AI as a Credence Good: Quality Competition, Limited Verification, and the Industrial Organization of Disclosure

Xufeng Zhang *, Han Li , Shenghui Bao

Posted Date: 11 March 2026

doi: 10.20944/preprints202603.0898.v1

Keywords: generative AI; credence goods; disclosure; quality competition; liability

 doi:10.20944/preprints202603.0898.v1

Article

AI as a Credence Good: Quality Competition, Limited Verification, and the Industrial Organization of Disclosure

Xufeng Zhang [1,*], Han Li [1,2] and Shenghui Bao [1,3]

[1] Resp AI Research Lab, CIIOE, Xiamen, 361000, China
[2] Lanzhou University, Lanzhou, 73000, China
[3] National University, Manila, 1008, Philippines
* Correspondence: xufeng@nau.edu

Abstract

In this paper we study competition between AI providers when users cannot fully verify model quality. Many AI services are not well described as standard search goods, and they are not pure experience goods either. Price, interface quality, and latency are usually observable, but reliability, hallucination risk, and the downstream cost of error are often only imperfectly observable even after use. Building on the emerging view that algorithmic advice can exhibit credence-good features, we embed that insight in a static industrial-organization model of vertical quality differentiation, costly certification, and limited user comprehension. Two firms choose whether to offer a low-quality or high-quality model. High-quality AI reduces error risk, but it is slower and costlier. A high-quality firm can purchase credible certification or disclosure, yet only a subset of users can interpret it. We characterize the pure-strategy equilibrium set, show how low-quality pooling can arise even when superior technology exists, and identify a quality-trap region in which the unique market equilibrium is low-quality pooling although an allocation with one high-quality provider is welfare superior. We then analyze policy. Standardized certification works through the demand side by increasing the fraction of users who can reward quality; minimum quality standards work directly but bluntly; liability shifts firms' cost incentives and weakly shrinks the region in which a low-quality industry outcome can be sustained. Contrary to a common rhetorical move in AI policy debates, these instruments are not interchangeable. The model also clarifies that our framework is a certification model rather than a full Grossman–Milgrom unraveling game: the key distortion comes from limited user comprehension of costly, truthful quality communication. The results offer a tractable industrial-organization foundation for current debates over hallucinations, model evaluation, AI documentation, and governance.

Keywords: generative AI; credence goods; disclosure; quality competition; liability

JEL Classification: D82; L13; L15; L51; O33

1. Introduction

Generative AI services are increasingly sold and adopted in market environments in which quality is hard to verify. A user can usually observe subscription price, rough speed, interface design, integration with other software, and perhaps a few headline benchmark claims. But the user often cannot observe the true reliability of the model before purchase, and may not be able to verify correctness after use either. A fluent answer may contain fabricated citations; a programming solution may fail only in edge cases; a legal or medical recommendation may look persuasive to a nonexpert despite being wrong. For a large class of economically important uses, AI quality therefore has a substantial credence-good component.

 doi:10.20944/preprints202603.0898.v1

That conceptual placement is not itself novel. In particular, Biermann et al. (2022) explicitly argue, and experimentally document, that algorithmic advice can be perceived as a credence good even after repeated use. What remains underdeveloped is the industrial-organization theory that follows once one takes that observation seriously. If AI quality is difficult to verify, what kind of market equilibrium should we expect? Under what conditions do firms invest in lower-hallucination systems, more robust retrieval, or stronger safeguards? How much can credible certification or documentation accomplish when only some users can decode it? And when do liability or minimum quality standards outperform disclosure-based governance?

These questions are central because contemporary AI markets display a salient vertical trade-off. More reliable systems are often more expensive to build and, at least in some applications, slower to run or more likely to abstain. Faster and cheaper models can look attractive on dimensions that users immediately notice, while their reliability deficit may be much harder to evaluate. That combination is precisely what makes industrial-organization analysis valuable. Competition alone does not guarantee that firms will optimize socially important dimensions of quality when those dimensions are weakly rewarded by demand.

Our starting point is classic. Akerlof (1970) established that quality uncertainty can degrade market outcomes, while Darby and Karni (1973) and the subsequent credence-good literature emphasized that some qualities remain difficult to evaluate even after consumption. AI fits that structure in an especially forceful way. Many answers are not self-verifying. Even when a user eventually discovers a mistake, the discovery may come too late, or only after the user has already made an important decision. Moreover, a large recent literature in machine learning and human-computer interaction emphasizes that model documentation, benchmark design, and evaluation accessibility matter because ordinary users and even expert organizations often struggle to diagnose AI failure modes (Bommasani et al. 2021; Jakesch et al. 2023; Ji et al. 2023; Liang et al. 2023; Mitchell et al. 2019).

We develop a static duopoly model in which firms first choose whether to offer a low-quality or high-quality model, then choose certification intensity if they adopt the high-quality technology, and finally compete in prices. The high-quality model has a higher marginal cost and an additional fixed cost; it also incurs a latency penalty relative to the low-quality model. Certification is truthful and credible but costly. Crucially, only a fraction of users can understand or act on certification. Those users correctly recognize the reliability gain from the high-quality product; the rest mostly respond to price and speed. We interpret certification broadly: standardized evaluation reports, third-party audits, model cards, or any institution that truthfully communicates reliability in a way that is at least partially legible to the market.

The model yields four main results.

First, we characterize the pure-strategy equilibrium set in threshold form. There is always a region in which a symmetric low-quality industry outcome is sustainable, a region in which a symmetric high-quality industry outcome is sustainable, and, when the fixed cost of quality lies between those regions, a region in which one firm differentiates upward and the other stays low quality. The existence of a premium reliable tier is therefore endogenous rather than imposed.

Second, certification need not be privately sufficient to support separation. The reason is not that certification is false or noisy; it is that only some users can convert truthful certification into willingness to pay. A high-quality firm buys certification only if the resulting demand response is strong enough to offset the cost and latency disadvantages of higher quality. This is the central market failure: quality may be real and socially valuable without being adequately rewarded in demand.

