 percent for the Floor Auditor, 62.3 to 70.0 for the Robust Auditor, 92.3 to 93.2 for the Ceiling Auditor. The same model, auditing another model's answer to the task it would itself have to solve, operates as if it were an abler solver than its own solving record shows.

The bootstrap decides how much of this survives, Auditor by Auditor. The Robust Auditor's shift excludes zero in the pooled interval, $[+0.034, +1.208]$, with the finding stratum-robust, since in the anchored regime, where the draw re-estimates the auditing ability jointly with its solving reference, the interval is $[+0.058, +0.571]$ with 99.1 percent of draws positive. The Floor Auditor's shift with the Ceiling Auditor's are positive but not significant, pooled intervals $[-0.175, +0.782]$ with $[-0.350, +0.799]$ and anchored-stratum intervals $[-0.153, +0.726]$ with $[-0.314, +0.402]$, exactly the borderline the asymptotic t near two predicted. The pooled intervals' long upper reaches have a named source. In roughly a third of draws the Auditor's contestant model draws zero agent weight, so the solving reference stays

at its point value while the auditing ability re-estimates, a regime in which the verdict layer alone reads the shift higher. The claim this paper carries is therefore the Robust Auditor's, stated with its stratum-robust interval, while the other two shifts are reported as positive without significance.

One dispersion artifact is disclosed rather than hidden. In two to three of the 1,000 draws a contestant drew zero agent weight with its auditing ability wandering a near-flat direction of the refit, which inflates moment-based dispersion measures for the Ceiling Auditor's shift while the percentile endpoints, the pre-committed inference object, move by at most 0.013 when those draws are dropped.

6 The Market for Checking

The estimates of Section 5 are the inputs to a second model. A principal who delegates the solving takes the estimated audit technology as given, then decides what checking to buy, so the propositions of this section describe the market rather than the Auditor. They follow the estimates because the market runs on them. The framing is costly state verification with the verification technology estimated rather than assumed, a delegation problem in the lineage of Townsend (1979) with monitoring conceded to be noisy, so the audit is soft evidence, a noisy signal of the state, rather than the hard evidence that mechanism design with partial verification takes as its primitive (Green and Laffont, 1986; Hart et al., 2017). There is no strategic auditee. The agent does not respond to the audit, so the analysis is a decision problem, which the experimental design makes literal. Every interval in this section is conditional on the four Auditors, which the bootstrap treats as fixed technologies rather than draws from a population. The claims are about these four models on these tasks. Any extension to an Auditor the experiment did not run is the model's rather than the data's.

6.1 The principal's problem

A principal needs one delivered answer on task t in arm r. She values a correct delivery at one, bears a loss $\ell_0 > 0$ on a wrong delivery, holds an outside option worth zero. A production draw from the pool costs the production price Π_s, the money price plus the valued time price of one agent response, Panel A of Table 12. A verdict by Auditor j in pass p costs $\Pi_{jp} = p_{jp} + w\,\tau_{jp}$, the total price of the call at the principal's opportunity cost of time w (Acton, 1975; Becker, 1965). The pool share of correct candidates is π_{tr} from (2). The Auditor's operating point on that pool is (C, F) from Section 4, which the principal takes

as technology.[13] Table 13 reports how the total bill of an audited response splits between production and verification, on money and on time.

Three policies span the relevant menu. Trust delivers the first draw unaudited, worth $V_{\text{trust}} = \pi_{tr}(1 + \ell_0) - \ell_0 - \Pi_s$. Refusal delivers nothing, worth zero. The audit policy sends each draw to Auditor j in pass p, delivers on an accepting verdict, redraws otherwise. Under the audit policy a draw clears with probability $A = \pi_{tr}C + (1 - \pi_{tr})F$, the delivered answer is correct with probability $R = \pi_{tr}C/A$, the expected outlay is $(\Pi_s + \Pi_{jp})/A$.

Proposition 4 (The value of a verdict)**.** *Hold the operating point (C, F) fixed with $C > F$.*

(i) The audit policy beats trust exactly when

$$(1 + \ell_0)\,\pi_{tr}(1 - \pi_{tr})\,(C - F) \;\geq\; \Pi_{jp} \;+\; (1 - A)\,\Pi_s. \tag{18}$$

(ii) The left side is the value of the verdict. It vanishes as π_{tr} approaches zero or one, peaks at interior uncertainty, scales with the stake $1 + \ell_0$, scales with the verdict's separation $C - F$. The right side is the verdict's full price, the call itself plus the redraw bill its rejections cause.

(iii) The set of pool shares at which the audit policy beats trust is an interval.

Proposition 4 is the option value of a costly look, rendered for a binary state. Its comparative statics are the ones Gabaix et al. (2006) derive for costly information operations, value falling in the decisiveness of the prior while rising in the look's precision, with the Gaussian closed form replaced by the exact binary-state expression that the setting demands. The three policies tile the pool-share line into refusal at the bottom, audit in an interior band, trust at the top. The red herring moves tasks toward the band, since it drags the pool share from 0.638 toward 0.499, exactly while it compresses the separation $C - F$ that the band's value rides on. That squeeze, demand for checking rising as the checking technology degrades, is the trap configuration of Goldsmith-Pinkham (2026), here as an inequality over estimated objects. The wage does its work inside the prices. Because Π_{jp} is linear in w, the ranking of Auditors can flip as the wage rises. Table 12 reports each input's break-even wage, the wage at which its time price equals its money price, between \$0.33 and \$7.18 an hour. For a principal whose time is dearer than about seven dollars an hour, the time price

[13] The three regimes of Bergemann et al. (2026), approve or fall back, restrict the permission set, reshape rewards, nest. They describe the first as broadly analogous to trusted-monitor control (Greenblatt et al., 2024). The menu here is their first regime with a redraw as the fallback, while the Refusal policy is their default action. Their finding that giving a biased monitor more discretion need not help the principal has a direct analogue here, since the Auditor's criterion (ℓ_j, w_j, μ_j) plays the role of the monitor's bias, which is why the frontier prices each Auditor at its own criterion.

is the larger share of every input's total price, so demand responds more to the time price than to the money price by the share identity of Acton (1975). Cheap fast checking then dominates dear slow checking even at equal verdict quality. This is how the Anchor Auditor, at three tenths of a cent a verdict, is the cheapest policy at the 70 percent target on the control pool in both blocks of Table 4. It is also how the Floor Auditor's Auditor verifying pass, at a tenth of a cent a verdict, stays on the lower envelope of the control pool at both operating points, at 0.758 by the model and 0.740 at the data, clearing the 75 percent target by the model and missing it at the data, a role its solving record alone would never justify.

6.2 Reliability and its price

Write the delivered reliability of the audit policy as $R = \pi_{tr}C/\big(\pi_{tr}C + (1-\pi_{tr})F\big)$. The frontier the paper prices is the minimum expected outlay at which a policy reaches a reliability target. The next result says when the frontier exists at all, with the camouflage displacement as the pivot.

Proposition 5 (The camouflage cap)**.** *Fix Auditor j in pass p on a pool whose wrong candidates share the displacement δ^C while its correct candidates share the wrong-side displacement δ^W, under Assumptions 1 to 3, with discrimination ψ.*

(i) For every criterion, $\operatorname{sign}(R-\pi_{tr}) = \operatorname{sign}(C-F) = \operatorname{sign}(\psi-\delta^C-\delta^W)$. *The audit raises the reliability of the accepted stream above the pool share exactly when its signal discrimination clears the two displacements together.*

(ii) If $\psi > \delta^C+\delta^W$, reliability rises strictly as the criterion tightens, approaching one only as the clearance rate approaches zero with the expected outlay growing without bound. If $\psi < \delta^C+\delta^W$, reliability falls strictly as the criterion tightens. A harsher audit then anti-selects, the accepted stream is worse than the unaudited pool at every criterion, worst at the strictest.

(iii) A panel of J Auditors with conditionally independent signals at common discrimination, facing the common displacements, has effective separation $\sqrt{J}\,(\psi-\delta^C-\delta^W)$. Panel size rescales the separation without changing its sign.

Corollary 2 (Verification debt)**.** *On a pool whose two displacements together, $\delta^C+\delta^W$, weakly exceed the best affordable signal discrimination, no acceptance policy at any budget, by any panel of any size, raises delivered reliability above the pool share. When the excess is strict, a stricter criterion strictly lowers it. Unverified error then accumulates at the pool's rate.*

The configuration is approached but not crossed in the Auditor verifying pass. The Floor Auditor's Pass-2 margin on the red-herring pool clears the two displacements by a realized discrimination of only +0.29, Table 21. Its Pass-1 verdicts on that pool do cross it, with a realized discrimination of −0.41 and an accepted stream less reliable than the untouched pool, 0.467 against the pool share 0.499 in the model of Table 28 with 0.430 at the measured operating point of Table 30. That pass is the mixture of Section 4.4, so the anti-selection there runs partly through concordant error, the Auditor's own wrong re-solve landing on the candidate's wrong answer, with the rest through camouflage. The corollary formalizes verification debt, the accumulation of unchecked error when the price of solving falls faster than the reliability of checking improves. The term is due to Goldsmith-Pinkham (2026). The formalization as an inequality over estimated objects is this paper's. It also gives the market version of the adversarial-estimation limit, in which the best feasible discriminator's verdicts concentrate at one half as the generator improves (Kaji et al., 2023). In this model that limit is the displacement climbing toward the discrimination, since the wrong answers that survive an improving generator's own filters are exactly the ones that read as right. The same limit is the one chain-of-thought monitoring meets when the monitored model is trained against the monitor, which teaches it to hide its intent while it keeps exploiting the task (Baker et al., 2025; Korbak et al., 2025). Here no one trains against the Auditor, yet the displacement climbs toward the discrimination as the producer improves. Aggregation is no escape, by part (iii), because the displacement is a property of the artifact, common to every Auditor, while only the signal noise averages out. Escape routes must move $\psi - \delta^C - \delta^W$ itself. The key does so by construction, which is why the Anchor Auditor's with-key agreement of 97.6 percent collapses to the without-key operating points of Table 21, a fall that prices the key as an input. Re-posing the task to the checker in a form the counterfeit was not written for is the other route, since camouflage tuned to one task text need not survive a change in that text. The experiment does not run it. The Auditor always sees the text the solver saw, so in the data the red herring enters the audit as a source of displacement, Section 5.2, not as an escape from it. The re-posed task is a counterfactual for the estimated model to price, while the key's price is measured in that fall.

