The Correlation Cascade: Optimal Evaluation Design When AI Pipelines Self-Assess

Anonymous

March 19, 2026

Abstract

When an AI system both produces and evaluates financial analysis, its errors are correlated across roles. This paper provides a quantitative framework for measuring the cost of this correlated self-evaluation. The value of evaluator independence, an explicit function of six observable parameters, determines optimal pipeline length, gate placement, and when independent audit is cost-justified. The main result (the Correlation Cascade) shows that the feasible pipeline length satisfies a sharp upper bound, $N^* < 1/(qm) - 1$, that depends on the product of the defect rate and the correlation-adjusted miss rate. At moderate correlation levels, the feasible pipeline is roughly half as long as the independent-evaluator benchmark. Experiments with five models from four providers reveal that the effective correlation ρ is not a fixed property of training data but a function of evaluation protocol design: four out of 21 planted errors escape detection under a structured response format but are caught at 100% under free-form evaluation across five models from four providers. The binding constraint on pipeline quality is not what information the evaluator receives but how the evaluator is permitted to reason. An illustrative calibration to credit underwriting, using assumed parameters, demonstrates the framework's applicability to financial risk management and regulatory design.

Keywords: AI financial analysis, self-evaluation, correlated errors, inspection economics, credit underwriting, independent audit

JEL Codes: G21, G28, D82, D83, L15

1 Introduction

When a bank deploys an AI system to underwrite credit, who audits the AI? Increasingly, the answer is: another instance of the same AI. Large language models now generate financial analysis, evaluate that analysis for errors, and certify results for downstream consumption. The same architecture that builds a credit model also stress-tests it. The same system that drafts a compliance report also reviews it. The economic question is whether this self-evaluation is reliable, and if not, how much reliability it sacrifices relative to independent audit.

The cost of correlated self-evaluation in AI financial analysis pipelines is measured by the *value of evaluator independence*: the difference in pipeline value between a system with access to a truly independent evaluator and one constrained to self-evaluate. The value of independence is an explicit, closed-form function of six observable parameters, and it determines optimal pipeline length, gate placement, and whether independent audit is cost-justified.

An AI pipeline processes financial analysis through sequential stages: data ingestion, feature engineering, model estimation, stress testing, documentation. Between stages, quality gates check for defects. When a single model serves as both generator and evaluator, its errors are correlated across roles. Shared training data creates blind spots that no amount of prompt engineering can eliminate. A credit model that fails to account for geographic concentration risk will not flag the omission when the same model audits the work, because the blind spot is embedded in the model's training, not in its instructions.

The framework builds on classical inspection economics (Dorfman, 1943). The standard result places quality gates before expensive operations to catch defects early. Lemma 1 extends this rule by introducing a *correlation discount*: each gate's value is reduced by the factor $(1 - \rho)$, where ρ measures the correlation between generator and evaluator errors. When $\rho = 0$, the classical rule applies. When $\rho > 0$, gates catch fewer defects per dollar spent, and the optimal pipeline is shorter.

The *Correlation Cascade* (Proposition 1) gives the sharpest result. In a pipeline with homogeneous stages and gates at every boundary, the probability that the output is defect-free decays exponentially in pipeline length, with a decay rate that depends on the product of the defect rate q and the miss rate $m = 1 - p(1 - \rho)$. The optimal pipeline length satisfies a sharp upper bound: $N^* < 1/(qm) - 1$. At empirically estimated correlation levels ($\rho \approx 0.15$), the feasible pipeline is 53% as long as the independent-evaluator benchmark. Each ten percentage points of correlation reduces feasible pipeline length by 15–20%.

The value of evaluator independence (Proposition 2) formalizes what this correlation

costs. At $\rho_{\min} = 0.15$, the value of independence represents 46% of the independent-evaluator pipeline value. Even modest correlation ($\rho_{\min} = 0.05$) destroys over a fifth of pipeline value. Section 5 calibrates the framework to credit underwriting using assumed parameters, illustrating how the value of independence translates into dollar magnitudes for specific financial applications.

An experimental methodology reveals that the *evaluation protocol* is a dominant determinant of ρ. A conceptual error about CAPM diversification escapes detection in all 30 trials when the evaluator must respond in a structured JSON format. The same error is caught in 24 out of 25 trials when the evaluator responds in free-form natural language, across five models from four different providers. The effective correlation ρ is not a fixed property of training data; it is a function of how the evaluator is asked to respond. Protocol design is an actionable lever for pipeline quality, connecting to the theory's context variable c, which should be interpreted broadly to include response format constraints.

For a sufficiently capable evaluator, detection rates are statistically invariant to the amount of *upstream reasoning* provided (Proposition 3). The 120-billion-parameter model detects errors at 85–87% regardless of whether it receives no upstream reasoning, partial reasoning, or the generator's full chain of thought ($p = 0.83$ for the null). The information channel does not operate for capable evaluators. But the format channel operates powerfully: the same model's detection of conceptual errors drops from near-perfect to zero when forced into a structured response format. The binding constraint is not what information the evaluator receives, but how the evaluator is permitted to reason.

Related literature. Three literatures inform the analysis. First, inspection economics (Dorfman, 1943) provides the baseline framework. The classical literature on sequential inspection (Raz, 1986; Mandroli et al., 2006) determines optimal gate ordering when inspectors have exogenous, independent error rates. Endogenizing the correlation between inspector and producer changes both optimal pipeline length and the conditions under which inspection is valuable.

Second, the mechanism design and self-regulation literature analyzes when internal audit is credible. Tirole (1986) shows that rigid rules can dominate discretion when audit judgment is contaminated by collusion. The contamination in this paper is stochastic (shared training biases) rather than strategic (collusion), but the prescription is similar: Proposition 7 shows that fixed thresholds outperform naive adaptive ones when evaluator confidence is correlated with generator errors, with a crossover correlation $\bar{\rho}$ above which commitment to rigid rules dominates evaluator discretion.

Third, an emerging empirical literature documents that LLMs perform worse at evalu-

ating solutions than at generating them (Oh et al., 2024), exhibit systematic self-preference bias in evaluation tasks (Panickssery et al., 2024), and cannot reliably self-correct reasoning errors without external feedback (Huang et al., 2024). The model risk management literature in finance (Cont and Bianchi, 2011) analyzes the cost of model misspecification but does not address the specific correlation structure that arises when generator and evaluator share a training distribution. These findings motivate ρ but leave pipeline design without a formal framework.

The framework also relates to the ML ensemble diversity literature (Kuncheva and Whitaker, 2003), which analyzes how correlated errors reduce ensemble accuracy. The distinction is that ensemble methods aggregate votes across independent classifiers, while this paper studies sequential production with binary quality gates, yielding different formal objects (multiplicative quality decay rather than vote aggregation) and different design implications (optimal pipeline length rather than optimal ensemble size).

Search theory (Stigler, 1961; Weitzman, 1979) informs the idea-screening stage of the pipeline, where generating multiple candidates and selecting the best is a real-options problem. The information design literature (Kamenica and Gentzkow, 2011) shares the formal structure of a designer controlling information flow, but in a production context (optimizing detection) rather than a communication context (persuasion). Neither literature addresses correlated self-evaluation.

Roadmap. Section 2 presents the model: environment, production technology, gate technology with context design, and the designer's problem. Section 3 derives the main results: the modified Dorfman-Savage rule, the Correlation Cascade, the value of evaluator independence, and context design propositions. Section 4 describes the experiments that estimate $\rho_{\min}$ and test the context channel. Section 5 develops the credit underwriting application with calibrated scenarios. Section 6 discusses extensions, limitations, and the relationship between the empirical findings and the theory. Section 7 concludes.

2 Model

2.1 Environment

An AI pipeline produces financial analysis through N sequential stages indexed $i = 1, \ldots, N$. A single underlying model operates as both generator and evaluator. A pipeline **designer** commits ex ante to the pipeline's architecture (number of stages, gate placements, context allocations) before production begins. The model is not strategic: it follows instructions,

but its errors are stochastic and governed by a latent bias profile shared across all roles.

The framework applies to any setting where an AI system both produces and evaluates financial analysis. Credit underwriting is the primary application (Section 5). Three additional domains motivate the generality:

- *Algorithmic compliance and audit:* A fintech firm uses AI to execute trades and to audit those trades for regulatory compliance. When the same model architecture evaluates its own trading decisions, systematic compliance failures escape detection.
- *Robo-advisory self-monitoring:* AI portfolio managers that self-evaluate allocation decisions. Shared biases from common training data amplify herding across portfolios that appear independently managed.
- *AI-assisted due diligence in M&A:* AI pipelines that both analyze acquisition targets and evaluate the quality of their own analysis. Correlated blind spots create systematic valuation errors that internal checks cannot catch.

2.2 Production technology

Each stage $i \in \{1, \ldots, N\}$ transforms the artifact and may introduce a **binary defect** (present or absent) with stage-specific probability $q_i > 0$. A defect at stage i renders all downstream work worthless. Stage i has production cost $k_i > 0$, with costs weakly increasing: $k_1 \leq k_2 \leq \cdots \leq k_N$.

Assumption 1 (Cross-stage independence)**.** *Defects are introduced independently across stages: $\theta_i \perp \theta_j$ for $i \neq j$, where $\theta_i \in \{0, 1\}$ is the defect indicator at stage i.*

This assumption isolates the effect of evaluation correlation from production correlation. Section 6 discusses the correlated-defect case and argues the qualitative results strengthen.

Assumption 2 (Stage-specific detectability)**.** *A defect introduced at stage i is detectable only at the gate immediately following stage i (i.e., gate i). Downstream gates $j > i$ cannot detect defects originating at stage i.*

Stage-specific detectability captures the idea that each stage transforms the artifact in a way that embeds upstream errors into the structure. Once a credit model is estimated using a feature set that omits geographic concentration risk (a stage-2 defect), the resulting model coefficients appear internally consistent at stage 3 and beyond. The omission is detectable only by someone evaluating the feature engineering step itself. In practice, downstream gates can sometimes detect upstream defects indirectly (e.g., a stress test revealing anomalous

results traceable to a missing risk factor). Relaxing this assumption yields a quality function in which Q depends on the union of detection opportunities across all downstream gates, producing a weaker cascade effect: the upper bound on pipeline length is higher because multiple gates can catch the same defect. The qualitative comparative statics in ρ survive because each detection opportunity is still discounted by $(1 - \rho)$.

Proposition 6 in Section 3 formally establishes that the key comparative statics survive under continuous quality degradation.

Let $V > 0$ denote the value of a defect-free completed pipeline.

2.3 Gate technology with context design

Between consecutive stages i and $i + 1$, the designer may place a **quality gate**. Each gate uses the same model in evaluator mode. The designer controls two variables at each potential gate location $i \in \{1, \ldots, N - 1\}$:

1. **Gate placement** $g_i \in \{0, 1\}$: whether to place a gate after stage i.
2. **Context fraction** $c_i \in [0, 1]$: the fraction of upstream reasoning (prior stage outputs, intermediate work, chain-of-thought) that the evaluator at gate i receives.

