UNIVERSITY OF MANNHEIM
Business School

CHAIR OF
BUSINESS ADMINISTRATION
AND ACCOUNTING

Prof. Dr. Dirk Simons
Dr. Sebastian Kronenberger
Yasmin Kuhlmann
Castle, Room O 242
68161 Mannheim, Germany
Phone: +49 621 181-1467
office-simons.bwl@uni-mannheim.de

University of Mannheim · Chair of Business Administration and Accounting
Castle · 68161 Mannheim, Germany

Chairman Demetrios Logothetis
Public Company Accounting Oversight Board
1666 K Street N. W.
Washington, D.C. 20006-2803

Mannheim, May 15, 2026

Re.: PCAOB Release No. 2026-001 - Requests for Public Comment on the PCAOB's Strategic Priorities of March 31, 2026.

Dear Chairman Logothetis,

We welcome the opportunity to answer the Board's request for input regarding the PCAOB's 2026-2030 strategic priorities.

In this letter, we provide insights into our recent research at the University of Mannheim Business School, Germany. Particularly, in our study

Audit Quality and Auditing Standards under Artificial Intelligence,

we conduct an economic analysis of the impact of artificial intelligence (AI) on audit quality and how variations in auditing standards influence market outcomes. Our research speaks directly to the invitation for comment on what standard-setting projects the Board should pursue. Our intention is to raise awareness that audit procedures are shaped not only by the stringency of effort requirements, as embodied for instance in AS 1201, or PCAOB inspection intensity. A second important dimension is whether the standards define non-AI (human) components of the audit, such as minimum partner hours or a blacklist of tasks that cannot be delegated to AI.

In short, we show that **the PCAOB faces an important strategic trade-off for future auditing standards**: Specifying a human component of audit effort raises audit quality, whereas remaining silent on the human-AI mix improves market value more substantially.

We provide a summary of our research in this comment letter. The full working paper is in the appendix to this memorandum and also submitted to the 2026 Conference on Auditing and Capital Markets in Washington, D.C. on September 10-11, 2026.

AI reshapes corporate processes and workflows. This is particularly important in auditing as AI supported pattern recognition tremendously improves fraud detection. However, an important regulatory challenge is the question how AI procedures should be addressed in terms of standard-setting. One option is to continue the existing practice and view AI as one tool of many, which is justified to use, as long as the principles of a good audit are not violated and all auditing standards are being followed. This is what we call *a total effort standard*. In contrast, auditing standards could dictate additional rules explicitly emphasizing that certain procedures cannot be outsourced to AI, emphasizing the human component in the audit. These rules could include, e.g., minimum partner-level hours, a list of prohibited AI tasks, AI explainability such as the documentation and disclosure of the used AI agents, prompts, outputs, and how these outputs were checked manually. We term this approach *the human effort standard*. The impact of these two options varies, depending on the focus the PCAOB envisions. Specifying human audit effort tasks, which boosts audit quality but leaves potential welfare gains on the table or remaining silent about human and AI involvement, which improves market value (investor welfare).

Two features of AI drive our results. First, AI is cheaper to deploy than human effort at the margin. After an initial investment, AI is faster in producing high-quality output and requires fewer working hours. Second, AI affects expertise building. A limited use of AI frees time for human effort, which improves the client-specific expertise, but an over-reliance on AI reduces the ability to gain knowledge and lowers the overall expertise. We term this decrease in expertise *"deskilling"*.

The key difference between the two regulatory instruments is how they influence the AI investment. Under a total-effort standard, the auditor responds to tighter regulation by substituting toward cheap AI effort. This may drive AI adoption well beyond the expertise-maximizing level, eroding audit quality through deskilling. Strikingly, audit quality with AI can fall below the status quo without AI. From a capital market perspective, this is not necessarily bad, because operating under high levels of AI cuts cost and improves overall welfare for investors. Under a human-effort standard, a minimum of human work is always preserved, which does not boost AI investment as much as the total effort standard and avoids deskilling. Thus, audit quality is higher, but the full potential of AI, in terms of expertise building and cost saving, is underexploited, resulting in smaller investor welfare gains and audit quality below its feasible maximum.

We believe these results are important for the PCAOB, as the strategic planning under the new leadership determines the path moving forward. How the Board chooses to address AI in standard-setting is a first-order lever shaping both the degree of auditors' AI adoption and its consequences for the capital market.

UNIVERSITY OF MANNHEIM
Business School

CHAIR OF
BUSINESS ADMINISTRATION
AND ACCOUNTING

We thank the Board for the opportunity to contribute to this consultation. We would be happy to discuss our work with PCAOB staff. Questions about this letter or the attached working paper may be directed to Sebastian Kronenberger (kronenberger@uni-mannheim.de).

Sincerely,

Dirk Simons
Professor of Accounting
University of Mannheim Business School

Sebastian Kronenberger
Assistant Professor of Accounting
University of Mannheim Business School

Yasmin Kuhlmann
PhD Candidate in Accounting
University of Mannheim Business School

Audit Quality and Auditing Standards under Artificial Intelligence*

Sebastian Kronenberger† Yasmin Kuhlmann‡

May 12, 2026

Abstract

We study auditing standards when auditors can adopt Artificial Intelligence (AI) as an assisting technology. In our model, a representative auditor chooses a mix of human and AI effort to verify financial statements. AI effort is cheaper and initially frees up resources for higher-value judgment tasks. However, its excessive use leads to deskilling: engagement-specific expertise erodes, which reduces the effectiveness of each unit of audit effort. We compare two auditing standards regimes, (i) a total-effort minimum and (ii) a human-effort minimum. The choice between them generates a regulatory trade-off. Under total-effort standards, a higher minimum inadvertently fuels AI over-adoption because the auditor does not fully internalize the social cost of audit failures. The resulting deskilling causes audit quality to fall below the no-AI benchmark as standards tighten, even as the AI world delivers strictly higher welfare through lower compliance costs. By contrast, human-effort standards preserve a minimum of human work, which avoids deskilling. Thus, audit quality is higher, but the full potential of AI remains underexploited, resulting in smaller welfare gains and audit quality below its feasible maximum. Our analysis suggests that regulators should not merely ask *how much* audit work to require, but *what kind* of work.

Keywords: Artificial Intelligence; Auditing Standards; Audit Quality; Welfare
JEL classification: D82, M42, L51, O33

*We thank Volker Laux and Thomas Simon for helpful feedback and suggestions. We further appreciate workshop participants' insightful comments at the University of Mannheim. This work was supported by the Graduate School of Economic and Social Sciences of the University of Mannheim. We gratefully acknowledge funding by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) Project-ID 403041268 - TRR 266 Accounting for Transparency. All remaining errors are our own.

†Sebastian Kronenberger, University of Mannheim. kronenberger@uni-mannheim.de.
‡Yasmin Kuhlmann, University of Mannheim. yasmin.kuhlmann@uni-mannheim.de.

1 Introduction

Artificial Intelligence (AI) has substantially shifted modern day information input and workforce processes in organizations. The prompt to ChatGPT-4.1 "for which tasks is AI best?"[1] delivers the top 3 answers: 1. Pattern recognition, 2. Language and text-heavy work, 3. Repetitive, rule-based tasks. The chat-bot's own assessment is also supported scientifically (Eloundou et al. (2023); Noy and Zhang (2023)). The task of financial statement verification matches well with these top three answers. Hence, it is not surprising that Frey and Osborne (2017) rank accountants and auditors in the top quartile of computerizable jobs.[2] For the same prompt, ChatGPT-4.1 also points out, where AI is not best. It lists "deep human judgment and values", "novel situations", and "high-stakes accountability". In general, the answer illustrates the dilemma in terms of the auditor's tasks as judgment and accountability being essential for financial statement verification.

Due to the importance of judgment in auditing, the general assessment is that AI does not replace human workforce, but supports the workers in their tasks (Law and Shen (2025)). However, a crucial concern is the negative consequences from delegating tasks to a black box algorithm. In medical diagnostic settings, which are structurally similar to auditing, studies find an over-reliance on computer-aided detection tools (Budzyń et al. (2025)). The use of generative AI can also demotivate workers for tasks that do not involve AI (Liu et al. (2025)). Correcting the hallucinations from AI can lead to a slower task completion of skilled open-source developers compared to non-AI use (Becker et al. (2025)). Further, the outsourcing of basic tasks might impair the ability to develop the knowledge to use judgment, leading to de-skilling of workers (Munoko et al. (2020); Gillespie et al. (2025)).[3] At the same time, the auditor's audit quality is shaped by the strictness of the regulatory requirements. As absolute assurance is prohibitively costly and de facto unachievable, auditing standards specify the auditor's due professional care level that determines a minimum threshold for compliance.

Hence, in this context we ask the following questions: How does AI shape the optimal audit effort decision and how can standards be adjusted to account for AI use? We find that the choice between two designs of such a standard, a minimum on total audit work versus a minimum on human work, generates a regulatory trade-off. The first can erode audit quality through the auditor's deskilling

[1] The prompt was posed in general terms without an accounting or auditing context specified, so the top-three answers reflect a generic assessment of AI's strengths.

[2] The authors calculate a 94 % probability for the job of an accountant or auditor to be computerizable, on one level with industrial truck operators and waiters/waitresses.

[3] Gillespie et al. (2025) survey KPMG employees globally about AI use. 48% report that they personally experienced de-skilling. 60% reported that they relied on AI as a student instead of learning how to do the task themselves.

once stringent, while the second constrains deskilling but limits how far AI's benefits can be realized.

We develop a model, where an auditor is hired by a client to verify the financial statements. The auditor can use what we label "human" effort and "AI" effort, which can be thought of as prompting the inputs or analyzing the outputs of the AI tool. AI effort is less costly than human effort and both contribute to detection. Similar to Gao and Zhang (2019), the auditor needs client-specific expertise to evaluate the audit risk of the client. Both human and AI effort affect expertise building. A limited use of AI frees time for human effort, which improves the client-specific expertise, but an over-reliance on AI reduces the ability to gain knowledge and lowers the overall expertise. Next, we consider an audit standard setter, such as the PCAOB, which implements a minimum level of audit effort. We consider two types of audit standards, (i) a total-effort floor, which requires a minimum amount of overall audit work regardless of whether it is performed by humans or AI, and (ii) a human-effort floor, which specifically mandates a minimum level of human involvement. This distinction is practically relevant as current auditing standards (e.g., Public Company Accounting Oversight Board (2010)) prescribe procedures that implicitly require human judgment, but as AI tools become more capable, regulators face the question of whether to measure compliance in terms of total work performed or to explicitly protect the human component. In practice, a total-effort standard would not (or only to a small extent) specify how AI relates to the audit process as long as the overall principles of due-care are met. A human effort standard could add additional requirements, such as minimum partner-level hours, a list of prohibited AI tasks, AI explainability such as the disclosure of the used AI agents, prompts, outputs, and how these outputs were used in terms of human checks. Compliance with these requirements can be checked in inspections based on the documented audit evidence and audit process. In case of a misstated financial statement, investors can sue the auditor for compensatory damage payments. Our analysis shows that the choice between a total- and a human-effort standard has first-order consequences for AI adoption, audit quality, and welfare.

We find that the choice of the regulatory instrument fundamentally determines how AI reshapes audit outcomes. If held liable, the auditor's litigation damage is only a fraction of total investments. Thus, she internalizes only a share of the social audit failure cost. This gap between private and social incentives is amplified by AI. Since AI effort is cheap, the auditor can satisfy regulatory requirements at low cost, even when the resulting audit mix might be less effective at detecting fraud. Under a total-effort standard, which does not distinguish between human and AI work, the auditor responds to tighter regulation by substituting toward cheap AI effort. This drives

AI adoption well beyond the expertise-maximizing level, eroding audit quality through deskilling. Strikingly, audit quality with AI can fall below the status quo without AI. Under a human-effort standard, a minimum of human work is always preserved, which avoids deskilling. Thus, audit quality is higher, but the full potential of AI rests underexploited, resulting in smaller welfare gains and audit quality below its feasible maximum. Our analysis thus suggests that one should not merely ask *how much* audit work to require, but *what kind* of work.

We contribute to analytical studies in auditing about the use of technology, an issue underrepresented in the extant literature (Ye (2023)). Cao et al. (2025) analyze the impact of federated blockchains to verify transactions in audit networks. Hwang et al. (2025) study the incentives of clients and their auditors to coordinate for mutually beneficial investments in technologies that enhance audit quality.[4] Different to the focus in these studies, we emphasize the optimal use of readily available AI as assisting technology tool and how audit standards would need to change to adjust to the new work environment. Empirically, Law and Shen (2025) document that the use of AI in audit offices increases the number of jobs and changes the demanded skills towards more cognitive skills. In contrast, Fedyk et al. (2022) find that auditors are replaced by AI, but both studies align in terms of their positive assessment of average short term audit outcomes, such as an increase in measures for audit quality and a decrease in audit fees. Relatedly, Choi and Xie (2026) provide early field evidence on the consequences of AI implementation. They document the benefits of AI, i.e. productivity gains, and the risk of AI, i.e. classification errors and overreliance on machine outputs. These characteristics are in line with the assumptions of our model. Overall, the authors find a higher financial reporting quality under the AI regime, but acknowledge that the results are not causal due to the endogenous AI adoption choice and the omission of adoption cost. Our model complements the findings by including these dimensions and demonstrating its effect on output measures, namely audit quality and welfare. We show that the specifics of auditing standards can influence the output measures and overreliance on AI can lead to a lower audit quality as compared to a world without AI.

We further contribute to the economics of standard setting in auditing. Dye (1993) shows that an increase in the tightness of auditing standards does not necessarily improve the average audit quality. The result is driven by low wealth auditors, whose cost of non-compliance decreases with a tighter standard. Similarly, Ye and Simunic (2013) analyze the optimal design of the tightness and vagueness of auditing standards. They show that the optimal standards should have no vagueness

[4]Whether the auditor possesses a high or low (unspecified) audit technology also plays a major role, e.g., in Caskey and Laux (2017) and Kronenberger and Laux (2022).

if one can set the tightness of the standards optimally. However, vague standards can be optimal otherwise. Most related to our paper, Gao and Zhang (2019) show that audit standards can limit the use of expertise and cause a compliance culture, which reduces incentives to invest in expertise building.[5] We extend their insights to expertise building in the presence of modern day technology. Our central contribution to this literature is to identify a regulatory trade-off in which AI-induced deskilling interacts with the regulatory instrument in a way which prior frameworks cannot capture. In Gao and Zhang (2019), tighter standards crowd out expertise investment by creating a compliance culture. In our setting, the crowding-out operates through technology substitution, where a total-effort standard incentivizes the auditor to meet the floor with cheap AI effort, which erodes the expertise needed to use judgment effectively. A human-effort standard breaks this channel by requiring a minimum of human work. This preserves learning-by-doing, which sustains expertise but underexploits the full potential of AI, resulting in smaller welfare gains and audit quality below its feasible maximum.

Further results show that a change in litigation damages and baseline expertise interact with the regulatory instrument in policy-relevant ways. Higher litigation increases AI adoption before reaching the expertise-maximizing effort mix, but decreases it past the peak. Thus, litigation acts as a self-correcting force on AI adoption, partially offsetting the over-adoption induced by stringent total-effort requirements. The effect of baseline expertise is regime-dependent. More experienced auditors invest less in AI as their expertise increases further. For less experienced auditors, the reaction depends on litigation. In high litigation regimes, an increase in baseline expertise lowers AI investment. In low litigation regimes, an increase in baseline expertise boosts AI investment.

2 Model

The model consists of two players, a representative auditor and a representative client firm. The firm has access to an investment project that requires an initial investment $I > 0$. The project can be good (success) or bad (failure), $i \in \{G, B\}$. A good project generates return $G > I$, while a bad project generates return B, which is normalized to zero. The prior probability that the project is good is $p \in (0, 1)$, so the project's net present value without any audit is $W_0 \equiv pG - I > 0$, and the default decision is to invest. The auditor is thus hired to prevent investment into bad projects. As in Schwartz (1997) or Laux and Newman (2010), the firm does not have private information

[5] Other studies in this field include Schwartz (1998), Ewert (1999), or Simunic et al. (2017).

about i and always sends the auditor a favorable report. The auditor performs an audit and issues an audit report $r \in \{g, b\}$, where $r = g$ denotes an unqualified opinion, i.e. the firm's favorable report is accepted, and $r = b$ denotes a qualified opinion, i.e. the firm's favorable report is rejected. The firm invests only if $r = g$.

Audit effort and costs. The auditor can supply audit effort through two technologies, human effort and AI-supported effort if AI was adopted. Let $q_h \geq 0$ be human audit effort and $q_{AI} \geq 0$ be AI audit effort. Total effort then is

$$q = q_h + q_{AI}. \tag{1}$$

We interpret one unit of q_{AI} as one unit of AI-supported audit activity, which includes both automated analysis and the auditor's time spent interacting with AI tools (e.g., prompt design and review of AI outputs).[6] Human and AI effort incur quadratic costs, $C_h(q_h) = \frac{c_h}{2} q_h^2$, $C_{AI}(q_{AI}) = \frac{c_{AI}}{2} q_{AI}^2$, with $c_h > c_{AI} > 0$, so AI effort is technologically cheaper at the margin than human effort. We parameterize AI adoption by the AI share

$$R = \frac{q_{AI}}{q} \in [0, 1], \tag{2}$$

so that $q_{AI} = Rq$ and $q_h = (1 - R)q$ whenever total effort $q > 0$. Given R, total effort q translates into the mixed cost parameter $c_{\text{eff}}(R) = c_h(1 - R)^2 + c_{AI}R^2$, so that the total effort cost can be written as $C_h(q_h) + C_{AI}(q_{AI}) = \frac{1}{2} c_{\text{eff}}(R) q^2$. Because $c_h > c_{AI}$, the mixed cost parameter $c_{\text{eff}}(R)$ is strictly convex and U-shaped in R, with a unique cost-minimizing mix $R^{\text{cost}} = \frac{c_h}{c_h + c_{AI}}$. This is the AI share that minimizes the marginal cost of an additional unit of total effort, absent any other considerations. AI adoption itself requires an engagement-independent investment (e.g., integration into systems and training), summarized by a quadratic adoption cost $C_R(R) = \frac{c_R}{2} R^2$, with $c_R > 0$. The convexity of C_R reflects that deeper integration of AI into the audit process requires progressively larger investments in infrastructure, training, and quality assurance.[7] We assume a perfectly competitive market for audit services, such that the analysis can focus on the auditor's cost-minimizing behavior without modeling strategic fee-setting.[8]

[6] We later capture complementarity between human and AI work, as would arise from CES or Cobb-Douglas aggregation at the production stage, through $\phi(R)$ which scales how effective audit effort is at detection rather than at the effort-aggregation stage. This preserves analytical tractability without losing the substantive interaction between the two inputs.

[7] The corresponding investment in human capital, e.g. CPA training, supervised on-the-job learning, professional education, is taken as predetermined and embedded in the baseline expertise $e(0)$. Our model focuses on the additional adoption of AI.

[8] This assumption is common in the analytical auditing literature, see, e.g., Dye (1993); Schwartz (1997).

Audit expertise. Engagement-specific expertise depends both on the auditor's baseline expertise and on the AI share used on the engagement. Let $e(0) \in (0,1)$ denote the auditor's baseline expertise in the absence of AI, for example, from CPA training and accumulated experience. Our model reflects that using AI changes engagement-specific audit expertise $e(R)$ in a non-linear way. On the one hand, AI initially comes with efficiency gains in routine tasks but on the other hand, it may come with a loss of expertise once the auditor lacks insights from routine tasks and over-relies on AI. As a consequence, the auditor's ability might be impaired, i.e., the performance of judgment tasks that go beyond sample analysis and require profound knowledge of the client's business. As such, in our model low to moderate AI use frees human resources for higher-value judgment tasks and can increase expertise, whereas high AI use leads to *deskilling* and reduces expertise. We model this impact of AI on audit expertise via a logistic map to keep $e(R) \in [0,1]$,

$$e(R) = \frac{1}{1 + \exp\big(-g(R)\big)}, \tag{3}$$

with

$$g(R) = \log\Big(\frac{e(0)}{1-e(0)}\Big) + \alpha R - \beta R^2, \qquad \alpha > 0,\ \beta > 0. \tag{4}$$

This specification has an intuitive decomposition. The first term is the logistic transformation of the baseline expertise such that if $R = 0$, then expertise is $e(0)$. The linear term αR captures efficiency gains from introducing AI, while the quadratic term $-\beta R^2$ captures the deskilling effect of excessive AI reliance. We also map audit expertise into the effectiveness of audit effort, which is represented by

$$\phi(R) = \frac{e(R)}{e(0)},$$

so that $\phi(0) = 1$, $\phi(R) > 1$ if moderate AI use raises expertise relative to the baseline level and $\phi(R) < 1$ when high AI use leads to deskilling relative to the baseline level. Intuitively, $\phi(R)$ later scales how powerful one unit of effort is in reducing audit risk. Because the logistic transformation of the expertise function is strictly increasing in $g(R)$, the qualitative shape of $e(R)$ is determined by $g(R)$. We provide more intuition for this in Definition 1 and Figure 1.

Definition 1. *The function $g(R)$ is concave, with first- and second-order derivative*

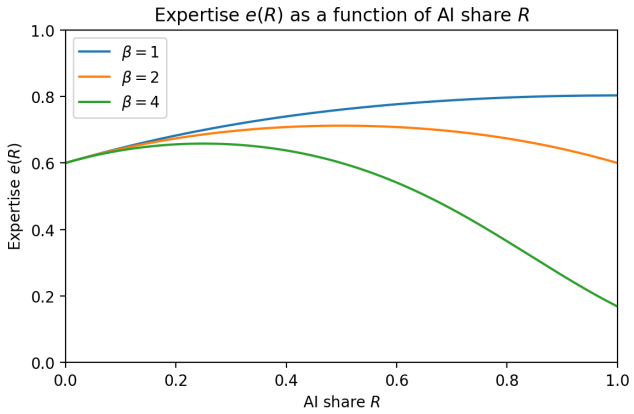
$$\frac{\partial g(R)}{\partial R} = \alpha - 2\beta R, \qquad \frac{\partial^2 g(R)}{\partial R^2} = -2\beta < 0.$$

Hence $g(R)$, and therefore $e(R)$ and $\phi(R)$, attain a unique global maximum at $R_{peak} = \frac{\alpha}{2\beta} \in (0,1)$.