Tables 28 and 29, in Appendix C, price every audit policy at the estimated technology. For each policy on each pool they report two numbers. The delivered reliability R is the correct share of what the policy delivers. The outlay per delivered answer is the money price of one production draw plus the money prices of the policy's calls, that sum divided by the clearance rate, since a rejected draw is redrawn. Both numbers pool the model's operating point over the pool's candidates at the pool's mean share, the construction of the measured operating points in Table 21, so the model and the data are compared like for like. Table 4

takes the frontier from them, the cheapest policy that reaches each reliability target, at the model's operating points and at the measured ones. On the control pool the cheapest policy that reaches a target of 70 percent is the Anchor Auditor's Pass 2, the Auditor verifying pass. It delivers 0.716 for \$0.0237. The cheapest that reaches 90 percent is the Ceiling Auditor's two passes under unanimous acceptance. It delivers 0.916 for \$0.0569, 3.2 times the \$0.0176 outlay of an unaudited draw. No policy in the menu reaches 95 percent at the point estimates. In the bootstrap of Table 4 the 90 percent target is attained in 76 percent of draws and the 95 percent target in 5 percent. On the red-herring pool the cheapest policy that reaches 70 percent is the Anchor and Ceiling Auditors' Pass-2 verdicts under unanimous acceptance, at 0.712 for \$0.0433. The highest reliability any policy reaches there is 0.822, the Ceiling Auditor's two passes at \$0.0677. No policy reaches 85 percent under this construction at the point estimates. In the bootstrap the 80 percent target is attained in 70 percent of draws, the 85 percent target in 23 percent and the 90 percent target in 1 percent. The Floor Auditor's Pass 1, the Auditor solving pass, delivers 0.467 on that pool against 0.499 in the model and 0.430 at the data, as above. Combining different Auditors' Pass-2 verdicts adds at most 3.6 points of reliability over the best single-Auditor pass on the control pool and nothing on the red-herring pool, while combining the Ceiling Auditor's own two passes is what reaches 90 percent, 7.2 points over its Pass 1. The Anchor and Ceiling pair is still the cheapest policy at two targets, 0.85 on the control pool and 0.70 on the red-herring pool. The four-Auditor unanimous panel delivers 0.880 on the control pool for \$0.0747, less reliability at a higher outlay than the Ceiling Auditor's two passes, because the four readings of one response share the response-level draw of (16). Under conditionally independent readings the same panel would deliver 0.928. Consensus of two production draws delivers 0.843 for \$0.0608 on the control pool, again below the Ceiling Auditor's two passes at a higher outlay, because two wrong final answers match with probability χ_{tr}, which is the correlated error that scoring one agent's answer against another's cannot see, the case peer-prediction mechanisms exclude by assumption (Bergemann et al., 2026; Qiu et al., 2026). The upper block of Table 4 takes the frontier at the model's operating points, the lower block at the measured ones of Table 30. At the measured points four of the upper block's eight entries fall short of their targets, by 0.006 to 0.030, the control frontier runs through the Anchor and Ceiling Auditors alone and the red-herring pool tops out at 0.788. The menu varies the Auditor, the pass and the panel at the estimated criteria. It never varies an Auditor's own acceptance threshold, so Proposition 5 describes the technology's frontier, not a choice the principal makes. The cheapest policy at every attainable target is the same at three wages, zero, \$7.18 an hour and \$20 an hour, where \$7.18 an hour is the highest break-even wage of Table 12. Every outlay rises with the wage. Table 31 also reports the frontier under two other aggregation

conventions. Averaging one delivered answer per task lowers the highest attainable reliability to 0.833 on the control pool and 0.728 on the red-herring pool. Evaluating at the mean task raises it to 0.947 and 0.852. Under every convention and wage no policy reaches 95 percent on the control pool or 90 percent on the red-herring pool at the point estimates. I do not price a targeting counterfactual that audits only the tasks where (18) holds, in the spirit of Duflo et al. (2018), since the experiment randomizes the red herring and not the audit assignment.

Table 4: The price of reliability: the cheapest policy that reaches each reliability target

	Control pool		Red-herring pool	
Target	Policy (R delivered)	Outlay (USD)	Policy (R delivered)	Outlay (USD)
At the model's operating points				
Highest attainable R	Ceiling P1+P2 (0.916) [0.867, 0.956]	0.0569	Ceiling P1+P2 (0.822) [0.728, 0.890]	0.0677
$R \geq 0.70$	Anchor P2 (0.716)	0.0237 [0.0176, 0.0355]	Anchor+Ceiling (0.712)	0.0433 [0.0302, 0.0781]
$R \geq 0.75$	Floor P2 (0.758)	0.0257 [0.0227, 0.0389]	Ceiling P1+P2 (0.822)	0.0677 [0.0344, 0.0784]
$R \geq 0.80$	Anchor P1+P2 (0.804)	0.0340 [0.0238, 0.0601]	Ceiling P1+P2 (0.822)	0.0677 [0.0426, 0.0774]
$R \geq 0.85$	Anchor+Ceiling (0.851)	0.0404 [0.0274, 0.0649]	–	–
$R \geq 0.90$	Ceiling P1+P2 (0.916)	0.0569 [0.0362, 0.0650]	–	–
$R \geq 0.95$	–	–	–	–
$R \geq 0.99$	–	–	–	–
At the measured operating points of Table 30				
Highest attainable R	Ceiling P1+P2 (0.914)	0.0560	Ceiling P1+P2 (0.788)	0.0643
$R \geq 0.70$	Anchor P2 (0.715)	0.0237	Ceiling P1 (0.713)	0.0442
$R \geq 0.75$	Anchor P1 (0.752)	0.0270	Ceiling P1+P2 (0.788)	0.0643
$R \geq 0.80$	Ceiling P2 (0.845)	0.0349	–	–
$R \geq 0.85$	Ceiling P1 (0.861)	0.0399	–	–
$R \geq 0.90$	Ceiling P1+P2 (0.914)	0.0560	–	–
$R \geq 0.95$	–	–	–	–
$R \geq 0.99$	–	–	–	–

The cheapest audit policy reaching each reliability target on each pool, with the reliability it delivers in parentheses and its outlay per delivered answer in dollars.

6.3 The assignment of checking

The last question is who should check whom. Order Auditors by ability. The prices of Table 12 span an order of magnitude at a common task list, from \$0.0026 for a two-pass audit by the Floor Auditor to \$0.0278 by the Robust Auditor, while ability commands no clean premium in them. The Pass 2 audit technology (11) makes an Auditor's discrimination rise with its ability at loading λ, while Proposition 5 makes its usable range end where the displacement catches its discrimination. The assignment problem binds in the menu

configuration where the abler Auditor is dearer. When the abler Auditor is also cheaper it simply dominates.

Proposition 6 (Assignment by camouflage)**.** *Let $\lambda > 0$, let two Auditors share the technology (11) with abilities $a_H > a_L$, hence discriminations $\psi_H > \psi_L$, at prices $\Pi_H > \Pi_L$, on a pool whose wrong candidates share the displacement δ^C and whose correct candidates share the wrong-side displacement δ^W, both common to the two Auditors, with $\tilde{\delta}^C = \delta^C + \delta^W$ their sum.*

(i) The pools an Auditor can improve are nested by ability, $\{\tilde{\delta}^C < \psi_L\} \subset \{\tilde{\delta}^C < \psi_H\}$.

(ii) As the reliability target falls to the pool share, the cheap Auditor attains it at the lower outlay. As the target rises to one, the able Auditor attains it at the lower outlay, for any bounded price gap, because the outlay's logarithm grows in proportion to $(\psi - \tilde{\delta}^C)^{-2}$.

(iii) Define the switching target as the supremum of the targets the cheap Auditor attains at the weakly lower outlay. By part (ii) it is interior, while above it the able Auditor is strictly cheaper at every target. At $\tilde{\delta}^C = \psi_L$ the cheap Auditor's delivered reliability equals the pool share at every criterion, so every target above the pool share belongs to the able Auditor alone.

Comparative advantage in checking belongs to the cheap, not to the able, until camouflage binds. On clean pools the audit flows down the price ladder, to the Auditors whose verdicts cost a few tenths of a cent, the delegation of monitoring that makes verified output scale at all. On corrupted pools the cheap Auditor's margin closes first, so scarce ability earns its price exactly where the counterfeits are good. The assignment mirrors knowledge hierarchies, in which routine problems flow to cheap workers while exceptions climb to scarce ability (Garicano, 2000). Here the exception is not a rare task but an adversarial pool, with the boundary located at $\delta^C + \delta^W = \psi_L$, an estimable line on which both of the Floor Auditor's red-herring loadings in Table 2 sit. The proposition compares the two Auditors at common displacements, the artifact interpretation of Assumption 1, so it is the pure capability comparison. The estimation lets the red-herring loadings differ by Auditor and Table 4 prices each Auditor at its own, which is the data's comparison. The market's price schedule prices verdicts, while what the principal buys is discrimination net of camouflage, whose price Table 4 reports as the outlay per delivered answer at each reliability target. Propositions 5(ii) and 6 treat the criterion as the principal's instrument. The experiment did not vary it. Table 4 with Tables 28 to 31 therefore price each Auditor at its own estimated criterion, so the targets a tunable criterion would reach are the model's counterfactual rather than the data's.

6.4 Self-audit and the improvement loop

The recursive-self-improvement argument requires a model to check its own work at scale (Cunningham et al., 2026). The design excluded the diagonal on purpose, flagging the 552 calls where an Auditor met its own response, 549 of which survive the earlier drops, then dropping every one of them, so the evidence speaks to self-audit only by extrapolation, which I flag rather than hide.

Corollary 3 (The self-audit gate)**.** *Under self-audit the re-solving branch of Assumption 4 draws from the same solving cell as the candidate, so its concordance probability is the cell's own repeat probability* χ^{self}*. On a wrong candidate the re-solving branch of the Pass-1 self-audit rejects with probability* $1-(1-q)\chi^{\text{self}}$*, which falls toward* $1-\chi^{\text{self}}$ *exactly on the cells the model reliably fails.*

At the estimated criteria the branch without solving in the Auditor solving pass accepts every candidate on all but the tasks the pool almost never solves, Section 5.2, so the whole Pass-1 self-audit catches a wrong own answer with probability $\vartheta_j[1-(1-q)\chi^{\text{self}}]$, which Table 5 reports. The agent solving phase measures the entrenchment directly. Table 5 reports, for the three Auditors that solved the tasks as agents, the probability that two of the model's own wrong no-explanation draws on the same task version carry the same final answer, by pool. It runs from 0.226 for the Floor Auditor on the control pool to 0.693 for the Robust Auditor on the red-herring pool, against the concordance the same Auditors face on other agents' wrong answers, 0.431 to 0.527. A fresh own draw therefore reproduces the entrenched answer more often than a stranger's draw would for the Robust Auditor and less often for the Floor Auditor. A companion study reports that 48.0 percent of red-herring waves repeat an identical wrong answer (Akimitsu, 2026b), a share of waves rather than the pairwise probability the model needs, which the table supplies. The table brackets the diagonal's Pass-1 catch, the probability that a self-audit rejects the model's own wrong answer, between the own repeat probability and the concordance measured on others, at the estimated re-solving weights. On the red-herring pool the catch is 0.400 for the Floor Auditor, 0.362 for the Robust Auditor and 0.718 for the Ceiling Auditor at the own repeat probability, against 0.385, 0.424 and 0.721 at the others' concordance. Those are mean-task values. A model's own wrong answers concentrate on the tasks it reliably fails, where Corollary 3 sends the catch to $\vartheta_j(0)(1-\chi^{\text{self}})$, the entrenched-cells column of the table, 0.567, 0.197 and 0.461 on the control pool and 0.364, 0.188 and 0.358 on the red-herring pool for the Floor, Robust and Ceiling Auditors. Weighting the catch over the cells where each model's own wrong waves fell, with the re-solving weight and the branch without solving evaluated on each cell, gives 0.595, 0.286 and 0.556 on the control pool and 0.401, 0.238

and 0.621 on the red-herring pool. Part of that catch is not a discrimination. Below the flip point of the branch without solving, the branch rejects every answer. The share of own wrong waves on such cells is 0.053, 0.000 and 0.059 on the control pool and 0.041, 0.000 and 0.170 on the red-herring pool for the Floor, Robust and Ceiling Auditors. On the cells above the flip point the catch is 0.573, 0.286 and 0.556 on the control pool and 0.377, 0.238 and 0.589 on the red-herring pool. The Pass-2 diagonal adds the displacement of the model's own style, which estimates from Auditors ruling on other agents' responses bound only from below, so the table's full-response column, 0.384 to 0.594, is an upper bound. A chain of delegated steps multiplies its reliabilities, so a verification term below one compounds at the chain's length (Kremer, 1993). The loop's gate is then the diagonal's $\psi - \delta^C - \delta^W$, an object this design brackets rather than measures, because the diagonal itself was excluded by design. On the cells where its own wrong answers concentrate, a model checking its own work catches them with probability 0.188 to 0.567 at the entrenched limit and 0.238 to 0.589 on the cells where the branch without solving still delivers, so self-audit does not lift the cap it faces. The market for checking binds hardest exactly where the recursive-self-improvement argument leans on it most. The weak-to-strong oversight agenda asks a weaker trusted system to regulate a stronger one (Christiano et al., 2018; Bowman et al., 2022). Fudenberg and Liang (2025) study a principal who limits what an AI of unknown alignment is told, while Xiao and Phuong (2026) have a weak trusted model read a stronger model's reasoning to catch collusion. The capability gradient of Section 3, Table 10, is the empirical content of that arrangement in this setting. The weak Auditor gains from the producer's derivation on correct answers and is fooled on wrong ones, so the arrangement inherits the camouflage cap of Proposition 5.