Third, there exists a quality-trap region in which the unique pure-strategy equilibrium is low-quality pooling even though an allocation with one high-quality provider and one low-quality provider is welfare superior. This region emerges when the fixed cost of building a reliable system is too high for private incentives but not too high for social efficiency. The wedge is driven by limited user comprehension of certification. In equilibrium, market demand underweights reliability relative to its realized social value.

 doi:10.20944/preprints202603.0898.v1

Fourth, policy tools work through distinct margins. Standardized certification and better model evaluation raise the fraction of users who can reward quality and can lower the effective cost of quality communication. Minimum quality standards remove low-quality actions entirely, which can be desirable in uniformly high-stakes applications but is blunt in heterogeneous markets. Liability operates differently again: by raising the expected private cost of lower-quality AI, it shifts firms' incentives directly rather than relying on user inference. We show that liability weakly shrinks the region in which low-quality pooling is sustainable and weakly expands the region in which symmetric high-quality supply is sustainable. The effect on the separating region is more subtle and generally ambiguous.

Two clarifications are important. First, our paper does not claim originality for the proposition that AI can have credence-good properties. Rather, our contribution is to embed that premise in a tractable oligopoly model of quality choice, certification, and policy. Second, our framework is not a full unraveling model in the Grossman–Milgrom sense. We do not study a revelation game in which every type chooses whether and how much to disclose and users draw Bayesian inferences from silence. Instead, we study costly certification by higher-quality providers when users have limited ability to process that certification. Grossman–Milgrom style insights remain useful for comparison, but the core distortion here is different: certification is truthful yet insufficiently effective because comprehension is limited.

The paper contributes to several literatures. Relative to classic vertical-differentiation models, we endogenize certification and impose a speed–quality trade-off. Relative to credence-good models, we analyze platform competition rather than expert fraud or overtreatment. Relative to the disclosure literature, we study a market in which quality communication is costly and only partially understood. Relative to AI-governance work, we supply a micro-founded industrial-organization explanation for why markets may reward fluency, speed, and price more strongly than reliability.

The rest of the paper proceeds as follows. Section 2 positions the paper in the relevant literatures. Section 3 presents the model. Section 4 characterizes equilibrium, the quality trap, and liability. Section 5 develops the welfare and policy analysis. Section 6 concludes. Proofs are collected in the Appendix.

2. Related Literature and Industrial-Organization Perspective

Our analysis sits at the intersection of five literatures.

The first is the economics of quality uncertainty. Akerlof (1970) showed that when buyers cannot observe quality, market trade can unravel or become distorted toward low quality. That framework remains foundational, but AI markets often depart from the canonical lemons setting in an important way. Quality is not necessarily a once-and-for-all hidden attribute that buyers learn after purchase. Instead, some dimensions of quality remain opaque after use. A user may consume the output yet still be unsure whether it was correct or whether a substantially better output was feasible. This makes AI closer to the environment emphasized by Darby and Karni (1973), in which quality can remain difficult to verify even post-consumption.

The second is the credence-good literature. Wolinsky (1993), Emons (1997), and the survey by Dulleck and Kerschbamer (2006) analyze markets in which sellers know more than buyers about the quality or appropriateness of treatment. Those papers center on diagnosis, fraud, undertreatment, and overtreatment. AI differs from the canonical doctor or mechanic example because the platform may not tailor treatment to individual users in the same way. Even so, the structural analogy is strong: a provider can supply advice of varying quality, users often cannot tell whether the advice was right, and the market outcome depends on what can be inferred from institutions that reduce information asymmetry.

The third literature is vertical differentiation and reputation. Mussa and Rosen (1978) and Shapiro (1983) provide canonical approaches to quality competition when higher quality costs more. Those frameworks are highly relevant, but AI suggests two extensions. First, higher quality may come with a convenience or latency penalty, not merely a higher production cost. Second, quality may be only

 doi:10.20944/preprints202603.0898.v1

imperfectly rewarded even after repeated use, because many users do not receive clean feedback. We therefore use a static model with costly certification rather than a pure reputation model. This choice is not only for tractability; it is also conceptually appropriate for markets in which even ex post learning is limited.

The fourth literature is disclosure, certification, and partial consumer understanding. Grossman (1981) and Milgrom (1981) derive the classic unraveling logic under truthful disclosure. Subsequent work has made clear that the result depends heavily on the institutional details of disclosure and on consumer sophistication. Fishman and Hagerty (2003) show that when not all customers understand disclosure, voluntary disclosure may fail and mandatory disclosure can improve welfare. Board (2009) show that competition can itself undermine disclosure incentives. Lizzeri (1999) studies intermediaries that reveal only coarse information. Dranove and Jin (2010) survey the theory and evidence on disclosure and certification.

Our paper belongs closest to this literature, but with two important differences. First, our application is AI reliability rather than conventional product quality. Second, we model certification rather than a full revelation game. We do not ask whether silence causes Bayesian consumers to infer the worst possible type. Instead, we ask when truthful quality certification is privately valuable in a market where only a subset of users can interpret it. That difference matters for how one should read our results. When certification fails to sustain quality, the problem is not a failure of truthfulness but a failure of comprehension and market reward.

The fifth literature is product safety, signaling, and liability. Daughety and Reinganum (1995) and Daughety and Reinganum (2008a,b) analyze safety choice, signaling, and disclosure when consumers cannot directly observe a product's safety. AI reliability is not identical to safety, but it is analytically close: lower error rates are costly to produce, socially valuable, and often imperfectly observable. This makes the safety-liability analogy fruitful. In both settings, one can distinguish policies that improve market information from policies that alter firms' incentives directly.