For $R < R_{peak}$, expertise is increasing in the AI share, $\frac{\partial e(R)}{\partial R} > 0$ and for $R > R_{peak}$, expertise is decreasing in the AI share, $\frac{\partial e(R)}{\partial R} < 0$. Excessive AI adoption therefore leads to deskilling, captured by a decline in $e(R)$. Throughout the analysis we maintain the parameter restriction $\alpha < 2\beta$.

Intuitively, R_{peak} describes the expertise-maximizing AI mix, as for R below this point, the information gains from AI outweigh deskilling concerns, while beyond R_{peak}, additional AI primarily crowds out learning-by-doing and reduces expertise. The parameter restriction $\alpha < 2\beta$ ensures that the expertise-maximizing AI share lies strictly inside the unit interval, such that the deskilling region $R > R_{\text{peak}}$ is economically relevant. Figure 1 shows the expertise function for different values of β, i.e. different severity of deskilling.

Figure 1: Auditor expertise $e(R)$ as a function of the AI share R.



We set parameters to $e(0) = 0.6$, $\alpha = 2$, $R_{\text{peak}} = \alpha/2\beta$.

Our definition of deskilling as a function of the AI share R fundamentally captures the composition of work. This does not imply that an auditor, who audits with a certain level of q before AI, now adds AI to this level and suddenly experiences deskilling. In contrast, the use of AI substitutes for human effort. For the same engagement, the total audit effort level required is still the same q, where a fraction of the human work is replaced by AI, $q = q_h + q_{AI}$. As a result, the engagement can be completed faster at lower marginal cost. However, as the human effort is reduced, information necessary to apply judgment might be missing, leading to deskilling.

Information structure. The key informational friction in the model is that the auditor is uncertain about the engagement's inherent audit risk denoted by parameter γ, where γ is a random variable over $[0, 1]$ with mean γ_0. The cdf and pdf are $F(\tilde{\gamma})$ and $f(\tilde{\gamma})$. In other words, the audit risk may vary across engagements. In line with Gao and Zhang (2019), an auditor with more expertise has a better ability in assessing uncertain audit risk, i.e., assessing the conditional probability that the client's report about the investment is misstated conditional on the project being bad.

Specifically, with $Pr(\tau = in) = e(R)$ the auditor with expertise $e(R)$ is informed about audit risk ($\tau = in$) or uninformed about audit risk ($\tau = un$) with probability $1 - e(R)$. The information set Ω_τ reflects all the information and professional judgment available to the auditor. That the expert auditor has better judgment about audit risk than her non-expert counterpart is captured by the assumption that Ω_{in} is finer than Ω_{un} in the Blackwell sense. As such, the posterior $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ is a random variable with cdf $F_\tau(\cdot)$ over $[0, 1]$.[9] This general Blackwell structure nests many specific information technologies. For analytical tractability, we adopt the polar case in which the informed auditor learns the risk realization perfectly, i.e., $\gamma \sim U[0, 1]$, while the uninformed auditor receives no signal and the posterior is equal to the unconditional mean γ_0. Expectations taken before the realization of the information state are therefore weighted by expertise as a probability $e(R)$

$$\mathbb{E}[\cdot \mid R] = e(R)\,\mathbb{E}[\cdot \mid \text{informed}] + \big(1 - e(R)\big)\,\mathbb{E}[\cdot \mid \text{uninformed}].$$

Audit technology and failure probability. We now specify how audit effort translates into detection. In line with Dye (1993), a good project is never incorrectly rejected, so the audit technology has no Type I error. The key margin is the probability of failing to detect a bad project (Type II error), which depends on the engagement's perceived misstatement risk $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$, total audit effort q, and effort effectiveness $\phi(R)$. Formally, conditional on the project state $i \in \{G, B\}$, misstatement risk, and total effort with AI share R, the probability of a clean opinion on a good project is $\Pr(r = g \mid i = G,\, q, R) = 1$, and the probability of a clean opinion on a bad project is $\Pr(r = g \mid i = B,\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau], q, R) = \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\big(1 - \phi(R)\,q\big)$. As such, issuing a clean opinion on a bad project becomes less likely as effective effort $\phi(R)\,q$ increases. The product $\phi(R)\,q$ combines the quantity of work performed q with the quality of that work $\phi(R)$. When AI erodes expertise, so that $\phi(R) < 1$, each unit of effort becomes less effective at detecting misstatements. This interaction between effort quantity and effort quality is the channel through which deskilling affects audit outcomes. Because $q \in [0, 1]$ and $\phi(R)$ is bounded, we restrict attention to parameter regions where $\phi(R)\,q \in [0, 1]$ so that the failure probability remains in $[0, 1]$. Combining the prior probability of a bad project, the engagement-level misstatement risk, and the detection technology, the ex-ante audit risk, i.e., the probability that the auditor issues a clean opinion on a bad project, is

$$AR(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], q, R) \;=\; \underbrace{(1 - p)\;\cdot\;\mathbb{E}[\tilde{\gamma}|\Omega_\tau]}_{\text{inherent risk \& control risk}} \;\cdot\; \underbrace{\big(1 - \phi(R)\,q\big)}_{\text{detection risk}}. \tag{5}$$

[9] Since Ω_{in} is finer than Ω_{un}, the informed posterior is a mean-preserving spread of the uninformed posterior. For example, if the expert's judgment is perfect while the non-expert has no clue at all, then $\mathbb{E}[\tilde{\gamma}|\Omega_{in}] = \gamma$, while $\mathbb{E}[\tilde{\gamma}|\Omega_{un}] = \gamma_0$.

This expression mirrors the standard audit risk model $AR = IR \times CR \times DR$. Inherent and control risk $(1-p)\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ reflect the prior probability of a bad project and the engagement's misstatement probability as assessed by the auditor given her information. Detection risk $(1 - \phi(R)\, q)$ is the probability that the auditor's effective effort fails to uncover the misstatement, i.e., it falls in total effort q and in effort effectiveness $\phi(R)$, and is thus the channel through which audit standards and AI-induced deskilling affect audit outcomes. Ex-ante audit quality is then $\mathcal{A} = 1 - \mathbb{E}[AR]$, where the expectation is taken over information states and engagement risk.

Audit standards. We consider two types of audit standards, (i) a total-effort floor $Q \in [0, 1]$, requiring $q \geq Q$, and (ii) a human-effort floor $Q_h \in [0, 1]$, requiring $q_h \geq Q_h$. Given AI share R, the human-effort floor is equivalent to a minimum total effort $q \geq Q_h/(1 - R)$ whenever $R < 1$. The auditor always complies with the applicable standard, choosing a cost-minimizing effort level subject to the relevant constraint. A regulator evaluates a given pair (q, R) based on welfare, as introduced in eq. (9) later. The distinction between the two standards is economically significant because they constrain the auditor's effort substitution margin differently. A total-effort floor Q can be satisfied by any mix of human and AI effort, leaving the auditor free to meet the requirement entirely with cheap AI work. A human-effort floor Q_h, by contrast, directly limits AI substitution by requiring a minimum amount of human involvement. As we show later in the equilibrium analysis, this difference has first-order consequences for equilibrium AI adoption, audit quality, and welfare.

In practice, a total-effort standard would not (or only to a small extent) specify how AI relates to the audit process as long as the overall principles of due-care are met. A human effort standard could add additional requirements, such as minimum partner-level hours, a list of prohibited AI tasks, AI explainability such as the disclosure of the used AI agents, prompts, outputs, and how these outputs were used in terms of human checks. The compliance with these requirements can be checked in inspections based on the documented audit evidence and audit process. In our main analysis we assume perfect inspections. Section 4.1 discusses how relaxing this assumption impacts the differential inspectability of total and human effort and how this shapes the effective stringency of each instrument.

Litigation and damage payment. An undetected misstatement that causes investor losses exposes the auditor to litigation. We assume that whenever an audit failure occurs, the auditor faces litigation with probability π. The probability reflects that courts could still hold auditors liable despite compliance with minimum effort levels. Then, the damage paid is D which is smaller

than the initial investment, $D \leq I$.

Payoffs and competitive fees. Given audit effort q, AI share R, and perceived engagement risk $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$, the auditor's and firm's per-engagement payoffs are respectively

$$U_A = \omega - (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-q)\,\pi\,D - C(q), \tag{6}$$

$$U_F = p(G-I) - (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-q)\,\big(I - \pi D\big) - \omega, \tag{7}$$

where ω is the audit fee, $C(q) = \frac{1}{2}c_{\text{eff}}(R)\,q^2$ is the effort cost. The auditor receives the fee, bears expected litigation cost $(1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-q)\pi D$ when held liable after an audit failure, and incurs effort cost. The firm earns the net project value $p(G-I)$, suffers the uncompensated portion of the failure loss $I - \pi D$ per failure, and pays the fee. Under perfect competition the audit fee is bid down to expected cost, driving auditor payoff to zero such that[10]

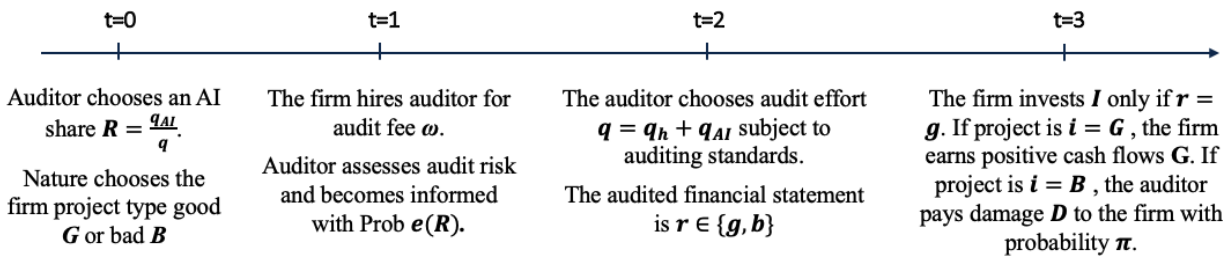
$$U_A = 0 \quad \Longrightarrow \quad \omega = (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-q)\,\pi\,D + C(q). \tag{8}$$

Substituting (8) into (7), the litigation-recovery terms πD cancel between the firm's direct loss and the fee, yielding

$$U_F = \underbrace{pG - I}_{W_0} + \underbrace{(1-p)\left[1 - \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-q)\right] I - C(q)}_{\text{net audit benefit}}. \tag{9}$$

Because the competitive fee absorbs the auditor's expected litigation cost, the damage transfer πD from auditor to investor is fully priced into the fee. Litigation is therefore purely redistributive and drops out of total welfare, which equals U_F (since $U_A = 0$). The social loss per audit failure is simply I, and the regulator's welfare evaluation depends only on the failure loss, effort costs, and adoption costs, not on the damage payment D. The timeline is as presented in Figure 2.

Figure 2: Sequence of events.



[10] Note that the effort cost $C(q) = \frac{1}{2}c_{\text{eff}}(R)\,q^2$ depends on the AI share through $c_{\text{eff}}(R)$. As such, the equilibrium fee ω varies with R where substitution toward cheaper AI effort lowers the cost component of the fee, partially offset by changes in the litigation component through $\phi(R)$.

3 Analysis

3.1 Benchmark: Audits without AI

To isolate the role of AI, we begin with a benchmark in which the auditor has no access to AI technology. The AI share is fixed at $R = 0$, so all effort is human, $q_h = q$, effort effectiveness is $\phi(0) = 1$, and there is no adoption cost. Stripping out these forces leaves a single distortion, the gap between the auditor's private damage exposure πD and the full social loss I per audit failure, which drives a misalignment between the auditor's private and the socially optimal effort. The benchmark isolates how this litigation gap, together with the tightness of the regulatory floor, jointly determine effort, audit quality and welfare in the absence of any AI-induced deskilling, and thus provides the reference point against which the introduction of AI is later evaluated.

At date 1, after the information state is realized, the auditor forms an engagement-specific assessment of misstatement risk $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ and chooses effort $q(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ subject to the regulatory floor $q(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \geq Q_h$ to minimize her expected cost per engagement. This cost comprises two components, the convex effort cost $\frac{c_h}{2}q^2$, and the expected damage payment $(1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-q)\,\pi D$. The latter depends jointly on the probability of a bad project, the engagement's misstatement risk, and the audit failure probability $1-q$. Effort reduces the probability of an audit failure, lowering the auditor's expected damage payment. Because she pays only $D \leq I$ when held liable, while the social loss per failure is I, her marginal private benefit of effort falls short of the marginal social benefit. This $D \leq I$ wedge is the sole source of misalignment between the auditor's incentives and those of a regulator in the benchmark. To isolate the role of the regulatory floor, consider first the auditor's unconstrained optimum. Setting the FOC of her cost with respect to q to zero yields

$$q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = \frac{(1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\pi D}{c_h}, \tag{10}$$

where the numerator captures the auditor's marginal private benefit of effort, πD, i.e., the reduction in expected damage from lowering the failure probability. The unconstrained optimum $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ is thus strictly increasing in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$, as a higher perceived misstatement risk raises the expected loss from audit failure, so the auditor finds it privately optimal to work harder on riskier engagements even without regulatory compulsion. Imposing the regulatory floor Q_h produces the constrained effort choice

$$q^{*}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h) = \max\big\{q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]),\, Q_h\big\}, \tag{11}$$

which partitions the engagement pool into binding and slack regions. Because $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ is increasing in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$, there exists a unique risk threshold

$$\bar{\gamma}(Q_h) = \frac{c_h\, Q_h}{(1-p)\,\pi D}, \tag{12}$$

such that the floor binds if and only if $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}(Q_h)$. On engagements below this threshold the auditor is pushed up to the floor while above it, her unconstrained optimum already exceeds Q_h and the standard is slack. The threshold $\bar{\gamma}(Q_h)$ is increasing in Q_h, because a tighter floor expands the binding engagement pool, and decreasing in damage payment D, as a larger damage payment raises the auditor's unconstrained effort, reducing the set of engagements where the floor needs to intervene.

Welfare and the optimal standard. Under the competitive fee condition, the audit fee absorbs the auditor's expected litigation cost, so the damage transfer πD from auditor to investor is fully priced into the fee and cancels from total welfare.[11] Litigation is therefore purely redistributive and the social loss per audit failure is simply I. Welfare as a function of the standard is therefore

$$W(Q_h) = \underbrace{pG - I}_{W_0} + \mathbb{E}_{\tilde{F}}\Big[\underbrace{(1-p)\left(1 - \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\left(1 - q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)\right)\right)}_{\text{prob. misstatement detected or absent}} \cdot \underbrace{I}_{\text{loss averted}}\Big] - \mathbb{E}_{\tilde{F}}\Big[\underbrace{\tfrac{c_h}{2}\left(q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)\right)^2}_{\text{effort cost}}\Big], \tag{13}$$

where $\tilde{F}$ is the mixture distribution combining the informed state in which the auditor observes the realization of the engagement's risk perfectly with probability $e(0)$, and the uninformed state in which she must rely on the unconditional prior with probability $1 - e(0)$. Because the standard is slack for $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}(Q_h)$, only the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}(Q_h)\}$ contributes to the FOC dW/dQ_h. In this region optimal effort equals the required floor, i.e., $q^* = Q_h$, so the social loss per failure is I. Differentiating (13) with respect to Q_h yields

$$\frac{dW}{dQ_h} = \underbrace{(1-p)\, I \int_0^{\bar{\gamma}} \mathbb{E}[\tilde{\gamma}|\Omega_\tau] d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])}_{\text{marginal social benefit on binding engagements}} - \underbrace{c_h\, Q_h\, F\big(\bar{\gamma}(Q_h)\big)}_{\text{marginal effort cost on binding engagements}}. \tag{14}$$

Each unit of additional mandated effort on a binding engagement reduces the probability of an undetected misstatement, averting social loss I. Due to the quadratic effort cost and the linear failure probability in effort q, welfare is strictly concave in Q_h, so the first-order condition is necessary and sufficient for a unique interior optimum.

[11] See the "triangle effect" in Laux and Newman (2010).

Lemma 1. *In the benchmark without AI, there exists a unique welfare-maximizing human-effort standard $Q_h^*(0)$ characterized by the equality*

$$(1-p)\,I\int_0^{\bar{\gamma}^*}\mathbb{E}[\tilde{\gamma}|\Omega_\tau]d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = c_h\,Q_h^*(0)\,F(\bar{\gamma}^*), \tag{15}$$

where $\bar{\gamma}^* \equiv \bar{\gamma}\big(Q_h^*(0)\big)$. The optimal standard $Q_h^*(0)$ is increasing in the failure loss I, decreasing in the marginal effort cost c_h, decreasing in the prior probability p that the project is good, and increasing in $\tilde{F}$ in the sense of first-order stochastic dominance.

The optimality condition in eq. (15) equates the aggregate marginal benefit of tighter regulation to its aggregate marginal cost. The benefit is the expected reduction in social loss I across binding engagements where the floor raises effort above the auditor's unconstrained optimum. The cost is the increase in effort expenditure across those same engagements. Because the competitive fee already prices in the auditor's litigation exposure, the damage payment D does not appear in the welfare FOC. As such the regulator's problem depends only on the gross failure loss I and the cost of effort. A larger investment amount I raises the social loss per failure and pushes $Q_h^*(0)$ upward. A larger effort cost c_h raises the marginal cost of mandated effort without affecting the benefit, pushing $Q_h^*(0)$ downward. A higher p shrinks the probability that a favorable opinion on a bad project causes harm, reducing the benefit of tighter regulation. A first-order stochastic dominance shift in the mixture distribution $\tilde{F}$ toward higher engagement risks increases $\int_0^{\bar{\gamma}^*}\mathbb{E}[\tilde{\gamma}|\Omega_\tau]d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ for any threshold $\bar{\gamma}$, raising the failure-loss benefit and therefore $Q_h^*(0)$. This benchmark establishes both the optimal regulatory target and the distortion the standard is correcting. When AI becomes available, these two objects are no longer separable. AI changes the cost of compliance, the effectiveness of effort, and the auditor's response to the standard itself. The next subsection introduces AI as an exogenous technology and traces through each of these channels before the adoption decision is endogenized.

3.2 Audits with a Fixed AI Share

Before fully endogenizing the auditor's AI adoption decision, it is useful to treat the AI share R as exogenous and ask how its presence reshapes the auditor's effort choice and the regulator's optimal standard. The value of this intermediate step lies in the clean decomposition it provides. With R fixed, the two channels through which AI reshapes audit outcomes can be isolated in closed form, (i) the cost channel, by which cheaper AI effort lowers the marginal cost of compliance, and

(ii) the expertise channel, by which the AI share shifts effort effectiveness and the quality of risk assessment. This allows us to abstract from the adoption distortion that arises once R is chosen strategically in response to the standard.

At date 1, after the information state is realized, the auditor forms a posterior belief about engagement-specific risk $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ and chooses total effort q subject to the exogenously pre-committed effort mix $q_{AI} = Rq$ and $q_h = (1-R)q$. For a given pair $(R, \mathbb{E}[\tilde{\gamma}|\Omega_\tau])$, she minimizes

$$C^A(q, R, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = \frac{1}{2} c_{\text{eff}}(R)\, q^2 + (1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau] \left(1 - \phi(R)q\right) \pi D, \tag{16}$$

where $c_{\text{eff}}(R) = c_h(1-R)^2 + c_{AI}R^2$ and $\phi(R) = e(R)/e(0)$. Comparing (16) to the benchmark cost, AI enters through two modifications. First, the cost curvature $c_{\text{eff}}(R)$ replaces c_h, where any interior mix with $R > 0$ lowers the marginal cost of supplying an additional unit of total effort because $c_{AI} < c_h$. Second, the audit failure probability $1 - \phi(R)q$ replaces $1 - q$, where effort effectiveness $\phi(R)$ scales how powerfully each unit of total effort reduces the probability of missing a misstatement. For $R < R_{\text{peak}}$, expertise is rising and $\phi(R) > 1$, so each unit of effort is more effective than in the no-AI benchmark. For $R_{\text{peak}} < R < R_{\text{desk}}$ expertise has passed its peak but remains above the baseline $e(0)$, so $\phi(R) > 1$ still. For $R > R_{\text{desk}}$, deskilling in the strict sense has set in, i.e., $\phi(R) < 1$, where each unit of effort accomplishes less than in the no-AI benchmark. Minimizing (16) over effort $q \geq 0$ yields the unique unconstrained optimum

$$q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R) = \frac{(1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, \pi D\, \phi(R)}{c_{\text{eff}}(R)}. \tag{17}$$

The numerator once again captures the marginal private benefit of effort, $\pi D\phi(R)$, the marginal reduction in expected damage from lowering the audit failure probability, scaled by effectiveness. As in the benchmark in the absence of AI, $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)$ is strictly increasing in perceived engagement risk $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$.

Lemma 2. *Under a total-effort standard Q, the auditor's constrained effort choice is*

$$q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q) = \max\{Q,\, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)\},$$

with a unique binding threshold $\bar{\gamma}(Q)$ defined implicitly by $q^{**}(\bar{\gamma}(Q), R) = Q$. Under a human-effort standard Q_h, the constraint $q_h \geq Q_h$ translates via $q_h = (1-R)q$ into a total-effort requirement

$q \geq Q_h/(1-R)$, *so the constrained choice becomes*

$$q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q_h) = \max\{Q_h/(1-R),\, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)\},$$

with binding threshold $\hat{\tilde{\gamma}}(Q_h)$ *defined by* $q^{**}(\hat{\tilde{\gamma}}(Q_h), R) = Q_h/(1-R)$.