Table 5: The self-audit gate, bracketed

					Pass-1 self-audit catch					Pass-2
					mean task		own wrong		entrenched	
Auditor	Pool	$q_j(0,r)$	χ_{jr}^{self}	χ_{jr}	own χ^{self}	others' χ	waves	raw (n)	cells	upper bound
Floor	Control	0.207	0.226	0.441	0.601	0.476	0.595	0.500 (20)	0.567	0.554
Floor	Red herring	0.099	0.503	0.527	0.400	0.385	0.401	0.391 (23)	0.364	0.384
Robust	Control	0.623	0.679	0.431	0.456	0.514	0.286	0.240 (25)	0.197	0.449
Robust	Red herring	0.409	0.693	0.521	0.362	0.424	0.238	0.345 (29)	0.188	0.386
Ceiling	Control	0.923	0.417	0.433	0.764	0.763	0.556	0.500 (2)	0.461	0.594
Ceiling	Red herring	0.835	0.546	0.523	0.718	0.721	0.621	0.143 (7)	0.358	0.483

Model counterfactual of each contestant Auditor auditing its own model's responses at the mean task, by pool, bracketed by its own repeat probability and the concordance probability.

7 Conclusion

The expansion of machine reasoning into the economy rests on a single presumption, that checking an answer is cheaper than producing it while remaining accurate enough to catch the errors that matter. Delegation with audit rests on it, as do autonomous agents and the promise of self-improving systems. This paper puts a price on that presumption on the output it governs, in a setting built so that checking a delivered answer is a genuine problem rather than a formality. The presumption holds only in part, failing in the place it is most relied upon.

The cheap way for one machine to check another is to use its explanation. The check is cheap because it trusts what it reads. That trust is the flaw. A check that reads a fluent derivation confirms answers already right far better than it catches answers that are wrong, while it is disarmed by the very material that misled the producer. Error in the work and error in its audit then move together, so the cheaper the check, the more its own failures line up with the failures it exists to catch. Cheapness and credulity are not separable properties of machine verification. They are one property seen from two sides.

For a market that meters machine work by the token while it certifies quality by benchmark, the lesson is that the one attribute delegation depends on, verified correctness, goes unpriced exactly as it grows scarce. As the price of producing answers falls faster than the reliability of checking them rises, unverified error accumulates as verification debt, a liability the market does not record. Where the counterfeits are good, scarce ability rather than cheap capacity must do the checking, a hierarchy of trust the economics of machine reasoning will have to price, the guardian problem of Hurwicz (2008) with the guardians metered by the token. The binding constraint on delegating reasoning to machines is not the cost of the reasoning but the cost and the reliability of its verification. Treating verification as the scarce input it is, rather than a free byproduct of cheap computation, is the first requirement of a market that can sell machine-produced knowledge and be believed.

References

Acton, Jan Paul (1975) "Nonmonetary Factors in the Demand for Medical Services: Some Empirical Evidence," *Journal of Political Economy*, 83 (3), 595–614.

Akerlof, George A. (1970) "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 84 (3), 488–500.

Akimitsu, Piyush (2026a) "GERB: Graduate Economic Reasoning Benchmark," Dataset, Zenodo, https://doi.org/10.5281/zenodo.21541072.

——— (2026b) "Wrong and More Confident: A Field Experiment on Large Language Models Taking a Graduate Economics Exam," Preprint, arXiv:2607.23424 [econ.GN], https://arxiv.org/abs/2607.23424.

Baker, Bowen et al. (2025) "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation," arXiv:2503.11926.

Becker, Gary S. (1965) "A Theory of the Allocation of Time," *Economic Journal*, 75 (299), 493–517.

Bergemann, Dirk, Andrew Koh, and Stephen Morris (2026) "Mechanism Design for Alignment and Control," Cowles Foundation Discussion Paper No. 2557, Yale University, and arXiv:2609.01595.

Bowman, Samuel R., Jeeyoon Hyun, Ethan Perez et al. (2022) "Measuring Progress on Scalable Oversight for Large Language Models," arXiv:2211.03540.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond (2025) "Generative AI at Work," *Quarterly Journal of Economics*, 140 (2), 889–942, 10.1093/qje/qjae044.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2011) "Robust Inference with Multiway Clustering," *Journal of Business & Economic Statistics*, 29 (2), 238–249.

Card, David, Stefano DellaVigna, Chenxi Jiang, and Dmitry Taubinsky (2024) "Understanding Expert Choices Using Decision Time," NBER Working Paper 32515, National Bureau of Economic Research, Revised July 2026.

Chan, David C., Matthew Gentzkow, and Chuan Yu (2022) "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *Quarterly Journal of Economics*, 137 (2), 729–783.

Christiano, Paul, Buck Shlegeris, and Dario Amodei (2018) "Supervising Strong Learners by Amplifying Weak Experts," arXiv:1810.08575.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian et al. (2021) "Training Verifiers to Solve Math Word Problems," *arXiv preprint arXiv:2110.14168*, https://arxiv.org/abs/2110.14168.

Cunningham, Tom, Lukas Althoff, Basil Halperin et al. (2026) "The Economics of Recursive Self-Improvement," Working paper, Elasticity Institute, with support from METR.

DeepSeek-AI et al. (2025) "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, 645, 633–638, 10.1038/s41586-025-09422-z.

Demirer, Mert, Andrey Fradkin, and Nadav Tadelis (2026) "The Emerging Market for Intelligence: How Firms Buy and Sell AI," *Journal of Economic Perspectives*, 40 (3), 23–46.

Dogonowski, Robert R. (2026) "Verified Output and the Microeconomics of AI Adoption: Why AI Productivity Depends on Verification, Not Generation Alone," Working paper, 8R Advisory, Copenhagen. Available at SSRN.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan (2018) "The Value of Regulatory Discretion: Estimates from Environmental Inspections in India," *Econometrica*, 86 (6), 2123–2160.

Dulleck, Uwe and Rudolf Kerschbamer (2006) "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic Literature*, 44 (1), 5–42.

Egan, Mark L., Dominic Russel, Claire Shi, and Vincent Yao (2026) "Competition in Generative AI: Evidence from Consumer Demand," NBER Working Paper 35953, National Bureau of Economic Research, https://www.nber.org/papers/w35953.

Feinstein, Jonathan S. (1990) "Detection Controlled Estimation," *Journal of Law and Economics*, 33 (1), 233–276.

Firth, David (1993) "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80 (1), 27–38.

Fudenberg, Drew and Annie Liang (2025) "Friend or Foe: Delegating to an AI Whose Alignment is Unknown," arXiv:2509.14396.

Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg (2006) "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model," *American Economic Review*, 96 (4), 1043–1068.

Garicano, Luis (2000) "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 108 (5), 874–904.

Goldfarb, Avi and Catherine Tucker (2019) "Digital Economics," *Journal of Economic Literature*, 57 (1), 3–43.

Goldsmith-Pinkham, Paul (2026) "Research in the Time of AI," Essay, "A Causal Affair" (Substack), https://paulgp.substack.com/p/research-in-the-time-of-ai, Accessed 2026-08.

Green, David M. and John A. Swets (1966) *Signal Detection Theory and Psychophysics*, New York: John Wiley & Sons.

Green, Jerry R. and Jean-Jacques Laffont (1986) "Partially Verifiable Information and Mechanism Design," *Review of Economic Studies*, 53 (3), 447–456.

Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan, and Fabien Roger (2024) "AI Control: Improving Safety Despite Intentional Subversion," in *Proceedings of the 41st International Conference on Machine Learning*, 235 of Proceedings of Machine Learning Research, 16295–16336.

Hart, Sergiu, Ilan Kremer, and Motty Perry (2017) "Evidence Games: Truth and Commitment," *American Economic Review*, 107 (3), 690–713.

Horton, John J., Apostolos Filippas, and Benjamin S. Manning (2023) "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" NBER Working Paper 31122, National Bureau of Economic Research, 10.3386/w31122.

Hurwicz, Leonid (2008) "But Who Will Guard the Guardians?" *American Economic Review*, 98 (3), 577–585.

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot (2023) "An Adversarial Approach to Structural Estimation," *Econometrica*, 91 (6), 2041–2063, 10.3982/ECTA18707.

Korbak, Tomek et al. (2025) "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety," arXiv:2507.11473.

Kremer, Michael (1993) "The O-Ring Theory of Economic Development," *Quarterly Journal of Economics*, 108 (3), 551–575.

Lacker, Jeffrey M. and John A. Weinberg (1989) "Optimal Contracts under Costly State Falsification," *Journal of Political Economy*, 97 (6), 1345–1363.

List, John A. (2025) "The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs," NBER Working Paper 33456, National Bureau of Economic Research, https://www.nber.org/papers/w33456, Also circulated as Natural Field Experiments No. 00804, The Field Experiments Website.

Macmillan, Neil A. and C. Douglas Creelman (2005) *Detection Theory: A User's Guide*, Mahwah, NJ: Lawrence Erlbaum Associates, 2nd edition.

Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar (2024) "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," *arXiv preprint arXiv:2410.05229*, https://arxiv.org/abs/2410.05229.

Naeini, Saeid Alavi, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati (2023) "Large Language Models are Fixated by Red Herrings: Exploring Creative Problem Solving and Einstellung Effect using the Only Connect Wall Dataset," in *Advances in Neural Information Processing Systems (NeurIPS)*, https://arxiv.org/abs/2306.11167.

Qiu, Tianyi Alex, Micah Carroll, and Cameron Allen (2026) "Truthfulness Despite Weak Supervision: Evaluating and Training LLMs Using Peer Prediction," in *International Conference on Learning Representations*, arXiv:2601.20299.

Rosen, Sherwin (1974) "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82 (1), 34–55.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou (2023) "Large Language Models Can Be Easily Distracted by Irrelevant Context," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, https://arxiv.org/abs/2302.00093.

Townsend, Robert M. (1979) "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, 21 (2), 265–293.

Xiao, Frank and Mary Phuong (2026) "Bootstrapped Monitoring: Leveraging Transparent Reasoning to Oversee Stronger AI Agents," arXiv:2606.11998.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng et al. (2023) "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, https://arxiv.org/abs/2306.05685.