The context fraction determines the evaluator's signal quality through two reduced-form objects.

Assumption 3 (Detection probability)**.** *The gate detection probability $p(c)$ satisfies:*

- $p(0) = p_{\min} > 0$ *(surface errors are detectable even without context),*
- $p(1) = p_{\max} < 1$ *(detection is imperfect even with full context),*
- $p'(c) \geq 0$ *(more context weakly improves detection),*
- $p''(c) \leq 0$ *(diminishing returns to context).*

Assumption 4 (Error correlation)**.** *The error correlation $\rho(c)$ satisfies:*

- $\rho(0) = \rho_{\min} \geq 0$ *(irreducible baseline correlation from shared training data),*
- $\rho(1) = \rho_{\max} < 1$ *(maximum correlation under full context),*
- $\rho'(c) \geq 0$ *(more context weakly increases correlation),*
- $\rho''(c) \geq 0$ *(correlation is convex in context).*

The maintained assumptions use weak inequalities to accommodate both the general case where context affects detection and correlation, and the empirically relevant special case where both are context-invariant ($p'(c) = 0$, $\rho'(c) = 0$). The context-invariant case receives separate treatment in Proposition 3.

Remark 1 (Micro-foundations)**.** *These reduced-form properties can be motivated by a signal model in which the evaluator observes* $s_i = \theta_i + (1 - c_i)\varepsilon_i^{\text{indep}} + c_i\varepsilon_i^{\text{shared}}$*, where* $\varepsilon^{\text{indep}}$ *is independent noise and* $\varepsilon^{\text{shared}}$ *is correlated with the generator's assessment error. The specific functional forms of* $p(c)$ *and* $\rho(c)$ *depend on the noise distributions and detection threshold, which we do not specify. The results require only the ordinal properties in Assumptions 3–4. The signal model bears a formal resemblance to mixed-signal models in financial economics, but the application here is to production inspection, not to information asymmetry in markets.*

Effective detection and miss probability. The per-gate *effective detection rate* and *miss probability* are:

$$D(c) \equiv p(c)(1 - \rho(c)), \qquad m(c) \equiv 1 - D(c). \tag{1}$$

The effective detection rate is the raw detection probability discounted by the independence factor $(1 - \rho)$.

False positives. The evaluator generates false positives with probability $f(c)$, where $f(0) = f_{\max}$, $f(1) = f_{\min} \geq 0$, $f'(c) < 0$, and $f''(c) > 0$. More context reduces false alarms.

Gate cost. Each active gate incurs cost:

$$G_i(c_i) = g + f(c_i) \cdot k_i \tag{2}$$

where $g > 0$ is the fixed compute and latency cost and $f(c_i) \cdot k_i$ is the expected cost of false rejection (rework triggered by a false alarm). The rework cost k_i is the cost of repeating stage i. Under Assumption 2, a false alarm at gate i requires re-running only stage i, not downstream stages, because downstream work has not yet begun when gate i fires. In settings where rework involves re-running stages i through a checkpoint, k_i should be replaced by the cumulative re-run cost $\sum_{j=i}^{N} k_j$, which strengthens the case for providing more context (to reduce $f(c_i)$) at later gates.

2.4 Designer's problem

The designer chooses gate placements $\{g_i\}_{i=1}^{N-1}$ and context fractions $\{c_i\}_{i=1}^{N-1}$ to maximize:

$$\max_{\{g_i,c_i\}} \; V \cdot Q(\mathbf{g},\mathbf{c}) - \sum_{i=1}^{N} k_i - \sum_{i=1}^{N-1} g_i \cdot G_i(c_i) \tag{3}$$

where:

$$Q(\mathbf{g},\mathbf{c}) = \prod_{i=1}^{N} [1 - q_i + q_i \cdot g_i \cdot D(c_i)] \tag{4}$$

is the probability that the final output is defect-free.

Error type composition. Let $\phi \in [0,1]$ denote the **fraction of deep errors** among all defect types. A deep error is one where the evaluator's miss probability depends on ρ (shared training-data blind spots affect detection). A surface error ($1-\phi$ fraction) has miss probability independent of ρ: the evaluator detects the error regardless of shared biases. The effective miss rate for a randomly drawn defect is $m = (1-\phi)m_{\text{surface}} + \phi \cdot m_{\text{deep}}(\rho)$, where m_{deep} is increasing in ρ and m_{surface} is not. In the experiments, surface errors are detected at 100%, medium errors at 78%, and deep errors at 73%, suggesting $\phi \approx 1/3$ for the tested battery.

Notation. Throughout the paper, $D(c) \equiv p(c)(1-\rho(c))$, $m(c) = 1 - D(c)$, and $D_{\max} \equiv \max_c D(c)$.

3 Results

3.1 Gate placement with correlated evaluation

Lemma 1 (Modified Dorfman-Savage with correlation discount)**.** *Consider the gate-placement problem with context fractions c_i fixed at some common level $c \in [0,1]$. Gate i has positive net value if and only if:*

$$q_i \cdot D(c) \cdot W_i > G_i(c) \tag{5}$$

where $W_i = \sum_{j=i+1}^{N} k_j + V \cdot \prod_{j \neq i}[1 - q_j \cdot m(c)]$ is the downstream value protected by gate i. As $\rho \to 0$, the condition reduces to the classical Dorfman-Savage rule.

Proof. The designer's objective (3) is separable in gate placements when context fractions

are fixed. The marginal contribution of activating gate i (setting $g_i = 1$ versus $g_i = 0$) is:

$$\Delta_i = V \cdot \left[\prod_{j \neq i}(1 - q_j m)\right] \cdot q_i D(c) + \sum_{j=i+1}^{N} k_j \cdot q_i D(c) - G_i(c)$$

where the first term is the increase in expected output value from catching a stage-i defect (detection saves the pipeline from a defect that would otherwise render the output worthless with probability $q_i m$), and the second term is the expected savings from avoiding wasted downstream production costs. Collecting terms, $\Delta_i = q_i D(c) W_i - G_i(c)$, where $W_i \equiv \sum_{j=i+1}^{N} k_j + V \cdot \prod_{j \neq i}[1 - q_j m]$ is the downstream value protected by gate i. Gate i should be placed if and only if $\Delta_i > 0$. □

Correlation acts as a tax on every gate. The effective detection rate $D(c) = p(c)(1-\rho(c))$ is the raw detection probability discounted by the independence factor. When generator and evaluator share biases, each gate catches fewer defects per dollar spent.

Corollary 1 (Gate ordering principle)**.** *For fixed context, gate i is more valuable than gate $j > i$ whenever $q_i W_i > q_j W_j$ and $G_i(c) \leq G_j(c)$. Since $W_i > W_j$ (earlier gates protect more downstream investment), earlier gates tend to dominate, rationalizing placement of the most stringent quality checks early in any sequential analysis pipeline.*

3.2 The Correlation Cascade

Proposition 1 (Correlation Cascade)**.** *Consider a pipeline of N stages with gates at every boundary. Stages are homogeneous: $q_i = q$, $k_i = k$ for all i, and each stage contributes productive value $v > 0$, so $V(N) = Nv$. Let D denote the effective detection rate at the chosen context level and $m = 1 - D$ the miss rate.*

(a) The probability that the final output is defect-free is $Q(N) = (1 - qm)^N$.

(b) The designer's payoff is $\Pi(N) = Nv(1 - qm)^N - Nk - (N - 1)G$.

(c) There exists a unique $N^* \geq 1$ maximizing Π, and N^* is strictly decreasing in m (hence decreasing in ρ for any fixed p).

(d) ***Quality ceiling bound.*** *$N^* < 1/(qm) - 1$. This bound depends on the product $qm = q(1 - p(1 - \rho))$, so both defect frequency and evaluation correlation jointly determine the maximum feasible pipeline length.*

Proof. Define $\alpha \equiv 1 - qm$ and $\beta \equiv qm = 1 - \alpha$. The marginal value of extending from N to $N + 1$ stages is:

$$\Delta(N) = v\alpha^N[1 - (N + 1)\beta] - k - G. \tag{6}$$

Define $h(N) \equiv v\alpha^N[1 - (N + 1)\beta]$. Taking the derivative:

$$h'(N) = v\alpha^N[\ln \alpha \cdot (1 - (N + 1)\beta) - \beta].$$

When $h(N) > 0$, we have $1 - (N + 1)\beta > 0$. Since $\ln \alpha < 0$ and $\beta > 0$, both terms in the bracket are negative, so $h'(N) < 0$: the marginal value is strictly decreasing wherever positive. Therefore $\Delta(N)$ crosses zero at most once, establishing uniqueness of N^*.

Upper bound (quality ceiling). $\Delta(N) > 0$ requires $1 - (N + 1)\beta > 0$, i.e., $N < 1/\beta - 1 = 1/(qm) - 1$.

Comparative static $\partial N^*/\partial m < 0$. Higher m increases β and decreases α. Both effects reduce $\Delta(N)$ for every N: the upper bound $1/\beta - 1$ shrinks, and the decay rate α^N steepens. Since Δ is a decreasing function that shifts down uniformly, its unique zero crossing moves left: N^* decreases. Since $\partial m/\partial \rho = p > 0$ for fixed p, N^* is decreasing in ρ. □

Each new stage adds value v but introduces a defect with probability q. The gate catches the defect with effective probability $D = p(1 - \rho)$. When correlation is high, D is small, so most defects survive the gate. As stages accumulate, the probability that at least one defect survives grows. The pipeline hits diminishing returns not because the model lacks capability, but because correlated evaluation cannot keep up with error accumulation.

Quantitative calibration. Table 1 shows how the quality ceiling varies with ρ for fixed $q = 0.15$, $p = 0.85$.

Table 1: Upper bound on feasible pipeline length as a function of error correlation

ρ	$m = 1 - 0.85(1 - \rho)$	qm	Upper bound $\lfloor 1/(qm) - 1 \rfloor$	Ratio to $\rho = 0$
0.00	0.150	0.0225	43	1.00
0.10	0.235	0.0353	27	0.63
0.15	0.278	0.0416	23	0.53
0.20	0.320	0.0480	19	0.46
0.30	0.405	0.0608	15	0.36
0.50	0.575	0.0863	10	0.24

At $\rho = 0.15$ (the experimental estimate from Section 4), the upper bound on feasible pipeline length drops to 53% of the independent-evaluator benchmark. At $\rho = 0.30$, it

drops to 36%. The relationship is approximately linear in ρ for moderate values: each ten percentage points of correlation reduces the feasible pipeline by roughly 15–20%.

3.3 Value of evaluator independence

Correlated evaluation is qualitatively worse than independent evaluation whenever $\rho > 0$. Proposition 2 provides the quantitative cost as an explicit function of observables; Section 4 provides the methodology for estimating $\rho_{\min}$.