Lemma 2 shows the first instance at which the two instruments diverge. A total-effort floor Q and a human-effort floor $Q_h = (1-R)Q$ impose the same total-effort requirement for a given R, but only the human-effort floor has a requirement that tightens endogenously as R rises. We now derive welfare-optimal standards under the fixed AI share R. As in the benchmark, the competitive fee absorbs the auditor's expected litigation cost, so damage payments cancel from total welfare and the social loss per audit failure is I. With AI, the audit failure probability is $\mathbb{E}[\tilde{\gamma}|\Omega_\tau](1-\phi(R)q)$. Welfare for fixed R and total-effort standard Q is therefore

$$W(Q,R) = \underbrace{pG - I}_{W_0} + \mathbb{E}_{\tilde{F}_R}\Big[(1-p)\left(1 - \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\left(1-\phi(R)q^*\right)\right) I \;-\; \tfrac{c_{\text{eff}}(R)}{2}\, q^{*2}\Big], \tag{18}$$

where $q^* \equiv q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)$, $\tilde{F}_R$ places probability $1-e(R)$ on the uninformed state $\mathbb{E}[\tilde{\gamma}|\Omega_{un}] = \gamma_0$ and probability $e(R)$ on the informed state $\mathbb{E}[\tilde{\gamma}|\Omega_{in}] \sim F$, and the adoption cost $C_R(R)$ is omitted as it is constant in Q when R is exogenous. Because the standard is slack for $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}_R(Q)$, only the binding region contributes to the FOC dW/dQ. In this region the optimal effort is equal to the required floor, i.e., $q^* = Q$, so the social loss per failure is I. Differentiating (18) with respect to Q yields

$$\frac{dW}{dQ} = \underbrace{(1-p)\,\phi(R)\,I \int_0^{\bar{\gamma}_R} \mathbb{E}[\tilde{\gamma}|\Omega_\tau] d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])}_{\text{marginal social benefit, scaled by effort effectiveness}} - \underbrace{c_{\text{eff}}(R)\, Q\, F_R\big(\bar{\gamma}_R(Q)\big)}_{\text{marginal effort cost}}. \tag{19}$$

Comparing eq. (19) to the respective FOC in the benchmark in eq. (14), two differences emerge. First, I is now scaled by effort effectiveness $\phi(R)$ as each unit of mandated effort averts social loss $\phi(R)\,I$ rather than I, depending on where R sits relative to R_{peak} and R_{desk}. For $R < R_{\text{desk}}$, the benefit of tighter standards is amplified and for $R > R_{\text{desk}}$, it is attenuated. Second, the cost term carries $c_{\text{eff}}(R)$ rather than c_h, which makes compliance with the effort floor cheaper under AI. The absence of D from the benefit coefficient follows from the competitive fee absorbing litigation costs, as in the benchmark. Strict concavity of $W(Q,R)$ in Q continues to hold, so the first-order condition characterizes a unique optimum.

Lemma 3. *For a fixed AI share $R \in [0,1)$, the unique welfare-maximizing total-effort standard $Q^*(R)$ satisfies the following equality*

$$(1-p)\,\phi(R)\,I\int_0^{\bar{\gamma}_R^*}\mathbb{E}[\tilde{\gamma}|\Omega_\tau]d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = c_{\mathit{eff}}(R)\,Q^*(R)\,F_R(\bar{\gamma}_R^*), \tag{20}$$

where $\bar{\gamma}_R^* \equiv \bar{\gamma}_R\big(Q^*(R)\big)$ and the unique welfare-maximizing human-effort standard is

$$Q_h^*(R) = (1-R)\,Q^*(R). \tag{21}$$

Lemma 3 shows that the optimality condition in eq. (20) has the same structure as the benchmark condition in eq. (15), with two modifications. The benefit side carries $\phi(R)$ scaling the social loss I, and the cost side uses $c_{\text{eff}}(R)$ reflecting the cheaper AI-assisted effort. The human-effort standard $Q_h^*(R) = (1-R)Q^*(R)$ expresses the total-effort requirement in human-work units. Setting $R = 0$ in (20) recovers the benchmark optimum from Lemma 1, since $\phi(0) = 1$ and $c_{\text{eff}}(0) = c_h$. Thus, the adoption of AI influences the benefits and the cost of a tighter effort standard. First, AI lowers the marginal cost of compliance as AI-assisted effort is cheaper, $c_{\text{eff}}(R) < c_h$. As such, the regulator can mandate a higher audit effort level all else equal.

Second, AI affects the benefits of tighter audit standards, depending on the degree of AI adoption. For $R < R_{\text{desk}}$ and $\phi(R) > 1$ the benefit of tighter standards is amplified because each unit of effort is more effective at detecting misstatements. AI also impacts the information available to the auditor. A higher level of expertise $e(R)$, as a consequence of the freed up capacity by AI, raises the probability that the auditor receives an informative signal about engagement-specific risk, so a larger share of binding engagements is selected on the basis of posterior risk assessments rather than the unconditional prior γ_0. The integral $\int_0^{\bar{\gamma}_R}\mathbb{E}\tilde{\gamma}|\Omega_\tau]\,d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ then aggregates the marginal benefit across engagements whose risk has been more sharply identified, so mandated effort is directed where its social return is highest. Together, the effectiveness and information channels amplify the benefit of tighter standards. In contrast, for $R > R_{\text{desk}}$ and $\phi(R) < 1$ the benefit is attenuated, as effort is less effective and less targeted due to the lack of expertise induced by deskilling. These effects work against the cost channel.

For the human-effort standard, $Q_h^*(R) = (1-R)Q^*(R)$ carries the additional factor $(1-R) < 1$. Because AI now performs a share R of the work that previously had to be done by humans, less human effort is needed to achieve the same total audit output. Even when the cost and effectiveness channels push the optimal floor $Q^*(R)$ above its benchmark level, the optimal human-effort floor

$Q_h^*(R)$ can therefore lie below $Q_h^*(0)$ for intermediate values of R. This comparison proves essential in the next subsection, where R is endogenized.

3.3 Endogenous AI Adoption

We now close the model with AI adoption itself as an endogenous choice of the auditor. This final step introduces the adoption friction. When the auditor selects R in response to the regulatory standard, the standard no longer merely constrains effort but also shapes the composition of that effort by changing the marginal value of substituting toward AI. Whether this feedback raises or lowers audit quality and welfare as compared to a world without AI depends critically on which instrument the regulator uses. This subsection characterizes the equilibrium along three dimensions of interest: the AI share R^*, audit quality $\mathcal{A}$, and social welfare W.

Optimal AI share. At date 0, before the engagement's risk is realized, the auditor anticipates the date-1 audit effort policy $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)$ derived in Section 3.2 and the mixture distribution $\tilde{F}_R$ that governs engagement-specific risk assessments at expertise level $e(R)$. Taking both as functions of R, she chooses the AI share to minimize total expected cost per engagement,

$$R^*(Q) \in \arg\min_{R\in[0,1]} \mathbb{E}_{\tilde{F}_R}\Big[C^A\big(q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q), R, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\big)\Big] + C_R(R), \tag{22}$$

where $C_R(R) = \frac{c_R}{2}R^2$ is the AI adoption cost incurred at date 0. The same problem applies under a human-effort standard Q_h, replacing $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)$ with $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q_h)$ throughout. In equilibrium, any AI share that fails to minimize expected cost would require a higher fee than a cost-minimizing rival and would therefore not be chosen. The adoption problem in eq. (22) embeds the date-1 best response $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)$, which depends on R through both $c_{\text{eff}}(R)$ and $\phi(R)$, and the distribution $\tilde{F}_R$, which depends on R through $e(R)$. We now establish that the optimal AI share $R^*(Q)$ is well-defined, interior, and characterized by a first-order condition.

Lemma 4. *The adoption objective $J(R) \equiv \mathbb{E}_{\tilde{F}_R}[\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)] + \frac{c_R}{2}R^2$, where $\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q) \equiv C^A\big(q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q), R, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\big)$ denotes the minimized per-engagement cost, is continuously differentiable in R on $(0,1)$. Provided adoption costs rise fast enough with R to offset the non-convexity introduced by the hump-shaped expertise function, formally*

$$c_R > -\min_{R\in[0,1]} \frac{\partial^2}{\partial R^2}\,\mathbb{E}_{\tilde{F}_R}\big[\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)\big], \tag{23}$$

$J(R)$ is strictly convex, and the optimal $R^*(Q)$ is unique, interior, and characterized by

$$\underbrace{\tfrac{1}{2}c'_{\mathit{eff}}(R)\,\mathbb{E}_{\tilde{F}_R}\big[q^{*2}\big]}_{\textit{cost-substitution}} - \underbrace{(1-p)\,\pi D\,\phi'(R)\,\mathbb{E}_{\tilde{F}_R}\big[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,q^*\big]}_{\textit{effectiveness}} + \underbrace{e'(R)\,\Delta\Lambda(R,Q)}_{\textit{information}} + \underbrace{c_R R}_{\textit{adoption cost}} = 0, \quad (24)$$

where $\Delta\Lambda(R,Q) \equiv \mathbb{E}_F[\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q)] - \Lambda(\gamma_0,R,Q)$ is the informed-minus-uninformed gap in per-engagement cost. The derivation, including explicit expressions for $c'_{\mathit{eff}}(R)$, $\phi'(R)$, and $e'(R)$, is provided in Appendix A.7.

Each term in the FOC presented in eq. (24) corresponds to a distinct economic force. The cost-substitution term is negative for R below the pure cost-minimizing mix and positive beyond it as shifting toward AI reduces total cost curvature and raises it thereafter. The effectiveness term is positive for $R < R_{\text{peak}}$ where $\phi'(R) > 0$, where each unit of effort becomes more effective as expertise rises, and negative for $R > R_{\text{peak}}$ where $\phi'(R) < 0$ and deskilling erodes the detection value of effort. The information term captures how shifting R changes the expertise-weighted mixture distribution. When $e'(R) > 0$, more probability mass shifts to the informed state where the auditor observes $\mathbb{E}[\tilde{\gamma}|\Omega_{in}] \sim F$. The difference in the auditor's minimized per engagement cost between the informed and uninformed state is negative, $\Delta\Lambda < 0$, because informed auditors target effort more precisely and incur lower expected costs. The unique equilibrium AI share $R^*(Q)$ is the interior point at which these forces balance.

We now ask how $R^*(Q)$ responds to tighter total-effort floors. A tighter standard Q expands the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \le \bar{\gamma}_R(Q)\}$ on which the auditor must supply a minimum of Q units of total effort regardless of the engagement's risk profile. On those engagements the only cost-reduction lever is to shift the fixed total effort toward cheaper AI, raising the marginal value of adoption and making the floor and the AI share strategic complements. To formalize this, let $J(R) \equiv \mathbb{E}_{\tilde{F}_R}\big[\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q)\big] + \frac{1}{2}c_R R^2$ denote the auditor's total expected cost as a function of the AI share. Since $R^*(Q)$ minimizes $J(R)$, applying the implicit function theorem to the adoption FOC in eq. (24) yields

$$\frac{dR^*}{dQ} = -\frac{\partial^2 J/\partial R\,\partial Q}{\partial^2 J/\partial R^2}, \quad (25)$$

where the denominator is positive by strict convexity, so the sign of dR^*/dQ equals the sign of $-\partial^2 J/\partial R\,\partial Q$. On the slack region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}_R(Q)\}$, effort does not depend on the floor Q and the cross-partial is zero by the envelope theorem. Only the binding region contributes, where the

optimal audit effort $q^* = Q$ and the cross-partial reduces to

$$\frac{\partial^2 J}{\partial R\,\partial Q} = c'_{\text{eff}}(R) Q F_R\big(\bar{\gamma}_R(Q)\big) - (1-p)\,\pi D\,\phi'(R) \int_0^{\bar{\gamma}_R} \mathbb{E}[\tilde{\gamma}|\Omega_\tau] d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]). \tag{26}$$

The first term captures how a tighter floor amplifies the cost-substitution motive for AI adoption. On binding engagements the auditor must supply exactly Q units of effort, and the cost of doing so is governed by $c_{\text{eff}}(R)$. A marginal increase in Q raises this cost on all binding engagements and the sensitivity of that cost to the AI share is $c'_{\text{eff}}(R)$. When the auditor sits below the cost-minimizing share R^{cost}, the cost curvature is still declining in AI share R as $c'_{\text{eff}} < 0$. So tightening the floor makes AI adoption more attractive. The second term captures the effectiveness of audit effort. A marginal increase in Q also changes the social benefit of audit effort on binding engagements, and the sensitivity of that benefit to R runs through $\phi'(R)$, the rate at which effort effectiveness changes with the AI share. When the auditor is below R_{peak}, expertise is still rising ($\phi' > 0$), so this term enters with negative sign.

Total-effort floors and AI adoption reinforce each other. As the floor tightens, a larger share of engagements becomes binding, and on each binding engagement the only lever for lowering the cost of mandated effort is to substitute toward cheaper AI. A standard meant to raise audit work therefore also shifts the composition of that work toward machines.

Proposition 1. *Under the conditions laid out in Lemma 4, the equilibrium AI share $R^*(Q)$ is continuously differentiable in Q, with*

$$\frac{dR^*}{dQ} = -\frac{c'_{eff}(R^*) Q F_R(\bar{\gamma}_R) - (1-p)\,\pi D\,\phi'(R^*) \int_0^{\bar{\gamma}_R} \mathbb{E}[\tilde{\gamma}|\Omega_\tau] d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])}{\partial^2 J/\partial R^2\big|_{R=R^*}}.$$

For $R^* < \min(R^{cost}, R_{peak})$, both terms in the numerator are negative, so $dR^*/dQ > 0$, meaning that tighter total-effort standards induce strictly greater AI adoption. For $R^* > R_{peak}$, the effectiveness term reverses sign and the comparative static is ambiguous, but $dR^*/dQ > 0$ continues to hold whenever the cost channel dominates.

3.4 Comparing regulatory instruments

3.4.1 Implications for AI adoption

When we compare the two regulatory instruments, a total effort standard or a human effort standard, there is an important structural difference. A total-effort floor Q is independent of who performs the work, while a human-effort floor Q_h depends on the auditor's AI share through $q_h = (1 - R)q$. This creates two consequences. First, a unit-conversion effect, i.e., implementing the same total work as in the no-AI benchmark requires a human-effort floor of $(1 - R)$ times the total-effort benchmark. Second, a conservative-adoption effect, since the human-effort floor transforms into the total-effort requirement $q \geq Q_h/(1-R)$, any increase in R the auditor contemplates at date 0 tightens the constraint she faces at date 1. This creates a self-limiting force on AI adoption absent from the total-effort standard.

At low values of Q_h the floor is non-binding for most engagements and the auditor's AI share equals the unregulated optimum. However, as Q_h rises and the floor begins to bind, AI adoption is penalized. Each additional unit of R amplifies the total-effort requirement, raising compliance costs rather than reducing them. As Q_h rises, $R_h^*(Q_h)$ continues to increase across the full range $Q_h \in [0, 1]$, but remains strictly bounded below the expertise-maximizing share R_{peak}. However, this effect keeps AI adoption sufficiently costly such that the auditor never reaches the deskilling region, even as the floor tightens to its maximum. On the downside, the auditor also never reaches the full expertise-maximizing AI share R_{peak}. Thus, the human-effort floor imposes an efficiency cost that goes beyond preventing deskilling as it also prevents the auditor from fully exploiting the beneficial region $R < R_{\text{peak}}$ where AI enhances expertise. This cost is the price of the self-limiting property that makes human-effort floors effective at curbing deskilling.

Proposition 2. *Under the conditions of Lemma 4, there exists a threshold $Q_h^{bind} \in [0, 1)$ such that the equilibrium AI share $R_h^*(Q_h)$ under a human-effort floor satisfies:*

(i) ***Non-binding interval.*** *$R_h^*(Q_h) = 0$ for all $Q_h \in [0, Q_h^{bind}]$ and $R_h^*(Q_h) > 0$ for $Q_h \in (Q_h^{bind}, 1]$.*

(ii) ***Comparative static and monotonicity.*** *On the binding region $(Q_h^{bind}, 1]$, R_h^* is continuously differentiable, with dR_h^*/dQ_h given by the implicit function theorem applied to (24). The derivative is strictly positive whenever*

$$(1 - p)\,\pi D\left[\phi(R_h^*) + (1 - R_h^*)\,\phi'(R_h^*)\right] \int_0^{\bar{\gamma}^h} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, d\tilde{F}_R \;>\; 2\, c_{AI}\, R_h^*\, \tilde{Q}\, \tilde{F}_R(\bar{\gamma}^h), \tag{C1}$$

where $\tilde{Q} \equiv Q_h/(1-R_h^*)$ *and* $\bar{\gamma}^h \equiv \bar{\gamma}_{R_h^*}(\tilde{Q})$. *Combined with (i),* R_h^* *is weakly monotone on* $[0,1]$ *whenever* (C1) *holds throughout* $(Q_h^{bind}, 1]$.

(iii) ***Strict boundedness below*** R_{peak}**.** $R_h^*(Q_h) < R_{peak} \equiv \alpha/(2\beta)$ *for every* $Q_h \in [0,1]$, *provided*

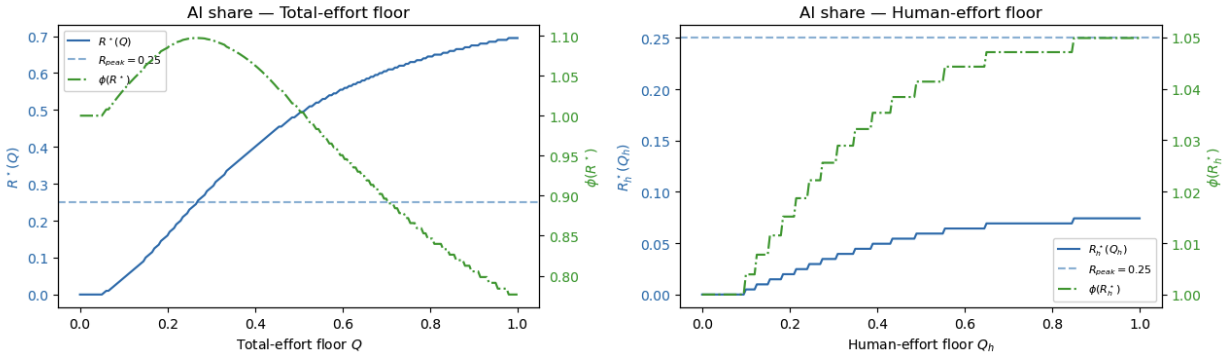
$$(1-p)^2\,\phi(R_{peak})^2\,\pi^2 D^2\,\mathbb{E}_{\tilde{F}_{R_{peak}}}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]]^2 \;<\; 4\,c_{AI}\,c_R\,R_{peak}^2\,(1-R_{peak}). \tag{C2}$$

The full cross-partial expression and derivation are provided in Appendix A.10.

Proposition 2 exploits the fact that under the human-effort standard the auditor's date-0 adoption problem is isomorphic to the total-effort problem at an effective total-floor $\tilde{Q}(R, Q_h) \equiv Q_h/(1-R)$. Thus, the binding thresholds coincide and the per-engagement cost Λ agrees on each region. The human-floor adoption FOC therefore equals the total-floor FOC (24) evaluated at $Q = \tilde{Q}$, plus the contribution of the conservative-adoption effect that captures how a marginal increase in R tightens the total-effort requirement $q \geq Q_h/(1-R)$. Two sufficient conditions, labeled (C1) and (C2), pin down strict monotonicity on the binding region and strict boundedness below R_{peak}, respectively. Condition (C2) is substantive and holds strictly for low damage payments but fails for large damage payments, where the cost-substitution motive becomes strong enough to push the auditor's equilibrium AI adoption toward the deskilling region $R > R_{\text{peak}}$.

Figure 3 visualizes Propositions 1 and 2. The left panel plots $R^*(Q)$ (solid blue) as a function of the total-effort floor. At $Q = 0$, the auditor's unregulated AI share is zero. Absent a binding floor, the cost savings from AI adoption are too small to justify the adoption cost and the effort effectiveness loss, so the auditor relies entirely on human effort. As Q rises, the floor begins to bind on a growing share of engagements and $R^*(Q)$ increases. On binding engagements the auditor must supply a minimum of Q units of effort regardless of the engagement's risk, and the only lever for reducing the cost of this mandated effort is to shift toward cheaper AI. Beyond this transition, tighter regulation and AI adoption are self-reinforcing: each increment in Q enlarges the binding pool, raises the marginal value of cheap AI effort, and increases R^*, which in turn pushes the auditor deeper into the deskilling region where effort effectiveness $\phi(R)$ is declining. At $Q = 1$, the equilibrium AI share reaches its maximum far above both R_{peak} and the deskilling threshold, in a region where expertise has deteriorated significantly. The right panel plots $R_h^*(Q_h)$ as a function of the human-effort floor. On the non-binding interval the auditor's AI share is zero, once the floor binds, R_h^* rises but remains strictly bounded below R_{peak} across the full range. AI adoption remains sufficiently costly that the auditor never reaches the deskilling region, even as the floor tightens to its maximum.

Figure 3: Equilibrium AI adoption under effort floors.



Left: equilibrium AI share $R^*(Q)$ as a function of the total-effort floor Q. Right: equilibrium AI share $R_h^*(Q_h)$ as a function of the human-effort floor Q_h. Twin axes show effort effectiveness $\phi(R^*)$ and $\phi(R_h^*)$; the dashed horizontal line marks the expertise-maximizing share R_{peak}. We set parameters such that $p = 0.6$, $I = 1$, $\gamma_0 = 0.35$, $c_h = 1$, $c_{AI} = 0.2$, $e(0) = 0.6$, $\pi = 1$, $\alpha = 2$, $\beta = 4$, $c_R = 0.2$, $D = 0.1$.

3.4.2 Implications for audit quality

A further implication of Proposition 2 is that regulating human effort alone raises total effort. We now formalize how audit quality is affected by the tightening of audit standards in a context of endogenous AI shares. Define ex-ante audit quality under a total-effort floor Q as

$$\mathcal{A}(Q) = 1 - (1-p)\,\mathbb{E}\Big[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\big(1 - \phi(R^*(Q))\,q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(Q), Q)\big)\Big], \tag{27}$$

where the expectation integrates over both the informed and uninformed states, weighting the informed state by expertise $e(R^*(Q))$. Differentiating with respect to total effort floor Q yields

$$\frac{d\mathcal{A}}{dQ} = \underbrace{\frac{\partial \mathcal{A}}{\partial Q}\Big|_R}_{\substack{\text{direct binding} \\ \text{effect}>0}} + \underbrace{\frac{\partial \mathcal{A}}{\partial R}\Big|_Q \cdot \frac{dR^*(Q)}{dQ}}_{\text{adoption feedback}}. \tag{28}$$

The direct binding effect $\partial\mathcal{A}/\partial Q|_R > 0$ is unambiguously positive. Holding R fixed, a tighter floor raises effort on binding engagements and reduces the audit failure probability $\mathbb{E}[\tilde{\gamma}|\Omega_\tau](1 - \phi(R)q)$ by $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\phi(R)$ per engagement. The adoption feedback $(\partial\mathcal{A}/\partial R|_Q)(dR^*(Q)/dQ)$ has an ambiguous sign depending on where $R^*(Q)$ sits relative to R_{peak}. For $R < R_{\text{peak}}$, higher expertise sharpens risk targeting and both forces of the adoption feedback reinforce the direct binding effect. For $R > R_{\text{peak}}$, deskilling worsens risk targeting, shifts more engagements toward the uninformed state, and raises the expected miss probability. When this feedback is large enough to dominate the direct binding effect, which occurs at high values of Q once $R^*(Q)$ has moved deep into the deskilling region, the audit quality effect of tightening the standard turns negative and the

benchmark eventually overtakes the AI world's audit quality level.