Recent AI work motivates the model's primitives. Bommasani et al. (2021) argue that foundation models propagate common capabilities and common failure modes across downstream applications. Mitchell et al. (2019) propose model cards as standardized reporting tools. Liang et al. (2023) emphasize broad, comparable evaluation across tasks and metrics. Ji et al. (2023) survey hallucination across natural-language generation. Jakesch et al. (2023) show that users rely on flawed heuristics when judging AI-generated language. Most directly, Biermann et al. (2022) provide experimental evidence that users can perceive algorithmic advice as a credence good. Our paper takes that idea into oligopoly theory.

From an industrial-organization perspective, the key point is that AI quality competition need not look like standard quality competition. A reliable model is often slower, more expensive, or more restrictive. A less reliable model can be cheap and fast. When quality is hard to verify, the faster and cheaper model can capture demand even if the reliable model creates larger realized surplus. That makes market power, competitive intensity, certification technology, and legal rules central determinants of AI quality.

3. Model

3.1. Firms, quality, and certification

There are two firms, indexed by $i \in \{1,2\}$, located at the endpoints of a Hotelling line $[0,1]$. Users are uniformly distributed along the line. Horizontal differentiation captures interface familiarity, integration with existing software, prompt libraries, switching frictions, and other nonquality reasons why users may prefer one provider over another. Let $t > 0$ denote the Hotelling transportation parameter.

Each firm chooses one of two technologies:

$$q_i \in \{L, H\}.$$

 doi:10.20944/preprints202603.0898.v1

The low-quality technology L is normalized as the baseline. The high-quality technology H reduces expected error and hallucination risk. We denote the realized per-user value of that reliability gain by $v > 0$. The high-quality technology also imposes a latency or convenience penalty $\ell > 0$ and a marginal production cost premium $c > 0$. In addition, any firm that adopts H pays a fixed cost $F > 0$.

If a firm adopts H, it may purchase certification intensity $x \geq 0$. Certification is truthful and verifiable, and we interpret it broadly: audited evaluation reports, standardized benchmark disclosure, model cards with externally verifiable content, or third-party certification. Certification is costly:

$$K(x) = \frac{k}{2}x^2, \qquad k > 0.$$

Only a fraction of users can understand and exploit certification. Let that fraction be

$$\lambda(x) = \lambda_0 + \rho x,$$

where $\lambda_0 \in [0, 1)$ is baseline market comprehension and $\rho > 0$ is the effectiveness of certification. We assume parameters such that $\lambda(x) \in [0, 1]$ on the equilibrium path.

This reduced form captures two realistic features of AI markets. First, even truthful quality communication may be difficult for ordinary users to process. Second, standardized, comparable certification can make high-quality systems more legible.

3.2. Users and Demand

Users differ horizontally by location $z \in [0, 1]$. Consider the asymmetric case in which firm 1 offers H with certification intensity x and firm 2 offers L. A user who can understand certification correctly perceives the reliability benefit v from the high-quality product. A user who cannot understand certification does not attach that value ex ante. All users perceive the latency penalty ℓ.

Accordingly, the average *perceived* vertical advantage of the high-quality product is

$$\Delta(x) = \lambda(x)v - \ell. \tag{1}$$

A user at location z has utilities

$$U_1 = \bar{u} + \Delta(x) - p_1 - tz, \qquad U_2 = \bar{u} - p_2 - t(1 - z),$$

where $\bar{u}$ is large enough to guarantee full market coverage.

The indifferent user satisfies

$$\bar{u} + \Delta(x) - p_1 - tz = \bar{u} - p_2 - t(1 - z),$$

so

$$z^* = \frac{1}{2} + \frac{\Delta(x) - p_1 + p_2}{2t}.$$

Hence demand for firm 1 is

$$D_1 = \frac{1}{2} + \frac{\Delta(x) - p_1 + p_2}{2t}, \qquad D_2 = 1 - D_1. \tag{2}$$

Two points are worth stressing. First, the market response depends on perceived quality, not actual quality. Users who do not understand certification underweight reliability in their purchase decision even though they will later enjoy the realized reliability gain if they consume the high-quality product. Second, the model intentionally compresses user heterogeneity into a tractable reduced form. One can reinterpret v as the expected value of error reduction across users, while $\lambda(x)$ captures the mass of users with enough verification ability, sophistication, or organizational support to act on certification. The same logic survives if one lets users differ in risk sensitivity, the cost of being wrong,

 doi:10.20944/preprints202603.0898.v1

or the value of speed; what matters is that only a subset of the market internalizes reliability at the point of purchase.

3.3. Timing

The game has three stages.

1. Firms simultaneously choose quality $q_i \in \{L, H\}$.
2. Any firm choosing H selects certification intensity $x_i \geq 0$.
3. Firms compete in prices.

We solve for subgame-perfect equilibrium.

3.4. Regularity Conditions

Let

$$a \equiv \rho v, \qquad D \equiv 9tk - a^2.$$

We assume

Assumption 1. $D = 9tk - a^2 > 0$.

Assumption 1 guarantees strict concavity of the certification problem.
Let

$$B \equiv 3t - c - \ell + \lambda_0 v.$$

We focus on the economically relevant interior region in which the premium provider has positive demand in the asymmetric subgame and the low-quality rival does not disappear. This is ensured by:

Assumption 2.

$$0 < B + ax^* < 6t,$$

where x^* denotes the equilibrium certification level derived below.

Assumption 2 rules out corner solutions in the pricing stage.

4. Equilibrium Analysis

4.1. Pricing and Certification in the Asymmetric Subgame

Suppose firm 1 chooses H and firm 2 chooses L. Firm profits are

$$\Pi_H = (p_1 - c)D_1 - F - \frac{k}{2}x^2, \qquad \Pi_L = p_2 D_2.$$

Using (2), we obtain the pricing equilibrium.