Proposition 2 (Value of evaluator independence)**.** *Let $V_{\text{self}}(\rho_{\min})$ be the value of the optimally designed single-model pipeline and V_{ind} be the value of a pipeline with a truly independent evaluator ($\rho(c) = 0$ for all c). Define the* ***value of evaluator independence****:*

$$\Lambda \equiv V_{\text{ind}} - V_{\text{self}}(\rho_{\min}). \tag{7}$$

Then:

(a) $\Lambda > 0$ *whenever* $\rho_{\min} > 0$*.*

(b) Λ *is strictly increasing in* $\rho_{\min}$*.*

(c) Λ *is increasing in the optimal pipeline length* N^{**}*: longer pipelines benefit more from independence because the cascade compounds over more stages.*

(d) Λ *is increasing in* ϕ *(the fraction of deep errors): deep errors are exactly the errors where correlation matters.*

Proof. Part (a). With $\rho_{\min} > 0$, the single-model pipeline has $D(c) = p(c)(1 - \rho(c)) < p(c)$ for all c. An independent evaluator achieves $D_{\text{ind}}(c) = p(c)$. Strictly higher effective detection at every gate means strictly higher quality or strictly lower costs, so $V_{\text{ind}} > V_{\text{self}}$.

Part (b). $V_{\text{self}}(\rho_{\min})$ is decreasing in $\rho_{\min}$: increasing $\rho_{\min}$ shifts $\rho(c)$ upward for all c (since $\rho(c) \geq \rho_{\min}$), reducing $D(c)$ for all c and reducing quality or requiring shorter pipelines. V_{ind} is unaffected. Therefore $\Lambda = V_{\text{ind}} - V_{\text{self}}$ is increasing.

Part (c). The quality ratio $Q_{\text{self}}/Q_{\text{ind}} = \prod_{i=1}^{N}(1 - q_i m^*)/(1 - q_i m_{\text{ind}})$ is the product of N terms, each less than 1 (since $m^* > m_{\text{ind}}$). This product is strictly decreasing in N, so the gap in pipeline value grows with pipeline length.

Part (d). Surface errors ($1 - \phi$ fraction) have the same detection rate regardless of ρ. Deep errors (ϕ fraction) are where correlation reduces effective detection. Higher ϕ means more defects are affected by correlation, widening the gap. □

Closed-form value of independence for homogeneous pipelines. Under homogeneous stages (Proposition 1 setting), let N^*_{ind} and N^*_{self} be the optimal pipeline lengths under independent and self-evaluation respectively, and let $\Pi^*(N; D)$ denote the optimized payoff for a pipeline of length N with effective detection rate D. Then:

$$\Lambda = \Pi^*(N^*_{\text{ind}}; p) - \Pi^*(N^*_{\text{self}}; p(1 - \rho_{\min})) \tag{8}$$

where $\Pi^*(N; D) = Nv(1 - q(1 - D))^N - Nk - (N - 1)G$. This expression depends on six observable quantities: $\rho_{\min}$, p, q, v, k, and G.

Table 2 reports Λ for a range of $\rho_{\min}$ values with baseline parameters $p = 0.85$, $q = 0.15$, $v = 10$, $k = 1$, $G = 0.55$.

Table 2: Value of evaluator independence as a function of $\rho_{\min}$

$\rho_{\min}$	D_{self}	N^*_{self}	Π^*_{self}	N^*_{ind}	Π^*_{ind}	$\Lambda/\Pi^*_{\text{ind}}$
0.00	0.850	30	105.6	30	105.6	0.0%
0.05	0.808	24	82.2	30	105.6	22.2%
0.10	0.765	19	67.2	30	105.6	36.4%
0.15	0.723	16	56.8	30	105.6	46.2%
0.20	0.680	14	49.1	30	105.6	53.5%
0.30	0.595	11	38.7	30	105.6	63.4%
0.50	0.425	8	27.0	30	105.6	74.4%

At $\rho_{\min} = 0.15$ (the experimental estimate from the blind spot analysis in Section 4), the value of independence represents 46% of the independent-evaluator pipeline value. The self-evaluating pipeline is optimally limited to 16 stages (versus 30 under independence), and its total value is roughly half the independent benchmark.

Sensitivity to defect rate. The magnitude of Λ depends on q. For low-defect-rate pipelines ($q = 0.05$), Λ is actually larger because low defect rates permit much longer pipelines, and the cumulative cost of correlation compounds over many more stages. Table 3 reports Λ across three defect rates.

Table 3: Value of independence sensitivity to defect rate

$\rho_{\min}$	Λ at $q = 0.05$	Λ at $q = 0.15$	Λ at $q = 0.30$
0.10	115.4	38.4	19.3
0.15	146.8	48.8	24.4
0.30	200.9	66.9	33.5

Corollary 2 (Single-model domination threshold)**.** *The single-model pipeline is strictly dominated by an independent-evaluator pipeline whenever $\Lambda > C_{\text{ind}}$, where C_{ind} is the cost of accessing an independent evaluator. This defines a threshold $\rho_{\min}^{\text{crit}}$ implicitly by $\Lambda(\rho_{\min}^{\text{crit}}) = C_{\text{ind}}$. By the intermediate value theorem and strict monotonicity of Λ in $\rho_{\min}$, the threshold $\rho_{\min}^{\text{crit}}$ exists and is unique.*

3.4 Context-invariant detection

Proposition 3 (Context-invariant detection)**.** *Suppose detection is context-invariant ($p(c) = p$ for all c) but correlation has an irreducible floor ($\rho_{\min} > 0$) and $\rho'(c) \geq 0$. Then:*

(a) The effective detection rate simplifies to $D(c) = p(1 - \rho(c))$, which is weakly decreasing in c.

(b) If $\rho'(c) > 0$ and detection stakes dominate false-positive savings ($q_i W_i p \rho'(c) > |f'(c)| k_i$ for all c), the optimal context is $c^* = 0$: minimize correlation, since detection is unaffected. When $f(c) \approx 0$ (as the experiments suggest for capable models with zero false-positive rate), $c^* = 0$ holds unambiguously.

(c) If $\rho'(c) = 0$ (both detection and correlation are context-invariant), then $D(c) = p(1-\rho_{\min})$ is constant in c, and context affects the designer's objective only through the false-positive channel. The optimal context is $c^* = 1$ (minimize false positives).

(d) In either case, the pipeline problem reduces to: how many gates, with irreducible miss rate $m = 1 - p(1 - \rho_{\min})$, are worth placing? The optimal pipeline length depends only on $\rho_{\min}$, not on context design.

(e) The value of evaluator independence is $\Lambda = V_{\text{ind}} - V_{\text{self}}$, where V_{self} uses the constant effective detection rate $D = p(1 - \rho_{\min})$ and V_{ind} uses $D_{\text{ind}} = p$. Λ is fully determined by $\rho_{\min}$.

Proof. Part (a). With $p(c) = p$ constant, $D(c) = p(1 - \rho(c))$. Since $\rho'(c) \geq 0$, $D'(c) = -p\rho'(c) \leq 0$.

Part (b). With $p' = 0$, the marginal effect of context at gate i is:

$$\frac{\partial \Omega_i}{\partial c} = -q_i W_i p \rho'(c) + |f'(c)| k_i.$$

When $q_i W_i p \rho'(c) > |f'(c)| k_i$ for all c, $\partial \Omega_i / \partial c < 0$ everywhere and $c^* = 0$.

Part (c). With $\rho' = 0$ and $p' = 0$, $D(c) = p(1 - \rho_{\min})$ is constant. The objective becomes $\Omega_i(c) = q_i W_i p(1 - \rho_{\min}) - f(c) k_i$. Since $f'(c) < 0$, Ω_i is increasing in c, so $c^* = 1$.

Part (d). In both cases, the effective detection rate at the optimally chosen context is $D^* = p(1-\rho_{\min})$. The miss rate is $m^* = 1-p(1-\rho_{\min})$. The optimal pipeline length follows from Proposition 1 with this m^*, and depends on $\rho_{\min}$ through m^*.

Part (e). The independent evaluator has $D_{\text{ind}} = p$ (no correlation discount). The value of independence is $\Lambda = V_{\text{ind}}(D_{\text{ind}} = p) - V_{\text{self}}(D^* = p(1-\rho_{\min}))$, which depends on p and $\rho_{\min}$. □

The experiments in Section 4 support this proposition. For the 120-billion-parameter model, correct detection rates are 85%, 80%, and 87% at $c = 0$, $c = 0.5$, and $c = 1$ respectively ($p = 0.83$ for the null on context effect). Raw detection rates are 87–88% at all context levels. The zero false-positive rate across all conditions simplifies the analysis further: $f(c) \approx 0$, so gate cost reduces to the fixed cost g.

3.5 Interior optimal context

Assumption 5 (Sufficient correlation for interior optimum)**.** *$p'(c) > 0$, $\rho'(c) > 0$, and $p_{\max}\rho'(1) > p'(1)(1-\rho_{\max})$.*

Proposition 4 (Interior optimal context)**.** *Under Assumption 5, the effective detection rate $D(c) = p(c)(1-\rho(c))$ is strictly concave with a unique interior maximum $c^D \in (0,1)$, and the optimal context fraction c_i^* at gate i satisfies:*

(a) $c_i^* \in (0,1)$: the optimal context is strictly interior.

(b) c_i^* is decreasing in W_i (downstream value): high stakes push toward detection-maximizing context.

(c) c_i^* is increasing in k_i (stage cost): expensive rework pushes toward more context to reduce false alarms.

(d) c_i^* is decreasing in the marginal correlation sensitivity $\rho'(c_i^*)$.

Proof. Global concavity of D. Since $p'(c) > 0$, $p''(c) < 0$, $\rho'(c) > 0$, and $\rho''(c) \geq 0$:

$$D''(c) = p''(c)(1-\rho(c)) - 2p'(c)\rho'(c) - p(c)\rho''(c) < 0.$$

Each term is non-positive: $p'' < 0$ and $(1-\rho) > 0$; $p' > 0$ and $\rho' > 0$; $p > 0$ and $\rho'' \geq 0$. So D is strictly concave, hence single-peaked with a unique interior maximum c^D.

The objective at gate i is $\Omega_i(c) = q_i W_i D(c) - f(c)k_i$ (dropping the constant g). The FOC is:

$$q_i W_i D'(c_i^*) = f'(c_i^*)k_i. \tag{9}$$

Since $f'(c) < 0$, the right side is negative, so $D'(c_i^*) < 0$: the optimum is past the detection peak, $c_i^* > c^D$.

SOC: $\partial^2\Omega_i/\partial c^2 = q_i W_i D''(c_i^*) - f''(c_i^*)k_i < 0$ since $D'' < 0$ and $f'' > 0$.