Appendix

A The Audit Protocol

A.1 Verdict parsing and the grade ruling

Table 6: Reduced-form headlines before and after the grade ruling

	Machine grade	Ruled grade
Usable Auditor calls	32,551	31,979
Graded candidate responses	4,190	4,116
Candidate wrong-answer share	0.4247	0.4074
Sound verdict rate, pooled	0.6790	0.6936
Sound verdict rate, Auditor solving pass	0.6714	0.6814
Sound verdict rate, Auditor verifying pass	0.6864	0.7057
Realized discrimination d', Floor Auditor	0.581	0.648
Realized discrimination d', Robust Auditor	0.999	1.087
Realized discrimination d', Anchor Auditor	1.194	1.304
Realized discrimination d', Ceiling Auditor	1.764	1.911
Solve-minus-verify pass gap	+0.0136	+0.0232

Reduced-form headlines under the original machine grade and under the corrected grade ruling.

Table 7: The run at a glance: measurement integrity and the price of a verdict

Auditor	Strict parse	Finish = stop	Mean cost (USD/call)	Mean latency (s/call)	Mean tokens (per call)	Accept P1	Accept P2
Floor Auditor (GPT-5.4 Mini)	100%	100%	0.0013	2.1	1,133	0.62	0.73
Robust Auditor (Claude Sonnet 4)	100%	100%	0.0140	8.2	1,866	0.77	0.85
Anchor Auditor (Grok 4.3)	99%	100%	0.0030	8.3	1,944	0.73	0.86
Ceiling Auditor (DeepSeek V4 Pro)	96%	96%	0.0103	107.0	7,580	0.71	0.71

Parse rates, finish rates, mean price, latency, tokens, and Pass-1 and Pass-2 acceptance shares by Auditor, over all 33,520 calls.

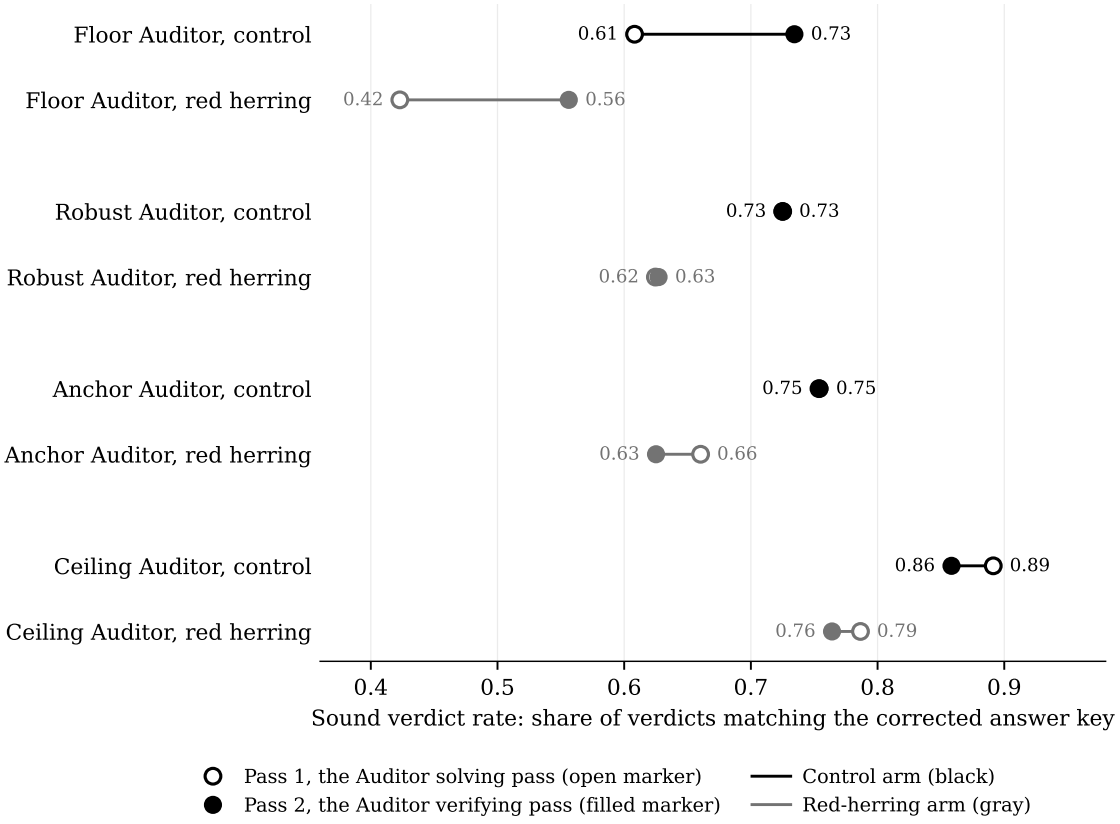
B Reduced-Form Exhibits

Table 8: The confusion economy: the sound verdict rate and error rates by pass

	Sound verdict rate		False accept		False reject	
	Pass 1	Pass 2	Pass 1	Pass 2	Pass 1	Pass 2
All	0.681	0.706	0.532	0.604	0.175	0.078
Control arm	0.742	0.768	0.507	0.559	0.136	0.065
Red-herring arm	0.620	0.643	0.549	0.635	0.224	0.095
Floor Auditor (GPT-5.4 Mini)	0.516	0.646	0.628	0.613	0.384	0.175
Robust Auditor (Claude Sonnet 4)	0.675	0.676	0.614	0.716	0.122	0.048
Anchor Auditor (Grok 4.3)	0.707	0.690	0.524	0.715	0.130	0.031
Ceiling Auditor (DeepSeek V4 Pro)	0.839	0.812	0.330	0.374	0.059	0.056

Sound verdict rate, false-accept rate, and false-reject rate by Auditor, arm, and pass, on the usable calls.

Figure 2: The sound verdict rate by Auditor, by arm, by pass



Sound verdict rate by Auditor and arm, Pass 1 against Pass 2.

Table 9: The within-response pass contrast and acceptance

	Δ_{ij} (Pass 2 − Pass 1, sound verdict rate)			Accept_{ijp}	
	(1)	(2)	(3)	(4)	(5)
Constant / β_0	+0.023 (0.013)*	+0.052 (0.018)***	+0.035 (0.017)**	.	.
Wrong_i (β_1/α_2)	.	-0.272 (0.038)***	-0.242 (0.038)***	-0.286 (0.053)***	-0.350 (0.058)***
RH_i (β_2)	.	.	+0.065 (0.016)***	.	-0.089 (0.020)***
$\text{Wrong}_i \times \text{RH}_i$ (β_3)	.	.	-0.071 (0.020)***	.	+0.130 (0.032)***
P_p (α_1)	.	.	.	+0.097 (0.019)***	+0.071 (0.018)***
$P_p \times \text{Wrong}_i$ (α_3)	.	.	.	-0.022 (0.026)	-0.015 (0.025)
$P_p \times \text{RH}_i$	.	.	.	.	+0.058 (0.016)***
$P_p \times \text{Wrong}_i \times \text{RH}_i$	.	.	.	.	-0.026 (0.024)
Auditor FE	No	Yes	Yes	Yes	Yes
Task FE	No	Yes	Yes	Yes	Yes
N	15,797	15,797	15,797	31,979	31,979

Within-response Pass-2-minus-Pass-1 verdict contrast and acceptance, by candidate state and arm.

Table 10: The pass contrast by Auditor

		mean Δ_{ij}				
Auditor	N	all	correct	wrong	β_1 (Wrong)	β_3 (Wrong × RH)
Floor Auditor	4,046	+0.130 (0.022)***	+0.209 (0.036)***	+0.015 (0.032)	-0.159 (0.066)**	-0.176 (0.057)***
Robust Auditor	3,998	+0.001 (0.019)	+0.074 (0.025)***	-0.102 (0.037)***	-0.289 (0.060)***	+0.009 (0.034)
Anchor Auditor	4,032	-0.016 (0.022)	+0.100 (0.025)***	-0.183 (0.037)***	-0.429 (0.049)***	-0.053 (0.030)*
Ceiling Auditor	3,721	-0.027 (0.016)*	+0.003 (0.015)	-0.077 (0.034)**	-0.082 (0.042)**	-0.064 (0.045)

The within-response pass contrast estimated separately for each Auditor.

Table 11: The price of checking: Auditor resources, Pass 2 against Pass 1

	Mean per call			φ_1 by Auditor			
Resource	Pass 1	Pass 2	φ_1 (pooled)	Floor Auditor	Robust Auditor	Anchor Auditor	Ceiling Auditor
Completion tokens	2191.0	1364.8	-0.580 (0.035)***	-0.77***	-0.21***	-0.70***	-0.70***
Reasoning tokens	1855.4	1156.6	-0.624 (0.036)***	.	.	-0.59***	-0.71***
Total tokens (input + output)	2688.6	2998.3	+0.368 (0.035)***	+0.86***	+0.55***	+0.20***	-0.19***
Money price (USD)	0.0069	0.0066	+0.041 (0.033)	+0.35***	+0.10***	+0.04	-0.38***
Time price (s)	33.3	21.0	-0.420 (0.023)***	-0.29***	-0.16***	-0.65***	-0.63***

Log money, time, and token price of a call, Pass 2 against Pass 1, with Auditor and task fixed effects.

Table 12: The price account of a verdict: money prices and time prices

	Price per call		Verdict	Price per sound verdict		Break-even
	Money (USD)	Time (s)	accuracy	Money (USD)	Time (s)	w (USD/hr)
Panel A. The production input, the 4,116 audited responses of the 38 agents						
Agent production	0.0176	60.0	0.587[a]	0.0300[a]	102.3[a]	1.05
Panel B. The verification input, by Auditor and pass						
Floor Auditor (GPT-5.4 Mini)						
Pass 1	0.0011	2.4	0.516	0.0022	4.7	1.67
Pass 2	0.0015	1.7	0.646	0.0023	2.6	3.27
both passes	0.0026	4.1	.	.	.	2.32
Robust Auditor (Claude Sonnet 4)						
Pass 1	0.0132	8.9	0.675	0.0195	13.1	5.35
Pass 2	0.0147	7.3	0.676	0.0217	10.9	7.18
both passes	0.0278	16.2	.	.	.	6.18
Anchor Auditor (Grok 4.3)						
Pass 1	0.0029	11.0	0.707	0.0042	15.5	0.96
Pass 2	0.0030	5.5	0.690	0.0043	7.9	1.95
both passes	0.0059	16.4	.	.	.	1.29
Ceiling Auditor (DeepSeek V4 Pro)						
Pass 1	0.0108	117.7	0.839	0.0128	140.3	0.33
Pass 2	0.0073	69.9	0.812	0.0090	86.2	0.37
both passes	0.0181	187.6	.	.	.	0.35

Mean money price and time price of a verdict, by Auditor and pass.

Table 13: The input shares of verified knowledge

	Money bill (USD/response)			Time bill (s/response)		
Auditor, audit policy	Verification input	Total bill	Verification share (%)	Verification input	Total bill	Verification share (%)
Floor Auditor (GPT-5.4 Mini)						
Pass 1	0.0011	0.0187	6.0	2.4	62.4	3.9
Pass 2	0.0015	0.0191	7.9	1.7	61.7	2.7
both passes	0.0026	0.0202	13.0	4.1	64.1	6.3
Robust Auditor (Claude Sonnet 4)						
Pass 1	0.0132	0.0307	42.8	8.9	68.9	12.8
Pass 2	0.0147	0.0322	45.5	7.3	67.4	10.9
both passes	0.0278	0.0454	61.3	16.2	76.2	21.3
Anchor Auditor (Grok 4.3)						
Pass 1	0.0029	0.0205	14.3	11.0	71.0	15.5
Pass 2	0.0030	0.0205	14.4	5.5	65.5	8.4
both passes	0.0059	0.0235	25.1	16.4	76.5	21.5
Ceiling Auditor (DeepSeek V4 Pro)						
Pass 1	0.0108	0.0284	38.0	117.7	177.7	66.2
Pass 2	0.0073	0.0249	29.3	69.9	129.9	53.8
both passes	0.0181	0.0356	50.7	187.6	247.7	75.8

Production and verification factor bills per audited response, the input shares of verified knowledge.