Lemma 1. *Given an asymmetric quality profile (H, L) and certification intensity x, the unique Nash equilibrium in prices is*

$$p_H^*(x) = t + \frac{\Delta(x) + 2c}{3}, \tag{3}$$

$$p_L^*(x) = t + \frac{c - \Delta(x)}{3}. \tag{4}$$

The corresponding market shares are

$$s_H(x) = \frac{3t + \Delta(x) - c}{6t}, \tag{5}$$

$$s_L(x) = \frac{3t - \Delta(x) + c}{6t}. \tag{6}$$

 doi:10.20944/preprints202603.0898.v1

Profits are

$$\Pi_H(x) = \frac{(3t + \Delta(x) - c)^2}{18t} - F - \frac{k}{2}x^2, \tag{7}$$

$$\Pi_L(x) = \frac{(3t - \Delta(x) + c)^2}{18t}. \tag{8}$$

Lemma 1 already shows the central trade-off. Higher quality is rewarded only to the extent that certification-induced comprehension makes reliability salient enough to overcome the latency penalty and the marginal cost premium. A firm can be objectively better and yet only weakly differentiated in demand.

Using (1), equation (7) becomes

$$\Pi_H(x) = \frac{(B + ax)^2}{18t} - F - \frac{k}{2}x^2.$$

The high-quality firm's certification problem is therefore quadratic.

Proposition 1. *Under Assumption 1 and $B > 0$, the high-quality firm's profit in the (H, L) subgame is strictly concave in x. The unique optimal certification level is*

$$x^* = \frac{aB}{D}. \tag{9}$$

The resulting gross operating profit of the high-quality firm, net of certification cost but before subtracting the fixed quality cost F, is

$$\bar{\Pi}_H \equiv \max_x \left\{ \frac{(B + ax)^2}{18t} - \frac{k}{2}x^2 \right\} = \frac{kB^2}{2D}. \tag{10}$$

Proposition 1 yields immediate comparative statics. Certification is more valuable when baseline user comprehension λ_0 is higher, when certification is more legible (ρ is higher), and when the underlying reliability benefit v is larger. It is less valuable when quality is slower (ℓ is larger), when producing quality is more expensive (c is larger), and when certification itself is expensive (k is larger). These are not just engineering facts; they are determinants of equilibrium industrial structure.

4.2. Quality-Stage Equilibrium

When both firms choose the same technology, there is no incentive to certify in equilibrium because certification no longer changes relative demand. Hence

$$\Pi^{LL} = \frac{t}{2}, \qquad \Pi^{HH} = \frac{t}{2} - F.$$

Define two threshold values:

$$F_H \equiv \bar{\Pi}_H - \frac{t}{2}, \tag{11}$$

$$F_L \equiv \frac{t}{2} - \Pi_L(x^*). \tag{12}$$

The threshold F_H is the largest fixed cost consistent with a profitable deviation from (L, L) to high quality. The threshold F_L is the largest fixed cost consistent with high quality being a best response to a high-quality rival.

Proposition 2. *Under Assumptions 1 and 2, the pure-strategy quality-stage equilibria are characterized as follows.*

 doi:10.20944/preprints202603.0898.v1

(i) (L, L) *is a pure-strategy equilibrium if and only if*

$$F \geq F_H.$$

(ii) (H, H) *is a pure-strategy equilibrium if and only if*

$$F \leq F_L.$$

(iii) *An asymmetric pure-strategy equilibrium,* (H, L) *or* (L, H)*, exists if and only if*

$$F_L \leq F \leq F_H.$$

If $F_L > F_H$*, the asymmetric interval is empty and the model exhibits a multiplicity region*

$$F \in [F_H, F_L]$$

in which both symmetric profiles, (L, L) *and* (H, H)*, are pure-strategy equilibria.*

Proposition 2 is the core equilibrium classification. It clarifies that three qualitatively distinct industry structures may arise: low-quality pooling, high-quality pooling, or vertical differentiation with a premium reliable provider and a low-cost rival. The model also permits multiplicity among symmetric equilibria when the incentive to deviate upward from (L, L) is weaker than the incentive to stay at H once both firms are already there. That possibility matters for welfare because it shows that equilibrium selection cannot be assumed away.

Several implications follow.

First, limited user comprehension expands the low-quality region. Since $\bar{\Pi}_H$ depends positively on the effectiveness of certification, lower λ_0 or lower ρ reduce F_H and thereby make (L, L) easier to sustain.

Second, the market can fail to create a premium reliable tier even when quality is technically feasible. If the fixed cost F exceeds F_H, no firm wants to be the first mover into a high-quality niche.

Third, stronger competition in the sense of easier substitution is not guaranteed to improve reliability incentives. What matters is not the number of firms per se, but whether the market reward to reliability is strong enough to support costly differentiation.

Corollary 1. *In the interior solution,* x^* *and* F_H *are increasing in* λ_0*,* ρ*, and* v*, and decreasing in* c*,* ℓ*, and* k*.*

Corollary 1 gives a compact comparative-static rationale for standardized evaluation, interoperable certification, and more legible model documentation. Those policies do not merely "improve transparency" in a loose sense; they expand the parameter region in which market incentives support reliable AI.

4.3. A Quality Trap

We now turn to welfare. The crucial distinction is between *perceived* quality, which governs demand, and *realized* quality, which governs surplus. When users who do not understand certification consume the high-quality product, they still enjoy the lower error rate ex post even though they did not fully value it ex ante.

Consider again the asymmetric allocation with one high-quality firm and one low-quality firm. Relative to (L, L), welfare equals the incremental benefit from users served by the high-quality firm, minus certification cost, minus the fixed cost of developing high quality, minus the extra Hotelling distortion from asymmetric market shares. Using (5), the welfare gain from an asymmetric allocation is

$$W^{HL}(x) - W^{LL} = s_H(x)(v - \ell - c) - \frac{k}{2}x^2 - F - t\left(s_H(x) - \frac{1}{2}\right)^2. \tag{13}$$

 doi:10.20944/preprints202603.0898.v1

Define the welfare threshold for an asymmetric allocation:

$$F_W \equiv \sup_{x \geq 0} \left\{ s_H(x)(v - \ell - c) - \frac{k}{2}x^2 - t\left(s_H(x) - \frac{1}{2}\right)^2 \right\}. \tag{14}$$

with $\{x \geq 0 : 0 \leq s_H(x) \leq 1\}$. Thus F_W is the largest fixed cost for which an allocation with one high-quality firm is socially preferable to (L, L).