Proposition 3. ***(Audit quality under human-effort floors.)*** *Under the conditions of Proposition 2 and the mild structural condition $R_{peak} \leq 2R^{cost}$, the audit quality under a human-effort floor weakly exceeds the no-AI benchmark:*

$$\mathcal{A}^{(h)}(Q_h) \;\geq\; \mathcal{A}_{bench}(Q_h) \qquad \text{for every } Q_h \in [0,1], \tag{29}$$

with strict inequality on the binding region $(Q_h^{bind}, 1]$. Here $\mathcal{A}_{bench}(Q_h)$ is the audit quality from Section 3.1 evaluated at floor Q_h (i.e., under $R = 0$, $\phi(0) = 1$, and $q = q_h$), and $\mathcal{A}^{(h)}(Q_h)$ is the equilibrium audit quality of Section 3.3 under a human-effort floor with adoption $R_h^*(Q_h)$. The derivation is in Appendix A.11.

The same force that sustains Proposition 3 has a counterpart. Proposition 2(iii) pins R_h^* strictly below R_{peak}, so the auditor never reaches the region where AI is most cost-effective and most expertise-enhancing. The full potential of AI thus remains underexploited, resulting in smaller welfare gains and audit quality below its feasible maximum. The human-effort floor preserves quality by foregoing resources that AI could otherwise free up.

By contrast under a total-effort floor as Figure 3 showed, the complementarity between regulation and adoption eventually undermines audit quality. As the floor tightens, the auditor shifts further toward AI and enters the deskilling region, where each additional unit of effort is less effective at detecting misstatements than in the no-AI benchmark. When the loss of detection power outweighs the direct gain from higher mandated effort, audit quality falls below the benchmark level. Proposition 4 formalizes this result.

Proposition 4. ***(Audit quality crossover under total-effort floors.)*** *Under the conditions of Lemma 4, the baseline-corner property $R^*(0) = 0$, and the two substantive conditions*

$$R^*(1) \;>\; R_{desk}, \tag{C3}$$

$$\bar{\gamma}_{\max}\,(1-p)\,\pi D \;\leq\; \min\left\{ c_h,\; \frac{c_{\mathit{eff}}(R^*(1))}{\phi(R^*(1))} \right\}, \tag{C4}$$

there exists a threshold $Q^{quality} \in [0,1)$ such that the equilibrium audit quality under a total-effort floor lies strictly below the no-AI benchmark:

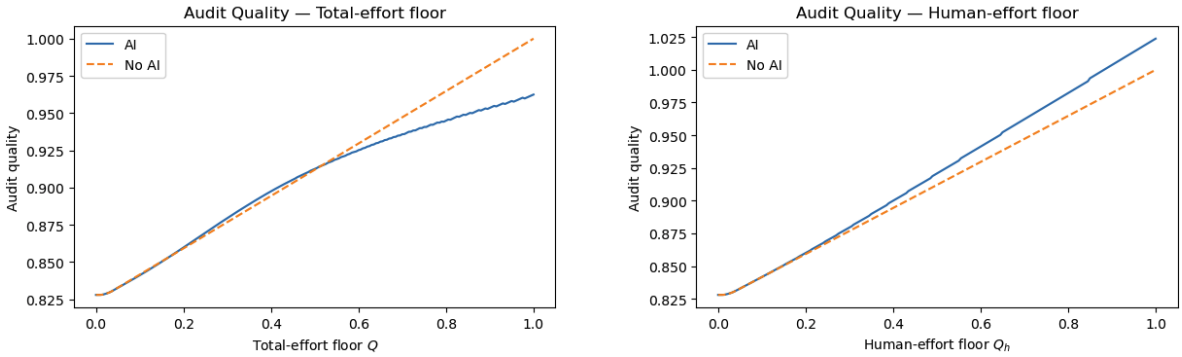
$$\mathcal{A}(Q) \;<\; \mathcal{A}_{bench}(Q) \qquad \text{for every } Q \in (Q^{quality}, 1]. \tag{30}$$

Here $\mathcal{A}(Q)$ is the equilibrium audit quality of Section 3.3 under a total-effort floor with adoption $R^*(Q)$, and $\mathcal{A}_{bench}(Q)$ is the audit quality from Section 3.1 evaluated at floor Q (i.e., under $R = 0$, $\phi(0) = 1$, and $q = q_h$). The derivation is in Appendix A.12.

Condition (C3) asserts that equilibrium adoption at the maximum total-effort floor lies strictly beyond the deskilling threshold R_{desk}. Condition (C4) is a mild saturation condition: at the maximum floor $Q = 1$, the floor binds on every engagement in both worlds.

Figure 4 visualizes Propositions 3 and 4. The left panel reveals that welfare dominance for tighter standards does not necessarily extend to audit quality under a total-effort floor. At $Q = 0$, the AI world (solid blue) and no-AI benchmark (dashed orange) coincide because the auditor adopts no AI when the floor is non-binding. As Q rises from zero, both quality curves increase, but the AI curve initially rises more steeply: without the deskilling feedback, each increment in the floor translates directly into higher detection probability at the full effectiveness $\phi = 1$. In the AI world, the rising $R^*(Q)$ then progressively erodes $\phi(R^*(Q))$ as the auditor moves deeper into the deskilling region, so the quality gain per unit of additional mandated effort is attenuated. The benchmark audit quality overtakes the AI world at an intermediate threshold: even moderate AI adoption pushes the auditor past R_{peak} and into the deskilling region, so that effort effectiveness begins to deteriorate before the cost savings from AI have had much chance to accumulate. The right panel delivers the sharpest contrast: audit quality $\mathcal{A}^{(h)}(Q_h)$ in the AI world (solid blue) lies strictly above the no-AI benchmark (dashed orange) at every value of Q_h. Two reinforcing mechanisms drive this result. First, the bounded $R_h^*(Q_h)$, which stays strictly below R_{peak}, keeps effort effectiveness $\phi(R_h^*)$ close to or above its baseline value throughout, so each unit of effort retains larger detection power than under a total-effort floor. Second, the denominator amplification means the auditor supplies total effort $Q_h/(1 - R_h^*)$, which exceeds Q_h, as the AI component provides additional detection capacity at low marginal cost on top of the mandated human component. The deskilling feedback that drives the crossover in the left panel is never triggered. A regulator focused solely on quality metrics would conclude that AI is harmful under moderate and high binding total-effort floors, while a welfare-minded regulator would recognize that the cost savings from AI dominate even as quality deteriorates: welfare and audit quality point in opposite directions under total-effort regulation, while under human-effort regulation the regulator faces no such tension at stringent standards but underexploits AI resource savings and thus forgoes potential welfare and audit quality gains.

Figure 4: Audit quality under effort floors.



Left: audit quality $\mathcal{A}(Q)$ in the AI world (solid) and the no-AI benchmark $\mathcal{A}_{\text{bench}}(Q)$ (dashed). Right: $\mathcal{A}^{(h)}(Q_h)$ versus $\mathcal{A}^{(h)}_{\text{bench}}(Q_h)$. We set parameters such that $p = 0.6$, $I = 1$, $\gamma_0 = 0.35$, $c_h = 1$, $c_{AI} = 0.2$, $e(0) = 0.6$, $\pi = 1$, $\alpha = 2$, $\beta = 4$, $c_R = 0.2$, $D = 0.1$. Same calibration as Figure 3.

3.4.3 Implications for welfare

We now turn to the welfare comparison. Total welfare as a function of the effort floor is defined as

$$W(Q) \equiv W\big(Q, R^*(Q)\big) - C_R\big(R^*(Q)\big), \tag{31}$$

where $W(Q, R)$ is the fixed-R welfare from (18), which already embeds W_0, and $C_R(R^*(Q))$ is the adoption cost subtracted as a real resource expenditure. Differentiating yields

$$\frac{dW}{dQ} = \underbrace{\frac{\partial W}{\partial Q}\Big|_R}_{\text{direct binding effect}} + \underbrace{\Big(\frac{\partial W}{\partial R}\Big|_Q - C_R'\big(R^*(Q)\big)\Big)\frac{dR^*(Q)}{dQ}}_{\text{adoption wedge}}. \tag{32}$$

The first term is the direct binding effect from eq. (19). The second is the adoption wedge, where $\partial W/\partial R|_Q - C_R'(R^*(Q))$ measures the net social value of a marginal increase in AI adoption, combining welfare effects of changing $c_{\text{eff}}(R)$ and $e(R)$ with the direct adoption cost $c_R R^*$. When the optimal share $R^*(Q)$ lies in the deskilling region, $\partial W/\partial R|_Q$ is negative and the adoption wedge is negative overall as further tightening the floor induces more AI adoption, raises adoption costs, erodes expertise, and increases failure losses, partially or fully offsetting the direct binding effect.

The auditor's adoption choice minimizes her own cost, but welfare depends on social cost, which exceeds private cost whenever an audit failure harms third parties. The two differ by a wedge that scales with the audit-quality gap between the AI world and the benchmark. When quality improves, welfare improves by more than the auditor's own savings, and when quality deteriorates, the wedge works against welfare and the auditor's cost savings must be large enough to compensate. Proposition 5 formalizes the results plotted in Figure 5.

Proposition 5. ***(Welfare dominance.)***

(a) ***Total-effort floor.*** *Under the conditions of Lemma 4 together with* $R^*(0) = 0$*,*

$$W(Q) \;\geq\; W_{bench}(Q) \qquad \text{for every } Q \in [0,\, Q^{quality}], \tag{33}$$

with strict inequality on $(Q^{bind}, Q^{quality}]$*. If conditions (C3) and (C4) hold and*

$$\frac{c_h - c_{eff}(R^*(1))}{2} \;>\; (1-p)\,\gamma_0\left(1-\phi(R^*(1))\right) I \;+\; \frac{c_R}{2}\,R^*(1)^2, \tag{C5}$$

then there exists $Q^{wel} \in [Q^{quality}, 1)$ *such that* $W(Q) \geq W_{bench}(Q)$ *for every* $Q \in [Q^{wel}, 1]$*, with strict inequality on* $(Q^{wel}, 1]$*.*

(b) ***Human-effort floor.*** *Under the conditions of Proposition 3,* $W^{(h)}(Q_h) \geq W_{bench}(Q_h)$ *for every* $Q_h \in [0, 1]$*, with strict inequality on* $(Q_h^{bind}, 1]$*.*

The derivation is in Appendix A.15.

The asymmetry between parts (a) and (b) reflects the same economic mechanism that drives Propositions 3 and 4. Under the human-effort floor, the conservative-adoption effect caps R_h^* strictly below R_{peak}, so audit quality is preserved and welfare dominance follows on every engagement at every floor level. Under the total-effort floor, equilibrium adoption can move into the deskilling region where audit quality drops below the benchmark and welfare dominance then requires that the cost savings on binding engagements exceed the deskilling-induced detection loss plus the adoption cost. Condition (C5) is precisely this comparison, evaluated at the maximum floor $Q = 1$ where the binding pool covers every engagement under (C4). Corollary 1 establishes the ordering of the welfare gap between the no-AI benchmark and the AI world across both the total effort floor and the human effort floor.

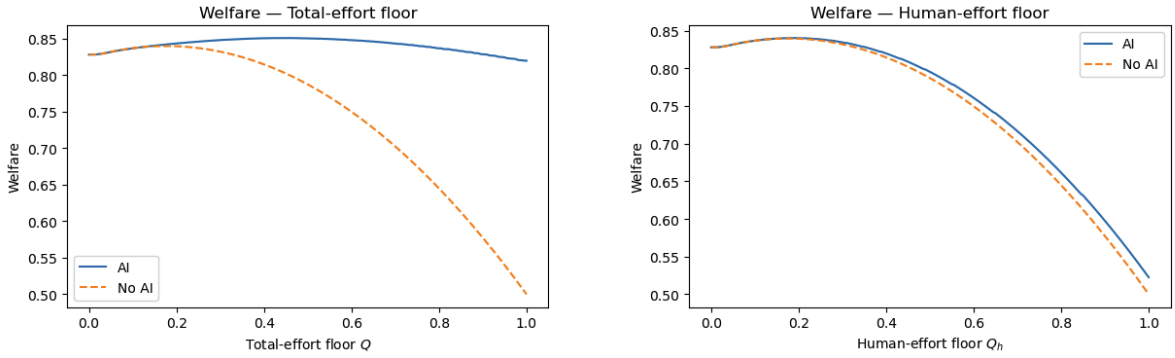
Corollary 1. ***(Welfare-gap ordering across instruments — baseline observation.)*** *At the baseline calibration of Figure 5, the welfare gain of the AI world over the no-AI benchmark is strictly larger under the total-effort floor than under the human-effort floor at every common floor level* $Q = Q_h$ *on which both gaps are positive:*

$$W(Q) - W_{bench}(Q) \;>\; W^{(h)}(Q) - W_{bench}(Q) \qquad \text{at every } Q = Q_h \geq \max\{Q^{wel},\, Q_h^{bind}\},$$

with the gap widening as the floor tightens.

Figure 5 visualizes Proposition 5 and Corollary 1. The left panel plots expected welfare under a total-effort floor. At $Q = 0$ the two worlds coincide at the margin, and the small welfare gap at low Q reflects the region where the AI world adopts a positive share and incurs adoption cost $C_R(R^*)$ before real cost savings materialize. As Q rises, the benchmark welfare drops steeply because mandated effort is expensive at the full human cost c_h and the quadratic effort cost grows rapidly on binding engagements, while the AI world's welfare declines more gradually because compliance costs grow at the lower rate $c_{\text{eff}}(R^*) \ll c_h$. The two lines separate early and the gap widens as Q increases, revealing that even modest regulatory pressure suffices to make AI cost-effective: once the floor begins to bind, cheaper AI-assisted effort generates cost savings that quickly compensate for the adoption cost. The right panel plots welfare under a human-effort floor. As under the total-effort floor, the two worlds coincide for low Q_h, and both curves decline as the floor tightens, with the benchmark dropping more steeply for the same reason. The AI world's welfare declines more gradually because the auditor retains access to cheaper AI-assisted effort, and the conservative-adoption effect of Proposition 2 simultaneously limits R_h^*, which contains the adoption cost and keeps effort effectiveness closer to its peak. Comparing the two panels, the total-effort floor sustains the larger welfare gain because it allows the auditor to reach and surpass R_{peak}, fully exploiting the expertise-enhancing potential of AI, whereas the conservative-adoption penalty under the human-effort floor caps the beneficial AI share. The widening gap between the two regimes is the visual content of Corollary 1.

Figure 5: Expected welfare under effort floors.



Left: expected welfare $W(Q)$ in the AI world (solid) and the no-AI benchmark $W_{\text{bench}}(Q)$ (dashed) as functions of the total-effort floor Q. Right: $W^{(h)}(Q_h)$ versus $W^{(h)}_{\text{bench}}(Q_h)$ as functions of the human-effort floor Q_h. We set parameters such that $p = 0.6$, $I = 1$, $\gamma_0 = 0.35$, $c_h = 1$, $c_{AI} = 0.2$, $e(0) = 0.6$, $\pi = 1$, $\alpha = 2$, $\beta = 4$, $c_R = 0.2$, $D = 0.1$. Same calibration as Figure 3.

Taken together, the comparison reveals a trade-off that is relevant for the current policy debate. Total-effort floors allow the auditor to reach the expertise-maximizing AI share R_{peak} and beyond, fully exploiting the upside potential of AI while still setting a quality floor, but at the cost of enabling over-adoption and deskilling when the floor is stringent. Human-effort floors eliminate the deskilling risk and preserve a quality advantage at every level of the standard, but they do so by capping the AI share well below R_{peak}, thereby sacrificing the region where AI most improves audit effectiveness. Regulators choosing human-effort floors should therefore be aware that they are not merely preventing downside risk but also limiting the upside potential of AI adoption.

3.5 Comparative Statics

The preceding analysis established that tighter effort floors induce greater AI adoption and traced the resulting effects on audit quality and welfare. We now ask how the *level* of equilibrium AI adoption responds to two parameters that are central to the liability regime and to the auditor's expertise: the damage payment D and the baseline expertise $e(0)$. Both comparative statics follow from the same implicit function theorem argument as Proposition 1, applied to the adoption FOC (24) with D or $e(0)$ as the varying parameter.

Corollary 2. *Under the conditions of Lemma 4 with $R^* < R^{cost}$:*

(i) ***Damages and AI adoption.*** *If $R^* < R_{peak}$, then $dR^*/dD > 0$, i.e. higher damages induce strictly greater AI adoption. If $R^* > R_{peak}$, then $dR^*/dD < 0$ provided the binding-region effect dominates the slack-region effect; this dominance condition is automatically satisfied when the slack region is empty, which holds under the baseline calibration with $D = 0.1$.*

(ii) ***Baseline expertise and AI adoption.*** *Suppose $R^* < R_{peak}$. If $e(R^*) > \frac{1}{2}$, then $dR^*/de(0) < 0$ provided the effectiveness channel dominates the information channel. If $e(R^*) < \frac{1}{2}$, the sign is negative (substitution) when the effectiveness-to-information ratio exceeds the logistic-curvature term at $e(R^*)$, and positive (complementarity) when the inequality is reversed. When $R^* > R_{peak}$, the effectiveness channel reverses sign and the comparative static admits an analogous regime classification in $e(R^*)$.*

The dominance conditions and the explicit effectiveness-to-information ratio are stated in Appendix A.16.

From a policy perspective, Corollary 2 reveals an interaction between liability policy, regulatory instrument choice, and the auditor's position in the adoption space. Strengthening the litigation regime (raising D) encourages AI adoption below R_{peak} and, when the binding-region effect dominates, discourages it above R_{peak}, so damages act as a self-correcting force that pushes the auditor toward the expertise-maximizing mix. Whether the self-correction is socially desirable depends on the regulatory instrument in place. Under human-effort floors, where the conservative-adoption effect prevents over-adoption, higher damages steer the auditor toward a more effective technology mix and improve both welfare and quality. Under total-effort floors, where adoption is already excessive at stringent standards, the self-correcting force from part (i)(b) partially counteracts the over-adoption tendency, though the net effect depends on the auditor's distance from R_{peak}.

Part (ii) reveals that the interaction between baseline expertise and AI adoption is governed by the effectiveness-to-information ratio (A.36). Experienced auditors ($e > 1/2$) substitute away from AI as their expertise rises, because both the effectiveness and information channels reinforce substitution. Less experienced auditors ($e < 1/2$) exhibit substitution or complementarity depending on the liability environment. In high-liability regimes (large D, small $|\Delta\Lambda|$), the effectiveness channel dominates and substitution persists, while in low-liability regimes (small D, large $|\Delta\Lambda|$), the information channel dominates and complementarity emerges. In the deskilling region ($R^* > R_{\text{peak}}$), for moderately skilled auditors ($e(R^*) > 1/2$) higher baseline expertise increases AI adoption, because it dampens both the deskilling penalty and the information sensitivity to R. Investments in auditor training therefore have regime-dependent effects whose direction is pinned down by (A.36): for already-skilled auditors they moderate AI adoption, while for less-skilled auditors in low-liability environments they may accelerate it.

4 Discussion

4.1 Imperfect Inspection

The analysis so far assumes that the auditor always complies with the applicable floor. In practice, the regulator inspects only a fraction of engagements, so compliance incentives depend jointly on the floor and on the intensity of inspection.

Suppose the regulator inspects any given engagement with probability $f \in [0, 1]$. If an inspected engagement is found to violate the applicable floor, the auditor pays a penalty $\tilde{x}$, drawn independently across engagements from a uniform distribution on $[0, \bar{x}]$ with $\bar{x} > 0$. Penalty het-

erogeneity captures the fact that sanctions depend on engagement-specific factors like severity, prior history, and publicity, that are not contracted upon in advance. The auditor learns her realization of $\tilde{x}$ at the time of effort choice but after the fee has been set, so the fee reflects only the expected penalty, not its realization. In this environment the auditor's compliance decision reduces to a comparison between the cost savings from under-compliance and the expected penalty $f \cdot \mathbb{E}[\tilde{x} \mid \text{violation detected}]$. Because $\tilde{x}$ is uniform, a threshold in penalty realizations separates engagements on which the auditor complies from those on which she does not. Under a total-effort floor Q, the auditor faces a one-dimensional compliance decision on q, while under a human-effort floor Q_h, a one-dimensional compliance decision on q_h. In either case the set of compliant engagements is characterized by a cutoff in $\tilde{x}$, and the probability of compliance is strictly increasing in f. Raising f at a fixed nominal floor therefore tightens the *effective* floor, where inspection strength and auditing standard stringency act as substitutes, comparable to Gao and Zhang (2019).

For each instrument in isolation, the comparative statics of Propositions 3, 4, and 5 in the nominal floor therefore translate directionally into the same comparative statics in f at a fixed nominal floor. Tightening the standard and intensifying inspection operate through the same channel: both raise the effective floor faced by the auditor and shift effort and expertise choices accordingly. The welfare ranking established in Section 3.3 is preserved as long as inspection intensity is held comparable across the two instruments, with the magnitude of the welfare gap scaling continuously in f. Two implications follow. First, the recommendation of a human-effort floor derived in Section 3.3 presumes that inspection practice is capable of distinguishing human from AI work with sufficient reliability. Second, this is a technological rather than fundamental contingency. The regulator can support f_h through complementary investments in inspection practice such as AI explainability standards, maintenance of prompt and output logs, reviewer sign-off protocols, and documented human-check trails.

4.2 Audit Quality and the Objectives of Audit Oversight

The preceding analysis identifies a regulatory trade-off across instruments. A total-effort standard fully exploits AI's cost-saving potential but can erode audit quality through deskilling once stringent, while a human-effort standard preserves quality but caps AI adoption below the expertise-maximizing share, leaving the full potential of AI underexploited. This trade-off in turn reveals a striking tension between two objectives that audit regulators might reasonably pursue, maximizing audit quality and maximizing social welfare. A regulator who focuses exclusively on audit quality

would reach different conclusions about the desirability of AI in auditing and about the appropriate stringency of standards than one who weighs the full social cost of the audit engagement. In this section we discuss the implications of both the instrument trade-off and the resulting quality-welfare divergence for the design of audit oversight, with particular attention to the stated objectives of the PCAOB.

