When Good Screeners Go Bad

A Correlation-Driven Paradox in Two-Stage Selection Pipelines and Human-AI Collaboration

Prasanna Parasurama
Emory University

Panos Ipeirotis
New York University

June 19, 2025

Abstract

Two-stage selection pipelines—where a cheap screener S decides who moves on to an expensive, high-quality assessor H—have become a default design in hiring, credit, medical triage, and content moderation. Intuition says that a higher level of agreement between the screener and the assessor must be helpful, or at least not harmful. We show the opposite: once the screener becomes *too similar* to the assessor, the pipeline can flip from helpful to harmful.

To formalize this, we consider a two-stage selection pipeline where each candidate is characterized by a latent true quality Q, a screener score S, and an assessor score H. We model (Q,S,H) as a trivariate Gaussian with unit variances and correlations $Corr(S,Q) = \theta_S$, $Corr(H,Q) = \theta_H$, $Corr(S,H) = \theta$. For fixed acceptance thresholds τ_S and τ_H, we derive a closed-form expression of expected quality for selected cases $\mathbb{E}[Q \mid S > \tau_S, H > \tau_H]$ and prove a simple rule:

$$S \text{ is } \textit{helpful} \iff \theta < \theta_S/\theta_H.$$

The ratio $\theta^\star := \theta_s/\theta_h$ is the screener's *normalized predictive power for* Q. If the screener predicts H better than this benchmark, it merely duplicates H's judgment, gatekeeps on the same mistakes, and lowers the final expected quality. Random shortlisting, skipping, or reversing S's decision then typically improves the expected quality of the selected cases, even when S is positively predictive of true quality Q—a counter-intuitive "good screener gone bad" effect.

We consider the algorithmic and managerial implications for human-AI collaboration, discuss how proxy targets, feature overlap, and imprinting can silently push θ above $\theta^\star$, and provide design levers to keep θ below $\theta^\star$. We also connect this phenomenon to the *fidelity paradox* in knowledge distillation in machine learning: when student models match teacher models too closely, they lose generalization. The main takeaway is that the predictive performance of the screener in predicting Q (θ_S) is an insufficient diagnostic metric for the overall performance of the pipeline. The inter-stage correlation θ is a key diagnostic that should be monitored as carefully as accuracy.

keywords: *Two-stage selection pipelines, Human-AI Collaboration, Knowledge Distillation, Fidelity Paradox*

1 Introduction

A common engineering fix for high-cost decision environments is to insert a *cheap screener* that quickly removes most low-value cases before an *expensive expert assessor* gives the shortlisted cases a thorough look. Applicant tracking systems pre-filter résumés for human interview panels; triage tests decide who proceeds to an MRI; rule-based triggers flag credit applications for a full bureau pull; keyword filters queue suspected posts for human moderators. In each example, the upstream screener S and the downstream assessor H aim at predicting some true latent quality Q (e.g., future job performance, a true medical state, creditworthiness, or policy compliance).

The folk intuition behind two-stage selection pipelines is that a higher level of agreement between the screener and the assessor must be helpful, or at least not harmful. Agreement is usually measured by the accuracy of the screener at predicting the decisions of the expert assessor, and greater agreement with the expert is often seen as "validation" of the screener. That is especially true in cases where we do not have direct observation of the true quality outcomes for the counterfactual cases (i.e., cases rejected by the screener or the expert assessor). This paper shows that such intuition can be fatally flawed.

When the screener S and the expert H become *too similar*, the screener no longer contributes novel signal—it merely duplicates the expert's judgment, but does so with lower fidelity, suppressing information gain. The unwitting result is that high-quality cases uniquely recognized by H are screened out, while low-quality cases that fool both stages slip through. Counterintuitively, when S and H are too similar, the overall quality of the selected pool *decreases*, yielding worse performance than random shortlisting, even when the screener is positively predictive of the true quality Q. We call this the "good screener gone bad" effect.

We formalize this effect in the simplest possible setting, where each candidate is characterized by a latent true quality Q, a screener score S, and an assessor score H. In the first stage, candidates whose S score exceeds a threshold τ_S are shortlisted. In the second stage, shortlisted candidates whose H score exceeds a threshold τ_H are finally selected. We model (Q,S,H) as a trivariate Gaussian with unit variances and correlations $Corr(S,Q) = \theta_S$, $Corr(H,Q) = \theta_H$, $Corr(S,H) = \theta$. θ_S captures the predictive ability of the screener S in predicting Q. θ_H captures the predictive ability of the expert assessor H in predicting Q. θ captures the level of agreement between the screener S and the expert assessor H. With fixed acceptance thresholds in each stage, we derive a closed-form expression for the expected latent quality of selected cases and prove a simple rule:

> *A screener is helpful while it predicts the expert more* poorly *than it predicts the true quality. Formally, the screener is guaranteed to improve outcomes iff* $\theta < \theta^\star$ *with* $\theta^\star = \theta_S/\theta_H$.