Now define

$$\bar{F} \equiv \max\{F_H, F_L\}.$$

Proposition 3. *If*

$$F_W > \bar{F},$$

then for every fixed cost

$$F \in (\bar{F}, F_W),$$

the unique pure-strategy equilibrium is (L, L), yet an asymmetric allocation with one high-quality firm and one low-quality firm is welfare superior to (L, L).

Proposition 3 formalizes the quality trap. The key wedge is simple. In demand, the reliability gain is weighted by $\lambda(x)$ because only some users understand certification. In welfare, the realized reliability gain accrues to every user who consumes the high-quality product. When the market underperceives quality, private incentives can be too weak to sustain it even though the social return is positive.

This is not merely a repackaged lemons result. There is no exogenous population of hidden product types. Quality is chosen endogenously. The failure is therefore one of investment incentives: the market supplies too little reliability because users do not sufficiently reward it. In AI applications where hallucinations are hard to detect and costly when they occur, that is precisely the distortion policymakers worry about.

4.4. Liability

We now add a liability rule. Let $d > 0$ denote the reduction in expected harm per query when a firm moves from L to H. Suppose the firm faces expected liability md per query avoided by using high quality, where $m \geq 0$ is a policy parameter. Then the effective marginal cost premium of high quality becomes

$$c(m) = c - md. \tag{15}$$

Liability does not directly change user comprehension; it changes private incentives by making unreliable AI more expensive to supply.

All the formulas above remain valid after replacing c with $c(m)$, provided the interior assumptions continue to hold. Define

$$B(m) = 3t - c(m) - \ell + \lambda_0 v.$$

Then

$$x^*(m) = \frac{aB(m)}{D}, \qquad \bar{\Pi}_H(m) = \frac{kB(m)^2}{2D},$$

and the thresholds become

$$F_H(m) = \bar{\Pi}_H(m) - \frac{t}{2}, \qquad F_L(m) = \frac{t}{2} - \Pi_L(x^*(m); c(m)).$$

Proposition 4. *For every admissible m such that the interior conditions hold, the following statements are true.*

(i) $x^*(m)$ *is weakly increasing in m.*

(ii) $F_H(m)$ *is strictly increasing in m.*

 doi:10.20944/preprints202603.0898.v1

(iii) $F_L(m)$ *is strictly increasing in* m.

Therefore liability weakly shrinks the set of fixed costs for which (L, L) *is a pure-strategy equilibrium and weakly expands the set of fixed costs for which* (H, H) *is a pure-strategy equilibrium.*

Proposition 4 is the clean liability result. Once unreliable AI generates expected legal or regulatory costs, the private profitability of quality rises. The low-quality industry outcome becomes harder to sustain, and the high-quality industry outcome becomes easier to sustain. This is strongest precisely when demand-side governance is weakest, because liability does not rely on user sophistication.

However, liability does *not* generally have a monotone effect on the width of the separating region $[F_L(m), F_H(m)]$.

Corollary 2. *Suppose the interior solution holds and define*

$$C(m) \equiv \lambda_0 v - \ell - c(m).$$

Then

$$\frac{d}{dm}\left(F_H(m) - F_L(m)\right) = \frac{dk\left[C(m)\left(18kt - a^2\right) + 3a^2t\right]}{D^2}. \tag{16}$$

Hence liability expands the separating region if and only if

$$C(m) > -\frac{3a^2t}{18kt - a^2}.$$

Otherwise the separating region contracts.

Corollary 2 is important for policy interpretation. Liability unambiguously pushes the industry away from low-quality pooling and toward high-quality outcomes, but it need not enlarge the set of parameters that generate vertical differentiation. In some environments, liability makes symmetric high-quality supply more attractive and thereby compresses the region in which one firm remains low quality. The policy lesson is that liability is primarily a tool for moving the market toward reliability, not necessarily for preserving a two-tier structure.

5. Welfare, Market Failure, and Policy

5.1. Why the Market Underprovides Reliability

The model produces underprovision of reliability for a structurally clear reason. The willingness to pay that firms face is governed by perceived quality, not actual quality. Users who cannot decode certification still enjoy the realized benefits of more reliable AI if they end up using it, but that surplus is not fully capitalized into demand. As a result, the private return to quality can be below the social return.

This wedge is especially likely to be large in AI markets for three reasons.

First, verification is costly. Checking citations, testing edge cases, or validating factual claims often requires outside expertise or time. Second, mistakes are unevenly distributed. Many low-quality outputs look fine until they fail in a high-consequence context. Third, AI systems are often adopted by organizations in which the buyer, the user, and the party bearing downstream harm are not the same. Even without adding explicit externalities to the model, these institutional features make it realistic that market demand underweights reliability.

The model also illustrates why price competition alone is not a sufficient remedy. If firms mostly compete on dimensions users can easily observe, then more competition may intensify the race along those dimensions rather than along reliability. In AI, those dimensions are often speed, convenience, and price. The outcome can be a competitive market that is nevertheless skewed toward low-quality provision.

 doi:10.20944/preprints202603.0898.v1

5.2. Certification Policy

A first policy family works by improving the information environment. Standardized benchmarking, common taxonomies of failure modes, third-party audits, certification labels, and clearer model cards map naturally into the model in three ways:

(a) they raise baseline user comprehension λ_0;
(b) they increase the effectiveness of certification ρ;
(c) they lower the private cost of certification k.

Corollary 1 implies that all three changes raise the private return to quality. In equilibrium language, they increase x^* and F_H, thereby shrinking the region in which low-quality pooling can survive.

This provides a precise economic rationale for current efforts to standardize AI evaluation and documentation (Liang et al. 2023; Mitchell et al. 2019). The point is not simply that more information is always better. Rather, standardized and legible certification changes market incentives by increasing the share of users who can reward reliability. In a market where quality is otherwise hidden behind fluency and speed, that can be the difference between a viable premium segment and a market that pools on low reliability.