The PCAOB was established by the Sarbanes-Oxley Act of 2002 with the statutory mission of protecting investors and furthering the public interest in the preparation of informative, accurate, and independent audit reports. In practice, the PCAOB's work has been evaluated primarily through the lens of audit quality. Following prominent cases of audit failure, the public urges the PCAOB to tighten inspection program evaluating audit procedures, strengthen enforcement actions to target deficient audits, and to adjust strategic plans in terms of improvements in the quality of audit services. The implicit premise is that higher audit quality, understood as a greater probability of detecting and reporting material misstatements, is unambiguously desirable and that regulatory interventions which raise quality serve the public interest. Our model suggests that this premise deserves scrutiny. Audit quality is not costless to produce as it requires real resources in the form of auditor effort, and in an AI-augmented world it also depends on the technology mix the auditor adopts. A regulator who maximizes audit quality without regard to cost will set standards that may be too stringent from a welfare perspective, because the marginal unit of quality improvement comes at an effort cost that exceeds its social value. AI creates a gap between the cost of producing effort and the effectiveness of that effort. A quality-maximizing regulator who ignores this gap may set standards that are too tight, inducing over-adoption of AI that erodes the very expertise the standard was meant to harness.

When AI is available, the auditor in our model can produce a given level of detection probability at lower cost, but the quality metric does not reflect this saving. A regulator who evaluates outcomes solely through the quality lens sees only that AI-augmented audits eventually detect fewer misstatements than pure-human audits used to at the same nominal standard, and concludes that AI is harmful. A welfare-minded regulator sees additionally that the resources saved by AI-assisted compliance are substantial and that the net social value of the audit engagement is higher in the AI world despite the quality shortfall. This disconnect is particularly acute under total-effort floors. As shown in Figures 5 and 4, the welfare advantage of the AI world is large and growing for moderate floors, while the quality advantage of the benchmark emerges only at higher effort floors. A quality-focused regulator observing the quality difference might respond by restricting AI

adoption or raising the standard further, both of which would reduce welfare. The welfare gains induced by the total-effort standard substantially reduce the cost of auditing, which might come at the expense of lower audit quality.

5 Conclusion

This paper studies the optimal regulation of audit effort when auditors can adopt AI as an assisting technology. We build a model in which an auditor chooses a mix of human and AI effort, where AI is cheaper but its excessive use erodes the engagement-specific expertise that makes effort effective. A regulator can respond to AI adoption through two instruments, a total-effort standard, which mandates a minimum amount of audit work regardless of how it is performed, and a human-effort standard, which explicitly requires a minimum level of human involvement.

Our central finding is that the choice of regulatory instrument has first-order consequences for AI adoption, audit quality, and welfare. Under a total-effort standard, tighter regulation inadvertently encourages over-adoption of AI. As the effort floor rises, the cost advantage of cheap AI effort makes compliance easier, driving AI adoption well beyond the expertise-maximizing level. The resulting deskilling erodes audit quality to the point where quality in the AI world falls below the no-AI benchmark. Under a human-effort standard, this channel is blocked. The human effort floor naturally pushes adoption back toward the expertise-maximizing mix as the standard tightens, and audit quality remains strictly above the no-AI benchmark at every level of the standard. However, this protection itself reflects the central trade-off: by sacrificing the region where AI most improves audit effectiveness, the human-effort standard delivers smaller welfare gains and quality below its feasible maximum.

We also shed light on regulators with a predominantly quality-focused approach to audit oversight and discuss how this may systematically undervalue the efficiency gains from AI and overvalue regulatory stringency. A welfare-oriented complement to quality-based evaluation would provide a more complete basis for standard-setting in an era of rapid technological change.

A Appendix

Throughout the appendix we use the notation established in the main text. In particular, $c_{\text{eff}}(R) = c_h(1-R)^2 + c_{AI}R^2$, $\phi(R) = e(R)/e(0)$, $\pi = 1$, and the social loss per audit failure is I (since the competitive fee absorbs the auditor's litigation cost, making the damage transfer purely redistributive; see eq. (9)). For a given standard and AI share, $\tilde{F}_R$ denotes the mixture distribution that places probability $1 - e(R)$ on $\mathbb{E}[\tilde{\gamma}|\Omega_{un}] = \gamma_0$ and probability $e(R)$ on $\mathbb{E}[\tilde{\gamma}|\Omega_{in}] \sim F$, and the truncated first moment $\equiv \int_0^{\bar{\gamma}_R} m\, d\tilde{F}_R(m)$ and the truncated mass $F_R(\bar{\gamma}) \equiv \tilde{F}_R(\bar{\gamma})$.

A.1 Proof of Definition 1 (Expertise Function)

Proof. The expertise index is $g(R) = \log\big(e(0)/(1-e(0))\big) + \alpha R - \beta R^2$ with $\alpha, \beta > 0$.

Part (a). Differentiating twice:

$$g'(R) = \alpha - 2\beta R, \qquad g''(R) = -2\beta < 0.$$

Hence g is strictly concave in R.

Part (b). The logistic map $e(R) = 1/(1+\exp(-g(R)))$ is a strictly increasing, continuously differentiable function of $g(R)$. By the chain rule,

$$e'(R) = e(R)\big(1-e(R)\big)\, g'(R). \tag{A.1}$$

Since $e(R) \in (0,1)$ for all $R \in [0,1]$, the factor $e(R)(1-e(R)) > 0$, so $\text{sign}(e'(R)) = \text{sign}(g'(R))$.

Part (c). $g'(R) = 0$ at $R_{\text{peak}} = \alpha/(2\beta)$. Because g is strictly concave, $g'(R) > 0$ for $R < R_{\text{peak}}$ and $g'(R) < 0$ for $R > R_{\text{peak}}$. By Part (b), $e(R)$ is strictly increasing on $[0, R_{\text{peak}})$ and strictly decreasing on $(R_{\text{peak}}, 1]$. Since $\phi(R) = e(R)/e(0)$ is a positive scalar multiple of $e(R)$, it inherits the same monotonicity and the unique maximum at R_{peak}. We require $\alpha/(2\beta) \in (0,1)$, i.e. $\alpha < 2\beta$. □

A.2 Derivation of the Benchmark Effort FOC (Equation (10)) and Binding Threshold (Equation (12))

Proof. In the benchmark ($R = 0$), $\phi(0) = 1$ and $c_{\text{eff}}(0) = c_h$. The auditor's per-engagement cost is

$$C^A(q, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = \tfrac{c_h}{2}\, q^2 + (1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, (1-q)\, \pi D,$$

which is strictly convex in q since $\partial^2 C^A / \partial q^2 = c_h > 0$. Differentiating with respect to q:

$$\frac{\partial C^A}{\partial q} = c_h\, q - (1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, \pi D.$$

Setting to zero yields the unique unconstrained minimum

$$q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = \frac{(1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, \pi D}{c_h}. \tag{A.2}$$

Monotonicity in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$. $\partial q^{**} / \partial \mathbb{E}[\tilde{\gamma}|\Omega_\tau] = (1-p)\pi D / c_h > 0$, so q^{**} is strictly increasing in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$.

Binding threshold. Imposing $q^{**}(\bar{\gamma}) = Q_h$ in (A.2) and solving for $\bar{\gamma}$:

$$\bar{\gamma}(Q_h) = \frac{c_h\, Q_h}{(1-p)\, \pi D},$$

which is Equation (12). The denominator is strictly positive for any $D > 0$, so the threshold is well-defined for any $Q_h \geq 0$. Inspection shows $\bar{\gamma}(Q_h)$ is linear and increasing in Q_h (tighter floor expands the binding pool) and decreasing in D (larger damage raises unconstrained effort, shrinking the binding pool). □

A.3 Proof of Lemma 1 (Benchmark Optimal Standard)

Proof. The proof has three parts: derivation of dW/dQ_h, strict concavity, and comparative statics.

Part (a): Derivation of dW/dQ_h (Equation (14)).

Welfare (13) integrates over the mixture $\tilde{F}$. We separate the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}(Q_h)\}$ where $q^* = Q_h$ from the slack region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}(Q_h)\}$ where $q^* = q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$. On the slack region, effort does not depend on Q_h, so only the binding region contributes to dW/dQ_h.

On the binding region, welfare net of W_0 and the baseline $(1-p)I$ equals $-h(Q_h, \mathbb{E}[\tilde{\gamma}|\Omega_\tau])$

where

$$h(Q_h, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \equiv (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1-Q_h)\,I \;+\; \tfrac{c_h}{2}\,Q_h^2,$$

since the social loss per failure is I under the competitive fee (eq. (9)). Differentiating h with respect to Q_h:

$$\frac{\partial h}{\partial Q_h} = -(1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,I + c_h\,Q_h. \tag{A.3}$$

The upper limit $\bar{\gamma}(Q_h)$ also varies with Q_h. By the Leibniz rule, there is a boundary contribution $(-h(Q_h,\bar{\gamma}) + h^{\text{slack}}(Q_h,\bar{\gamma}))\cdot\bar{\gamma}'(Q_h)\cdot\tilde{f}(\bar{\gamma})$, where h^{slack} is the welfare integrand on the slack region evaluated at $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] = \bar{\gamma}(Q_h)$. At the threshold $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] = \bar{\gamma}(Q_h)$, $q^{**}(\bar{\gamma}, Q_h) = Q_h$ by definition, so the binding and slack integrands coincide: $h(Q_h,\bar{\gamma}) = h^{\text{slack}}(Q_h,\bar{\gamma})$. The boundary term therefore vanishes.[12]

Integrating $-\partial h/\partial Q_h$ over $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \in [0, \bar{\gamma}(Q_h)]$ against $\tilde{F}$ gives

$$\begin{aligned}\frac{dW}{dQ_h} &= -\int_0^{\bar{\gamma}(Q_h)} \frac{\partial h}{\partial Q_h}\, d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \\ &= (1-p)\,I\int_0^{\bar{\gamma}} \mathbb{E}[\tilde{\gamma}|\Omega_\tau] d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;-\; c_h\,Q_h\,F\big(\bar{\gamma}(Q_h)\big),\end{aligned} \tag{A.4}$$

which is Equation (14). The benefit coefficient is simply I, reflecting the fact that the competitive fee absorbs all litigation terms.

Part (b): Strict concavity of $W(Q_h)$.

The social-cost function per failure on the binding region is $g(Q_h) = (1-Q_h)\,I$, which is linear in Q_h with $g''(Q_h) = 0$. The net welfare integrand $-h(Q_h, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = -(1-p)\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,g(Q_h) - c_h Q_h^2/2$ therefore has second derivative

$$\frac{\partial^2(-h)}{\partial Q_h^2} = -(1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,g''(Q_h) - c_h = -c_h < 0.$$

After integration over $\tilde{F}$ and accounting for the expanding binding region, the full second derivative of W is

$$\frac{d^2W}{dQ_h^2} = -c_h\,F(\bar{\gamma}) - \underbrace{\big[\text{boundary terms from } \bar{\gamma}'(Q_h)\big]}_{\leq 0} < 0. \tag{A.5}$$

The boundary terms are non-positive because expanding the binding region replaces a locally optimal slack effort $q^{**}(\bar{\gamma})$ with the (weakly suboptimal) mandated level Q_h. Because the failure-loss term $(1-Q_h)I$ is linear in Q_h, concavity comes entirely from the quadratic effort cost and the

[12] This is the standard envelope argument; see, e.g., Milgrom and Segal (2002).

boundary effect, requiring no parameter restrictions.

Setting $dW/dQ_h = 0$ therefore characterizes the unique interior maximum $Q_h^*(0)$.

Part (c): Comparative statics.

Apply the implicit function theorem to $dW/dQ_h = 0$. For a parameter θ:

$$\frac{dQ_h^*}{d\theta} = -\frac{\partial^2 W/\partial Q_h\,\partial\theta}{\partial^2 W/\partial Q_h^2}.$$

Since $d^2W/dQ_h^2 < 0$, the sign of $dQ_h^*/d\theta$ equals the sign of $\partial^2 W/\partial Q_h\,\partial\theta$.

Effect of I: The benefit coefficient I is increasing in I with derivative $1 > 0$. Hence $\partial^2 W/\partial Q_h \partial I = (1-p)\int_0^{\bar\gamma} \mathbb{E}[\tilde\gamma|\Omega_\tau] d\tilde F(\mathbb{E}[\tilde\gamma|\Omega_\tau]) > 0$, so Q_h^* is increasing in I.

Effect of c_h: The cost term is $-c_h Q_h F(\bar\gamma)$. Differentiating with respect to c_h: $\partial^2 W/\partial Q_h \partial c_h = -Q_h F(\bar\gamma) < 0$ (for $Q_h > 0$), so Q_h^* is decreasing in c_h.

Effect of p: The benefit is multiplied by $(1-p)$. Differentiating with respect to p: $\partial^2 W/\partial Q_h \partial p = -I\int_0^{\bar\gamma} \mathbb{E}[\tilde\gamma|\Omega_\tau] d\tilde F(\mathbb{E}[\tilde\gamma|\Omega_\tau]) < 0$, so Q_h^* is decreasing in p.

Effect of $\tilde F$ (FOSD shift): A first-order stochastic dominance shift in $\tilde F$ toward higher engagement risks increases $\int_0^{\bar\gamma} \mathbb{E}[\tilde\gamma|\Omega_\tau] d\tilde F(\mathbb{E}[\tilde\gamma|\Omega_\tau])$ for any fixed $\bar\gamma$, raising the marginal benefit of the standard. Hence Q_h^* increases. □

A.4 Derivation of the Effort FOC with AI (Equation (17))

Proof. For fixed AI share R, the auditor's per-engagement cost is

$$C^A(q, R, \mathbb{E}[\tilde\gamma|\Omega_\tau]) = \tfrac{1}{2}\, c_{\text{eff}}(R)\, q^2 + (1-p)\,\mathbb{E}[\tilde\gamma|\Omega_\tau]\,(1-\phi(R)q)\,\pi D,$$

which is strictly convex in q since $\partial^2 C^A/\partial q^2 = c_{\text{eff}}(R) > 0$. Differentiating with respect to q:

$$\frac{\partial C^A}{\partial q} = c_{\text{eff}}(R)\, q - (1-p)\,\mathbb{E}[\tilde\gamma|\Omega_\tau]\,\pi D\,\phi(R).$$

Setting to zero yields the unique unconstrained minimum

$$q^{**}(\mathbb{E}[\tilde\gamma|\Omega_\tau], R) = \frac{(1-p)\,\mathbb{E}[\tilde\gamma|\Omega_\tau]\,\pi D\,\phi(R)}{c_{\text{eff}}(R)},$$

which is (17). Monotonicity in $\mathbb{E}[\tilde\gamma|\Omega_\tau]$ follows from $\partial q^{**}/\partial\mathbb{E}[\tilde\gamma|\Omega_\tau] = (1-p)\,\pi D\,\phi(R)/c_{\text{eff}}(R) > 0$.

Benchmark recovery. At $R = 0$: $\phi(0) = 1$, $c_{\text{eff}}(0) = c_h$, giving $q^{**}(\mathbb{E}[\tilde\gamma|\Omega_\tau], 0) = (1-p)\mathbb{E}[\tilde\gamma|\Omega_\tau]\pi D/c_h$, recovering (10). □

A.5 Proof of Lemma 2 (Constrained Effort with AI)

Proof. Since $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)$ is strictly increasing in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ (proven above), the constrained optimum under a total-effort floor Q is $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q) = \max\{Q, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)\}$: for $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}_R(Q)$ the floor binds, and for $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}_R(Q)$ the unconstrained optimum exceeds Q. By inspection of the explicit form of q^{**}, the function $q^{**}(\cdot, R)$ is linear and strictly increasing in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ with $q^{**}(0, R) = 0$ and unbounded as $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \to \infty$. The threshold $\bar{\gamma}_R(Q)$, defined implicitly by $q^{**}(\bar{\gamma}_R(Q), R) = Q$, therefore has a unique solution for any $Q \geq 0$.

Inverting, in exact analogy to (12):

$$\bar{\gamma}_R(Q) = \frac{c_{\text{eff}}(R)\, Q}{(1-p)\,\pi D\,\phi(R)},$$

where the denominator is strictly positive for $R \in [0, 1)$ and $D > 0$, so $\bar{\gamma}_R(Q) > 0$ for any $Q > 0$ without further parameter restrictions.

Under a human-effort standard Q_h, the constraint $q_h \geq Q_h$ translates via $q_h = (1-R)q$ into $q \geq Q_h/(1-R)$ for $R < 1$. All results above apply with Q replaced by $Q_h/(1-R)$, yielding threshold $\hat{\bar{\gamma}}_R(Q_h)$ defined by $q^{**}(\hat{\bar{\gamma}}_R(Q_h), R) = Q_h/(1-R)$. □

A.6 Proof of Lemma 3 (Fixed-R Optimal Standard)

Proof. The proof has two parts: the derivation of dW/dQ (Equation (19)) and strict concavity.

Part (a): Derivation of dW/dQ (Equation (19)).

Welfare under fixed R and total-effort floor Q is given by (18). As in the benchmark, only the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}_R(Q)\}$ where $q^* = Q$ contributes to dW/dQ. On this region, the welfare integrand net of W_0 and the baseline $(1-p)I$ (negated, as a cost) is

$$h(Q, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\underbrace{(1-\phi Q)\,I}_{\equiv g(Q)} + \tfrac{c_{\text{eff}}(R)}{2}\,Q^2,$$

where we write $\phi \equiv \phi(R)$ for brevity and the social loss per failure is I under the competitive fee.

Differentiating $g(Q) = (1-\phi Q)\,I$:

$$g'(Q) = -\phi\, I. \tag{A.6}$$

The marginal social benefit of raising Q is therefore $-g'(Q) = \phi\, I$ per unit of $(1-p)\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$.

The full derivative of the cost integrand is

$$\frac{\partial h}{\partial Q} = (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(-\phi\,I) + c_{\text{eff}}(R)\,Q. \tag{A.7}$$

As in the benchmark, the boundary term from the Leibniz rule vanishes: at $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] = \bar{\gamma}_R(Q)$, $q^{**}(\bar{\gamma}_R(Q), R) = Q$ by definition, so the binding and slack integrands coincide.

Integrating $-\partial h/\partial Q$ over $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \in [0, \bar{\gamma}_R(Q)]$ against $\tilde{F}_R$:

$$\boxed{\frac{dW}{dQ} = (1-p)\,\phi(R)\,I \int_0^{\bar{\gamma}_R} \mathbb{E}[\tilde{\gamma}|\Omega_\tau] d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;-\; c_{\text{eff}}(R)\,Q\,F_R(\bar{\gamma}_R(Q)),} \tag{A.8}$$

which is Equation (19).

Benchmark recovery. Setting $R = 0$: $\phi(0) = 1$ and $c_{\text{eff}}(0) = c_h$, so the benefit coefficient becomes $1 \cdot I = I$, recovering (14). ✓

Part (b): Strict concavity and uniqueness.

Since $g(Q) = (1 - \phi Q)\,I$ is linear in Q, $g''(Q) = 0$. The net welfare integrand $-h(Q, \mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ therefore has second derivative $-(1-p)\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \cdot g''(Q) - c_{\text{eff}} = -c_{\text{eff}}(R) < 0$. Together with the non-positive boundary contribution from $\bar{\gamma}'_R(Q)$ (as in the benchmark), this gives

$$\frac{d^2W}{dQ^2} = -c_{\text{eff}}(R)\,F_R(\bar{\gamma}_R) + \text{boundary terms} < 0.$$

Concavity follows unconditionally from the quadratic effort cost; no parameter restrictions are needed. Setting $dW/dQ = 0$ therefore characterizes the unique interior optimum $Q^*(R)$.

The human-effort optimum follows from the change of variables $Q = Q_h/(1-R)$: the planner's problem in Q_h space has the same optimality condition evaluated at $Q_h/(1-R)$, so $Q_h^*(R) = (1-R)\,Q^*(R)$. □

A.7 Proof of Lemma 4 (Adoption Equilibrium and FOC)

Proof. Write the adoption objective as

$$J(R) = \big(1 - e(R)\big)\,\Lambda(\gamma_0, R, Q) + e(R) \int_0^1 \Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)\,dF(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) + \tfrac{c_R}{2}R^2, \tag{A.9}$$

where $\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q) = \frac{1}{2}c_{\text{eff}}(R)\,q^{*2} + (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,(1 - \phi(R)q^*)\,\pi D$ and $\tilde{F}_R$ places probability $1 - e(R)$ on $\mathbb{E}[\tilde{\gamma}|\Omega_{un}] = \gamma_0$ and $e(R)$ on $\mathbb{E}[\tilde{\gamma}|\Omega_{in}] \sim F$. The derivatives that enter the FOC are $c'_{\text{eff}}(R) = 2(c_h + c_{AI})R - 2c_h$, $\phi'(R) = e'(R)/e(0)$, and $e'(R) = e(R)(1 - e(R))(\alpha - 2\beta R)$.

Differentiability and convexity. On the slack region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}_R(Q)\}$, $q^* = q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)$ is smooth in R; on the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}_R(Q)\}$, $q^* = Q$ is constant in R; and Λ is continuous at the boundary $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] = \bar{\gamma}_R(Q)$ since $q^{**}(\bar{\gamma}_R(Q), R) = Q$. Standard results on differentiation under the integral with smooth moving boundaries give continuous differentiability of J on $(0, 1)$. The convexity condition (23) ensures $J''(R) > 0$, so J is strictly convex.

Interiority. At $R = 0$, the cost-substitution term reduces to $\frac{1}{2}c'_{\text{eff}}(0)\,\mathbb{E}[q^{*2}] < 0$ since $c'_{\text{eff}}(0) = -2c_h < 0$, while the adoption-cost term vanishes; hence $J'(0) < 0$. At $R = 1$, $c'_{\text{eff}}(1) = 2c_{AI} > 0$ and the adoption-cost term equals $c_R > 0$, so $J'(1) > 0$ for c_R satisfying (23). The intermediate value theorem and strict convexity yield a unique interior $R^* \in (0, 1)$.