The critical ratio $\theta^\star$ is the screener's *normalized* predictive power for Q. Once $\theta > \theta^\star$, S excels at imitating H relative to its own ability to predict Q, and the pipeline is typically better off with random shortlisting, skipping, or even reversing the screener's decision.

Our theoretical results have direct implications for human-AI collaboration, where an algorithm takes the role of the screener. We discuss some practical scenarios where θ becomes too high by design (proxy target, feature overlap, imprinting), and design levers to mitigate the detrimental effect. Our results also resonate with the *fidelity paradox* observed in knowledge distillation: student networks that match teacher outputs too closely often lose generalization performance compared with students allowed moderate disagreement (Nagarajan et al., 2023; Guo et al., 2024). Both settings warn that excessive agreement between consecutive decision modules can harm the end goal. The shared lesson is that the proxy decision-maker must complement, not out-mimic, the expert.

Roadmap: Section 2 introduces the formal model. Section 3 proves the critical condition θ^* that separates a helpful screener from a harmful one, and benchmarks a 2-stage selection pipeline against random shortlisting. Section 4 provides some practical scenarios where θ can exceed θ^*, and methods to mitigate the negative effect. Section 5 situates the findings within the knowledge-distillation literature. We conclude with limitations and directions for future work.

2 Model Setup

We consider a selection pipeline where candidates have a latent true quality Q (e.g., true job performance, actual disease state, policy compliance). This quality is estimated by a screener with score S and a high-quality assessor with score H. We model the vector $[Q,S,H]^T$ as a multivariate normal distribution with a mean vector of zeros and a positive-semi-definite covariance matrix Σ. All marginal variances are normalized to 1 without loss of generality (via linear rescaling).

$$\begin{pmatrix} Q \\ S \\ H \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \theta_S & \theta_H \\ \theta_S & 1 & \theta \\ \theta_H & \theta & 1 \end{pmatrix}\right)$$

Here, $Corr(S,Q) = \theta_S$, $Corr(H,Q) = \theta_H$, and $Corr(S,H) = \theta$. Figure 1 illustrates the correlation structure. Intuitively:

- θ_S captures the predictive ability of the screener S in predicting Q.
- θ_H captures the predictive ability of the expert assessor H in predicting Q.
- θ captures the level of agreement between the screener S and the expert assessor H.

The selection process operates in 2 stages with fixed cut-off scores $\tau_S \in \mathbb{R}$ and $\tau_H \in \mathbb{R}$:

1. Candidates are screened, and those with $S > \tau_S$ are shortlisted.
2. Shortlisted candidates are assessed by the expert, and only those with $H > \tau_H$ are finally selected.

Figure 1: Correlation structure of the model

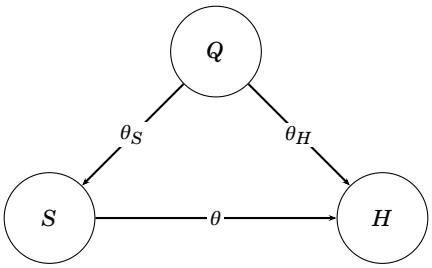
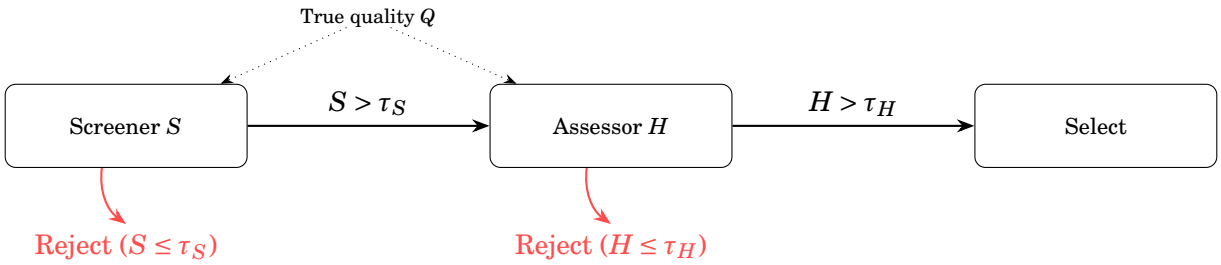


Figure 2: Two–stage pipeline



Notes: This figure illustrates the 2-stage selection pipeline. In the first stage, the screener shortlists candidates with scores $S > \tau_S$. In the second stage, from the shortlist, the assessor selects candidates with scores $H > \tau_H$. Dotted arrows indicate that both stages are noisy estimates of the latent true quality Q.