At the same time, the model disciplines optimism about disclosure-based governance. Certification works only through the users who can understand it. If λ_0 remains small, if certification is costly, or if the quality improvement comes with a large latency penalty, then transparency alone may not be enough. In such cases, policymakers should not expect a disclosure mandate to replicate the effects of liability or minimum standards.

This point also clarifies how our analysis relates to unraveling. In a full Grossman–Milgrom environment, nondisclosure can itself reveal bad news if consumers reason sharply enough. Our model does not rely on that mechanism. The problem is not merely that some firms choose silence. The deeper problem is that even truthful certification has limited reach because users do not uniformly understand it. For AI governance, that distinction is practical. A model card that exists but cannot be interpreted by most buyers will not reliably discipline the market.

5.3. Minimum Quality Standards

A second policy family imposes minimum quality standards. In the model, a standard that bans low-quality AI forces the industry into (H, H) as long as firms continue to operate. Relative to (L, L), the welfare change is

$$W^{HH} - W^{LL} = v - \ell - c - 2F. \tag{17}$$

Hence a minimum standard is socially desirable if and only if

$$v - \ell - c > 2F.$$

This criterion yields a clear policy distinction. Minimum standards are especially attractive when applications are uniformly high stakes and verification is hard. In medicine, legal compliance, safety-critical coding, or other domains where undetected hallucinations can generate substantial harm, forcing low-quality systems out of the market may be efficient. In such environments, the bluntness of a standard is not a major drawback because the social value of reliability is large for almost every user.

The picture changes when applications are heterogeneous. Many AI uses are low stakes, easily checked, or complementary to strong human oversight. In those domains, a blanket standard can overcorrect by suppressing fast and cheap tools that are socially useful despite being less reliable. Our model therefore suggests that minimum standards should be use-case specific when possible. A standard that is optimal for professional legal research need not be optimal for casual drafting or brainstorming.

 doi:10.20944/preprints202603.0898.v1

In industrial-organization terms, minimum standards dominate when the regulator wants to compress the market onto a single high-quality technology. Certification policies dominate when the regulator wants to preserve segmentation while making reliable quality commercially sustainable.

5.4. Liability, Safe Harbors, and the Allocation of Responsibility

A third policy family uses liability. Proposition 4 shows that liability works through a fundamentally different channel from certification. Certification raises demand-side rewards to quality. Liability changes the firm's effective cost comparison between H and L.

This difference matters when users are unsophisticated. If firms know much more about residual hallucination risk than buyers do, then demand-side tools can be weak because users cannot accurately map disclosure into expected harm. Liability is attractive precisely because it does not depend on that mapping. It directly internalizes part of the harm generated by lower-quality AI.

This does not mean maximal liability is always optimal. Excessive liability can chill beneficial deployment, raise entry barriers, or induce defensive overrefusal. Moreover, some AI harms are hard to attribute to a particular provider or to distinguish from user misuse. These are real concerns. But the model suggests a clear principle: liability is most valuable where the downstream cost of undetected error is large and traceable, and where user-side verification is especially poor.

A useful institutional implication follows. Policymakers can combine liability with safe harbors. For example, a provider that satisfies certain certification or audit requirements could face reduced liability exposure. In the model, such a regime would jointly improve comprehension and alter cost incentives. That combination is likely to outperform either instrument in isolation when both user understanding and firm incentives are distorted.

5.5. Heterogeneous Applications and Regulatory Targeting

The paper's baseline model suppresses some heterogeneity for tractability, but its policy logic points toward targeted regulation. In practice, users differ in at least three ways: their ability to verify outputs, the cost they incur when the AI is wrong, and the value they place on speed. These differences can be accommodated by interpreting v and ℓ as averages over application-specific segments.

Doing so yields several useful lessons.

First, markets with a large mass of low-stakes users may rationally sustain a low-quality tier. That is not by itself a market failure. The failure occurs when the same tier spills into high-stakes applications whose users cannot verify quality. This is one reason why enterprise settings often demand procurement rules, vendor audits, or ex ante certification.

Second, broad platform-level regulation can be too crude. The optimal policy mix depends on whether the relevant market segment is dominated by uninformed users, by high-stakes use cases, or by strong downstream monitoring. The model therefore supports application-contingent regulation more strongly than one-size-fits-all rules.

Third, public or industry-funded evaluation infrastructure can have large leverage. When third parties generate credible, legible comparisons across models, the effective values of λ_0 and ρ rise for the whole market. That can tilt competition toward reliable systems without requiring pervasive command-and-control regulation.

5.6. Dynamic Interpretation

Although our model is static, its logic has a natural dynamic reading. Repeated adoption and learning do not automatically eliminate the information problem if users cannot accurately diagnose AI errors. In that sense, a static certification model is not merely a shortcut for a reputation model; it captures an environment in which reputation itself may be slow or noisy to form.

This perspective helps explain why AI policy discussions have focused so heavily on evaluation and documentation. In standard experience-good markets, firms can often build reputation through repeated consumer feedback. In credence-good environments, reputation is weaker because users may not know whether they were served well. The same is true for AI when answers sound persuasive re-

 doi:10.20944/preprints202603.0898.v1

gardless of correctness. Certification and liability therefore become substitutes for missing reputational discipline.

6. Conclusion

We have developed a theory of AI competition when model quality is difficult for users to verify. The paper takes seriously the increasingly common observation that algorithmic advice can have credence-good features and derives the corresponding industrial-organization consequences.

Three broad conclusions follow.

First, markets can rationally pool on low-quality AI even when better technology exists. The core reason is that reliability is not fully priced when only some users understand certification. Second, there is a quality-trap region in which low-quality pooling is the unique pure-strategy equilibrium even though an allocation with one reliable provider is welfare superior. Third, policy instruments should be distinguished by the margins through which they operate. Certification improves the market reward to quality; minimum standards remove low-quality actions; liability alters firms' private cost comparison directly.