Table 14: The price schedule of verification

	log money price		log time price	
	(1)	(2)	(3)	(4)
P_p (Pass 2, the Auditor verifying pass)	-0.1524***	-0.2143***	-0.3416***	-0.3569***
	(0.0363)	(0.0796)	(0.0222)	(0.0566)
	[0.000]	[0.007]	[0.000]	[0.000]
Wrong_i (final answer wrong)	+0.1209***	+0.0783**	+0.1622***	+0.1575***
	(0.0225)	(0.0368)	(0.0244)	(0.0328)
	[0.000]	[0.033]	[0.000]	[0.000]
RH_i (red-herring arm)	+0.0172	+0.0263**	+0.0160	+0.0185
	(0.0116)	(0.0122)	(0.0140)	(0.0133)
	[0.137]	[0.031]	[0.254]	[0.164]
$\log L_i$ (explanation length)	+0.0593***	+0.2142***	+0.0333***	+0.1799***
	(0.0146)	(0.0439)	(0.0115)	(0.0366)
	[0.000]	[0.000]	[0.004]	[0.000]
$\log L_i \times P_p$ (explanation-length gradient)	+0.2262***	+0.2373***	-0.0924***	-0.0880***
	(0.0356)	(0.0350)	(0.0177)	(0.0160)
	[0.000]	[0.000]	[0.000]	[0.000]
s_t (task clean solve rate)	.	-0.5473***	.	-0.4385***
		(0.1454)		(0.1197)
		[0.000]		[0.000]
$s_t \times P_p$	.	+0.0925	.	+0.0228
		(0.1041)		(0.0755)
		[0.374]		[0.763]
Auditor FE	Yes	Yes	Yes	Yes
Task FE	Yes	No	Yes	No
N	31,979	31,979	31,979	31,979

Log money price and time price of a call on task difficulty and the red-herring arm, pooled across Auditors.

Table 15: The price schedule of verification, by Auditor

	Floor Auditor (GPT-5.4 Mini)	Robust Auditor (Claude Sonnet 4)	Anchor Auditor (Grok 4.3)	Ceiling Auditor (DeepSeek V4 Pro)
Panel A. The log money price: what the Auditor's call bills				
P_p (Pass 2, the Auditor verifying pass)	+0.0053	-0.0819***	-0.1026***	-0.4490***
	(0.0775)	(0.0250)	(0.0394)	(0.0743)
	[0.945]	[0.001]	[0.009]	[0.000]
Wrong_i (final answer wrong)	-0.0693***	-0.0092	+0.0889***	+0.5167***
	(0.0193)	(0.0120)	(0.0209)	(0.0651)
	[0.000]	[0.441]	[0.000]	[0.000]
RH_i (red-herring arm)	-0.0138	+0.0129**	+0.0368***	+0.0288
	(0.0205)	(0.0061)	(0.0121)	(0.0365)
	[0.500]	[0.036]	[0.002]	[0.429]
$\log L_i$ (explanation length)	+0.0786**	+0.0524***	+0.0669***	+0.0734**
	(0.0316)	(0.0094)	(0.0155)	(0.0295)
	[0.013]	[0.000]	[0.000]	[0.013]
$\log L_i \times P_p$ (explanation-length gradient)	+0.3927***	+0.2058***	+0.1792***	+0.0642
	(0.0659)	(0.0231)	(0.0384)	(0.0506)
	[0.000]	[0.000]	[0.000]	[0.205]
Panel B. The log time price: how long the Auditor's call takes				
P_p (Pass 2, the Auditor verifying pass)	-0.2421***	-0.1598***	-0.3795***	-0.5727***
	(0.0491)	(0.0192)	(0.0339)	(0.0527)
	[0.000]	[0.000]	[0.000]	[0.000]
Wrong_i (final answer wrong)	-0.0526**	+0.0135	+0.1195***	+0.5974***
	(0.0222)	(0.0114)	(0.0236)	(0.0699)
	[0.018]	[0.237]	[0.000]	[0.000]
RH_i (red-herring arm)	-0.0320*	+0.0140**	+0.0130	+0.0652
	(0.0185)	(0.0071)	(0.0176)	(0.0443)
	[0.084]	[0.049]	[0.460]	[0.141]
$\log L_i$ (explanation length)	+0.0468**	+0.0391***	+0.0472***	+0.0285
	(0.0216)	(0.0096)	(0.0163)	(0.0287)
	[0.030]	[0.000]	[0.004]	[0.321]
$\log L_i \times P_p$ (explanation-length gradient)	-0.0567*	-0.0031	-0.3218***	-0.0807*
	(0.0337)	(0.0148)	(0.0333)	(0.0421)
	[0.093]	[0.836]	[0.000]	[0.055]
Task FE	Yes	Yes	Yes	Yes
N, Panel A	8,092	7,996	8,148	7,743
N, Panel B	8,092	7,996	8,148	7,743

The price schedule of a call estimated separately for each Auditor.

Table 16: The unit price of a sound verdict

	log money unit price		log time unit price	
	(1)	(2)	(3)	(4)
P_p (Pass 2, the Auditor verifying pass)	-0.1357***	-0.3223**	-0.5771***	-0.8448***
	(0.0493)	(0.1355)	(0.0460)	(0.1189)
	[0.006]	[0.017]	[0.000]	[0.000]
s_t (task clean solve rate)	.	-2.3076***	.	-2.1862***
		(0.2971)		(0.3015)
		[0.000]		[0.000]
$s_t \times P_p$	.	+0.3171*	.	+0.4547***
		(0.1781)		(0.1604)
		[0.075]		[0.005]
Auditor FE	Yes	Yes	Yes	Yes
Task FE	Yes	No	Yes	No
Cells	472	472	472	472
Calls aggregated	31,979	31,979	31,979	31,979

Log unit price of one sound verdict, per Auditor-task-pass cell.

Table 17: Does the agent's metacognition predict which wrong explanations fool a verifier?

Agent trait x_{mt} (standardized)	$\hat{\theta}_1$	SE	N
Overconfidence	+0.0071	(0.0117)	6,058
Metacognition score	+0.0231	(0.0142)	6,058
Self-awareness score	+0.0195	(0.0130)	6,058
Persistence score	+0.0221	(0.0141)	6,058
Corruption	-0.0032	(0.0095)	6,058
Recovery	+0.0096	(0.0090)	6,058

Whether the agent's metacognition traits predict false accepts in the Auditor verifying pass, one trait at a time.

Table 18: The falsification margin: which wrong candidates pass audit

	Accept_{ij} (successful falsification, Pass 2, wrong candidates)			
	(1)	(2)	(3)	(4)
RH_i (red-herring arm)	+0.0816***	+0.0768***	+0.0651***	+0.0646***
	(0.0232)	(0.0216)	(0.0211)	(0.0210)
	[0.000]	[0.000]	[0.002]	[0.002]
$\log L_i$ (worked-solution length)	.	-0.0705***	-0.0614***	-0.0542***
		(0.0196)	(0.0199)	(0.0194)
		[0.000]	[0.002]	[0.005]
$\bar{a}^{bl}_{m(i)t(i)}$ (agent clean accuracy)	.	.	+0.0772**	+0.0585
			(0.0331)	(0.0367)
			[0.020]	[0.111]
$D^{bl}_{m(i)t(i)}$ (agent difficulty rating)	.	.	-0.0151**	-0.0068
			(0.0072)	(0.0061)
			[0.036]	[0.266]
$\text{Reasoning}_{m(i)}$	.	.	.	+0.0976***
				(0.0351)
				[0.005]
Auditor FE	Yes	Yes	Yes	Yes
Task FE	Yes	Yes	Yes	Yes
N	6,624	6,624	6,002	6,002

Which wrong candidates are accepted in the Auditor verifying pass, on response and agent covariates.

Table 19: The confirmation margin: which correct candidates clear audit

	Accept_{ij} (correct acceptance: Pass 2, correct candidates)			
	(1)	(2)	(3)	(4)
RH_i (red-herring arm)	-0.0314***	-0.0296***	-0.0316***	-0.0322***
	(0.0093)	(0.0094)	(0.0096)	(0.0097)
	[0.001]	[0.002]	[0.001]	[0.001]
$\log L_i$ (worked-solution length)	.	-0.0426***	-0.0331**	-0.0316**
		(0.0156)	(0.0150)	(0.0155)
		[0.006]	[0.027]	[0.042]
$\bar{a}^{bl}_{m(i)t(i)}$ (agent clean accuracy)	.	.	+0.1013***	+0.0949***
			(0.0221)	(0.0219)
			[0.000]	[0.000]
$D^{bl}_{m(i)t(i)}$ (agent difficulty rating)	.	.	-0.0011	+0.0012
			(0.0029)	(0.0038)
			[0.708]	[0.752]
$\text{Reasoning}_{m(i)}$	.	.	.	+0.0205
				(0.0155)
				[0.187]
Auditor FE	Yes	Yes	Yes	Yes
Task FE	Yes	Yes	Yes	Yes
N	9,463	9,463	9,106	9,106

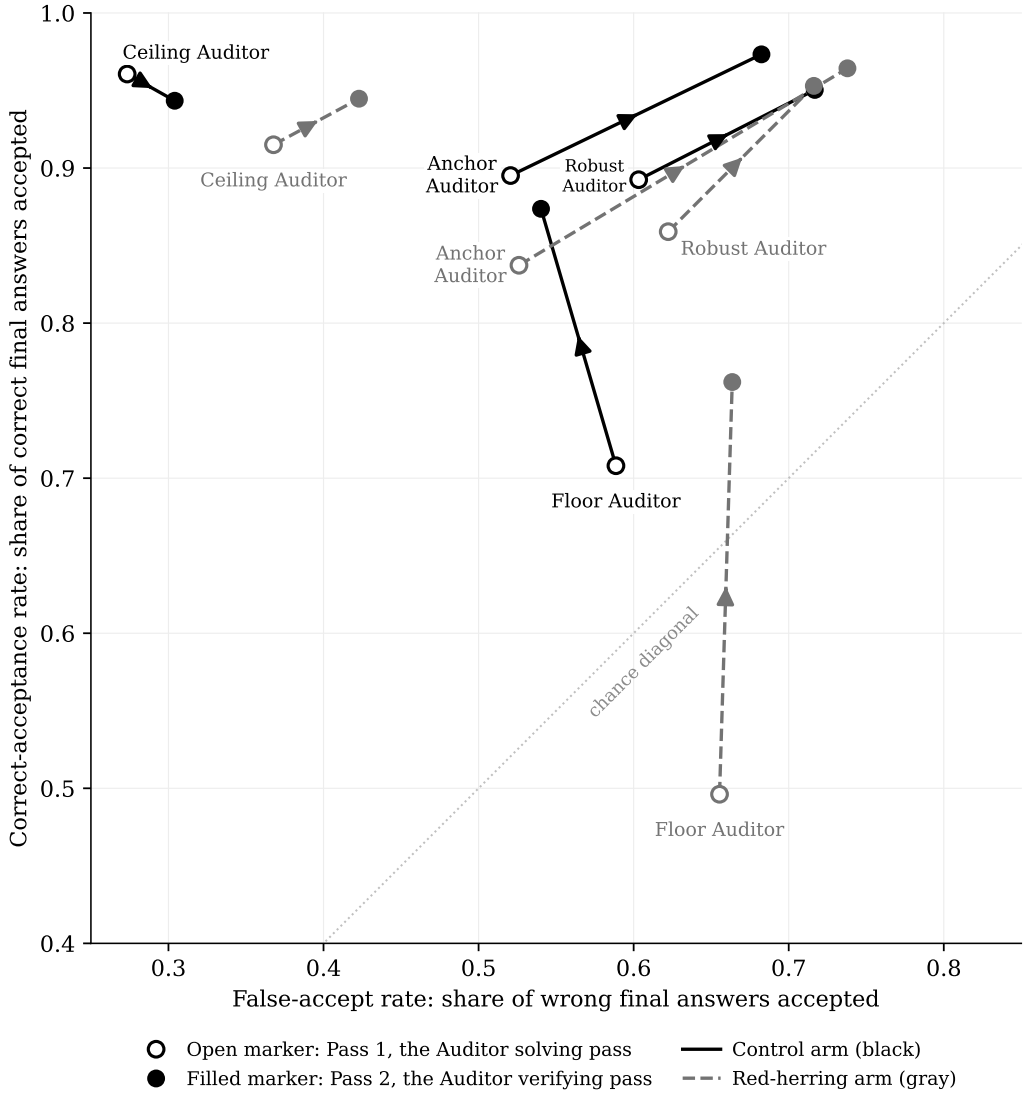
Which correct candidates are accepted in the Auditor verifying pass, the twin of the falsification margin.