The final selected group is the set of candidates for whom $\{S > \tau_S, H > \tau_H\}$.

$$\text{Select}(S,H) = \begin{cases} 1, & \text{if } S > \tau_S \text{ and } H > \tau_H \\ 0, & \text{otherwise.} \end{cases}$$

Figure 2 shows this setting graphically.

Assumptions: For the remainder of our analysis, we assume $\theta_H > \theta_S > 0$ and $0 < \theta < 1$. The condition $\theta_H > \theta_S > 0$ implies that the assessor H is more strongly correlated with the true quality Q than the screener S is, and both have a positive correlation with Q. This is a common scenario where a more expensive, higher-quality assessor is better than a cheaper screener.

3 Theoretical Results

3.1 When the screener is negatively informative about true quality Q

We first analyze how the screener score S relates to true quality Q fixing the assessor score H.

Proposition 1. *For a fixed assessor score H, the screener score S is positively informative about true quality Q if $\theta < \theta_S/\theta_H$. Conversely, it is negatively informative if $\theta > \theta_S/\theta_H$.*

Proof. See Appendix A.1. □

The above proposition shows a counterintuitive result—that S can be negatively informative about Q given H even when $\theta_S > 0$. Intuition says that higher S scores should be associated with higher Q scores, since $\theta_S > 0$. While this is true in the overall candidate pool, once we condition on H, S becomes negatively informative about Q if $\theta > \theta_S/\theta_H$. This is to say, given two candidates with the same H score, the candidate with the *lower* S score has a higher expected true quality when $\theta > \theta_S/\theta_H$.

The intuition here is similar to the "explaining away" effect (Pearl, 1988). If the screener S is a better predictor of assessor H than true quality Q (i.e., $\theta > \theta_S/\theta_H$), then S's primary role becomes explaining H. A high S score sufficiently "explains away" a high H score, reducing the need to infer a high Q. A low S score, however, leaves the high Q as the more likely explanation for the high H score.

3.2 When screening is worse than random shortlisting

Proposition 1 shows that the screener is negatively informative about Q at any given $H = h$ if $\theta > \theta_S/\theta_H$. Now, we consider the case where the screener is negatively informative about Q when aggregating over all $H > \tau_H$. We use random shortlisting (or no screening[1]) as a useful benchmark to determine when the screener is detrimental to the expected quality of the selected pool in the overall selection pipeline.

Proposition 2. *The screener is detrimental to the average quality of selected candidates, yielding worse performance than random shortlisting, if $\theta > \theta_S/\theta_H$ and the thresholds (τ_S, τ_H) lie within a non-empty region $\mathcal{T} \subset \mathbb{R}^2$.*

Proof. See Appendix A.2. □

This result shows that screening can be worse than random shortlisting (or no screening) even when the screener is positively informative about true quality Q (i.e., $\theta_S > 0$). A necessary, but not a sufficient condition, for this to happen is that the screener is more informative about the assessor than it is about true quality Q (i.e., $\theta > \theta_S/\theta_H$). A second condition is that the thresholds (τ_S, τ_H) lie within a non-empty region $\mathcal{T} \subset \mathbb{R}^2$, whose boundary is defined by:

$$\theta_S \phi(\tau_S) \Phi\left(\frac{\theta\tau_S - \tau_H}{\sqrt{1-\theta^2}}\right) = \theta_H \phi(\tau_H) \left[\frac{L(\tau_S, \tau_H; \theta)}{1 - \Phi(\tau_H)} - \Phi\left(\frac{\theta\tau_H - \tau_S}{\sqrt{1-\theta^2}}\right)\right] \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal PDF and CDF, and $L(a, b; \rho) = P(X > a, Y > b)$ is the bivariate normal survival function. The detrimental region $\mathcal{T}$ is where the LHS is strictly less than the RHS. Here, the LHS captures the negative effect due to S and the RHS captures the positive effect due to H.