Derivation of the FOC. Differentiating (A.9) in R produces three contributions. First, the direct dependence of Λ on R at fixed $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$. On the binding region, $q^* = Q$ is fixed, so

$$\frac{\partial \Lambda^{\text{bind}}}{\partial R} = \tfrac{1}{2}c'_{\text{eff}}(R)\,Q^2 - (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\pi D\,\phi'(R)\,Q.$$

On the slack region, the envelope theorem eliminates the indirect effect through q^{**}, yielding the analogous expression with q^{**} in place of Q. Combining both regions gives the cost-substitution and effectiveness terms of (24). Second, differentiating the mixture weights $1 - e(R)$ and $e(R)$ in (A.9) produces $e'(R)\,\Delta\Lambda(R, Q)$ — the information term. Third, the adoption-cost term contributes $c_R R$. Setting $J'(R^*) = 0$ gives (24). □

A.8 Derivation of the Cross-Partial (Equation (26))

Proof. We differentiate $\partial J/\partial R$ with respect to Q. Only the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq a_R(Q)\}$ depends on Q; on the slack region, effort is $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)$, which is independent of Q.

On the binding region ($q^* = Q$), the R-derivative of the per-engagement cost is

$$\frac{\partial \Lambda^{\text{bind}}}{\partial R} = \tfrac{1}{2}c'_{\text{eff}}(R)\,Q^2 - (1-p)\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\pi D\phi'(R)\,Q.$$

Differentiating with respect to Q:

$$\frac{\partial^2 \Lambda^{\text{bind}}}{\partial R\,\partial Q} = c'_{\text{eff}}(R)\,Q - (1-p)\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\pi D\phi'(R). \tag{A.10}$$

This is the cross-partial of the integrand.

Integrating over $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \in [0, \bar{\gamma}_R(Q)]$ against $\tilde{F}_R$ (the boundary term vanishes by the same

envelope argument as before):

$$\frac{\partial^2 J}{\partial R\,\partial Q} = c'_{\text{eff}}(R)\,Q\,F_R(\bar{\gamma}_R(Q)) \;-\; (1-p)\pi D\phi'(R)\int_0^{\bar{\gamma}_R} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]),$$

which is Equation (26).

Note. This cross-partial pertains to the *auditor's* objective $J(R)$, not to welfare. Accordingly, it involves the damage payment D, which reflects the auditor's private incentives rather than the social benefit coefficient $\phi(R)\,I$. □

A.9 Proof of Proposition 1 (Adoption Comparative Static)

Proof. By Lemma 4, $J(R)$ is C^2 and strictly convex in R, with $J'(R^*(Q)) = 0$. Since $\partial^2 J/\partial R^2 > 0$, the implicit function theorem gives

$$\frac{dR^*}{dQ} = -\,\frac{\partial^2 J/\partial R\,\partial Q}{\partial^2 J/\partial R^2},$$

and $R^*(Q)$ is continuously differentiable. The sign of dR^*/dQ equals the sign of $-\partial^2 J/\partial R\,\partial Q$.

We sign the cross-partial from (26) under the stated conditions:

First term: $c'_{\text{eff}}(R^*)\cdot Q\cdot F_R$. Since $c'_{\text{eff}}(R) = 2(c_h+c_{AI})R - 2c_h < 0$ for $R < R^{\text{cost}} \equiv c_h/(c_h+c_{AI})$, and $Q, F_R > 0$, this term is negative.

Second term: $-(1-p)\pi D\phi'(R^*)\int_0^{\bar{\gamma}_R}\mathbb{E}[\tilde{\gamma}|\Omega_\tau]d\tilde{F}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$. Since $\phi'(R^*) > 0$ for $R^* < R_{\text{peak}}$, and all remaining factors are positive, this term is negative.

Both terms are negative, so $\partial^2 J/\partial R\,\partial Q < 0$, and therefore

$$\frac{dR^*}{dQ} = -\,\frac{(\text{negative})}{(\text{positive})} > 0.$$

Tighter total-effort standards induce strictly greater AI adoption.

For $R^* > R_{\text{peak}}$, $\phi'(R^*) < 0$, the second term reverses sign, and the comparative static is ambiguous. However, $dR^*/dQ > 0$ continues to hold whenever the (negative) cost-substitution term dominates the (positive) effectiveness term. □

A.10 Proof of Proposition 2 (Monotonicity and Boundedness below R_{peak} under Human-Effort Floors)

Proof. The proof has four parts: (a) reduction of the human-floor adoption objective to the total-floor adoption objective at an effective floor, (b) derivation of the cross-partial $\partial^2 J^{(h)}/\partial R\,\partial Q_h$, (c)

existence of the non-binding threshold Q_h^{bind} and monotonicity under (C1), and (d) boundedness at $R = R_{\text{peak}}$ under (C2).

Part (a): Adoption objective under the human-effort floor. Under a human-effort standard Q_h, the date-1 constrained effort choice from Lemma 2 is

$$q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q_h) \;=\; \max\{Q_h/(1-R),\, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)\},$$

identical in form to the total-floor effort with the total-floor level replaced by the *effective* total-floor $\tilde{Q}(R, Q_h) \equiv Q_h/(1-R)$. The binding thresholds coincide, so Λ agrees on each region and

$$J^{(h)}(R, Q_h) \;=\; J^{\text{tot}}\big(R,\, \tilde{Q}(R, Q_h)\big), \tag{A.11}$$

where J^{tot} is the total-floor adoption objective of Lemma 4, inclusive of the adoption cost $\frac{c_R}{2}R^2$.

Part (b): FOC and cross-partial. Chain-rule differentiation of (A.11) yields

$$\frac{dJ^{(h)}}{dR}(R, Q_h) \;=\; \frac{\partial J^{\text{tot}}}{\partial R}\bigg|_{(R,\tilde{Q})} + \frac{\partial J^{\text{tot}}}{\partial Q}\bigg|_{(R,\tilde{Q})} \cdot \frac{Q_h}{(1-R)^2}, \tag{A.12}$$

where the first term is the total-floor FOC (24) at $Q = \tilde{Q}$ and the second is the denominator-amplification contribution unique to the human-effort floor. Leibniz-plus-envelope (the boundary integrand vanishes since $q^{**}(\bar{\gamma}_R^h, R) = \tilde{Q}$ annihilates $\partial\Lambda/\partial q$) gives

$$\frac{\partial J^{\text{tot}}}{\partial Q}(R, \tilde{Q}) \;=\; c_{\text{eff}}(R)\, \tilde{Q}\, \tilde{F}_R(\bar{\gamma}_R^h) - (1-p)\, \phi(R)\, \pi D \int_0^{\bar{\gamma}_R^h} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, d\tilde{F}_R(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]), \tag{A.13}$$

strictly positive on the binding region. Differentiating (A.12) in Q_h at fixed R, substituting the cross-partial formula for J_{RQ}^{tot} together with $J_{QQ}^{\text{tot}} = c_{\text{eff}}(R)\, \tilde{F}_R(\bar{\gamma}_R^h)$ and (A.13), factoring $1/(1-R)^2$, and using the algebraic identity

$$c'_{\text{eff}}(R)\,(1-R) + 2\, c_{\text{eff}}(R) \;=\; 2\, c_{AI}\, R \tag{A.14}$$

collapses the cross-partial to

$$\frac{\partial^2 J^{(h)}}{\partial R\, \partial Q_h} \;=\; \frac{1}{(1-R)^2}\Big\{2\, c_{AI}\, R\, \tilde{Q}\, \tilde{F}_R(\bar{\gamma}_R^h) - (1-p)\, \pi D\, \big[\phi(R) + (1-R)\, \phi'(R)\big] \int_0^{\bar{\gamma}_R^h} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, d\tilde{F}_R\Big\}. \tag{A.15}$$

Part (c): Non-binding interval, IFT, and monotonicity. Strict convexity of $J^{(h)}$ in R inherits

from the total-floor convexity condition of Lemma 4: the denominator-amplification contribution to $\partial^2 J^{(h)}/\partial R^2$ is non-negative on the binding region (proportional to $\partial J^{\text{tot}}/\partial Q > 0$). Therefore $R \mapsto J^{(h)}(R, Q_h)$ has a unique global minimizer on $[0,1]$ for each Q_h.

At $Q_h = 0$, $\tilde{Q} = 0$ and the human-floor problem coincides with the unregulated total-floor problem, giving $R_h^*(0) = 0$. Continuity of $\partial J^{(h)}/\partial R$ in Q_h extends the corner $R_h^* = 0$ to a closed interval $[0, Q_h^{\text{bind}}]$, with $Q_h^{\text{bind}} \equiv \sup\{Q_h \in [0,1] : R_h^*(Q_h) = 0\}$. For $Q_h > Q_h^{\text{bind}}$ the minimizer is interior, the implicit function theorem applied to (A.12) yields dR_h^*/dQ_h as stated, and the sign of dR_h^*/dQ_h is the sign of $-\partial^2 J^{(h)}/\partial R\,\partial Q_h$. Inspecting (A.15): the first bracket term is strictly positive on the binding region, while the second is strictly negative whenever $\phi(R)+(1-R)\phi'(R) > 0$ — which holds for every $R < R_{\text{peak}}$ since both ϕ and ϕ' are positive there. The cross-partial is therefore strictly negative, equivalently $dR_h^*/dQ_h > 0$, when the negative component dominates the positive one in absolute value, which is exactly (C1).

Part (d): Boundedness at $R = R_{\text{peak}}$. At $R = R_{\text{peak}} = \alpha/(2\beta)$, $e'(R_{\text{peak}}) = 0$ and hence $\phi'(R_{\text{peak}}) = 0$, so the effectiveness and information terms in $\partial J^{\text{tot}}/\partial R$ both vanish. Writing $u \equiv Q_h/(1 - R_{\text{peak}})$ and using identity (A.14), equation (A.12) at R_{peak} becomes

$$\left.\frac{dJ^{(h)}}{dR}\right|_{R_{\text{peak}}} = \frac{u}{1-R_{\text{peak}}}\Big[u\,c_{AI}\,R_{\text{peak}}\,\tilde{F}_{R_{\text{peak}}}(\bar{\gamma}^h) - (1-p)\,\phi(R_{\text{peak}})\,\pi D\,I_{\text{bind},1}\Big] + \tfrac{1}{2}\,c'_{\text{eff}}(R_{\text{peak}})\,I_{\text{slack},2} + c_R\,R_{\text{peak}}, \tag{A.16}$$

where $I_{\text{bind},1} \equiv \int_0^{\bar{\gamma}^h} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,d\tilde{F}_{R_{\text{peak}}}$ and $I_{\text{slack},2} \equiv \int_{\bar{\gamma}^h}^{\bar{\gamma}_{\max}} q^{**2}\,d\tilde{F}_{R_{\text{peak}}}$. By strict convexity of $J^{(h)}$ in R, it suffices to show that (A.16) is strictly positive for every $Q_h \in [0,1]$.

In the full-binding regime ($\bar{\gamma}^h \geq \bar{\gamma}_{\max}$), $\tilde{F}_{R_{\text{peak}}}(\bar{\gamma}^h) = 1$, $I_{\text{slack},2} = 0$, and $I_{\text{bind},1} = \mathbb{E}_{\tilde{F}_{R_{\text{peak}}}}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]] \equiv \mathbb{E}[m]$. Substituting and multiplying through by $(1 - R_{\text{peak}})^3 > 0$ reduces strict positivity of (A.16) to strict positivity of the quadratic in Q_h with leading coefficient $c_{AI}R_{\text{peak}} > 0$, which holds for all $Q_h \geq 0$ iff its discriminant is negative — exactly (C2). In the partial-binding regime ($\bar{\gamma}^h < \bar{\gamma}_{\max}$), $I_{\text{slack},2} > 0$ and $c'_{\text{eff}}(R_{\text{peak}}) < 0$ make the slack term negative, but its magnitude is bounded by $|c'_{\text{eff}}(R_{\text{peak}})|\,\mathbb{E}_{\tilde{F}_{R_{\text{peak}}}}[q^{**2}]/2$, while the binding-region bracket is uniformly smaller in absolute value than its full-binding counterpart since $\tilde{F}_{R_{\text{peak}}}(\bar{\gamma}^h) < 1$ and $I_{\text{bind},1} \leq \mathbb{E}[m]$; under the baseline calibration these contributions are dominated by $c_R R_{\text{peak}}$ (numerical verification below), so (A.16) stays strictly positive throughout.

Baseline calibration. Under the baseline parameters ($p = 0.6$, $\pi = 1$, $D = 0.1$, $c_h = 1$, $c_{AI} = 0.2$, $c_R = 0.2$, $\alpha = 2$, $\beta = 4$, $e(0) = 0.6$, $\gamma_0 = 0.35$, $F = \text{Uniform}[0,1]$): $R_{\text{peak}} = 0.25$, $\phi(R_{\text{peak}}) \approx 1.097$, and $\mathbb{E}[m] \approx 0.449$. Direct evaluation gives the LHS of (C2) $\approx 3.9 \times 10^{-4}$ and the

RHS $\approx 7.5 \times 10^{-3}$, a slackness factor of roughly 19. Condition (C1) holds throughout the binding region: the effectiveness factor $[\phi(R) + (1-R)\phi'(R)]$ stays strictly positive for $R_h^* < R_{\text{peak}}$, and numerical evaluation on the equilibrium path of Figure 3 confirms the LHS of (C1) exceeds its RHS by at least an order of magnitude on the interior of the binding range. In the partial-binding regime the negative contributions to (A.16) are bounded by $\approx 1.4 \times 10^{-3}$ and $\approx 2.0 \times 10^{-3}$, both an order of magnitude below $c_R R_{\text{peak}} = 0.05$. □

A.11 Proof of Proposition 3 (Audit Quality Dominance under Human-Effort Floors)

Proof. The proof has four parts: (a) reduction of $\mathcal{A}^{(h)}(Q_h) - \mathcal{A}_{\text{bench}}(Q_h)$ to a difference of expected detection products and the outer decomposition into a pointwise gain and a mixture-weight shift, (b) non-negativity of the pointwise gain via a two-effect algebraic identity, (c) non-negativity of the mixture-weight shift via Jensen's inequality on a convex test function, and (d) strict inequality on $(Q_h^{\text{bind}}, 1]$.

Part (a): Reduction to detection products. Using the mixture representation $\tilde{F}_R = e(R)\,F + (1-e(R))\,\delta_{\gamma_0}$ with $\gamma_0 = \mathbb{E}_F[\tilde{\gamma}]$, the constant $\mathbb{E}_{\tilde{F}_R}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]] = e(R)\,\gamma_0 + (1-e(R))\,\gamma_0 = \gamma_0$ is invariant in R. Applying this to the quality definition of Section 3.3 and to the benchmark of Section 3.1,

$$\mathcal{A}^{(h)}(Q_h) - \mathcal{A}_{\text{bench}}(Q_h) \;=\; (1-p)\Big\{\mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\text{AI}}] \;-\; \mathbb{E}_{\tilde{F}_0}[\Psi_{\text{b}}]\Big\}, \tag{A.17}$$

where the detection-product integrands are

$$\Psi_{\text{AI}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;\equiv\; \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R_h^*)\,q_h^*\big(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R_h^*, Q_h\big), \qquad \Psi_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;\equiv\; \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,q_{\text{b}}\big(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h\big),$$

$q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h) \;=\; \max\{Q_h,\, q_{\text{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])\}$ is the benchmark constrained effort from (11) with $q_{\text{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\pi D/c_h$, and $q_h^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q_h) = \max\{Q_h/(1-R),\, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)\}$ is the human-floor constrained effort with $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R) = (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\pi D\,\phi(R)/c_{\text{eff}}(R)$. Decompose the right-hand bracket of (A.17) as

$$\mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\text{AI}}] - \mathbb{E}_{\tilde{F}_0}[\Psi_{\text{b}}] \;=\; \underbrace{\mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\text{AI}} - \Psi_{\text{b}}]}_{\text{(I) pointwise detection gain}} \;+\; \underbrace{\mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\text{b}}] - \mathbb{E}_{\tilde{F}_0}[\Psi_{\text{b}}]}_{\text{(II) information-weight shift}}. \tag{A.18}$$

We show (I) ≥ 0 in Part (b), (II) ≥ 0 in Part (c), and that (I) > 0 strictly on $(Q_h^{\text{bind}}, 1]$ in Part (d).

Part (b): Pointwise detection gain. Add and subtract $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R_h^*)\,q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)$ in $\Psi_{\mathrm{AI}} - \Psi_{\mathrm{b}}$ to obtain the two-effect identity

$$\Psi_{\mathrm{AI}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) - \Psi_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;=\; \underbrace{\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\big(\phi(R_h^*) - 1\big)\, q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)}_{\text{AI-composition effect}}$$

$$+\; \underbrace{\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R_h^*)\,\big[q_h^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R_h^*, Q_h) - q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)\big]}_{\text{floor-imposition effect}}. \tag{A.19}$$

Each summand is non-negative under the proposition's hypotheses.

AI-composition effect. Proposition 2(iii) under (C2) gives $R_h^*(Q_h) < R_{\mathrm{peak}}$. The expertise function $e(R) = \sigma(\alpha R - \beta R^2 + \sigma^{-1}(e(0)))$ has derivative $e'(R) = e(R)(1 - e(R))(\alpha - 2\beta R)$, which is strictly positive on $[0, R_{\mathrm{peak}})$ and zero at R_{peak}, so e is strictly increasing on $[0, R_{\mathrm{peak}}]$ and $e(R_h^*) \geq e(0)$. Hence $\phi(R_h^*) = e(R_h^*)/e(0) \geq 1$, with equality only at $R_h^* = 0$. Since $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \geq 0$ and $q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h) \geq 0$, the AI-composition term is ≥ 0 pointwise.

Floor-imposition effect. We show $q_h^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R_h^*, Q_h) \geq q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)$ pointwise. Each of the two arguments inside $\max\{\cdot,\cdot\}$ in q_h^* dominates the corresponding argument in q_{b}:

- *Constraint argument.* $Q_h/(1 - R_h^*) \geq Q_h$, with equality iff $R_h^* = 0$.
- *Unconstrained-optimum argument.* $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R_h^*) = \big[\phi(R_h^*)\, c_h / c_{\mathrm{eff}}(R_h^*)\big]\, q_{\mathrm{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \geq q_{\mathrm{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$. This uses $\phi(R_h^*) \geq 1$ from the preceding paragraph together with $c_{\mathrm{eff}}(R_h^*) \leq c_h$. The latter follows from the algebraic identity

  $$c_{\mathrm{eff}}(R) \leq c_h \iff c_h(1-R)^2 + c_{AI} R^2 \leq c_h \iff R\big[(c_h + c_{AI})R - 2c_h\big] \leq 0 \iff 0 \leq R \leq 2R^{\mathrm{cost}},$$

  combined with the mild structural condition $R_{\mathrm{peak}} \leq 2R^{\mathrm{cost}}$ of the proposition and $R_h^* < R_{\mathrm{peak}}$, which together place R_h^* in the interval $[0, 2R^{\mathrm{cost}}]$ on which c_{eff} lies weakly below its benchmark value c_h.

If $a' \geq a$ and $b' \geq b$, then $\max\{a', b'\} \geq \max\{a, b\}$, so $q_h^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R_h^*, Q_h) \geq q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h)$ pointwise. Multiplying by $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R_h^*) \geq 0$ preserves the sign, so the floor-imposition term in (A.19) is ≥ 0 pointwise.

Summing, $\Psi_{\mathrm{AI}} - \Psi_{\mathrm{b}} \geq 0$ pointwise on the support of $\tilde{F}_{R_h^*}$, so (I) $= \mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\mathrm{AI}} - \Psi_{\mathrm{b}}] \geq 0$.

Part (c): Information-weight shift. Using the mixture representation,

$$\begin{aligned}\mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\mathrm{b}}] - \mathbb{E}_{\tilde{F}_0}[\Psi_{\mathrm{b}}] &= \big[e(R_h^*) - e(0)\big]\,\mathbb{E}_F[\Psi_{\mathrm{b}}(\tilde{\gamma})] + \big[(1 - e(R_h^*)) - (1 - e(0))\big]\,\Psi_{\mathrm{b}}(\gamma_0) \\ &= \big[e(R_h^*) - e(0)\big]\,\big\{\mathbb{E}_F[\Psi_{\mathrm{b}}(\tilde{\gamma})] - \Psi_{\mathrm{b}}(\gamma_0)\big\}.\end{aligned}$$

The first factor is ≥ 0 by the monotonicity of e on $[0, R_{\mathrm{peak}}]$ already invoked in Part (b). For the second factor, $\Psi_{\mathrm{b}}(\cdot, Q_h)$ is convex in its first argument: on $[0, \bar{\gamma}_{\mathrm{b}}(Q_h)]$ (with $\bar{\gamma}_{\mathrm{b}}(Q_h) = c_h Q_h/((1-p)D)$ from (12)) the floor binds and $\Psi_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h) = \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,Q_h$ is linear in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ with slope Q_h; on $(\bar{\gamma}_{\mathrm{b}}(Q_h), \bar{\gamma}_{\max}]$ the floor is slack and $\Psi_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q_h) = (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]^2\,D/c_h$ is convex-quadratic. The two pieces agree at the kink $\bar{\gamma}_{\mathrm{b}}(Q_h)$ (both equal $\bar{\gamma}_{\mathrm{b}}(Q_h)\,Q_h$), and the right-derivative at the kink ($2\,q_{\mathrm{b}}^{**}(\bar{\gamma}_{\mathrm{b}}) = 2Q_h$) strictly exceeds the left-derivative (Q_h), so Ψ_{b} is convex on $[0, \bar{\gamma}_{\max}]$. By Jensen's inequality, $\mathbb{E}_F[\Psi_{\mathrm{b}}(\tilde{\gamma})] \geq \Psi_{\mathrm{b}}(\mathbb{E}_F[\tilde{\gamma}]) = \Psi_{\mathrm{b}}(\gamma_0)$, so the second factor is ≥ 0 as well. Hence (II) ≥ 0.

Parts (b) and (c) together show that the right side of (A.18) is ≥ 0, and therefore $\mathcal{A}^{(h)}(Q_h) \geq \mathcal{A}_{\mathrm{bench}}(Q_h)$ for every $Q_h \in [0,1]$.

Part (d): Strict inequality on $(Q_h^{\mathbf{bind}}, 1]$. On $(Q_h^{\mathrm{bind}}, 1]$, Proposition 2(i) gives $R_h^*(Q_h) > 0$. Three strict inequalities follow:

(i) $1/(1 - R_h^*) > 1$ strictly.