Table 20: The two channels of falsification: answer versus explanation

	Accept_{ijp} (wrong candidates)			
	(1) Pass 1	(2) Pass 2	(3) Stacked	(4) Stacked, full FE
RH_i (red-herring arm)	+0.0691***	+0.0646***	+0.0566***	+0.0693***
	(0.0216)	(0.0210)	(0.0213)	(0.0216)
	[0.001]	[0.002]	[0.008]	[0.001]
$\log L_i$ (explanation length)	-0.0257*	-0.0542***	-0.0144	-0.0253*
	(0.0134)	(0.0194)	(0.0156)	(0.0137)
	[0.055]	[0.005]	[0.355]	[0.064]
$\text{Reasoning}_{m(i)}$	+0.0529***	+0.0976***	+0.0573***	+0.0522***
	(0.0189)	(0.0351)	(0.0185)	(0.0184)
	[0.005]	[0.005]	[0.002]	[0.005]
P_p (Pass 2, full response)	.	.	+0.1121***	+0.2561***
			(0.0207)	(0.0326)
			[0.000]	[0.000]
$\text{RH}_i \times P_p$	.	.	+0.0196	-0.0050
			(0.0142)	(0.0125)
			[0.167]	[0.692]
$\log L_i \times P_p$	.	.	-0.0541***	-0.0293**
			(0.0126)	(0.0116)
			[0.000]	[0.012]
$\text{Reasoning}_{m(i)} \times P_p$	.	.	+0.0352	+0.0461*
			(0.0242)	(0.0251)
			[0.146]	[0.066]
$\bar{a}^{bl}_{m(i)t(i)}$ (agent clean accuracy)	+0.0639**	+0.0585	+0.0613**	+0.0612**
	(0.0252)	(0.0367)	(0.0284)	(0.0286)
	[0.011]	[0.111]	[0.031]	[0.033]
$D^{bl}_{m(i)t(i)}$ (agent difficulty rating)	-0.0050*	-0.0068	-0.0059	-0.0059
	(0.0030)	(0.0061)	(0.0041)	(0.0041)
	[0.096]	[0.266]	[0.151]	[0.150]
Auditor FE	Yes	Yes	Yes	Yes
Task FE	Yes	Yes	Yes	Yes
Task × pass FE	.	.	No	Yes
Auditor × pass FE	.	.	No	Yes
N	5,802	6,002	11,804	11,804

Acceptance of wrong answers, each covariate split into its final-answer and derivation channels across passes.

Figure 3: The audit technology in ROC space



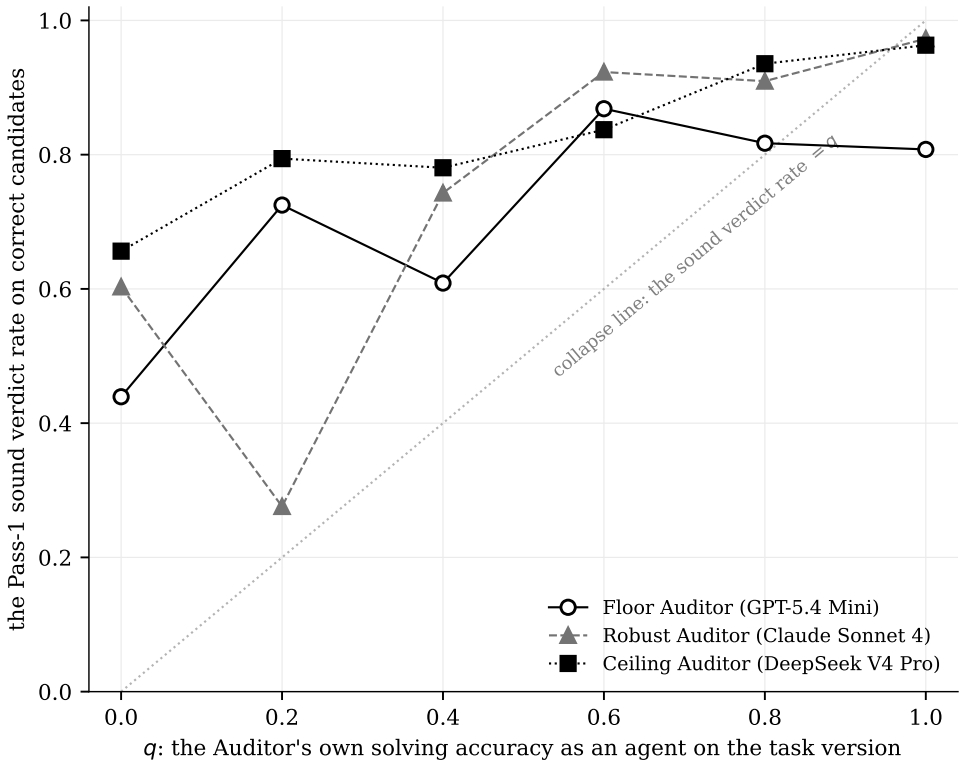
Each Auditor's operating point in ROC space, Pass 1 to Pass 2, by arm.

Table 21: Realized discrimination and criterion: the signal-detection decomposition

		ROC, Pass 1		ROC, Pass 2		Realized discrimination d'		Criterion c	
Auditor	Arm	FAR	CAR	FAR	CAR	Pass 1 → Pass 2	$\Delta d'$	Pass 1 → Pass 2	Δc
Floor Auditor	Control	0.589	0.708	0.540	0.874	+0.32 → +1.04	+0.72	-0.39 → -0.62	-0.24
	Red herring	0.655	0.496	0.664	0.762	-0.41 → +0.29	+0.70	-0.20 → -0.57	-0.37
Robust Auditor	Control	0.603	0.892	0.717	0.950	+0.98 → +1.08	+0.10	-0.75 → -1.11	-0.36
	Red herring	0.622	0.859	0.716	0.953	+0.76 → +1.10	+0.34	-0.69 → -1.12	-0.43
Anchor Auditor	Control	0.521	0.895	0.682	0.973	+1.20 → +1.46	+0.25	-0.65 → -1.20	-0.55
	Red herring	0.526	0.837	0.738	0.964	+0.92 → +1.17	+0.25	-0.52 → -1.22	-0.70
Ceiling Auditor	Control	0.273	0.961	0.304	0.943	+2.36 → +2.10	-0.26	-0.58 → -0.54	+0.04
	Red herring	0.368	0.915	0.423	0.945	+1.71 → +1.79	+0.08	-0.52 → -0.70	-0.18

Realized discrimination d' and criterion by Auditor and arm.

Figure 4: The collapse test: auditing a bare answer against one's own solving record



Pass-1 acceptance of correct bare answers against the Auditor's own solving accuracy on the task.

Table 22: The collapse test: the Pass-1 sound verdict rate on the Auditor's own solving record

	Correct candidates		Wrong candidates	
	(1)	(2)	(3)	(4)
φ_0 (constant)	+0.5173***	+0.6471***	+0.3324***	+0.5040***
	(0.0368)	(0.0522)	(0.0440)	(0.0693)
	[0.000]	[0.000]	[0.000]	[0.000]
φ_1 (q, own solving accuracy)	+0.4424***	+0.3579***	+0.3274***	+0.2456***
	(0.0446)	(0.0587)	(0.0539)	(0.0590)
	[0.000]	[0.000]	[0.000]	[0.000]
RH_i (red-herring arm)	.	-0.0683***	.	-0.0332
		(0.0211)		(0.0269)
		[0.001]		[0.218]
Auditor FE	No	Yes	No	Yes
Wald $(\varphi_0, \varphi_1) = (0, 1)$, p	[0.000]	[0.000]	.	.
N	7,020	7,020	4,630	4,630

Pass-1 sound verdict rate on the Auditor's own solving share, for the three Auditors that also solved as agents.

Table 23: The collapse test in both passes: the slope of the sound verdict rate on the Auditor's own solving share

	Correct candidates		Wrong candidates	
Specification	Pass 1	Pass 2	Pass 1	Pass 2
(2) q, red herring, Auditor FE	+0.358 (0.059)	+0.089 (0.022)	+0.246 (0.059)	+0.162 (0.056)
	[0.000]	[0.000]	[0.000]	[0.004]
(6) q, task-version FE, Auditor FE	+0.216 (0.061)	+0.020 (0.024)	+0.256 (0.068)	+0.146 (0.051)
	[0.000]	[0.412]	[0.000]	[0.004]
N	7,020	7,024	4,630	4,843

Slope of the sound verdict rate on the Auditor's own solving share, both passes, by candidate state.

Table 24: The clock test: does the cost meter read the verdict?

	Pooled (all Auditors)	Floor Auditor (GPT-5.4 Mini)	Robust Auditor (Claude Sonnet 4)	Anchor Auditor (Grok 4.3)	Ceiling Auditor (DeepSeek V4 Pro)
Panel A. Cost meter = the clock, log latency in seconds					
$\log y_i$ (cost meter)	+0.0280**	+0.0481***	+0.0587	-0.1220***	+0.0483**
	(0.0122)	(0.0146)	(0.0472)	(0.0359)	(0.0192)
	[0.022]	[0.001]	[0.214]	[0.001]	[0.012]
$\log L_i$ (worked-solution length)	-0.0047	-0.0192	-0.0040	-0.0403	+0.0106
	(0.0239)	(0.0259)	(0.0247)	(0.0297)	(0.0201)
	[0.844]	[0.458]	[0.873]	[0.175]	[0.598]
Panel B. Cost meter = the completion meter, log completion tokens					
$\log y_i$ (cost meter)	+0.0346***	+0.0347**	+0.4020***	-0.0986***	+0.0538***
	(0.0092)	(0.0135)	(0.0938)	(0.0360)	(0.0178)
	[0.000]	[0.010]	[0.000]	[0.006]	[0.003]
$\log L_i$ (worked-solution length)	-0.0025	-0.0140	-0.0128	-0.0420	+0.0119
	(0.0238)	(0.0262)	(0.0227)	(0.0305)	(0.0200)
	[0.918]	[0.593]	[0.572]	[0.169]	[0.552]
Auditor FE	Yes	No	No	No	No
Task version FE	Yes	Yes	Yes	Yes	Yes
N	12,725	2,986	3,417	3,490	2,832

Whether the cost meter predicts the accepted verdict's reliability, on the accepted Pass-2 stream.