Although higher S scores are "locally" worse for overall Q for every $H = h$ (Proposition 1), when we aggregate over all $H > \tau_H$, the overall quality can still increase due to the aggregation effect

[1] In terms of quality, random shortlisting is equivalent to no screening.

analogous to Simpson's Paradox. For example, when the screener is excessively selective (high τ_S), the screener disproportionately shortlists candidates with very high H scores, and these H scores strongly drive Q upwards. Next, we formally characterize the detrimental region $\mathscr{T}$ in four asymptotic regimes.

3.2.1 Characterization of the Detrimental Threshold Region $\mathscr{T}$

We analyze the behavior of the boundary condition in four asymptotic regimes to characterize the shape of $\mathscr{T}$. We assume the necessary condition $\theta > \theta_S/\theta_H$ holds.

1. **Very High Screener Selectivity ($\tau_S \to +\infty$, fixed τ_H):** The LHS of (1) decays proportionally to $\phi(\tau_S) \sim e^{-\tau_S^2/2}$. The RHS decays faster, proportionally to $\phi\left(\frac{\tau_S - \theta\tau_H}{\sqrt{1-\theta^2}}\right) \sim e^{-\tau_S^2/(2(1-\theta^2))}$. Since $1/(1-\theta^2) > 1$, the RHS decays faster than the LHS. Therefore, for sufficiently large τ_S, the LHS will be greater than the RHS. *Implication:* The detrimental condition is not met. For any fixed τ_H, the detrimental region is bounded from above by a finite $\tau_S^{\text{upper}}(\tau_H)$. Extremely selective screeners are always beneficial.

2. **Very Low Assessor Selectivity ($\tau_H \to -\infty$, fixed τ_S):** As $\tau_H \to -\infty$, the LHS of (1) approaches a positive constant, $\theta_S\phi(\tau_S)$. The RHS approaches zero because it is multiplied by $\phi(\tau_H) \to 0$. The LHS is therefore greater than the RHS. *Implication:* The detrimental condition is not met. When the assessor is non-selective, adding a screener is always beneficial (compared to pure random selection). The boundary $\tau_S^{\text{upper}}(\tau_H)$ must approach $-\infty$ as $\tau_H \to -\infty$.

3. **Very High Assessor Selectivity ($\tau_H \to +\infty$, fixed τ_S):** As $\tau_H \to \infty$, the distribution of H for selected candidates becomes concentrated at τ_H. The beneficial "selection bias" effect (where selecting for high S also selects for higher H) vanishes. The inequality effectively reduces to analyzing the conditional expectation at $H = \tau_H$. As shown in the proof of Proposition 1, this is negative when $\theta > \theta_S/\theta_H$. *Implication:* The detrimental condition is met for any finite τ_S. As $\tau_H \to \infty$, the boundary $\tau_S^{\text{upper}}(\tau_H)$ must approach $+\infty$.

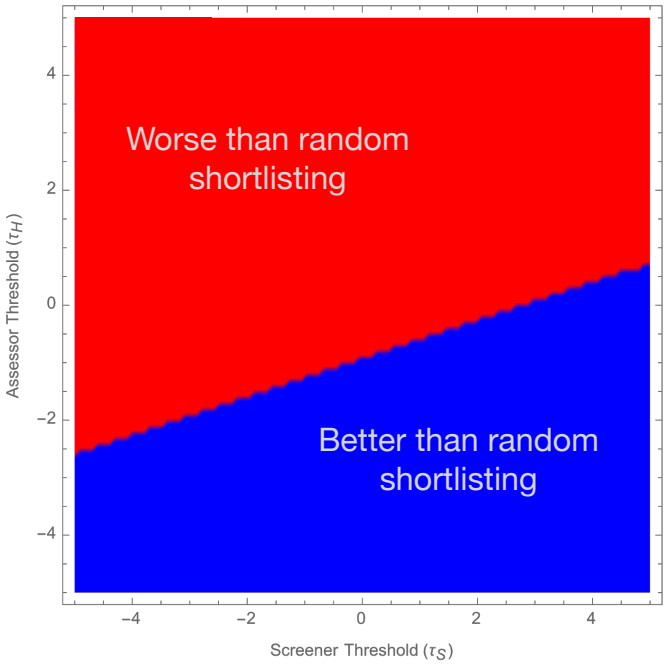
4. **Very Low Screener Selectivity ($\tau_S \to -\infty$, fixed τ_H):** As $\tau_S \to -\infty$, both sides of (1) approach zero, and the condition holds as an equality. *Implication:* The line $\tau_S = -\infty$ forms the lower boundary of the detrimental region.