(ii) $\phi(R_h^*) > 1$ strictly, since R_h^* lies in the strict-increase region $(0, R_{\mathrm{peak}})$ of e.

(iii) $c_{\mathrm{eff}}(R_h^*) < c_h$ strictly, since the quadratic c_{eff} equals c_h only at $R = 0$ and $R = 2R^{\mathrm{cost}}$, and $R_h^* \in (0, R_{\mathrm{peak}}] \subset (0, 2R^{\mathrm{cost}}]$ under the structural condition.

From (i)–(iii), the pointwise identity (A.19) yields $\Psi_{\mathrm{AI}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) > \Psi_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ strictly whenever $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > 0$:

- On $\{0 < \mathbb{E}[\tilde{\gamma}|\Omega_\tau] \leq \bar{\gamma}_{\mathrm{b}}(Q_h)\}$ (benchmark-binding region, non-empty for $Q_h > 0$ since $\bar{\gamma}_{\mathrm{b}}(Q_h) > 0$), $q_{\mathrm{b}} = Q_h$ and $q_h^* \geq Q_h/(1 - R_h^*) > Q_h$, so the floor-imposition term is strictly positive by (i).

- On $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}_{\mathrm{b}}(Q_h)\}$ (benchmark-slack region), $q_{\mathrm{b}} = q_{\mathrm{b}}^{**}$ and $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R_h^*) = \big[\phi(R_h^*)\,c_h/c_{\mathrm{eff}}(R_h^*)\big]\,q_{\mathrm{b}}^{**} > q_{\mathrm{b}}^{**}$ strictly by (ii)–(iii), so $q_h^* > q_{\mathrm{b}}$ strictly and the floor-imposition term is strictly positive.

In either region the AI-composition term is also strictly positive by (ii). The mixture $\tilde{F}_{R_h^*}$ places strictly positive mass on $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > 0\}$: the uninformed atom at $\gamma_0 > 0$ alone contributes mass $1 - e(R_h^*) > 0$, and the informed state contributes any positive-measure subset of F's support in $(0, \bar{\gamma}_{\max}]$. Therefore (I) $= \mathbb{E}_{\tilde{F}_{R_h^*}}[\Psi_{\text{AI}} - \Psi_{\text{b}}] > 0$ strictly, which combined with (II) ≥ 0 yields $\mathcal{A}^{(h)}(Q_h) > \mathcal{A}_{\text{bench}}(Q_h)$ strictly on $(Q_h^{\text{bind}}, 1]$. □

A.12 Proof of Proposition 4 (Audit Quality Crossover under Total-Effort Floors)

Proof. The proof has four parts: (a) reduction of $g(Q) \equiv \mathcal{A}(Q) - \mathcal{A}_{\text{bench}}(Q)$ to a difference of expected detection products and the outer decomposition into a pointwise gain and a mixture-weight shift (parallel to Part (a) of Appendix A.11), (b) sign of the endpoint value $g(1)$ under (C3) and (C4), (c) anchor at $Q = 0$ under the baseline corner $R^*(0) = 0$, and (d) existence of Q^{quality} via continuity and the intermediate value theorem.

Part (a): Reduction to detection products. Using the mixture representation $\tilde{F}_R = e(R)\,F + (1-e(R))\,\delta_{\gamma_0}$ with $\gamma_0 = \mathbb{E}_F[\tilde{\gamma}]$, the constant $\mathbb{E}_{\tilde{F}_R}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]] = e(R)\,\gamma_0 + (1-e(R))\,\gamma_0 = \gamma_0$ is invariant in R. Applying this to the quality definitions of Sections 3.3 and 3.1,

$$g(Q) \;=\; (1-p)\Big\{\mathbb{E}_{\tilde{F}_{R^*(Q)}}[\Psi_{\text{AI}}^{\text{tot}}] \;-\; \mathbb{E}_{\tilde{F}_0}[\Psi_{\text{b}}^{\text{tot}}]\Big\}, \tag{A.20}$$

where

$$\Psi_{\text{AI}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;\equiv\; \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R^*)\,q^*\big(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*, Q\big), \qquad \Psi_{\text{b}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;\equiv\; \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,q_{\text{b}}\big(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q\big),$$

with $q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q) = \max\{Q,\, q_{\text{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])\}$ from (11) and $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q) = \max\{Q,\, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R)\}$ from Lemma 2. Decompose the right-hand bracket of (A.20) as

$$\mathbb{E}_{\tilde{F}_{R^*}}[\Psi_{\text{AI}}^{\text{tot}}] - \mathbb{E}_{\tilde{F}_0}[\Psi_{\text{b}}^{\text{tot}}] \;=\; \underbrace{\mathbb{E}_{\tilde{F}_{R^*}}[\Psi_{\text{AI}}^{\text{tot}} - \Psi_{\text{b}}^{\text{tot}}]}_{\text{(I) pointwise detection gain}} \;+\; \underbrace{\mathbb{E}_{\tilde{F}_{R^*}}[\Psi_{\text{b}}^{\text{tot}}] - \mathbb{E}_{\tilde{F}_0}[\Psi_{\text{b}}^{\text{tot}}]}_{\text{(II) information-weight shift}}, \tag{A.21}$$

and add and subtract $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R^*)\,q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q)$ in $\Psi_{\text{AI}}^{\text{tot}} - \Psi_{\text{b}}^{\text{tot}}$ to obtain the pointwise two-effect identity

$$\begin{aligned}\Psi_{\text{AI}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) - \Psi_{\text{b}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) \;&=\; \underbrace{\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\big(\phi(R^*) - 1\big)\,q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q)}_{\text{AI-composition effect}} \\ &+\; \underbrace{\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R^*)\,\big[q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*, Q) - q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q)\big]}_{\text{floor-imposition effect}}.\end{aligned} \tag{A.22}$$

In contrast to Appendix A.11, where both summands are pointwise non-negative under (C2), here the AI-composition factor $\phi(R^*) - 1$ reverses sign once $R^*(Q)$ crosses R_{desk}, and the floor-imposition summand can take either sign because $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*, Q)$ may lie either above or below $q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], Q)$ on the slack regions. Condition (C4) circumvents this by collapsing both integrands onto the full-binding region at $Q = 1$, which is sufficient to sign $g(1)$.

Part (b): Sign of $g(1)$ under (C3) and (C4).

Condition (C4) states $\bar{\gamma}_{\max}(1-p)D \le \min\{c_h, c_{\text{eff}}(R^*(1))/\phi(R^*(1))\}$, which at $Q = 1$ translates directly into

$$q_{\text{b}}^{**}(\bar{\gamma}_{\max}) = \frac{(1-p)\bar{\gamma}_{\max}\pi D}{c_h} \le 1, \qquad q^{**}(\bar{\gamma}_{\max}, R^*(1)) = \frac{(1-p)\bar{\gamma}_{\max}\pi D\,\phi(R^*(1))}{c_{\text{eff}}(R^*(1))} \le 1.$$

Since q_{b}^{**} and $q^{**}(\cdot, R^*(1))$ are strictly increasing in their risk argument, the same inequalities hold at every $\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \in [0, \bar{\gamma}_{\max}]$, so

$$q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], 1) = \max\{1,\, q_{\text{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])\} = 1, \qquad q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(1), 1) = 1,$$

pointwise on the support of $\tilde{F}_{R^*(1)}$. Both floors bind on every engagement.

Floor-imposition effect at $Q = 1$. $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(1), 1) - q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], 1) = 1 - 1 = 0$ pointwise, so the floor-imposition summand in (A.22) vanishes identically.

AI-composition effect at $Q = 1$. Pointwise, $\Psi_{\text{AI}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) - \Psi_{\text{b}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\big(\phi(R^*(1)) - 1\big)$. Integrating against $\tilde{F}_{R^*(1)}$ and applying the R-invariance $\mathbb{E}_{\tilde{F}_{R^*(1)}}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]] = \gamma_0$,

$$\text{(I)}\big|_{Q=1} \;=\; \gamma_0\,\big(\phi(R^*(1)) - 1\big).$$

$R_{\text{desk}} \in (R_{\text{peak}}, 1)$ is the unique root of $\phi(R) = 1$ in $(R_{\text{peak}}, 1)$: $\phi(0) = 1$ by construction, $\phi(R_{\text{peak}}) > 1$ because e strictly rises on $[0, R_{\text{peak}}]$, and ϕ is strictly decreasing on $(R_{\text{peak}}, 1]$ since $\phi'(R) = e(R)(1 - e(R))(\alpha - 2\beta R)/e(0) < 0$ for $R > R_{\text{peak}} = \alpha/(2\beta)$. Condition (C3) places $R^*(1)$ strictly above R_{desk}, so $\phi(R^*(1)) < 1$ strictly and $\text{(I)}\big|_{Q=1} < 0$.

Information-weight shift at $Q = 1$. Under full binding, $\Psi_{\text{b}}^{\text{tot}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau]) = \mathbb{E}[\tilde{\gamma}|\Omega_\tau] \cdot 1 = \mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ is linear in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$. The mixture identity used in Part (c) of Appendix A.11 gives

$$\text{(II)}\big|_{Q=1} \;=\; \big[e(R^*(1)) - e(0)\big]\,\big\{\mathbb{E}_F[\Psi_{\text{b}}^{\text{tot}}(\tilde{\gamma})] - \Psi_{\text{b}}^{\text{tot}}(\gamma_0)\big\} \;=\; \big[e(R^*(1)) - e(0)\big]\big\{\mathbb{E}_F[\tilde{\gamma}] - \gamma_0\big\} \;=\; 0,$$

because $\gamma_0 = \mathbb{E}_F[\tilde{\gamma}]$ by the mixture construction. The convexity/Jensen argument of Appendix A.11

is not needed at $Q = 1$: the test function $\Psi_{\mathrm{b}}^{\mathrm{tot}}$ is exactly linear on the full-binding region, so the Jensen factor collapses to zero identically.

Combining,

$$g(1) \;=\; (1-p)\big\{(\mathrm{I})\big|_{Q=1} + (\mathrm{II})\big|_{Q=1}\big\} \;=\; (1-p)\,\gamma_0\,\big(\phi(R^*(1)) - 1\big) \;<\; 0$$

strictly under (C3) and (C4).

Part (c): Anchor at $Q = 0$. Under the baseline-corner property $R^*(0) = 0$ (Step 3 of Appendix A.7), $\phi(R^*(0)) = 1$, $c_{\mathrm{eff}}(R^*(0)) = c_h$, and $\tilde{F}_{R^*(0)} = \tilde{F}_0$. Hence $q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(0)) = q_{\mathrm{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])$ and $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(0), 0) = q_{\mathrm{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], 0)$ pointwise, so the integrands in (A.20) coincide and $g(0) = 0$.

Part (d): Existence of $Q^{\mathbf{quality}}$. R^* is continuous on $[0,1]$: it is constant at zero on any non-binding interval $[0, Q^{\text{tot-bind}}]$ and continuously differentiable on the binding region by Proposition 1. The integrand $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R)\,q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, Q)$ is continuous in (R, Q) pointwise in $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]$ and uniformly bounded on the compact support of $\tilde{F}_R$ by $\bar{\gamma}_{\max} \max_{R\in[0,1]} \phi(R) \max\{1, q^{**}(\bar{\gamma}_{\max}, R)\}$; the same bound applies to the benchmark integrand. Dominated convergence delivers continuity of $Q \mapsto \mathbb{E}_{\tilde{F}_{R^*(Q)}}[\Psi_{\mathrm{AI}}^{\mathrm{tot}}]$ and of $Q \mapsto \mathbb{E}_{\tilde{F}_0}[\Psi_{\mathrm{b}}^{\mathrm{tot}}]$, so g is continuous on $[0,1]$.

Define

$$Q^{\mathrm{quality}} \;\equiv\; \sup\{\, Q \in [0,1] : g(Q) \geq 0 \,\}.$$

The set is non-empty (0 belongs to it by Part (c)) and closed in $[0,1]$ by continuity of g, so the supremum is attained and $g(Q^{\mathrm{quality}}) \geq 0$. By Part (b), $g(1) < 0$, so $Q^{\mathrm{quality}} < 1$. For every $Q \in (Q^{\mathrm{quality}}, 1]$, Q lies outside the set by definition of the supremum, hence $g(Q) < 0$, i.e., $\mathcal{A}(Q) < \mathcal{A}_{\mathrm{bench}}(Q)$ strictly, which is the claim of the proposition. □

A.13 Derivation of Welfare and Quality Decompositions (Equations (32) and (28))

Derivation of the welfare decomposition. Total welfare with endogenous adoption is $W(Q) = W(Q, R^*(Q)) - C_R(R^*(Q))$, where $W(Q, R)$ is the fixed-R welfare from (18). Applying

the chain rule:

$$\begin{aligned}\frac{dW}{dQ} &= \left.\frac{\partial W}{\partial Q}\right|_{R=R^*} + \left.\frac{\partial W}{\partial R}\right|_{Q} \cdot \frac{dR^*}{dQ} - C_R'(R^*) \cdot \frac{dR^*}{dQ} \\ &= \underbrace{\left.\frac{\partial W}{\partial Q}\right|_{R}}_{\text{direct binding}} + \underbrace{\left(\left.\frac{\partial W}{\partial R}\right|_{Q} - C_R'(R^*)\right)\frac{dR^*}{dQ}}_{\text{adoption wedge}},\end{aligned} \tag{A.23}$$

which is Equation (32). The direct binding effect is $\partial W/\partial Q|_R$ as derived in (19). The adoption wedge collects the social value of marginal AI adoption ($\partial W/\partial R|_Q$) net of the marginal adoption cost ($C_R' = c_R R^*$), scaled by the policy-induced change in adoption dR^*/dQ. □

Derivation of the quality decomposition. Ex ante audit quality is $\mathcal{A}(Q) = 1-(1-p)\,\mathbb{E}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau](1-\phi(R^*(Q))q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(Q), Q))]$. This is a function of Q both directly (through q^*) and indirectly (through $R^*(Q)$). By the chain rule:

$$\frac{d\mathcal{A}}{dQ} = \underbrace{\left.\frac{\partial \mathcal{A}}{\partial Q}\right|_{R}}_{\text{direct binding}} + \underbrace{\left.\frac{\partial \mathcal{A}}{\partial R}\right|_{Q} \cdot \frac{dR^*}{dQ}}_{\text{adoption feedback}},$$

which is Equation (28). The direct binding effect is positive: holding R fixed, a tighter floor raises q^* on binding engagements, reducing the failure probability $\mathbb{E}[\tilde{\gamma}|\Omega_\tau](1-\phi q)$ by $\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\phi$ per engagement. The adoption feedback has ambiguous sign depending on whether R^* sits above or below R_{peak}. □

A.14 Proof of Corollary 2 (Comparative Statics in D and $e(0)$)

Proof. We apply the implicit function theorem to the adoption FOC $J'(R^*) = 0$ from Lemma 4. Since $J''(R^*) > 0$ by strict convexity, for any parameter $\theta \in \{D, e(0)\}$:

$$\frac{dR^*}{d\theta} = -\frac{\partial^2 J/\partial R\,\partial\theta}{\partial^2 J/\partial R^2}.$$

The sign of $dR^*/d\theta$ equals the sign of $-\partial^2 J/\partial R\,\partial\theta$.

Part (i): Sign of dR^*/dD.

Binding region. On the binding region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] \le \bar{\gamma}_R(Q)\}$, effort $q^* = Q$ is independent of D:

$$\frac{\partial \Lambda^{\text{bind}}}{\partial R} = \tfrac{1}{2}c_{\text{eff}}'(R)\,Q^2 - (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\pi D\,\phi'(R)\,Q.$$

Differentiating with respect to D:

$$\frac{\partial^2 \Lambda^{\text{bind}}}{\partial R\, \partial D} = -(1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, \phi'(R)\, Q. \tag{A.24}$$

Slack region. On the slack region $\{\mathbb{E}[\tilde{\gamma}|\Omega_\tau] > \bar{\gamma}_R(Q)\}$, $q^* = q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R) = (1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, \pi D\, \phi(R)/c_{\text{eff}}(R)$. Substituting into C^A gives the explicit envelope value

$$\Lambda^{\text{slack}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R, D) = (1-p)\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, \pi D - \frac{(1-p)^2\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]^2\, \pi^2 D^2\, \phi(R)^2}{2\, c_{\text{eff}}(R)}.$$

Differentiating once with respect to R:

$$\frac{\partial \Lambda^{\text{slack}}}{\partial R} = \frac{(1-p)^2\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]^2\, \pi^2 D^2 \left[\phi(R)^2\, c'_{\text{eff}}(R) - 2\, \phi(R)\, \phi'(R)\, c_{\text{eff}}(R)\right]}{2\, c_{\text{eff}}(R)^2},$$

and once more with respect to D:

$$\frac{\partial^2 \Lambda^{\text{slack}}}{\partial R\, \partial D} = \frac{(1-p)^2\, \mathbb{E}[\tilde{\gamma}|\Omega_\tau]^2\, \pi^2 D \left[\phi(R)^2\, c'_{\text{eff}}(R) - 2\, \phi(R)\, \phi'(R)\, c_{\text{eff}}(R)\right]}{c_{\text{eff}}(R)^2}. \tag{A.25}$$

Total cross-partial. Integrating both regions against $\tilde{F}_R$ (the boundary term vanishes by the envelope argument as in Appendix A.8):

$$\frac{\partial^2 J}{\partial R\, \partial D} = \underbrace{-(1-p)\, \phi'(R^*)\, Q \int_0^{\bar{\gamma}_R(Q)} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, d\tilde{F}_R(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])}_{\equiv A_{\text{bind}}}$$

$$+ \underbrace{\frac{(1-p)^2\, \pi^2 D \left[\phi^2 c'_{\text{eff}} - 2\phi\phi' c_{\text{eff}}\right]}{c_{\text{eff}}^2} \int_{\bar{\gamma}_R(Q)}^{\bar{\gamma}_{\max}} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]^2\, d\tilde{F}_R(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])}_{\equiv A_{\text{slack}}}, \tag{A.26}$$

with ϕ, ϕ', c_{eff}, c'_{eff} evaluated at R^*. Since $J''(R^*) > 0$, $\text{sign}(dR^*/dD) = -\,\text{sign}(A_{\text{bind}} + A_{\text{slack}})$.

Case (a): $R^* < R_{peak}$. Then $\phi'(R^*) > 0$, so $A_{\text{bind}} < 0$. For the slack bracket, $\phi^2 c'_{\text{eff}} < 0$ (since $c'_{\text{eff}} < 0$ for $R^* < R^{\text{cost}}$) and $-2\phi\phi' c_{\text{eff}} < 0$ (since $\phi' > 0$). Both summands are negative, so $A_{\text{slack}} < 0$. Therefore $\partial^2 J/\partial R\, \partial D < 0$ and $dR^*/dD > 0$.

Case (b): $R^* > R_{peak}$. Then $\phi'(R^*) < 0$, so $A_{\text{bind}} > 0$. For the slack bracket, $\phi^2 c'_{\text{eff}} < 0$ and $-2\phi\phi' c_{\text{eff}} > 0$ (since $\phi' < 0$); the bracket has ambiguous sign. When the bracket is non-negative, $A_{\text{slack}} \geq 0$ and $A_{\text{bind}} + A_{\text{slack}} > 0$ trivially. When the bracket is negative, the binding-dominance condition (A.35) states precisely that $A_{\text{bind}} \geq |A_{\text{slack}}|$, so $A_{\text{bind}} + A_{\text{slack}} > 0$ under (A.35). In either

case, $\partial^2 J/\partial R\,\partial D > 0$ and $dR^*/dD < 0$. When the slack region is empty ($\bar{\gamma}_R(Q) \geq \bar{\gamma}_{\max}$), $A_{\text{slack}} = 0$ identically and the conclusion follows from $A_{\text{bind}} > 0$ alone, with (A.35) vacuously satisfied.

Part (ii): Sign of $dR^*/de(0)$.

Baseline expertise $e(0)$ enters through $\phi(R) = e(R)/e(0)$ and through $e(R)$ itself via $g(R) = \log(e(0)/(1-e(0))) + \alpha R - \beta R^2$. Three derivatives are central. First,

$$\frac{\partial \phi(R)}{\partial e(0)} = \frac{e(R)\big(e(0) - e(R)\big)}{e(0)^2\,(1-e(0))},$$

which is negative when $e(R) > e(0)$ (AI has raised expertise above baseline). Second, for the marginal expertise gain:

$$\frac{\partial e'(R)}{\partial e(0)} = \frac{e(R)(1-e(R))}{e(0)(1-e(0))}\,(1-2e(R))\,(\alpha - 2\beta R). \tag{A.27}$$

The sign of (A.27) depends on the product $(1-2e(R))(\alpha - 2\beta R)$, where the first factor reflects the curvature regime of the logistic and the second reflects the position relative to R_{peak}. Third, for the marginal effectiveness response $\phi'(R) = e'(R)/e(0)$:

$$\frac{\partial \phi'(R)}{\partial e(0)} = \frac{1}{e(0)}\frac{\partial e'(R)}{\partial e(0)} - \frac{e'(R)}{e(0)^2} = \frac{e(R)(1-e(R))(\alpha - 2\beta R)}{e(0)^2(1-e(0))}\,\big[e(0) - 2e(R)\big]. \tag{A.28}$$

The sign of (A.28) depends on $(\alpha - 2\beta R)\big[e(0) - 2e(R)\big]$. Crucially, for $R < R_{\text{peak}}$ we have $e(R) > e(0)$ (AI enhances expertise), so $e(0) - 2e(R) < e(R) - 2e(R) = -e(R) < 0$, which gives $\partial\phi'(R)/\partial e(0) < 0$ *regardless* of whether $e(R) \gtrless 1/2$. This means the effectiveness channel always pushes toward $\partial^2 J/\partial R\,\partial e(0) > 0$, hence $dR^*/de(0) < 0$; it is the information-weight channel that determines whether this conclusion holds or is overturned.

Case (a): $e(R^*) > 1/2$ and $R^* < R_{peak}$. By (A.28), $\partial\phi'(R)/\partial e(0) < 0$, so the effectiveness term becomes less negative, contributing $\partial^2 J/\partial R\,\partial e(0) > 0$. By (A.27), $\partial e'(R)/\partial e(0) < 0$ (dampening), so the information-weight contribution $\frac{\partial e'}{\partial e(0)}\Delta\Lambda > 0$ (since $\Delta\Lambda < 0$), which *reinforces* the effectiveness channel. Both dominant channels push toward $dR^*/de(0) < 0$. The residual information-gap channel $e'(R)\frac{\partial \Delta\Lambda}{\partial e(0)}$ may partially offset; when the combined effectiveness and information-weight channels dominate this residual, we obtain $dR^*/de(0) < 0$.