The model is deliberately parsimonious, but it yields an implication of wider significance. In AI markets, competition can reward what is legible before purchase rather than what matters after use. Speed, convenience, and low price are legible. Reliability often is not. Where that gap is large, the case for policy is not a generic distrust of markets. It is a standard industrial-organization response to a market in which socially valuable quality is hard to verify.

Appendix A. Proofs

Proof of Lemma 1. Using (2), firm profits in the asymmetric subgame are

$$\Pi_H = (p_1 - c)\left[\frac{1}{2} + \frac{\Delta(x) - p_1 + p_2}{2t}\right] - F - \frac{k}{2}x^2,$$

$$\Pi_L = p_2\left[\frac{1}{2} + \frac{-\Delta(x) + p_1 - p_2}{2t}\right].$$

Differentiating with respect to own price yields the first-order conditions

$$\frac{\partial \Pi_H}{\partial p_1} = \frac{1}{2} + \frac{\Delta(x) - p_1 + p_2}{2t} - \frac{p_1 - c}{2t} = 0,$$

$$\frac{\partial \Pi_L}{\partial p_2} = \frac{1}{2} + \frac{-\Delta(x) + p_1 - p_2}{2t} - \frac{p_2}{2t} = 0.$$

Multiplying by $2t$ gives

$$t + \Delta(x) - 2p_1 + p_2 + c = 0,$$

$$t - \Delta(x) + p_1 - 2p_2 = 0.$$

Solving the linear system yields (3) and (4). Substituting those prices into demand gives (5) and (6). Substituting prices and shares into profits gives (7) and (8). Uniqueness follows from strict concavity of each firm's profit in its own price. □

Proof of Proposition 1. From Lemma 1,

$$\Pi_H(x) = \frac{(B + ax)^2}{18t} - F - \frac{k}{2}x^2.$$

Differentiating,

$$\frac{d\Pi_H}{dx} = \frac{a(B + ax)}{9t} - kx, \qquad \frac{d^2\Pi_H}{dx^2} = \frac{a^2}{9t} - k = -\frac{D}{9t} < 0$$

 doi:10.20944/preprints202603.0898.v1

by Assumption 1. Hence the objective is strictly concave. Setting the first derivative equal to zero gives

$$a(B + ax) = 9tkx,$$

so

$$aB = (9tk - a^2)x = Dx,$$

and therefore

$$x^* = \frac{aB}{D}.$$

Substituting into the objective,

$$\bar{\Pi}_H = \frac{(B + ax^*)^2}{18t} - \frac{k}{2}(x^*)^2 = \frac{kB^2}{2D}.$$

□

Proof of Proposition 2. In the symmetric low-quality profile, each firm earns

$$\Pi^{LL} = \frac{t}{2}.$$

A unilateral deviation to high quality yields profit

$$\bar{\Pi}_H - F.$$

Hence (L, L) is an equilibrium if and only if

$$\frac{t}{2} \geq \bar{\Pi}_H - F \iff F \geq \bar{\Pi}_H - \frac{t}{2} = F_H.$$

In the symmetric high-quality profile, each firm earns

$$\Pi^{HH} = \frac{t}{2} - F.$$

A unilateral deviation from H to L against a rival that remains high quality and chooses x^* yields profit

$$\Pi_L(x^*).$$

Hence (H, H) is an equilibrium if and only if

$$\frac{t}{2} - F \geq \Pi_L(x^*) \iff F \leq \frac{t}{2} - \Pi_L(x^*) = F_L.$$

Finally, an asymmetric profile is an equilibrium if and only if high quality is a best response to low quality and low quality is a best response to high quality. These conditions are

$$\bar{\Pi}_H - F \geq \frac{t}{2} \iff F \leq F_H,$$

and

$$\Pi_L(x^*) \geq \frac{t}{2} - F \iff F \geq F_L.$$

Combining them gives

$$F_L \leq F \leq F_H.$$

If $F_L > F_H$, the interval is empty and there is no asymmetric pure-strategy equilibrium. In that case (L, L) exists for all $F \geq F_H$ and (H, H) exists for all $F \leq F_L$, implying coexistence of the symmetric equilibria on $[F_H, F_L]$. □

 doi:10.20944/preprints202603.0898.v1

Proof of Corollary 1. From (9),

$$x^* = \frac{\rho v(3t - c - \ell + \lambda_0 v)}{9tk - (\rho v)^2}.$$

The denominator is positive by Assumption 1. The numerator is increasing in λ_0, ρ, and v, and decreasing in c and ℓ. The denominator is increasing in k, so x^* is decreasing in k. Since

$$F_H = \frac{kB^2}{2D} - \frac{t}{2},$$

the same directional results hold for F_H. □

Proof of Proposition 3. Let

$$\bar{F} = \max\{F_H, F_L\}.$$

If $F > \bar{F}$, then $F > F_H$ and Proposition 2 implies that no asymmetric pure-strategy equilibrium exists. Also, $F > F_L$, so (H, H) is not a pure-strategy equilibrium. Since $F > F_H$, (L, L) is a pure-strategy equilibrium. Hence (L, L) is the unique pure-strategy equilibrium for every $F > \bar{F}$.

Now suppose additionally that $F < F_W$. By the definition of F_W in (14), there exists some certification level x such that

$$s_H(x)(v - \ell - c) - \frac{k}{2}x^2 - t\left(s_H(x) - \frac{1}{2}\right)^2 - F > 0.$$

By (13), this is equivalent to

$$W^{HL}(x) - W^{LL} > 0.$$

Therefore an asymmetric allocation is welfare superior to low-quality pooling. Combining the two steps proves that for every

$$F \in (\bar{F}, F_W),$$

the unique pure-strategy equilibrium is (L, L) even though an asymmetric allocation is welfare superior. □

Proof of Proposition 4. Under liability m, the effective marginal cost premium becomes

$$c(m) = c - md.$$

Thus

$$B(m) = 3t - c(m) - \ell + \lambda_0 v = 3t - c + \lambda_0 v - \ell + md,$$

so

$$\frac{dB(m)}{dm} = d > 0.$$

From (9),

$$x^*(m) = \frac{aB(m)}{D},$$

hence

$$\frac{dx^*(m)}{dm} = \frac{ad}{D} > 0.$$

This proves part (i).