Table 25: The Floor Auditor's Pass-2 gain by call length

Panel A. The Floor Auditor's Pass-2 calls at fifty completion tokens, by candidate state

	Pass-2 gain Δ_{ij}	the Pass-2 sound verdict rate	Pass-2 acceptance	n
Long call, correct candidate	+0.197	0.780	0.780	1,422
Long call, wrong candidate	+0.059	0.438	0.562	1,052
Short call, correct candidate	+0.228	0.891	0.891	966
Short call, wrong candidate	−0.061	0.299	0.701	606

Panel B. The gain and the acceptance test at three thresholds, task fixed effects

Short call	Share short	Gain, long and short	Short − long gain, (a)	Extra acceptance, (b)
Under 50 tokens, Pass-2 call	0.389	+0.138 and +0.116	−0.063 (0.032)	+0.177 (0.034)
			[0.046]	[0.000]
short × Wrong, (b)				−0.037 (0.065)
				[0.565]
Under 30 tokens, Pass-2 call	0.376	+0.137 and +0.118	−0.057 (0.033)	+0.164 (0.034)
			[0.082]	[0.000]
short × Wrong, (b)				−0.058 (0.064)
				[0.364]
Under 100 tokens, Pass-2 call	0.706	+0.137 and +0.127	−0.025 (0.038)	+0.287 (0.046)
			[0.507]	[0.000]
short × Wrong, (b)				+0.088 (0.050)
				[0.082]
Under 50 tokens, Pass-1 call	0.303	+0.131 and +0.127	+0.107 (0.071)	
			[0.134]	

The Floor Auditor's Pass-2 verdict gain, split by whether the verifying call was short.

C The Model Appendix

Table 26: Model notation

Symbol	Object	Defined
m, j, t	agent, Auditor, task	Sec. 4.1
r, e, p	arm, explanation requirement, pass	Sec. 4.1
i, ω_i	candidate response, its true state	Sec. 4.1
a_m, a_A	agent ability, the Anchor Auditor's ability	Eq. (1)
d_t, h_t, g	task difficulty, red-herring load, explanation shift	Eq. (1)
$q_m(t, r, e)$	solving probability	Eq. (1)
π_{tr}	pool share of correct candidates	Eq. (2)
ξ, ε	audit signal, its noise	Eq. (3)
ψ, ψ_0, $\psi_{j2}(t, r)$	discrimination, its form without solving and its Pass 2 form	Eqs. (3), (11)
c_2, λ, κ	Pass 2 level, capability loading, difficulty loading	Eq. (11)
δ^C_{ip}	correct-side displacement	Eqs. (3), (4)
ρ_j, ρ_R, ρ_L	displacement loadings, ρ the Floor Auditor's base	Eq. (4)
δ^W_{ip}	wrong-side displacement	Eqs. (3), (5)
ρ_{Wj}	wrong-side loading of Auditor j	Eq. (5)
$\tilde{L}_i$, Reasoning$_m$	derivation length, reasoning indicator	Eq. (4)
ϑ_j	re-solving probability of the Auditor solving pass	Ass. 4
χ_{tr}	wrong-answer concordance probability	App. C
ℓ_j, ℓ_0	loss ratio of Auditor j, loss of the principal	Ass. 2, Sec. 6.1
μ_j	form premium	Ass. 3
t^y_{jp}, α_y, γ_y, σ_y	cost meter levels with loadings	Ass. 5
ζ, ν, η_y	engagement draw, learning dividend, engagement loads	Ass. 5
$\mathcal{D}(z)$	expected distance to the criterion	Prop. 3
$\xi^*_{jp}(t, r)$	chosen cutoff	Eq. (7)
C, F	correct-acceptance rate, false-accept rate	Eq. (8)
d', c	realized discrimination, measured criterion	Lem. 1
A, R	clearance rate, delivered reliability	Sec. 6.1
Π_s, Π_{jp}	total price of a draw, of a verdict	Sec. 6.1
p_{jp}, τ_{jp}, w	money price, time price, wage	Sec. 6.1

Λ is the logistic function, Φ the standard normal distribution function, ϕ its density.

Table 27: The concordance measure χ_{tr}: how concentrated are the wrong answers on a task version?

	Tasks	P(wrong) mean	Distinct wrong answers, mean	χ_{tr} mean	Range of χ_{tr} min to max
Panel A. By task type (59 tasks, both arms, 118 task versions)					
Categorical (a fixed vocabulary of answers)	27	0.484	2.8	0.788	0.347 to 1.000
Numeric (one number)	21	0.378	20.0	0.220	0.029 to 0.705
Numeric tuple (several numbers)	6	0.473	17.0	0.276	0.095 to 1.000
Symbolic (an expression)	5	0.933	73.1	0.108	0.019 to 0.335
Panel B. By arm (59 task versions each)					
Control arm	59	0.413		0.430	
Red-herring arm	59	0.554		0.522	
Within-task difference, red herring minus control	59			+0.092	paired $t = 3.64$, $p = 0.001$
Panel C. Does χ_{tr} depend on how often the task is failed?					
Slope of χ_{tr} on P(wrong), no controls	118			+0.035 (0.148)	$p = 0.812$, $R^2 = 0.001$
Slope with arm and task-type controls	118			+0.064 (0.111)	$p = 0.560$, $R^2 = 0.687$

The wrong-answer concordance χ_{tr} by answer format, by arm, and against the task failure rate.

Table 28: Delivered reliability and outlay per delivered answer: single-Auditor audits and the no-Auditor benchmarks

	Control pool		Red-herring pool	
Policy	Reliability R	Outlay (USD)	Reliability R	Outlay (USD)
Single Auditor, one pass				
Floor Auditor, Pass 1	0.659 [0.555, 0.762]	0.0305 [0.0273, 0.0348]	0.467 [0.351, 0.577]	0.0320 [0.0286, 0.0366]
Floor Auditor, Pass 2	0.758 [0.670, 0.838]	0.0257 [0.0238, 0.0279]	0.535 [0.421, 0.634]	0.0267 [0.0248, 0.0292]
Robust Auditor, Pass 1	0.711 [0.604, 0.814]	0.0391 [0.0366, 0.0419]	0.550 [0.439, 0.665]	0.0422 [0.0393, 0.0459]
Robust Auditor, Pass 2	0.716 [0.624, 0.800]	0.0375 [0.0357, 0.0396]	0.571 [0.467, 0.661]	0.0387 [0.0365, 0.0415]
Anchor Auditor, Pass 1	0.763 [0.665, 0.852]	0.0276 [0.0256, 0.0306]	0.602 [0.477, 0.711]	0.0303 [0.0281, 0.0334]
Anchor Auditor, Pass 2	0.716 [0.624, 0.796]	0.0237 [0.0226, 0.0250]	0.568 [0.466, 0.654]	0.0242 [0.0229, 0.0257]
Ceiling Auditor, Pass 1	0.844 [0.765, 0.903]	0.0394 [0.0358, 0.0440]	0.722 [0.611, 0.811]	0.0451 [0.0401, 0.0510]
Ceiling Auditor, Pass 2	0.833 [0.763, 0.890]	0.0346 [0.0314, 0.0391]	0.689 [0.589, 0.769]	0.0364 [0.0327, 0.0409]
No Auditor				
Trust, one unaudited draw	0.638 [0.543, 0.728]	0.0176 [0.0176, 0.0176]	0.499 [0.404, 0.583]	0.0176 [0.0176, 0.0176]
Consensus of two draws	0.843 [0.739, 0.912]	0.0608 [0.0528, 0.0695]	0.658 [0.509, 0.772]	0.0699 [0.0615, 0.0795]

Delivered reliability and outlay per delivered answer for each single-Auditor policy and the trust and consensus benchmarks, by pool.

Table 29: Delivered reliability and outlay per delivered answer: policies that combine verdicts

	Control pool		Red-herring pool	
Policy	Reliability R	Outlay (USD)	Reliability R	Outlay (USD)
Single Auditor, both passes, unanimous acceptance				
Floor Auditor	0.750	0.0409	0.503	0.0446
	[0.651, 0.842]	[0.0356, 0.0482]	[0.364, 0.629]	[0.0383, 0.0533]
Robust Auditor	0.764	0.0639	0.604	0.0710
	[0.663, 0.853]	[0.0585, 0.0703]	[0.488, 0.718]	[0.0642, 0.0793]
Anchor Auditor	0.804	0.0340	0.646	0.0383
	[0.707, 0.886]	[0.0310, 0.0382]	[0.516, 0.752]	[0.0347, 0.0429]
Ceiling Auditor	0.916	0.0569	0.822	0.0677
	[0.867, 0.956]	[0.0503, 0.0662]	[0.728, 0.890]	[0.0585, 0.0790]
Two Auditors, Pass 2, unanimous acceptance				
Floor and Robust	0.790	0.0486	0.582	0.0528
	[0.705, 0.863]	[0.0447, 0.0536]	[0.465, 0.685]	[0.0482, 0.0589]
Floor and Anchor	0.791	0.0316	0.579	0.0341
	[0.705, 0.862]	[0.0290, 0.0350]	[0.463, 0.681]	[0.0312, 0.0380]
Floor and Ceiling	0.861	0.0423	0.680	0.0485
	[0.794, 0.914]	[0.0377, 0.0488]	[0.567, 0.775]	[0.0425, 0.0564]
Robust and Anchor	0.759	0.0445	0.612	0.0466
	[0.670, 0.838]	[0.0416, 0.0479]	[0.505, 0.702]	[0.0431, 0.0513]
Robust and Ceiling	0.850	0.0578	0.713	0.0623
	[0.782, 0.903]	[0.0522, 0.0659]	[0.614, 0.794]	[0.0552, 0.0714]
Anchor and Ceiling	0.851	0.0404	0.712	0.0433
	[0.783, 0.904]	[0.0365, 0.0461]	[0.611, 0.794]	[0.0385, 0.0496]
Four Auditors, Pass 2				
Unanimous acceptance	0.880	0.0747	0.716	0.0886
	[0.815, 0.935]	[0.0660, 0.0878]	[0.598, 0.816]	[0.0765, 0.1061]
Majority of three	0.794	0.0573	0.624	0.0589
	[0.714, 0.860]	[0.0531, 0.0626]	[0.521, 0.712]	[0.0541, 0.0646]

Delivered reliability and outlay per delivered answer for the policies that combine verdicts, by pool.

Table 30: Model and measured operating points of every audit policy

	Control pool				Red-herring pool			
	Correct delivered C		Wrong delivered F		Correct delivered C		Wrong delivered F	
Policy	Model	Measured	Model	Measured	Model	Measured	Model	Measured
Single Auditor, one pass								
Floor Auditor, Pass 1	0.632	0.708	0.575	0.589	0.545	0.496	0.622	0.655
Floor Auditor, Pass 2	0.882	0.874	0.497	0.540	0.765	0.762	0.665	0.664
Robust Auditor, Pass 1	0.877	0.892	0.629	0.603	0.801	0.859	0.654	0.622
Robust Auditor, Pass 2	0.965	0.950	0.675	0.717	0.954	0.953	0.715	0.716
Anchor Auditor, Pass 1	0.891	0.895	0.486	0.521	0.815	0.837	0.537	0.526
Anchor Auditor, Pass 2	0.973	0.973	0.682	0.682	0.965	0.964	0.736	0.738
Ceiling Auditor, Pass 1	0.954	0.961	0.325	0.273	0.909	0.915	0.363	0.368
Ceiling Auditor, Pass 2	0.938	0.943	0.331	0.304	0.943	0.945	0.422	0.423
Single Auditor, both passes, unanimous acceptance								
Floor Auditor	0.580	0.645	0.340	0.394	0.454	0.424	0.451	0.517
Robust Auditor	0.851	0.858	0.463	0.513	0.773	0.827	0.506	0.529
Anchor Auditor	0.871	0.875	0.376	0.461	0.793	0.813	0.438	0.471
Ceiling Auditor	0.902	0.911	0.153	0.151	0.867	0.875	0.195	0.235
Two Auditors, Pass 2, unanimous acceptance								
Floor and Robust	0.860	0.857	0.404	0.447	0.745	0.748	0.537	0.514
Floor and Anchor	0.866	0.869	0.408	0.440	0.750	0.759	0.551	0.558
Floor and Ceiling	0.842	0.829	0.240	0.223	0.740	0.735	0.348	0.356
Robust and Anchor	0.942	0.942	0.532	0.586	0.926	0.932	0.591	0.616
Robust and Ceiling	0.912	0.901	0.284	0.271	0.907	0.910	0.362	0.375
Anchor and Ceiling	0.918	0.927	0.288	0.272	0.916	0.920	0.371	0.405
Four Auditors, Pass 2								
Unanimous acceptance	0.813	0.817	0.201	0.200	0.712	0.725	0.287	0.308
Majority of three	0.956	0.953	0.444	0.482	0.932	0.940	0.569	0.580

The estimated model's operating point next to its measured counterpart for every audit policy.