Part (a). At $c = 0$: $\partial\Omega_i/\partial c = q_i W_i D'(0) + |f'(0)|k_i > 0$ (since $D'(0) > 0$ under Assumption 5). At $c = 1$: $\partial\Omega_i/\partial c < 0$ under Assumption 5 for non-negligible W_i. By continuity, $c_i^* \in (0, 1)$.

Part (b). By the implicit function theorem on $H(c_i^*, W_i) \equiv q_i W_i D'(c_i^*) - f'(c_i^*)k_i = 0$:

$$\frac{dc_i^*}{dW_i} = -\frac{q_i D'(c_i^*)}{\text{SOC}} < 0.$$

The numerator $q_i D'(c_i^*)$ is negative (since $c_i^* > c^D$) and the SOC is negative, so the ratio is negative.

Part (c). $dc_i^*/dk_i = -(-f'(c_i^*))/\text{SOC} > 0$.

Part (d). Increasing $\rho'(c)$ makes $D'(c)$ more negative (since $D'(c) = p'(c)(1 - \rho(c)) - p(c)\rho'(c)$), shifting the FOC and reducing c_i^* by the implicit function theorem. □

Assumption 5 is more likely to hold for less capable evaluators, where additional context genuinely aids error identification, or for subjective evaluation tasks (novelty assessment, economic importance) where context provides qualitative information. For highly capable models, detection is context-invariant and Proposition 3 applies.

3.6 Context design extends the pipeline

Proposition 5 (Context design extends the optimal pipeline)**.** *Under Assumption 5, let N^* be the optimal pipeline length under full context sharing ($c_i = 1$), N^{**} the optimal length under optimized context, and N_{ind} the optimal length under independent evaluation. Then:*

$$N^* \leq N^{**} \leq N_{\text{ind}}. \tag{10}$$

*The first inequality is strict when $\rho_{\max} > \rho_{\min}$ and the improvement in effective detection from context optimization at the marginal stage exceeds the increase in false-positive costs (*Condition S*). The second inequality is strict when $\rho_{\min} > 0$.*

Proof. $N^{**} \geq N^*$*:* The full-context allocation $c_i = 1$ is feasible, so $\Pi^{**}(N) \geq \Pi^*(N)$ for all N. The optimized pipeline extends at least as far.

$N^{**} > N^*$ *under Condition S:* At $N = N^* + 1$, context optimization at the new gate achieves $D(c^D) > D(1)$ (by strict concavity of D and $c^D < 1$). The detection gain $q[D(c^D) -$

$D(1)]W_{N^*+1}$ exceeds the false-positive cost increase $[f(c^D) - f_{\min}]k$ by Condition S, so the net gate value at stage $N^* + 1$ turns positive. Therefore $N^{**} \geq N^* + 1$.

$N^{**} \leq N_{\text{ind}}$*:* Under independent evaluation, $D_{\text{ind}} = p_{\max} \geq p(c^D)(1 - \rho(c^D)) = D(c^D)$ (since $\rho(c^D) \geq \rho_{\min} > 0$). Higher detection rate means lower miss rate, allowing a longer optimal pipeline. □

When $p'(c) \approx 0$ (as the experiments suggest for capable models), $D(c^D) \approx D(0) = p(1 - \rho_{\min})$ and context design cannot extend the pipeline. The gap $N^{**} - N^*$ collapses to zero, and the pipeline design problem reduces to choosing N given the fixed miss rate $m = 1 - p(1 - \rho_{\min})$. Proposition 5's contribution is conditional on the context channel operating, which may hold for less capable models, subjective evaluation tasks, or settings where context genuinely provides new information.

3.7 Comparative statics

Table 4 summarizes the effects of key parameters on optimal pipeline design.

Table 4: Comparative statics summary

Parameter	Effect on N^*	Effect on Λ	Mechanism
$\rho_{\min} \uparrow$	$\downarrow$	$\uparrow$	Higher correlation reduces gate effectiveness
$p \uparrow$	$\uparrow$	Ambiguous	Higher capability extends pipeline
$q \uparrow$	$\downarrow$	Ambiguous	More defects shorten both pipelines
$v/k \uparrow$	$\uparrow$	$\uparrow$	Higher value-to-cost ratio amplifies cascade
$\phi \uparrow$	$\downarrow$	$\uparrow$	More deep errors amplify correlation channel
$N \uparrow$	–	$\uparrow$	Longer pipelines compound disadvantage

The comparative statics on N^* and Λ with respect to $\rho_{\min}$ follow from Propositions 1(c) and 2(b) respectively. The effect of q on Λ is ambiguous because higher q shortens both pipelines (reducing Λ through the pipeline-length channel) but raises the per-stage correlation cost (increasing Λ through the per-gate channel).

3.8 Robustness to continuous quality degradation

Proposition 6 (Correlation Cascade under continuous degradation)**.** *Consider a pipeline of N stages where each stage adds noise ϵ_i drawn independently with $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.*

The final output has quality index $\Theta = \sum_{i=1}^{N} \epsilon_i$ *and value* $V(\Theta) = \bar{V} - \lambda \mathrm{Var}(\Theta)$ *for some* $\lambda > 0$*. A gate after stage* i *detects and removes fraction* $D = p(1-\rho)$ *of the variance contributed by stage* i*. Then:*

(a) The residual variance after optimal gating is $\mathrm{Var}(\Theta_{\mathrm{res}}) = N\sigma^2(1-D)$*.*

(b) The designer's payoff is $\Pi(N) = \bar{V} - \lambda N\sigma^2(1-D) + Nv - Nk - (N-1)G$*, which is concave in* N *with a unique interior maximum* N^**.*

(c) $\partial N^*/\partial\rho < 0$*: higher correlation reduces the optimal pipeline length.*

(d) The value of evaluator independence is $\Lambda = \lambda N^{**}\sigma^2\rho p - \lambda(N^{**} - N^*_{\mathrm{self}})\sigma^2(1-p) > 0$ *whenever* $\rho > 0$*, where* N^{**} *and* N^*_{self} *are the optimal lengths under independent and correlated evaluation.*

Proof. Part (a). Each gate removes fraction D of stage-i variance. Residual variance per stage is $\sigma^2(1-D)$. By cross-stage independence, total residual variance is $N\sigma^2(1-D)$.

Part (b). Substituting into $V(\Theta)$: $\Pi(N) = \bar{V} + Nv - \lambda N\sigma^2(1-D) - Nk - (N-1)G$. The marginal value of extending from N to $N+1$ is $\Delta(N) = v - \lambda\sigma^2(1-D) - k - G$, which is constant in N. The optimal N^* is the largest integer satisfying $v - \lambda\sigma^2(1-D) - k - G > 0$, or $N^* = \lfloor (v-k-G)/(\lambda\sigma^2(1-D)) \rfloor$ when the payoff function includes diminishing marginal returns from additional stages (e.g., $v(N) = v/N$ or convex costs).

For concreteness, with per-stage value $v(N) = v \cdot \alpha^{N-1}$ (declining marginal value, $\alpha < 1$), the marginal stage value is $\Delta(N) = v\alpha^{N-1} - \lambda\sigma^2(1-D) - k - G$. Since $\alpha < 1$, $\Delta(N)$ is strictly decreasing, giving a unique N^*.

Part (c). $\partial(1-D)/\partial\rho = p > 0$, so residual variance per stage increases in ρ. Higher residual variance reduces the marginal value of each stage, so N^* decreases.

Part (d). Under independent evaluation, $D_{\mathrm{ind}} = p$ and residual variance per stage is $\sigma^2(1-p)$. Under correlated evaluation, residual variance per stage is $\sigma^2(1-p(1-\rho)) = \sigma^2(1-p+p\rho)$. For any fixed N, the value gap is $\lambda N\sigma^2 p\rho > 0$. The full value of independence also accounts for the difference in optimal pipeline lengths. □

The key comparative statics carry over: N^* is decreasing in ρ, $\Lambda > 0$ whenever $\rho > 0$, and Λ is increasing in ρ. The quantitative magnitudes differ because continuous degradation is less catastrophic than binary defects (individual noise contributions reduce value smoothly rather than rendering the output worthless), so the continuous model yields longer optimal pipelines. The binary model is a conservative approximation that produces sharper bounds.

Corollary 3 in Appendix A shows that idea screening in early pipeline stages can be modeled as a real-options problem, recovering a Weitzman (1979)-style reservation-value

result. Because the screening result does not interact with the correlation mechanism, it is relegated to the appendix.

3.9 Commitment value of fixed thresholds

The preceding analysis treats the gate as a binary mechanism: detect or miss, with probability $D(c)$. In practice, the evaluator observes a continuous quality signal and applies a threshold. When the evaluator can adjust the threshold based on its own confidence, the correlation structure creates a new channel through which shared biases reduce detection.

Proposition 7 (Commitment value of fixed thresholds)**.** *Extend the gate technology so that the evaluator at each gate observes two signals:*

(i) A quality signal $s \in \mathbb{R}$: higher s indicates higher perceived quality. Conditional on the true state:

- *No defect ($\theta = 0$): $s \sim F_0$ with mean μ_0.*
- *Independent defect (type-I, probability $1 - \rho$): $s \sim F_1$ with mean $\mu_1 < \mu_0$.*
- *Correlated defect (type-C, probability ρ): $s \sim F_C$ with mean $\mu_C \in [\mu_1, \mu_0]$. Shared reasoning makes the defective work appear higher-quality than a type-I defect.*

(ii) A confidence signal $z \in \mathbb{R}$: higher z means the evaluator is more confident the work is correct. Conditional on the true state:

- *No defect: $z \sim G_0$ with mean ν_0.*
- *Type-I defect: $z \sim G_1$ with mean $\nu_1 < \nu_0$ (the evaluator senses something is off).*
- *Type-C defect: $z \sim G_C$ with mean $\nu_C \geq \nu_0$ (the evaluator finds the generator's reasoning convincing because it shares the same training).*

Consider two threshold rules, both achieving false-positive rate α:

- ***Fixed threshold** $\bar{\tau}$: pass iff $s \geq \bar{\tau}$. The confidence signal z is ignored. The designer commits to $\bar{\tau}$ ex ante.*
- ***Naive adaptive threshold** $\tau(z)$: the evaluator adjusts the threshold using z, calibrated under the incorrect belief that $\rho = 0$. Since high z implies $\theta = 0$ under $\rho = 0$, the evaluator lowers the acceptance bar when confident: $\tau'(z) < 0$.*

Let $D_{\text{fixed}}(\alpha)$ and $D_{\text{naive}}(\alpha, \rho)$ denote the overall detection rates under each rule. Then:

(a) At $\rho = 0$: $D_{\text{naive}}(\alpha, 0) \geq D_{\text{fixed}}(\alpha)$. The naive rule is (weakly) optimal when its beliefs are correct.

(b) $D_{\text{naive}}(\alpha, \rho)$ is strictly decreasing in ρ for $\rho > 0$.