Next,

$$F_H(m) = \frac{kB(m)^2}{2D} - \frac{t}{2},$$

so

$$\frac{dF_H(m)}{dm} = \frac{kdB(m)}{D} > 0.$$

This proves part (ii).

 doi:10.20944/preprints202603.0898.v1

For part (iii), let

$$Y(m) \equiv B(m) + ax^*(m).$$

Using (9),

$$Y(m) = B(m)\left(1 + \frac{a^2}{D}\right) = \frac{9tk}{D}B(m),$$

so

$$\frac{dY(m)}{dm} = \frac{9tkd}{D} > 0.$$

From Lemma 1,

$$\Pi_L(x^*(m); c(m)) = \frac{(6t - Y(m))^2}{18t}.$$

Under Assumption 2, $Y(m) < 6t$, so

$$\frac{d\Pi_L}{dm} = -\frac{(6t - Y(m))Y'(m)}{9t} < 0.$$

Therefore

$$\frac{dF_L(m)}{dm} = -\frac{d\Pi_L}{dm} > 0.$$

Hence both thresholds move upward with liability. Because (L, L) requires $F \geq F_H(m)$, the low-quality region shrinks. Because (H, H) requires $F \leq F_L(m)$, the high-quality region expands. □

Proof of Corollary 2. Let

$$S(m) \equiv F_H(m) - F_L(m).$$

From the proof of Proposition 4,

$$F_H'(m) = \frac{kdB(m)}{D}, \qquad F_L'(m) = \frac{kd(6t - Y(m))}{D}.$$

Therefore

$$S'(m) = \frac{kd}{D}[B(m) - 6t + Y(m)].$$

Using

$$Y(m) = \frac{9tk}{D}B(m), \qquad B(m) = 3t + C(m),$$

we obtain

$$S'(m) = \frac{kd}{D}\left[3t + C(m) - 6t + \frac{9tk(3t + C(m))}{D}\right].$$

After simplification,

$$S'(m) = \frac{dk\left[C(m)\left(18kt - a^2\right) + 3a^2t\right]}{D^2},$$

which is (16). The sign condition follows immediately. □

References

Akerlof, George A. 1970. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84(3), 488–500. https://doi.org/10.2307/1879431.

Biermann, Jan, John J. Horton, and Johannes Walter. 2022. Algorithmic advice as a credence good. ZEW Discussion Paper 22-071, ZEW – Leibniz Centre for European Economic Research. https://doi.org/10.2139/ssrn.4326911.

Board, Oliver. 2009. Competition and disclosure. *The Journal of Industrial Economics* 57(1), 197–213. https://doi.org/10.1111/j.1467-6451.2009.00369.x.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card,

 doi:10.20944/preprints202603.0898.v1

Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, et al. 2021. On the opportunities and risks of foundation models. https://doi.org/10.48550/arXiv.2108.07258.

Darby, Michael R. and Edi Karni. 1973. Free competition and the optimal amount of fraud. *The Journal of Law and Economics 16*(1), 67–88. https://doi.org/10.1086/466756.

Daughety, Andrew F. and Jennifer F. Reinganum. 1995. Product safety: Liability, R&D, and signaling. *The American Economic Review 85*(5), 1187–1206.

Daughety, Andrew F. and Jennifer F. Reinganum. 2008a. Communicating quality: A unified model of disclosure and signalling. *The RAND Journal of Economics 39*(4), 973–989. https://doi.org/10.1111/j.1756-2171.2008.00046.x.

Daughety, Andrew F. and Jennifer F. Reinganum. 2008b. Imperfect competition and quality signalling. *The RAND Journal of Economics 39*(1), 163–183. https://doi.org/10.1111/j.1756-2171.2008.00008.x.

Dranove, David and Ginger Zhe Jin. 2010. Quality disclosure and certification: Theory and practice. *Journal of Economic Literature 48*(4), 935–963. https://doi.org/10.1257/jel.48.4.935.

Dulleck, Uwe and Rudolf Kerschbamer. 2006. On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature 44*(1), 5–42. https://doi.org/10.1257/002205106776162717.

Emons, Winand. 1997. Credence goods and fraudulent experts. *The RAND Journal of Economics 28*(1), 107–119. https://doi.org/10.2307/2555942.

Fishman, Michael J. and Kathleen M. Hagerty. 2003. Mandatory versus voluntary disclosure in markets with informed and uninformed customers. *Journal of Law, Economics, and Organization 19*(1), 45–63. https://doi.org/10.1093/jleo/19.1.45.

Grossman, Sanford J. 1981. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics 24*(3), 461–483. https://doi.org/10.1086/466995.

Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences 120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys 55*(12), 248:1–248:38. https://doi.org/10.1145/3571730.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Lizzeri, Alessandro. 1999. Information revelation and certification intermediaries. *The RAND Journal of Economics 30*(2), 214–231. https://doi.org/10.2307/2556078.

Milgrom, Paul R. 1981. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics 12*(2), 380–391. https://doi.org/10.2307/3003562.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. https://doi.org/10.1145/3287560.3287596.

Mussa, Michael and Sherwin Rosen. 1978. Monopoly and product quality. *Journal of Economic Theory 18*(2), 301–317. https://doi.org/10.1016/0022-0531(78)90085-6.

Shapiro, Carl. 1983. Premiums for high quality products as returns to reputations. *The Quarterly Journal of Economics 98*(4), 659–679. https://doi.org/10.2307/1881782.

Wolinsky, Asher. 1993. Competition in a market for informed experts' services. *The RAND Journal of Economics 24*(3), 380–398. https://doi.org/10.2307/2555964.