Table 31: The frontier under alternative aggregation conventions and wages

Target	Pooled stream $w = 0$ (Table 4)	Pooled stream $w =$ \$7.18/hr	One answer per task $w = 0$	Mean task $w = 0$
Control pool				
Highest attainable R	Ceiling P1+P2 (0.916) 0.0569	Ceiling P1+P2 (0.916) 0.8458	Ceiling P1+P2 (0.833) 0.3181	Ceiling P1+P2 (0.947) 0.0586
$R \geq 0.70$	Anchor P2 (0.716) 0.0237	Anchor P2 (0.716) 0.1746	Anchor P2 (0.701) 0.0240	Anchor P2 (0.761) 0.0245
$R \geq 0.75$	Floor P2 (0.758) 0.0257	Floor P2 (0.758) 0.1915	Ceiling P2 (0.765) 0.0508	Anchor P2 (0.761) 0.0245
$R \geq 0.80$	Anchor P1+P2 (0.804) 0.0340	Anchor P1+P2 (0.804) 0.2549	Four, unanimous (0.800) 0.1568	Anchor P1 (0.839) 0.0306
$R \geq 0.85$	Anchor+Ceiling (0.851) 0.0404	Anchor+Ceiling (0.851) 0.4332	–	Ceiling P2 (0.856) 0.0352
$R \geq 0.90$	Ceiling P1+P2 (0.916) 0.0569	Ceiling P1+P2 (0.916) 0.8458	–	Ceiling P1+P2 (0.947) 0.0586
$R \geq 0.95$	–	–	–	–
$R \geq 0.99$	–	–	–	–
Red-herring pool				
Highest attainable R	Ceiling P1+P2 (0.822) 0.0677	Ceiling P1+P2 (0.822) 1.0062	Ceiling P1+P2 (0.728) 0.1612	Ceiling P1+P2 (0.852) 0.0743
$R \geq 0.70$	Anchor+Ceiling (0.712) 0.0433	Anchor+Ceiling (0.712) 0.4642	Ceiling P1+P2 (0.728) 0.1612	Anchor+Ceiling (0.724) 0.0456
$R \geq 0.75$	Ceiling P1+P2 (0.822) 0.0677	Ceiling P1+P2 (0.822) 1.0062	–	Ceiling P1 (0.764) 0.0492
$R \geq 0.80$	Ceiling P1+P2 (0.822) 0.0677	Ceiling P1+P2 (0.822) 1.0062	–	Ceiling P1+P2 (0.852) 0.0743
$R \geq 0.85$	–	–	–	Ceiling P1+P2 (0.852) 0.0743
$R \geq 0.90$	–	–	–	–
$R \geq 0.95$	–	–	–	–
$R \geq 0.99$	–	–	–	–

The cheapest policy reaching each reliability target under alternative aggregation conventions and a positive wage.

Table 32: The red-herring arm gaps of the Auditor verifying pass, decomposed by channel

		Arm gap		Channels of the model gap			
Auditor	Rate	Measured	Model	Displacement	Pool	Difficulty	Interaction
Floor	F, wrong candidates	+0.123	+0.169	+0.103	−0.028	+0.077	+0.017
	C, correct candidates	−0.112	−0.117	−0.100	−0.024	−0.014	+0.021
Robust	F, wrong candidates	−0.001	+0.040	−0.027	−0.022	+0.120	−0.030
	C, correct candidates	+0.003	−0.011	−0.007	−0.008	0.000	+0.004
Anchor	F, wrong candidates	+0.055	+0.055	−0.013	−0.022	+0.128	−0.038
	C, correct candidates	−0.009	−0.008	−0.004	−0.007	0.000	+0.003
Ceiling	F, wrong candidates	+0.119	+0.090	+0.029	−0.044	+0.094	+0.011
	C, correct candidates	+0.001	+0.005	+0.027	−0.010	−0.013	+0.001

Each Auditor's red-herring arm gap in the verifying pass, measured and model, decomposed by channel.

Table 33: The cost behind a verdict

	Pooled	Pooled	Floor Auditor	Robust Auditor	Anchor Auditor	Ceiling Auditor
Panel A. Tokens and seconds against the task clean solve rate. Coefficient on s_t, the task clean solve rate						
Prompt-length control, log prompt tokens	No	Yes	Yes	Yes	Yes	Yes
log completion tokens	-0.6489***	-0.4820**	+0.6930	-0.3079*	-0.6311***	-1.8912***
	(0.1938)	(0.2092)	(0.5280)	(0.1653)	(0.1614)	(0.3044)
	[0.001]	[0.021]	[0.189]	[0.062]	[0.000]	[0.000]
log latency	-0.7077***	-0.6522***	+0.1359	-0.3597***	-0.6789***	-1.8906***
	(0.1245)	(0.1468)	(0.2017)	(0.1172)	(0.1846)	(0.2987)
	[0.000]	[0.000]	[0.501]	[0.002]	[0.000]	[0.000]
N	15,892	15,892	4,046	3,998	4,116	3,732
Panel B. A verdict against the same model's own answer. Log verdict meter minus log own-answer meter						
log completion tokens, gap			+0.8389***	+0.2621***	.	-0.2041***
			(0.1500)	(0.0413)	.	(0.0691)
			[0.000]	[0.000]	.	[0.003]
implied percent gap			+131%	+30%	.	-18%
log latency, gap			+0.3109***	+0.0573*	.	-0.4333***
			(0.0551)	(0.0298)	.	(0.0706)
			[0.000]	[0.055]	.	[0.000]
implied percent gap			+36%	+6%	.	-35%
N			4,046	3,998	.	3,668

Pass-1 completion-token and latency gradients in task solve rate, and verdict cost against the Auditor's own answer cost.

Table 34: Bootstrap against asymptotic inference: the reconciliation gate

Statistic	Point	95% interval	Bootstrap SD	Asymptotic SE	Ratio	Band	Outside
Panel A. The audit block, Hessian standard errors, informational band							
Ability shift, Floor Auditor	+0.279	[−0.175, +0.782]	0.2562	0.0546	4.69	[1, 10]	no
Ability shift, Robust Auditor	+0.344	[+0.034, +1.208]	0.3032	0.0546	5.55	[1, 10]	no
Ability shift, Ceiling Auditor	+0.129	[−0.350, +0.799]	1.6949	0.0618	27.43	[1, 10]	yes
Pass 2 difficulty κ	+0.218	[+0.134, +0.298]	0.0417	0.0080	5.21	[1, 10]	no
Re-solve difficulty slope ϑ_1	+0.211	[+0.101, +0.347]	0.0624	0.0157	3.97	[1, 10]	no
Pass 2 level c_2	+0.299	[+0.075, +0.476]	0.0994	0.0316	3.15	[1, 10]	no
Pass 2 capability λ	+0.145	[+0.062, +0.244]	0.0483	0.0173	2.80	[1, 10]	no
Correct-side displacement ρ_{Floor}	+0.283	[+0.053, +0.562]	0.1325	0.0620	2.14	[1, 10]	no
Reasoning-agent displacement ρ_R	+0.414	[+0.173, +0.727]	0.1431	0.0355	4.03	[1, 10]	no
Derivation-length displacement ρ_L	−0.139	[−0.299, +0.015]	0.0794	0.0217	3.66	[1, 10]	no
Wrong-side displacement $\rho_{W,\text{Floor}}$	+0.391	[+0.169, +0.625]	0.1191	0.0593	2.01	[1, 10]	no
Wrong-side displacement $\rho_{W,\text{Robust}}$	+0.076	[−0.211, +0.365]	0.1537	0.0828	1.86	[1, 10]	no
Wrong-side displacement $\rho_{W,\text{Anchor}}$	+0.058	[−0.325, +0.345]	0.1656	0.0895	1.85	[1, 10]	no
Wrong-side displacement $\rho_{W,\text{Ceiling}}$	−0.213	[−0.561, +0.080]	0.1666	0.0834	2.00	[1, 10]	no
Panel B. The audit cost equations, two-way clustered standard errors, hard band							
γ, money meter, Floor	+0.110	[−0.069, +0.238]	0.0781	0.0650	1.20	[0.5, 3]	no
$\theta_{\text{CW},2}$, money meter, Floor	+0.109	[+0.016, +0.238]	0.0556	0.0443	1.25	[0.5, 3]	no
γ, time meter, Floor	+0.054	[−0.074, +0.170]	0.0608	0.0482	1.26	[0.5, 3]	no
$\theta_{\text{CW},2}$, time meter, Floor	+0.163	[+0.074, +0.272]	0.0506	0.0325	1.56	[0.5, 3]	no
γ, money meter, Robust	−0.034	[−0.102, +0.050]	0.0381	0.0315	1.21	[0.5, 3]	no
$\theta_{\text{CW},2}$, money meter, Robust	+0.019	[−0.112, +0.120]	0.0581	0.0504	1.15	[0.5, 3]	no
γ, time meter, Robust	−0.032	[−0.094, +0.051]	0.0366	0.0288	1.27	[0.5, 3]	no
$\theta_{\text{CW},2}$, time meter, Robust	+0.072	[−0.047, +0.163]	0.0525	0.0397	1.32	[0.5, 3]	no
γ, money meter, Anchor	+0.018	[−0.116, +0.128]	0.0630	0.0574	1.10	[0.5, 3]	no
$\theta_{\text{CW},2}$, money meter, Anchor	+0.125	[−0.038, +0.322]	0.0941	0.0899	1.05	[0.5, 3]	no
γ, time meter, Anchor	+0.073	[−0.080, +0.187]	0.0683	0.0658	1.04	[0.5, 3]	no
$\theta_{\text{CW},2}$, time meter, Anchor	+0.220	[+0.043, +0.450]	0.1047	0.1008	1.04	[0.5, 3]	no
γ, money meter, Ceiling	−0.017	[−0.232, +0.175]	0.1004	0.0919	1.09	[0.5, 3]	no
$\theta_{\text{CW},2}$, money meter, Ceiling	+0.716	[+0.489, +1.006]	0.1307	0.1144	1.14	[0.5, 3]	no
γ, time meter, Ceiling	+0.044	[−0.135, +0.220]	0.0918	0.0795	1.16	[0.5, 3]	no
$\theta_{\text{CW},2}$, time meter, Ceiling	+0.844	[+0.604, +1.127]	0.1336	0.1147	1.16	[0.5, 3]	no

Bootstrap standard deviation against asymptotic standard error for every tracked statistic, with the gate's band.