(c) There exists $\bar{\rho} \in (0, 1)$ such that for all $\rho > \bar{\rho}$: $D_{\text{fixed}}(\alpha) > D_{\text{naive}}(\alpha, \rho)$. The fixed threshold strictly dominates the naive adaptive rule.

(d) The commitment value $\Gamma(\rho) \equiv D_{\text{fixed}}(\alpha) - D_{\text{naive}}(\alpha, \rho)$ is increasing in ρ for $\rho > \bar{\rho}$.

Proof. Decomposition by defect type. Under any threshold rule, the overall detection rate decomposes as:

$$D(\alpha, \rho) = (1 - \rho) \cdot D_I + \rho \cdot D_C$$

where $D_I = P(\text{reject} \mid \text{type-I})$ and $D_C = P(\text{reject} \mid \text{type-C})$.

Part (a). At $\rho = 0$, all defects are type-I. The naive evaluator's belief is correct ($\rho = 0$ is the truth), so the adaptive rule is Bayesian-optimal given both s and z. By sufficiency, using additional information weakly improves detection at any fixed false-positive rate.

Part (b). The naive adaptive threshold $\tau(z)$ is decreasing in z: higher confidence lowers the bar. Type-C defects generate $z \sim G_C$ with mean $\nu_C \geq \nu_0$, so they face systematically lower thresholds than type-I defects (which generate $z \sim G_1$ with mean $\nu_1 < \nu_0$). Combined with the fact that F_C is closer to F_0 than F_1 is (type-C defects are less anomalous in the quality signal), this gives $D_C^{\text{naive}} < D_I^{\text{naive}}$. Since $\partial D_{\text{naive}}/\partial \rho = D_C^{\text{naive}} - D_I^{\text{naive}} < 0$, the naive detection rate is strictly decreasing in ρ.

Part (c). The fixed threshold does not condition on z, so it does not differentially treat type-C and type-I defects through the confidence channel. Still, $D_C^{\text{fixed}} < D_I^{\text{fixed}}$ because F_C is closer to F_0, but the gap is smaller than under the naive rule because the fixed rule does not compound the disadvantage through confidence-based threshold adjustment. At $\rho = 1$ (all defects are type-C), $D_{\text{naive}}(\alpha, 1) = D_C^{\text{naive}} < D_C^{\text{fixed}} = D_{\text{fixed}}(\alpha)$: the naive rule's lenient threshold for high-confidence observations strictly underperforms the fixed threshold's uniform treatment. Since D_{naive} is continuous in ρ, decreasing (part (b)), and starts above D_{fixed} at $\rho = 0$ (part (a)), the intermediate value theorem gives a unique crossing $\bar{\rho}$.

Part (d). For $\rho > \bar{\rho}$:

$$\Gamma'(\rho) = \frac{\partial D_{\text{fixed}}}{\partial \rho} - \frac{\partial D_{\text{naive}}}{\partial \rho} = (D_C^{\text{fixed}} - D_I^{\text{fixed}}) - (D_C^{\text{naive}} - D_I^{\text{naive}})$$

Since $D_C^{\text{naive}} < D_C^{\text{fixed}}$ (naive is worse on type-C) and $D_I^{\text{naive}} > D_I^{\text{fixed}}$ (naive is better on type-I, using z correctly for independent defects), we have $(D_C^{\text{naive}} - D_I^{\text{naive}}) < (D_C^{\text{fixed}} - D_I^{\text{fixed}}) < 0$, so $\Gamma'(\rho) > 0$. □

The naive evaluator misinterprets its own confidence. High confidence (z high) is consistent with both "no defect" and "correlated defect," but the naive evaluator, believing $\rho = 0$, treats high confidence as evidence of correct work and lowers the acceptance bar. The evaluator is most lenient precisely on the defects where its judgment is most contaminated.

The problem is not strategic: the evaluator has no incentive to manipulate. The evaluator applies Bayes' rule correctly given its (incorrect) beliefs. The commitment value arises because the designer, who knows $\rho > 0$, can prevent the evaluator from acting on contaminated confidence by committing to a fixed threshold. The parallel to Tirole (1986) is exact: rigid rules dominate discretion when the auditor's judgment is contaminated, whether the contamination is strategic (collusion) or stochastic (shared training biases).

4 Experiments

The evaluation protocol (how the evaluator is asked to respond) dominates both the context channel (how much upstream reasoning the evaluator receives) and model identity (which model evaluates) as a determinant of ρ. The experiments below measure detection rates across context levels, models, and response formats, identifying which design variables affect ρ and quantifying the gains from protocol optimization.

4.1 Design

Stimuli. Twenty-one mathematical derivations with planted errors serve as the test battery: 7 surface errors (sign mistakes, arithmetic), 7 medium errors (incorrect formula application, wrong distributional assumptions), and 7 deep errors (conceptual mistakes requiring domain understanding). Nine correct derivations serve as false-positive controls.

All stimuli involve standard financial economics: CAPM, Black-Scholes, Modigliani-Miller, portfolio theory, fixed income pricing. Errors span the range from easily detectable (a sign error in a derivative) to conceptually subtle (misinterpreting a pricing equation as a diversification statement).

Conditions. Each error stimulus is evaluated at three context levels:

- $c = 0$: The evaluator receives only the derivation (no upstream reasoning).
- $c = 0.5$: The evaluator receives the derivation plus a summary of the generator's approach.

- $c = 1$: The evaluator receives the derivation plus the generator's full chain-of-thought reasoning.

Two models from the same provider are tested: a 120-billion-parameter model (120b) and a 20-billion-parameter model (20b). Temperature is fixed at 0.3 across all conditions. Each condition-stimulus pair is replicated 5 times. Total: 627 valid error-detection trials and 135 valid false-positive trials across 765 API calls.

Classification. Each response is classified as:

- *Correct detection:* The evaluator identifies the planted error by describing the specific mistake using language consistent with the ground truth.
- *Spurious detection:* The evaluator flags an issue but describes a different error than the one planted.
- *Miss:* The evaluator does not flag any error.

Classification uses keyword overlap between the evaluator's response and the ground truth description. This automated classifier has limitations (discussed below), but manual inspection of a subsample confirms that it correctly classifies the large majority of responses.

4.2 Results

4.2.1 Detection rates by context level (120b)

Table 5 reports correct detection rates for the 120b model.

Table 5: Correct detection rate by error type and context level (120b model)

Error Type	$c = 0$	$c = 0.5$	$c = 1.0$	All
Surface	35/35 (100%)	35/35 (100%)	35/35 (100%)	105/105 (100%)
Medium	28/35 (80%)	27/35 (77%)	27/35 (77%)	82/105 (78%)
Deep	26/35 (74%)	22/35 (63%)	29/35 (83%)	77/105 (73%)
All	**89/105 (85%)**	**84/105 (80%)**	**91/105 (87%)**	**264/315 (84%)**

Raw detection rates (correct plus spurious) are 87–88% across all context levels, with no statistically significant variation.

Statistical tests. A chi-squared test with Yates correction for the comparison of correct detection at $c = 0$ versus $c = 1$ yields $p = 0.83$. The 2-percentage-point difference (85% versus 87%) is not statistically significant. For deep errors alone, the pattern is non-monotonic (74% at $c = 0$, 63% at $c = 0.5$, 83% at $c = 1$; $p = 0.54$ for $c = 0$ versus $c = 1$). The non-monotonic pattern may reflect stimulus-specific effects or sampling variability with 7 deep stimuli.

Interpretation. For the 120b model, detection is approximately context-invariant. The theory's prediction that $\rho(c)$ is increasing in c receives no support from the 120b data. Proposition 3 characterizes optimal pipeline design under context-invariant detection, and the experiments confirm that this characterization applies to capable evaluators.

4.2.2 Cross-model comparison

Table 6 reports correct detection rates for the 20b model.

Table 6: Correct detection rate by error type and context level (20b model)

Error Type	$c = 0$	$c = 0.5$	$c = 1.0$	All
Surface	22/35 (63%)	24/35 (69%)	27/35 (77%)	73/105 (70%)
Medium	23/33 (70%)	21/35 (60%)	15/35 (43%)	59/103 (57%)
Deep	13/35 (37%)	15/35 (43%)	19/34 (56%)	47/104 (45%)
All	**58/103 (56%)**	**60/105 (57%)**	**61/104 (59%)**	**179/312 (57%)**

The 20b model detects errors at a similar raw rate (74–81%) but correctly identifies the planted error far less often (56–59% versus 85–87% for 120b; $p < 0.001$). The 20b model frequently flags a different error than the one planted, consistent with lower evaluation capability rather than independent evaluation.

Capability-controlled comparison. To separate capability from correlation, restrict to the 17 stimuli that 120b detects at $c = 0$ (3 or more out of 5 replications) and compare detection at $c = 1$:

Table 7: Capability-controlled cross-model comparison at $c = 1$

Evaluator	Correct detection	p-value
Same model (120b)	85/85 (100%)	
Cross model (20b)	60/84 (71%)	< 0.001

For errors that 120b reliably catches, 120b at $c = 1$ catches all of them correctly. The 20b model catches only 71%. This result does not test Proposition 2's prediction about the value of evaluator independence, because the two models are from the same provider and likely share substantial training data. The theory predicts independence benefits from evaluators with different training distributions, not merely different model sizes within the same provider.

Instead, the cross-model comparison confirms a different prediction: models from the same training distribution are *not* independent evaluators. The 20b model performs strictly worse than the 120b model, consistent with a smaller model sharing the larger model's blind spots while also having lower baseline capability. The 20b model adds noise, not independence. The compound detection result in Section 6 (Equation 12) predicts exactly this outcome: when evaluators share a common factor from training data, adding evaluators from the same provider yields diminishing returns. Testing models from genuinely different providers would validate Proposition 2 for the independence case.

4.2.3 Protocol-induced blind spots

The initial structured-format experiments identified stimulus D3 (a conceptual error about CAPM diversification) as a shared blind spot: zero detections in 30 trials across both models, all context levels, and all replications. The 120b model reported `"has_error": false` with 92–93% confidence. A second stimulus, D5 (incorrectly claiming that log-utility investors are risk-neutral), was detected in only 6 out of 60 trials.

Two follow-up experiments tested whether these failures reflect *training-data* blind spots or *protocol-induced* blind spots.

Cross-provider test. The D3 stimulus was presented to five models from four providers (gpt-oss-120b, gpt-oss-20b, Llama-3.3-70b, Gemma-3-27b, Mistral-Small-3.1) using a free-form natural-language prompt. Temperature and stimulus wording were held constant. Result: 24/25 correct detections across all models, including gpt-oss-120b (5/5), which had scored 0/15 under structured format.

Full-battery protocol comparison. All 21 error stimuli were evaluated by gpt-oss-120b under the free-form format (5 replications per stimulus, $c = 0$). Table 8 reports the comparison with the structured-format results.