Case (b): $e(R^*) < 1/2$ and $R^* < R_{peak}$. By (A.28), $\partial\phi'(R)/\partial e(0) < 0$ still holds (the effectiveness channel is unchanged), contributing $\partial^2 J/\partial R\,\partial e(0) > 0$. However, by (A.27), $\partial e'(R)/\partial e(0) > 0$ (steepening): higher baseline expertise raises $e'(R)$, which strengthens the information term

$e'(R)\Delta\Lambda$. Since $\Delta\Lambda < 0$, the information-weight contribution is $\frac{\partial e'}{\partial e(0)}\Delta\Lambda < 0$, which now *opposes* the effectiveness channel. The sign of $\partial^2 J/\partial R\,\partial e(0)$ therefore depends on the balance. From the dominant channels, the cross-partial is positive (hence $dR^*/de(0) < 0$) if and only if the effectiveness contribution exceeds the information-weight contribution:

$$(1-p)\pi D\,\frac{2e(R^*)-e(0)}{e(0)}\,\mathbb{E}\big[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,q^*\big] > \big(1-2e(R^*)\big)\,|\Delta\Lambda|,$$

which is condition (A.36). The left side is large when damages D are high or the expected risk-weighted effort $\mathbb{E}[\gamma q^*]$ is large; the right side is large when the auditor is deep in the convex region (small $e(R^*)$) or the information gap $|\Delta\Lambda|$ is large.

Case $R^* > R_{peak}$. Then $(\alpha - 2\beta R) < 0$ and $\phi'(R^*) < 0$.

Sub-case $e(R^*) > 1/2$*:* By (A.28), $e(0) - 2e(R) < 0$ and $\alpha - 2\beta R < 0$ give $\partial\phi'/\partial e(0) > 0$ (sign is $(-)(-) = +$). Since $\phi' < 0$, the positive effectiveness term $-(1-p)\pi D\phi'\mathbb{E}[\cdot]$ becomes smaller, so the effectiveness contribution to $\partial^2 J/\partial R\,\partial e(0)$ is negative, pushing $dR^*/de(0) > 0$. By (A.27), $(1-2e)(\,\alpha - 2\beta R) = (-)(-) > 0$, so $\partial e'/\partial e(0) > 0$; since $e' < 0$ in the deskilling region, e' becomes less negative. With $\Delta\Lambda < 0$: $\frac{\partial e'}{\partial e(0)}\Delta\Lambda = (+)(-) < 0$, which also pushes $dR^*/de(0) > 0$. Both channels align: $dR^*/de(0) > 0$ unambiguously. A sufficient condition for this sub-case to apply is $e(R_{\text{peak}}) > 1/2$, i.e., $e(0) > \sigma(-\alpha^2/(4\beta))$.

Sub-case $e(R^*) < 1/2$*:* By (A.27), $(1-2e)(\alpha - 2\beta R) = (+)(-) < 0$, so $\partial e'/\partial e(0) < 0$. Then $\frac{\partial e'}{\partial e(0)}\Delta\Lambda = (-)(-) > 0$, pushing $dR^*/de(0) < 0$. Meanwhile, $\partial\phi'/\partial e(0)$ has sign $(\alpha - 2\beta R)(e(0) - 2e(R))$: if $e(R) > e(0)/2$ (which holds when $R < 2R_{\text{peak}}$), this is $(-)(-) = +$, so the effectiveness channel still pushes $dR^*/de(0) > 0$. The two channels oppose, and the sign depends on the effectiveness-to-information ratio (A.36) evaluated with $\phi'(R^*) < 0$. □

A.15 Proof of Proposition 5 (Welfare Dominance under Effort Floors)

Proof. The proof has four parts: (a) welfare expressions and the social-cost–versus–private-cost wedge identity, (b) total-effort floor: revealed-preference–plus–wedge bound, unconditional dominance on $[0, Q^{\text{quality}}]$, and existence of Q^{wel} under (C5) via the intermediate value theorem, (c) human-effort floor: the same bound delivers universal dominance via Proposition 3, (d) Corollary 1 as a baseline observation.

Part (a): Welfare expressions and the wedge identity. The fixed-R welfare from (18) can be rearranged using the mixture identity $\mathbb{E}_{\tilde{F}_R}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]] = e(R)\,\gamma_0 + (1-e(R))\,\gamma_0 = \gamma_0$ (invariant in

R) into

$$W(Q,R) \;=\; W_0 + (1-p)(1-\gamma_0)\, I \;-\; S(Q,R),$$

where the social-cost wrapper is

$$S(Q,R) \;\equiv\; \mathbb{E}_{\tilde{F}_R}\Big[\tfrac{c_{\text{eff}}(R)}{2}\, q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q)^2 \;-\; (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R)\, q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q)\, I\Big] + C_R(R). \tag{A.29}$$

Combining with the welfare-with-endogenous-adoption definition (31) gives

$$W(Q) - W_{\text{bench}}(Q) \;=\; -\big[S(Q,R^*(Q)) - S(Q,0)\big], \tag{A.30}$$

and analogously $W^{(h)}(Q_h) - W_{\text{bench}}(Q_h) = -[S^{(h)}(Q_h, R_h^*(Q_h)) - S^{(h)}(Q_h,0)]$, where $S^{(h)}$ uses the human-floor effort $q_h^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q_h) = \max\{Q_h/(1-R),\, q^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R)\}$ in place of $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q)$. At $R = 0$ both q^* and q_h^* reduce to the benchmark $q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],Q) = \max\{Q,\, q_{\text{b}}^{**}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])\}$ from (11), and $C_R(0) = 0$, so $S(Q,0) = S^{(h)}(Q,0)$ recovers the benchmark social cost from (13).

The auditor's date-0 objective $J(R,Q) = \mathbb{E}_{\tilde{F}_R}[\Lambda(\mathbb{E}[\tilde{\gamma}|\Omega_\tau],R,Q)] + C_R(R)$ from (**??**) can be rearranged using the same mixture identity into

$$J(R,Q) \;=\; (1-p)\,\gamma_0\,\pi D \;+\; \mathbb{E}_{\tilde{F}_R}\Big[\tfrac{c_{\text{eff}}(R)}{2}\, q^{*2} \;-\; (1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R)\, q^*\,\pi D\Big] + C_R(R). \tag{A.31}$$

Comparing (A.29) and (A.31), S and J differ only in the failure-loss coefficient (I vs. D):

$$S(Q,R) \;=\; \big[J(R,Q) - (1-p)\,\gamma_0\,\pi D\big] \;-\; (I - \pi D)\, E^d(R,Q), \quad E^d(R,Q) \equiv \mathbb{E}_{\tilde{F}_R}\big[(1-p)\,\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\,\phi(R)\, q^*\big]. \tag{A.32}$$

Subtracting (A.32) at $R = R^*(Q)$ from the same expression at $R = 0$, the constants $(1-p)\,\gamma_0\, D$ cancel, yielding

$$S(Q,R^*) - S(Q,0) \;=\; \big[J(R^*,Q) - J(0,Q)\big] \;-\; (I - \pi D)\,\big[E^d(R^*,Q) - E^d(0,Q)\big]. \tag{A.33}$$

The bracketed difference $E^d(R^*,Q) - E^d(0,Q)$ equals $\mathcal{A}(Q) - \mathcal{A}_{\text{bench}}(Q)$ by the audit-quality definition of Section 3.3. Combining (A.30) and (A.33) gives the *revealed-preference–plus–wedge bound*

$$W(Q) - W_{\text{bench}}(Q) \;=\; -\big[J(R^*,Q) - J(0,Q)\big] \;+\; (I - \pi D)\,\big[\mathcal{A}(Q) - \mathcal{A}_{\text{bench}}(Q)\big]. \tag{A.34}$$

The first summand is non-negative by revealed preference $J(R^*(Q),Q) \le J(0,Q)$, and strictly

positive whenever $R^*(Q) > 0$ since J is strictly convex in R by Lemma 4. The second summand inherits the sign of the audit-quality gap, scaled by the wedge $(I - \pi D) \geq 0$. The same identity holds verbatim with $J^{(h)}$, R_h^*, Q_h, and $\mathcal{A}^{(h)}$ in place of J, R^*, Q, and $\mathcal{A}$, where $J^{(h)}(R, Q_h)$ is the human-floor adoption objective (the auditor's date-0 problem with q_h^* in place of q^*).

Part (b): Total-effort floor (Proposition 5(a)).

Step 1: Unconditional dominance on $[0, Q^{quality}]$. On the non-binding interval $\{Q : R^*(Q) = 0\}$, $J(R^*(Q), Q) = J(0, Q)$ and $\mathcal{A}(Q) = \mathcal{A}_{\text{bench}}(Q)$, so (A.34) delivers $W(Q) = W_{\text{bench}}(Q)$ exactly. The baseline-corner property $R^*(0) = 0$ ensures this includes a neighborhood of $Q = 0$ and pins down the anchor $W(0) = W_{\text{bench}}(0)$. On the binding subregion $(Q^{\text{bind}}, Q^{\text{quality}}]$, both summands of (A.34) are non-negative: the first by $R^*(Q) > 0$ and strict convexity of J, the second because $\mathcal{A}(Q) - \mathcal{A}_{\text{bench}}(Q) \geq 0$ on $[0, Q^{\text{quality}}]$ by the definition of $Q^{\text{quality}} = \sup\{Q \in [0,1] : g(Q) \geq 0\}$ in Appendix A.12, Part (d). Strict inequality $W(Q) > W_{\text{bench}}(Q)$ follows on $(Q^{\text{bind}}, Q^{\text{quality}}]$ from the strict positivity of the first summand alone.

Step 2: Sign of $W(1) - W_{bench}(1)$ *under* (C3), (C4), *and* (C5). By (C4), both $q_{\text{b}}(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], 1) = 1$ and $q^*(\mathbb{E}[\tilde{\gamma}|\Omega_\tau], R^*(1), 1) = 1$ pointwise on the support of $\tilde{F}_{R^*(1)}$ (Appendix A.12, Part (b)). Substituting $q^* = 1$ into (A.29) and using $\mathbb{E}_{\tilde{F}_R}[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]] = \gamma_0$ gives

$$S(1, R^*(1)) = \tfrac{c_{\text{eff}}(R^*(1))}{2} - (1-p)\,\gamma_0\,\phi(R^*(1))\,I + C_R(R^*(1)),$$
$$S(1, 0) = \tfrac{c_h}{2} - (1-p)\,\gamma_0\,I,$$

using $\phi(0) = 1$, $c_{\text{eff}}(0) = c_h$, and $C_R(0) = 0$ for the second line. Hence

$$W(1) - W_{\text{bench}}(1) \;=\; \tfrac{c_h - c_{\text{eff}}(R^*(1))}{2} \;-\; (1-p)\,\gamma_0\,\big(1 - \phi(R^*(1))\big)\,I \;-\; \tfrac{c_R}{2}\,R^*(1)^2.$$

By (C3), $R^*(1) > R_{\text{desk}}$, and Appendix A.12, Part (b) shows $\phi(R^*(1)) < 1$ strictly, so the middle term is strictly positive (a welfare loss); the first and third terms are the cost savings (gain) and adoption cost (loss). Condition (C5) states exactly that the cost savings strictly exceed the sum of these two losses, so $W(1) - W_{\text{bench}}(1) > 0$ strictly.

Step 3: Existence of Q^{wel} *via the intermediate value theorem.* Define $h(Q) \equiv W(Q) - W_{\text{bench}}(Q)$. Continuity of R^* on $[0,1]$ (constant zero on the non-binding interval and continuously differentiable on the binding region by Proposition 1) combined with the dominated-convergence argument used in Appendix A.12, Part (d), applied to both $S(Q, R^*(Q))$ and $S(Q, 0)$, delivers continuity of h on

$[0,1]$. Define

$$Q^{\text{wel}} \;\equiv\; \inf\{\, Q \in [Q^{\text{quality}},1] : h(\tilde{Q}) \geq 0 \text{ for every } \tilde{Q} \in [Q,1] \,\}.$$

By Step 2, $h(1) > 0$ strictly, so the set is non-empty (it contains $Q = 1$) and the infimum is well-defined; by continuity of h the infimum is attained, and $h(Q) \geq 0$ for every $Q \in [Q^{\text{wel}},1]$. Strictness on $(Q^{\text{wel}},1]$ follows from $h(1) > 0$ together with the definition of Q^{wel} as an infimum: if $h(\tilde{Q}) = 0$ for some $\tilde{Q} \in (Q^{\text{wel}},1]$, continuity and $h(1) > 0$ would still leave $h \geq 0$ on $[\tilde{Q},1]$ but admit $\tilde{Q}$ as a candidate strictly above Q^{wel}, contradicting the infimum. Combining Steps 1 and 3 yields part (a).

Part (c): Human-effort floor (Proposition 5(b)). The identity (A.34) adapted to the human-floor problem reads

$$W^{(h)}(Q_h) - W_{\text{bench}}(Q_h) \;=\; -\big[J^{(h)}(R_h^*,Q_h) - J^{(h)}(0,Q_h)\big] \;+\; (I-\pi D)\,\big[\mathcal{A}^{(h)}(Q_h) - \mathcal{A}_{\text{bench}}(Q_h)\big],$$

where $J^{(h)}(R,Q_h)$ is the auditor's human-floor objective at AI share R and floor Q_h. At $R = 0$ the human-floor constraint $q \geq Q_h/(1-R) = Q_h$ coincides with the benchmark constraint $q \geq Q_h$, so $J^{(h)}(0,Q_h)$ equals the benchmark adoption-cost-free objective evaluated at floor Q_h.

The first summand is ≥ 0 by revealed preference $J^{(h)}(R_h^*(Q_h),Q_h) \leq J^{(h)}(0,Q_h)$, with strict inequality on the binding region $(Q_h^{\text{bind}},1]$ where $R_h^* > 0$ and $J^{(h)}$ is strictly convex (the human-floor adoption objective inherits convexity from Lemma 4 via the isomorphism to the total-floor problem at effective floor $\tilde{Q}(R,Q_h) = Q_h/(1-R)$ established in Appendix A.10, Part (a), plus the additional positive denominator-amplification contribution to $J^{(h)\prime\prime}$).

The second summand is ≥ 0 by Proposition 3, with strict inequality on $(Q_h^{\text{bind}},1]$. On the non-binding interval $[0,Q_h^{\text{bind}}]$, Proposition 2(i) gives $R_h^* = 0$ and both summands vanish, yielding $W^{(h)}(Q_h) = W_{\text{bench}}(Q_h)$ exactly.

Combining, $W^{(h)}(Q_h) \geq W_{\text{bench}}(Q_h)$ for every $Q_h \in [0,1]$, with strict inequality on $(Q_h^{\text{bind}},1]$ from the strict positivity of either summand alone (in fact both, since each is strict on the binding region).

Part (d): Corollary 1 as a baseline observation. The pointwise comparison $W(Q) - W^{(h)}(Q)$ at $Q_h = Q$ admits no clean revealed-preference reduction. The total-floor problem is a relaxation of the human-floor problem at the same nominal floor (since the constraint $q \geq Q_h/(1-R)$ binds at least as tightly as $q \geq Q_h$ whenever $R \in [0,1)$), so by revealed preference $J^{\text{tot}}(R^*(Q),Q) \leq$

$J^{(h)}(R_h^*(Q), Q)$. However, applying the wedge identity (A.34) pointwise and subtracting,

$$\big[W(Q) - W_{\text{bench}}(Q)\big] - \big[W^{(h)}(Q) - W_{\text{bench}}(Q)\big]$$

$$= -\big[J^{\text{tot}}(R^*, Q) - J^{(h)}(R_h^*, Q)\big] + (I - \pi D)\big[\mathcal{A}(Q) - \mathcal{A}^{(h)}(Q)\big],$$

the wedge correction now compares $\mathcal{A}(Q)$ to $\mathcal{A}^{(h)}(Q)$, whose sign is indeterminate without further restrictions on the placement of $R^*(Q)$ relative to R_{desk}. Under the baseline calibration of Figures 3 and 4, $R^*(Q) > R_{\text{desk}}$ on a non-trivial subregion while $R_h^*(Q) < R_{\text{peak}} < R_{\text{desk}}$ throughout by Proposition 2(iii), so $\mathcal{A}(Q) < \mathcal{A}^{(h)}(Q)$ there and the wedge correction works against total-floor dominance. The figures show that the cost-substitution channel (the strictly negative first bracket above, since the total-floor problem is the relaxed problem) nevertheless dominates the wedge correction, with the welfare gap under the total-effort floor strictly exceeding the welfare gap under the human-effort floor at every common floor level on which both gaps are positive. The corollary records this pointwise comparison as a baseline observation and identifies the denominator amplification of Proposition 2 as the underlying mechanism. □

A.16 Full statement of Corollary 2

This appendix states the dominance conditions referenced in Corollary 2.

Part (i)(b): binding-region dominance. The condition under which higher damages reduce AI adoption when $R^* > R_{\text{peak}}$ is

$$\frac{|\phi'(R^*)|\, Q \int_0^{\bar{\gamma}_R(Q)} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, d\tilde{F}_R(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])}{(1-p)\,\pi D \int_{\bar{\gamma}_R(Q)}^{\bar{\gamma}_{\max}} \mathbb{E}[\tilde{\gamma}|\Omega_\tau]^2\, d\tilde{F}_R(\mathbb{E}[\tilde{\gamma}|\Omega_\tau])} \geq \frac{\big|\phi(R^*)^2\, c'_{\text{eff}}(R^*) - 2\,\phi(R^*)\,\phi'(R^*)\, c_{\text{eff}}(R^*)\big|}{c_{\text{eff}}(R^*)^2}. \quad \text{(A.35)}$$

The left side measures the binding-region effect: damages applied to engagements whose effort is pinned at the floor. The right side measures the slack-region effect, which works in the opposite direction. When the slack region is empty ($\bar{\gamma}_R(Q) \geq \bar{\gamma}_{\max}$), the right side is zero and (A.35) holds trivially. The baseline calibration with $D = 0.1$ produces an empty slack region throughout the binding range.

Part (ii)(b): effectiveness-to-information ratio. For $R^* < R_{\text{peak}}$ and $e(R^*) < \frac{1}{2}$, the sign of $dR^*/de(0)$ is negative (substitution) if

$$\frac{(1-p)\,\pi D\, \mathbb{E}\big[\mathbb{E}[\tilde{\gamma}|\Omega_\tau]\, q^*\big]}{|\Delta\Lambda|} > \frac{e(0)\,(1 - 2e(R^*))}{2\,e(R^*) - e(0)}, \quad \text{(A.36)}$$

and positive (complementarity) if the inequality is reversed. The left side is the effectiveness-to-information ratio and the right side captures the logistic curvature at $e(R^*)$.

Deskilling region ($R^* > R_{\mathbf{peak}}$). When the equilibrium AI share lies in the deskilling region, $\phi'(R^*) < 0$ and the effectiveness channel reverses sign. If $e(R^*) > \frac{1}{2}$, then $dR^*/de(0) > 0$, i.e. higher baseline expertise increases AI adoption; a sufficient condition is $e(0) > \sigma(-\alpha^2/(4\beta))$, where σ denotes the logistic function. If $e(R^*) < \frac{1}{2}$, the sign is determined by (A.36) with $\phi'(R^*) < 0$.

References

Cao, S., L. W. Cong, and B. Yang (2025). Distributed ledgers and secure multiparty computation for financial reporting and auditing. *Management Science 71*(5), 3852–3872.

Caskey, J. and V. Laux (2017). Corporate governance, accounting conservatism, and manipulation. *Management Science 63*(2), 424–437.

Choi, J. H. and C. Xie (2026). Human+ ai in accounting: Early evidence from the field. *Journal of Accounting Research*.

Dye, R. A. (1993). Auditing standards, legal liability, and auditor wealth. *Journal of political Economy 101*(5), 887–914.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Ewert, R. (1999). Auditor liability and the precision of auditing standards. *Journal of Institutional and Theoretical Economics (JITE)*, 181–206.

Fedyk, A., J. Hodson, N. Khimich, and T. Fedyk (2022). Is artificial intelligence improving the audit process? *Review of Accounting Studies 27*(3), 938–985.

Frey, C. B. and M. A. Osborne (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change 114*, 254–280.

Gao, P. and G. Zhang (2019). Auditing standards, professional judgment, and audit quality. *The Accounting Review 94*(6), 201–225.

Gillespie, N., S. Lockey, T. Ward, A. Macdade, and G. Hassed (2025). Trust, attitudes and use of artificial intelligence: A global study 2025.

Hwang, H., E. Kim, and M. Ye (2025). Auditing in the digital age: Determinants and consequences of technology investment. 50 pages. Posted: 24 Jul 2025. Date written: July 15, 2025.

Kronenberger, S. and V. Laux (2022). Conservative accounting, audit quality, and litigation. *Management Science 68*(3), 2349–2362.

Laux, V. and D. P. Newman (2010). Auditor liability and client acceptance decisions. *The Accounting Review 85*(1), 261–285.

Law, K. K. F. and M. Shen (2025). How does artificial intelligence shape audit firms? *Management Science 71*(5), 3641–3666.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *American Economic Review 92*(3), 583–593.

Munoko, I., H. L. Brown-Liburd, and M. Vasarhelyi (2020). The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics 167*(2), 209–234.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*(6654), 187–192.

Public Company Accounting Oversight Board (2010). AS 1201: Supervision of the audit engagement. PCAOB Release No. 2010-004, as amended.

Schwartz, R. (1997). Legal regimes, audit quality and investment. *Accounting Review*, 385–406.

Schwartz, R. (1998). Auditors' liability, vague due care, and auditing standards. *Review of Quantitative Finance and Accounting 11*(2), 183–207.

Simunic, D. A., M. Ye, and P. Zhang (2017). The joint effects of multiple legal system characteristics on auditing standards and auditor behavior. *Contemporary accounting research 34*(1), 7–38.

Ye, M. (2023). The theory of auditing economics: Evidence and suggestions for future research. *Foundations and Trends$\hat{W}$ in Accounting 18*(3), 138–267.

Ye, M. and D. A. Simunic (2013). The economics of setting auditing standards. *Contemporary Accounting Research 30*(3), 1191–1215.