Four stimuli (M3, M5, D3, D5) show large improvements under free-form evaluation, with gains of 60–100 percentage points. These four stimuli share a common feature: the planted error is conceptual rather than computational, requiring the evaluator to reason about the

Table 8: Detection rate by response format (120b model, $c = 0$, 5 reps)

Stimulus	Structured	Free-form	Δ	Error type
S1–S7 (surface)	35/35 (100%)	35/35 (100%)	0	
M1, M2, M4, M6	20/20 (100%)	20/20 (100%)	0	
M3	2/5 (40%)	5/5 (100%)	+60pp	chi-squared
M5	2/5 (40%)	5/5 (100%)	+60pp	OLS/R^2
D1, D2, D4, D6, D7	25/25 (100%)	25/25 (100%)	0	
D3	0/5 (0%)	3/3 (100%)	+100pp	CAPM diversif.
D5	2/5 (40%)	5/5 (100%)	+60pp	log utility
All (any error)	**86/105 (82%)**	**100/100 (100%)**	+18pp	

relationship between two correct-looking statements rather than spot a local mistake. Under the free-form format, the 120b model detects all errors: 100/100 valid trials flagged an error.

The protocol effect is not D3-specific. Four out of 21 stimuli (19%) show format sensitivity exceeding 30 percentage points. All four involve conceptual errors. No surface error shows any format sensitivity.

Mechanism. The structured format requires the evaluator to output a JSON object beginning with `"has_error": true/false`. This forces a binary classification before the model has completed its reasoning about the error. For conceptual errors (where the derivation is locally correct but globally flawed), the premature commitment to a classification short-circuits the reasoning process. The free-form format allows the model to reason through to the conceptual distinction before rendering a verdict.

A follow-up experiment tested whether placing reasoning *before* classification in the structured format restores detection. Three format variants were compared on stimulus M3 (chi-squared moments, format-sensitive under the original design): (i) classification-first JSON (`has_error` before `reasoning`): 2/5 detection; (ii) reasoning-first JSON (`reasoning` before `has_error`): 2/5; (iii) chain-of-thought followed by JSON (free-text reasoning, then structured conclusion): 5/5. Requiring the model to reason in free text before committing to a structured classification restores detection to the free-form baseline. The order of fields within a JSON schema does not help; what matters is whether the model reasons in unconstrained natural language before the classification decision.

Implication for the theory. The evaluation protocol (response format, reasoning constraints) operates as a component of the context variable c in the theoretical framework. The theory models c as the fraction of upstream reasoning shared with the evaluator, but

the experiments reveal that c should be interpreted more broadly: it encompasses any design choice that affects the evaluator's effective detection rate, including how the evaluator structures its response. The effective correlation ρ is not a fixed property of the training data; it is a function of the evaluation protocol. Protocol design reduces ρ: switching from structured to free-form format eliminates all detected blind spots in the test battery.

4.2.4 False positive rates

Nine correct derivations were evaluated by the 120b model at all three context levels with 5 replications per condition. The false-positive rate is zero across all 135 trials (95% CI: $[0.0\%, 7.9\%]$). The model does not flag correct derivations as erroneous, regardless of context. The theory's $f(c)$ term is empirically negligible for capable models on well-specified mathematical problems, simplifying the gate cost to $G_i = g$.

Sensitivity to false-positive rate. The zero observed false-positive rate has a 95% confidence interval upper bound of 7.9%. At $f = 0.079$, gate cost rises from $G = g = 0.50$ to $G = g + f \cdot k = 0.50 + 0.079 \times 1.0 = 0.579$ (in normalized units). Under baseline parameters ($q = 0.15$, $p = 0.85$, $\rho = 0.15$, $v = 10$, $k = 1$), the optimal pipeline length N^*_{self} decreases from 16 to 15 stages, and the pipeline payoff Π^* decreases by approximately 3%. The value of independence Λ increases slightly (from 46% to 47% of the independent benchmark) because higher gate costs penalize the correlated pipeline more heavily. The results are not sensitive to the false-positive rate within the confidence interval.

4.3 Estimating protocol-dependent correlation

Three dimensions of evaluation design (context level, model identity, response format) vary while stimuli remain fixed, identifying which design choices drive detection failures. The protocol varies evaluation conditions along these three dimensions (context level, model identity, response format) while holding stimuli fixed, identifying which design choices drive detection failures.

1. **Construct a test battery** of K stimuli containing known defects spanning surface, medium, and deep error types.

2. **Vary evaluation conditions** across context levels ($c = 0, 0.5, 1$), models (different sizes and providers), and response formats (structured JSON vs. free-form natural language). Replicate each condition R times.

3. **Identify condition-dependent failures:** stimuli where detection varies systematically across conditions reveal which design variables affect ρ.

4. **Decompose ρ:** failures that persist across all conditions within a format but vanish under a different format indicate protocol-induced correlation. Failures that persist across all formats indicate training-data correlation.

Results. Under the structured JSON format, four stimuli (D3, D5, M3, M5) show detection rates of 0–40%, while all other stimuli are detected at 100%. Under the free-form format, all 21 stimuli are detected at 100% (100/100 valid trials). No stimulus in the battery is undetectable across *both* formats, so the experiments provide no evidence of a training-data floor $\rho_{\min} > 0$ for the error types tested.

The effective correlation under structured format is substantial: $\hat{\rho}_{\text{structured}} \geq 4/21 \approx 0.19$ (counting stimuli with detection below 50% as protocol-induced blind spots). Under free-form format, $\hat{\rho}_{\text{free-form}} \approx 0$. The difference measures the cost of the structured evaluation protocol in correlation units. Protocol choice shifts ρ by approximately 0.19, which by Proposition 1 changes optimal pipeline length by roughly 40%.

4.4 Statistical properties

Protocol effect size. The protocol effect (0/30 under structured format vs. 24/25 under free-form) is statistically significant by Fisher's exact test ($p < 10^{-10}$). The effect size is large: the structured format reduces detection from approximately 96% to 0% for the D3 stimulus. This is not a marginal effect amenable to power analysis; it is a qualitative failure of the structured protocol for conceptual errors.

Scope of the protocol effect. The full-battery comparison (Table 8) shows format sensitivity in 4 of 21 stimuli (19%), all involving conceptual errors. A chi-squared test comparing the overall detection rate (structured: 82% vs. free-form: 100%) is significant ($p < 0.001$). The effect is concentrated in conceptual errors: 0 of 7 surface stimuli are format-sensitive, 2 of 7 medium stimuli, and 2 of 7 deep stimuli.

What a larger battery would provide. A battery with $K = 50$ deep stimuli evaluated under both formats would permit estimation of the fraction of errors that are format-sensitive, providing a distribution of protocol-induced ρ across error types rather than a single example. Constructing such a battery requires domain experts to design conceptually distinct errors spanning the full range of financial economics.

4.5 Limitations

Several limitations apply to the experimental evidence.

Mapping between experiments and theory. The theory models a sequential pipeline with binary defects at each stage and quality gates between stages. The experiments test detection of planted errors in standalone mathematical derivations. These are different objects. In a real pipeline, the evaluator assesses stage output in the context of the full pipeline history, potentially with access to all intermediate outputs. The experiments test whether an LLM can spot an error in an isolated derivation at three context levels.

Standalone error detection serves as a valid proxy for within-pipeline gate detection for two reasons. First, the key theoretical parameter ($\rho_{\min}$, the irreducible correlation floor) is a property of the model's training data, not of the evaluation context. An error that goes undetected across all context levels in standalone evaluation will also go undetected in a pipeline, because the blind spot is embedded in the model's training, not in the information environment. Second, Assumption 2 restricts detection to the gate immediately following the stage where the defect originates, so the relevant comparison is between a gate evaluating one stage's output (the pipeline setting) and an evaluator examining one derivation (the experiment setting).

The gap between standalone and pipeline evaluation is real, however. Pipeline-structured experiments would present the evaluator with a sequence of stage outputs and ask it to evaluate each stage in context, mimicking the sequential gate structure. Such experiments would test whether detection rates change when the evaluator has access to the full pipeline history, and whether the evaluator's attention is diluted across multiple stages. Pipeline-structured experiments represent a natural extension of the proof-of-concept design.

Stimulus domain. All stimuli involve mathematical derivations with planted errors. The theory applies more broadly to financial analysis evaluation, including subjective assessments (idea quality, novelty, economic importance). Whether the correlation mechanism operates similarly for less well-defined evaluation tasks is unknown.

Single-provider models. Both models are from the same provider and likely share training data. The cross-model comparison tests the effect of model size, not the effect of training-data independence. Proposition 2's prediction about evaluator independence remains untested for genuinely independent models from different providers.

Protocol effect concentrated in conceptual errors. The format sensitivity appears in 4 of 21 stimuli, all involving conceptual rather than computational errors. Whether the effect extends to other evaluation domains (subjective quality assessment, novelty checking) is unknown.

Detection quality classification. The keyword overlap classifier for correct versus spurious

detection is imperfect. Three stimuli (M7, D1, D4) show problematic classification. Manual validation of the classifier would strengthen confidence in the correct detection rates.

Sample size. With 7 stimuli per error type and 5 replications per condition, the experiments have adequate power to detect large effects (∼25 percentage points) but cannot detect moderate effects (∼10 percentage points). The null result on context effects for the 120b model is consistent with either no effect or a small effect that the experiment cannot detect.

5 Application: Credit Underwriting

The framework applies without modification to credit underwriting. This section maps the abstract objects to credit-specific quantities and computes illustrative dollar magnitudes using assumed parameters. The dollar figures demonstrate how the framework translates into concrete benchmarks, but they are not empirical estimates. A proper calibration would require production data from a financial institution on defect rates, detection probabilities, and stage costs.

5.1 Pipeline structure

An AI credit underwriting pipeline consists of five stages:

Table 9: Credit underwriting pipeline stages and defect types

Stage	Activity	What can go wrong (defect)
1	Data ingestion	Stale prices, missing variables, survivorship bias in historical defaults
2	Feature engineering	Omitting risk factors (e.g., geographic concentration risk)
3	Model estimation	Wrong functional form, overfitting, failure to capture nonlinear interactions
4	Stress testing	Missing tail events, underweighting correlated defaults, optimistic recovery assumptions
5	Documentation	Conclusions not supported by analysis, buried caveats, inconsistent risk metrics

Each stage builds on the previous one. A defect at stage 2 (omitting a relevant risk factor) renders stage 3 (model estimation) unreliable regardless of how well the model is estimated, because the model cannot capture a risk factor it does not have as input. This

sequential dependence is the sense in which a defect "renders downstream work worthless" in the binary model.

5.2 Mapping to model objects

Binary defects in credit. A defect is a risk assessment that materially misestimates default probability. The binary classification is defensible: a credit model either captures a risk factor or it does not. A model that omits geographic concentration risk will systematically misprice loans in concentrated portfolios, regardless of accuracy on other dimensions. The defect is the omission, not the magnitude of the resulting mispricing (which varies continuously). The model captures the question "does the pipeline catch the omission?" not "how large is the resulting error?"

Defect probabilities. Each stage introduces a new defect with some probability:

- Data ingestion (q_1): 0.05 for established products with mature data infrastructure; 0.15–0.25 for novel asset classes or new geographies.
- Feature engineering (q_2): 0.05 for vanilla lending; 0.15–0.20 for new products or borrower segments.
- Model estimation (q_3): 0.10–0.20, depending on model complexity and available data.
- Stress testing (q_4): 0.10–0.15. AI systems trained on historical data tend to underestimate tail events.
- Documentation (q_5): 0.05–0.10.

A reasonable average is $q \approx 0.10$ for established products and $q \approx 0.20$ for novel asset classes.

Irreducible correlation in credit. When the same AI system builds and audits the credit model, three sources generate $\rho_{\min} > 0$:

1. *Training data bias.* If the training data underrepresents defaults in a specific sector (e.g., commercial real estate in rising-rate environments), both the generator and evaluator will systematically underweight the risk. The evaluator cannot catch what it does not know is missing.
2. *Methodological blind spots.* If the model family (e.g., gradient boosting) systematically underweights tail dependencies, the same model evaluating its own output will not flag the limitation. This is the credit analog of the D3 blind spot in the experiments.

3. *Regime blindness.* AI systems trained on 2010–2024 data may underweight scenarios resembling 2007–2008. Both generator and evaluator share this temporal bias.

The empirical blind spot rate (1–2 blind spots out of 7 error types) implies $\rho_{\min} \approx 0.15$ for established products and $\rho_{\min} \approx 0.25$ for novel asset classes.

Value and cost parameters. For a \$10 million loan portfolio:

- *Per-stage value* (v): If the pipeline prevents one 5% mispricing on the portfolio, each of $N = 5$ stages contributes approximately \$100,000 of risk-assessment value. Normalizing: $v = 10$ (in units of \$10,000).
- *Per-stage cost* (k): Compute, data acquisition, and analyst time per stage. For an AI pipeline, $k \approx$ \$5,000–\$15,000 per stage, or $k = 1$ in normalized units.
- *Gate cost* (G): Fixed compute cost plus expected rework from false positives. With $g = 0.5$ (\$5,000 compute) and near-zero false-positive rates for capable models: $G \approx 0.55$.

5.3 Calibrated scenarios

The five stages in Table 9 represent a coarse decomposition. In practice, each stage contains multiple sub-steps (e.g., "model estimation" includes variable selection, coefficient estimation, and backtesting), so a finer decomposition could yield 15–30 stages. The optimal pipeline lengths computed below ($N^* = 9$–25) refer to this finer decomposition. The five-stage structure in Table 9 maps to the coarser organizational units within which multiple sub-stages and gates operate.

Scenario 1: Vanilla mortgages. Parameters: $q = 0.10$, $p = 0.85$, $\rho_{\min} = 0.15$, $v = 10$, $k = 1$, $G = 0.55$.

- Self-evaluation: $D_{\text{self}} = 0.85 \times 0.85 = 0.7225$, $m = 0.2775$, quality ceiling = 35 stages, optimal $N^*_{\text{self}} = 25$.
- Independent evaluation: $D_{\text{ind}} = 0.85$, $m = 0.15$, ceiling = 66 stages, $N^*_{\text{ind}} = 46$.
- Value of independence: $\Lambda = \Pi^*(46; 0.85) - \Pi^*(25; 0.7225) \approx 74$ normalized units $\approx$ **\$740,000** per \$10M portfolio.

Under these illustrative parameters, the value of evaluator independence for a $10 million vanilla mortgage portfolio is approximately $740,000. If an independent risk review (human expert or differently-trained model) costs less than this amount, the review is cost-justified. For a $100 million portfolio, the value of independence scales proportionally to $7.4 million. These figures illustrate the order of magnitude, not precise dollar estimates, because the underlying parameters (v, k, q, G) are assumed rather than estimated from production data.

Scenario 2: Crypto-collateralized lending. Parameters: $q = 0.20$, $p = 0.85$, $\rho_{\min} = 0.25$, $v = 15$ (higher information value because data is scarcer), $k = 1.5$ (higher cost for novel analysis), $G = 0.75$.

- Self-evaluation: $D_{\text{self}} = 0.85 \times 0.75 = 0.6375$, $m = 0.3625$, quality ceiling = 12 stages, optimal $N^*_{\text{self}} = 9$.
- Independent evaluation: $D_{\text{ind}} = 0.85$, $m = 0.15$, ceiling = 32 stages, $N^*_{\text{ind}} = 23$.
- Value of independence: $\Lambda \approx 71$ normalized units $\approx$ **$710,000** per $10M portfolio.

For novel asset classes, the value of evaluator independence is comparable in absolute terms to established products, but the self-evaluating pipeline is far more constrained: limited to 9 stages (versus 23 under independence). Independent audit is even more critical because the self-evaluating pipeline captures a smaller fraction of the potential value.

Scenario 3: When is self-evaluation adequate? Self-evaluation is adequate ($\Lambda < C_{\text{ind}}$) when:

1. Defect rates are low ($q \leq 0.05$): well-established products with mature data and proven methodologies.
2. Correlation is low ($\rho_{\min} \leq 0.05$): the AI system's training data covers the relevant risk factors comprehensively.
3. Pipelines are short ($N \leq 5$): the cascade has few stages to compound errors.
4. Independence is expensive (C_{ind} is large): no alternative model or expert is available at reasonable cost.

Using Corollary 2, $\rho^{\text{crit}}_{\min} \approx 0.08$ for credit underwriting with $C_{\text{ind}} = 50$ normalized units ($\approx$ $500,000). Given the experimental estimate of $\hat{\rho}_{\min} \geq 0.14$, independent audit is cost-justified for most credit underwriting pipelines.

5.4 Regulatory implications

Private firms under-invest in evaluation quality when defects impose external costs $E > 0$ on investors, counterparties, and the financial system. In credit underwriting:

- *Private cost of undetected defect:* The lender absorbs losses from mispriced loans. V captures this cost.
- *External cost of undetected defect:* Investors in securitized products, counterparties with credit exposure, and the financial system (through systemic risk channels) bear additional losses. E captures this cost.

The social planner sets $N^*_{\text{social}} > N^*_{\text{private}}$ and requires independent audit at a lower $\rho^{\text{crit}}_{\min}$ threshold than the private optimum. This provides a microfounded rationale for existing and proposed regulatory requirements:

- Model validation by independent parties (OCC SR 11-7 guidance).
- Stress testing by independent risk functions (Basel III/IV).
- Third-party review of AI/ML models in lending (proposed CFPB guidance).

The framework quantifies when these requirements are binding. For well-established products with low q and low $\rho_{\min}$, mandatory independent review may be unnecessarily costly. For novel products with high q and high $\rho_{\min}$, the social optimum requires even more stringent independent oversight than current regulations mandate.

5.5 Data requirements for empirical calibration

A proper calibration of Λ requires four types of data that are currently unavailable in the public domain:

1. *Stage-specific defect rates* (q_i): the frequency with which each pipeline stage introduces material errors. These could be estimated from internal quality audits at financial institutions that track error rates by pipeline stage.
2. *Detection probabilities* (p): the rate at which AI evaluators catch planted defects in production-realistic tasks. The blind spot protocol provides a template, but the test battery must be expanded to cover the full range of error types encountered in credit underwriting.

3. *Value and cost parameters* (v, k, G): the economic value of defect-free analysis, the cost per pipeline stage, and the cost per quality gate. These can be extracted from production cost accounting and from estimates of losses attributable to model errors.

4. *Correlation floor* ($\rho_{\min}$): the irreducible miss rate from shared training data. The blind spot protocol estimates this parameter, but the estimate must be based on a test battery tailored to credit underwriting rather than general financial economics.

Financial institutions with access to production error logs, cost data, and AI evaluation records are best positioned to calibrate the framework. The illustrative numbers in this section provide a template for such calibration, not a substitute for it.

6 Discussion

6.1 The null result on context effects

Context has no statistically significant effect on detection quality for the 120b model. The correct detection rate is 85% at $c = 0$, 80% at $c = 0.5$, and 87% at $c = 1$ ($p = 0.83$). The theory's prediction that $\rho(c)$ is increasing in c receives no support from the 120b data.

The null result on context (how much upstream reasoning the evaluator receives) contrasts sharply with the large protocol effect (how the evaluator responds). For capable models, the information channel is inoperative: the model detects errors equally well whether it sees no reasoning, partial reasoning, or full chain-of-thought. But the format channel is powerful: the same model's detection drops from near-perfect to zero when constrained to structured JSON output. The binding constraint is not what the evaluator knows but how the evaluator is permitted to reason. Proposition 3 characterizes optimal pipeline design under context-invariant detection; the experiments confirm this characterization holds within a fixed response format, while revealing that format choice is the decisive design variable.

The context channel ($p'(c) > 0$, Assumption 5) may operate for less capable models. The 20b model shows suggestive evidence: raw detection drops from 81% to 74% at $c = 1$ for medium errors, and correct detection is uniformly low (56–59%). These patterns are consistent with a model whose limited capability makes it more susceptible to anchoring on the generator's reasoning.

6.2 Why evaluation protocol matters

The cross-format experiment (Table 8) reveals that the D3 detection failure is protocol-induced, not training-data-induced. All five models catch the error in free-form mode. The

structured JSON format, which forces a binary `has_error` classification before the model completes its reasoning, suppresses detection of conceptual errors that require extended deliberation.

This finding has three implications for the theoretical framework.

First, the effective correlation ρ is endogenous to evaluation design, not exogenous to the model. The theory's context variable c should be interpreted broadly: it encompasses not only how much upstream reasoning the evaluator receives (the original interpretation) but also how the evaluator is permitted to structure its response. Protocol design is a dimension of c that the initial formalization did not anticipate but that the framework accommodates without modification.

Second, the protocol effect provides direct evidence for the mechanism underlying Proposition 7 (commitment value of fixed thresholds). The structured format is a rigid protocol that constrains the evaluator's reasoning. When this constraint interacts badly with the error type (conceptual errors requiring deliberation), the rigid format destroys detection. The commitment proposition predicts exactly this pattern: commitment to rigid rules helps when the evaluator's confidence is contaminated (suppressing over-confidence), but hurts when the rigid format itself prevents adequate reasoning.

Third, the distinction between protocol-induced and training-data-induced correlation has practical consequences. Protocol-induced correlation is fixable: the pipeline designer can switch from structured to free-form evaluation. Prompt design alone cannot fix training-data correlation. The experiments provide no evidence of irreducible training-data correlation ($\rho_{\min} > 0$) for the error types tested, but the test battery is small ($K = 7$ deep stimuli) and the absence of a detected floor does not prove $\rho_{\min} = 0$. Genuine training-data blind spots may exist for error types not represented in the battery.

Does optimal protocol design eliminate the cascade? If $\hat{\rho}_{\text{free-form}} \approx 0$ for capable models, the Correlation Cascade becomes quantitatively mild under the best available protocol. This observation does not make the framework irrelevant; it sharpens the prescriptions. The cascade characterizes the cost of suboptimal protocol choices, which are empirically common: structured evaluation is the default in production systems for parsing convenience. The value of independence Λ should be computed at the realized protocol, not the optimal one. If a firm uses structured evaluation (as most deployed systems do), $\rho \approx 0.19$ and the cascade binds. The framework quantifies the cost of protocol choices that firms make for engineering reasons without recognizing their effect on evaluation quality. Whether $\rho_{\min} = 0$ under the best possible protocol for the hardest possible errors remains an open question that a 21-stimulus battery cannot definitively answer.

6.3 Binary defects and continuous quality

The binary defect model treats each stage's output as either defect-free or fatally flawed. Real financial analysis involves continuous quality dimensions: a credit model can be slightly misspecified or severely misspecified, and the downstream impact varies accordingly.

Proposition 6 shows that the comparative statics survive under continuous quality degradation. The optimal pipeline length N^* remains decreasing in ρ, and the value of evaluator independence Λ remains positive whenever $\rho > 0$. The quantitative magnitudes differ: continuous degradation yields longer optimal pipelines because individual noise contributions are less catastrophic than binary defects. The binary model is a conservative approximation that produces sharper bounds at the cost of overstating the severity of individual defects.

6.4 Cross-stage defect correlation

Assumption 1 rules out correlation in defect introduction across stages. In practice, a data quality problem at stage 1 may make feature engineering errors at stage 2 more likely. When defects propagate ($\text{Corr}(\theta_i, \theta_j) > 0$ for $i < j$), the product structure $Q = \prod(1 - q_i m_i)$ breaks down, but the qualitative results strengthen. Cross-stage correlation means that a defect at stage 1 makes defects at subsequent stages more likely, so the cascade operates through both the evaluation channel (within-stage correlation ρ) and the production channel (cross-stage defect propagation). The optimal pipeline is even shorter than under Assumption 1.

6.5 Welfare and externalities

The designer optimizes private payoff: $V \cdot Q - \text{costs}$. When defects impose external costs $E > 0$ on downstream users (investors relying on flawed credit ratings, regulators relying on incomplete compliance reports), the social planner's problem replaces V with $V + E$. The private designer under-invests in evaluation quality:

$$N^*_{\text{social}} > N^*_{\text{private}} \quad \text{whenever } E > 0. \tag{11}$$

The gap $N^*_{\text{social}} - N^*_{\text{private}}$ is increasing in both E and $\rho_{\min}$. The social planner values quality more (higher effective $V' = V + E$), so each unit of quality lost from correlation is weighted by $V + E$ rather than V, and the planner's pipeline length is more sensitive to correlation. This welfare gap provides the economic rationale for mandatory independent audit in AI-driven financial analysis: when externalities are large, requiring independent evaluation corrects the private designer's under-investment in quality.

6.6 Non-stationarity of the correlation floor

The framework treats $\rho_{\min}$ as a fixed parameter, but in practice, the irreducible correlation floor changes over time. LLMs are retrained on expanded corpora, and retraining can shift, shrink, or create blind spots. A model retrained on data that includes explicit discussion of the CAPM diversification distinction (the D3 error) would likely detect the error, reducing $\rho_{\min}$. Conversely, a model retrained on data that reinforces a different misconception could introduce new blind spots, increasing $\rho_{\min}$ along that dimension while decreasing it along others.

Non-stationarity has three implications for the framework. First, the optimal pipeline length N^* is itself non-stationary. A pipeline designed for $\rho_{\min} = 0.15$ is suboptimal if retraining shifts $\rho_{\min}$ to 0.25, and over-conservative if $\rho_{\min}$ drops to 0.05. Pipeline designers must re-optimize after each model update. Second, the blind spot protocol must be re-run periodically, not just at initial deployment. The cost of the protocol (constructing and running the test battery) becomes part of the pipeline design problem: the designer balances the cost of frequent re-estimation against the cost of operating with a stale $\rho_{\min}$ estimate. Third, the value of evaluator independence (Λ) is a moving target. A procurement decision to invest in independent evaluation should account for the expected trajectory of $\rho_{\min}$, not just its current level. If foundation model providers are converging on similar training corpora (as appears to be the case empirically), $\rho_{\min}$ across providers may increase over time, reducing the value of cross-provider independence.

6.7 Commitment value of fixed thresholds

Proposition 7 formalizes the value of committing to fixed pass/fail thresholds. The result depends on the evaluator not knowing ρ: a naive evaluator that calibrates under $\rho = 0$ over-responds to its confidence signal, systematically lowering the acceptance bar for the defects where its judgment is most contaminated. The crossover correlation $\bar{\rho}$ above which fixed thresholds dominate is a policy-relevant threshold: below it, allow evaluator discretion; above it, commit to rigid rules.

The D3 experimental finding illustrates a subtlety. The structured JSON format is itself a commitment device: it constrains the evaluator's response. But the commitment proposition assumes the evaluator's quality signal s is unaffected by the threshold rule, while the format effect shows that the response protocol can degrade s itself (by preventing the model from reasoning through conceptual errors). Commitment to rigid evaluation criteria is valuable; commitment to rigid response formats that suppress reasoning is costly. The designer must distinguish between committing to *what* the evaluator decides (the threshold)

and committing to *how* the evaluator reasons (the format).

For financial applications, the commitment result supports the use of fixed, pre-specified quality criteria in AI-assisted credit analysis rather than allowing the model to adjust its own evaluation standards based on assessed confidence.

6.8 Compound detection with multiple evaluators

When K evaluators assess the same artifact at a single gate, the compound detection probability under conditional independence is:

$$D_{\text{compound}} = 1 - \prod_{k=1}^{K} m(c_k). \tag{12}$$

When the generator's bias has a common factor affecting all evaluators, Jensen's inequality implies $P(\text{all miss}) > \prod P(M_k)$: the independence benchmark overstates compound detection. This common-factor structure is the norm when evaluators share training data, and it limits the value of adding evaluators from the same provider. The experiments confirm the prediction: 20b from the same provider provides strictly worse evaluation than 120b, not complementary evaluation.

6.9 Limitations

Three limitations deserve emphasis.

First, the experiments test mathematical error detection only. Mathematical derivations have unambiguous ground truth, making detection easy to classify and measure. Financial analysis evaluation encompasses subjective dimensions (idea quality, novelty, economic importance) where detection is harder to define, false positives are more likely, and the correlation mechanism may operate differently.

Second, the calibrated dollar magnitudes in Section 5 are illustrative, not empirically validated. The parameters v, k, q, and G are set at plausible values, not estimated from production data. The framework's value is in providing a structure for calibration, not in the specific numbers.

Third, the framework treats the model as non-strategic. In practice, pipeline designers may have incentives to manipulate evaluation outcomes (choosing context levels that make defects harder to detect, for example). A strategic extension would connect to the adverse selection literature on self-certification.

7 Conclusion

Correlated self-evaluation in AI financial analysis pipelines costs $\Lambda = V_{\text{ind}} - V_{\text{self}}$, a value of independence that depends on six observable parameters: ρ, p, q, v, k, and G. The Correlation Cascade bounds the feasible pipeline length at $N^* < 1/(qm) - 1$, halving it at $\rho = 0.15$. Experiments with five models from four providers reveal that the effective correlation ρ is not a fixed property of training data but a function of evaluation protocol design: four out of 21 planted errors escape detection under a structured response format but are caught at 100% under free-form evaluation. The binding constraint on pipeline quality is not what the evaluator knows but how the evaluator reasons. Pipeline designers should optimize evaluation protocols alongside gate placement and context-sharing decisions, and test protocols using the blind spot methodology in Section 4.

References

Cont, R. and Bianchi, M. L. (2011). Model risk in the valuation of credit derivatives. *Quantitative Finance*, 11(2):163–177.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14(4):436–440.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Korinek, A. (2025). AI agents for economic research. NBER Working Paper No. 34202.

Mandroli, S. S., Shrivastava, A. K., and Ding, Y. (2006). A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes. *IIE Transactions*, 38(4):309–328.

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

Oh, S., Choi, J., and Lee, K. (2024). The generative AI paradox: What it can solve, it may not evaluate. arXiv preprint arXiv:2402.06204.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). Self-preference bias in LLM-as-a-judge. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Raz, T. (1986). A survey of models for allocating inspection effort in multistage production systems. *Journal of Quality Technology*, 18(4):239–247.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3):213–225.

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, and Organization*, 2(2):181–214.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica*, 47(3):641–654.

A Idea Screening as Real Options

Corollary 3 (Real options for idea screening)**.** *In early pipeline stages, each candidate analysis is a call option on a full implementation. Let m candidates be generated at cost c_s each, with qualities $\theta_1, \ldots, \theta_m$ drawn i.i.d. from F on $[0, 1]$. The pipeline selects $\theta^* = \max_i \theta_i$, with downstream defect probability $q(\theta) = 1 - \theta$. The value of screening m candidates:*

$$S(m) = \left(\mathbb{E}[\theta^{(m)}] - \mathbb{E}[\theta^{(1)}]\right) \cdot W_d - (m-1)c_s \tag{13}$$

where W_d denotes downstream value. For $F = \text{Uniform}[0, 1]$: $\mathbb{E}[\theta^{(m)}] = m/(m+1)$, so screening is valuable when $c_s < W_d/(2(m+1))$, i.e., cheap generation relative to downstream costs.

This recovers a Weitzman (1979)-style reservation-value result applied to the pipeline's idea generation stage. Because the screening result does not interact with the correlation mechanism that is the paper's central contribution, it is presented here rather than in the main text.

B Experimental Stimuli

Table 10 lists the 21 error stimuli with their error types and detection outcomes.

Table 10: Error stimuli and detection rates (120b model, pooled across context levels)

ID	Description	Type	Detection rate	Status
S1–S7	Sign, arithmetic, and notational errors	Surface	100%	Always caught
M1	False covariance claim	Medium	>90%	Always caught
M2	Wrong R^2 formula	Medium	>90%	Always caught
M3	Factoring expectation of a ratio	Medium	40%	Sometimes caught
M4	Incorrect distribution assumption	Medium	>90%	Always caught
M5	Boundary condition error	Medium	60%	Sometimes caught
M6	Cost of equity formula error	Medium	>90%	Always caught
M7	Classification uncertain	Medium	–	Problematic
D1	Classification uncertain	Deep	–	Problematic
D2	Tower property misapplication	Deep	>80%	Always caught
D3	CAPM diversification corollary	Deep	0%	**Blind spot**
D4	Classification uncertain	Deep	–	Problematic
D5	Log-utility "risk neutrality"	Deep	10%	Near-blind-spot
D6	Modigliani-Miller error	Deep	>80%	Always caught
D7	Intermediate detection	Deep	60%	Sometimes caught