In sum, $\mathscr{T}$ is the set of points (τ_S, τ_H) lying in the region $-\infty < \tau_S < \tau_S^{\text{upper}}(\tau_H)$. The upper boundary curve $\tau_S = \tau_S^{\text{upper}}(\tau_H)$ is a monotonically increasing function that passes from $(-\infty, -\infty)$ to $(+\infty, +\infty)$, partitioning the threshold plane. Therefore, a screener is detrimental when the assessor is sufficiently selective (high τ_H) but the screener itself is not excessively selective.

As shown in Proposition 2, the detrimental region $\mathscr{T}$ is non-empty if $\theta > \theta_S/\theta_H$ for any finite τ_H. A natural question is whether the region $\mathscr{T}$ is large enough for it to be a meaningful concern. To provide a sense of the relative size of the region $\mathscr{T}$, we numerically solve for the difference in expected quality between screening and random shortlisting for $(\theta = 0.6, \theta_S = 0.2, \theta_H = 0.8)$, and plot

it as a function of (τ_S, τ_H) in Figure 3. Note that for a reasonably selective H (e.g., $\tau_H = 0$), the screener is detrimental for a large interval of τ_S (i.e., $\tau_S < 3$).

Figure 3: Difference in expected quality of selected candidates between screening and random shortlisting



Notes: This figure shows the difference in expected quality of selected candidates between screening and random shortlisting, $\mathbb{E}[Q|S > \tau_S, H > \tau_H] - \mathbb{E}[Q|H > \tau_H]$. The parameters are: $(\theta = 0.6, \theta_S = 0.2, \theta_H = 0.8)$, satisfying condition $\theta > \theta_S/\theta_H$. The blue region shows where screening is beneficial (positive difference), while the red region shows where screening is detrimental (negative difference) compared to random shortlisting. The boundary curve $\tau_S^{\text{upper}}(\tau_H)$ separates these regions.

4 Managerial and Algorithmic Implications for Human-AI Collaboration

In many two-stage decision pipelines, the screener is often an algorithm. For example, in hiring, algorithmic screening systems filter résumés for human interview panels; in credit, rule-based triggers flag credit applications for a full bureau pull; in content moderation, keyword filters queue suspected posts for human moderators. Our theoretical results therefore have direct implications for human-AI decision pipelines.

As we showed in the previous section, once $\theta > \theta^\star$, the screener flips from being helpful to harmful. This critical ratio $\theta^\star = \theta_S/\theta_H$ is often not a design parameter. In practice, the inter-stage correlation $\theta = Corr(S, H)$ can creep above the critical threshold $\theta^\star$ under seemingly innocuous practices such as training the screener on assessor scores, feature convergence between the screener and as-

sessor, or the assessor imprinting on the screener. Below, we outline the algorithmic design choices that can push θ above the critical threshold, and strategies to mitigate the correlation creep.

4.1 Mechanisms of correlation creep

1. **Training S on H outcomes.** In many settings the screener S is trained on the observed "pass/fail" decisions of H for lack of ground-truth Q. The target then becomes a noisy proxy for H itself, driving θ upward. This is especially common in hiring, where an algorithmic screener is often trained on the "pass/fail" decisions of the hiring manager.

2. **Feature convergence between S and H.** Managers often align the screener's features with those deemed important by expert assessor H. Overlap in feature space tightens the correlation between scores.

3. **H imprinting on S (human-in-the-loop pipelines).** If the assessor dashboard surfaces the screener's score or highlights its top features, the human assessor may become anchored on S, again inflating θ.

4.2 Design levers to keep $\theta < \theta^\star$

1. **Feature diversification between S and H.** Train S on features that H cannot exploit due to time or cost. For example, in algorithmic hiring, S could be trained on public web footprint of job applicants rather than résumé keywords.

2. **Adversarial decorrelation objectives.** During training of S, add a regularization term $\lambda\big[Corr(S,H)\big]$ that penalizes correlation between S and H. This can also be done in an end-to-end manner using adversarial learning that forces S to capture variance in Q *orthogonal* to H (see, e.g., Zhang et al., 2018).

3. **Randomised smoothing.** Inject small random noise or dropout masks into S at inference time; the added noise could reduce deterministic alignment with H.

5 Connection to the Fidelity Paradox in Knowledge Distillation

The threshold phenomenon uncovered in two–stage pipelines echoes a seemingly unrelated observation from *knowledge distillation* (KD): students that mimic teachers *too faithfully* can suffer worse test accuracy than students that allow moderate disagreement. This "fidelity paradox" has been reported across vision and language benchmarks (Nagarajan et al., 2023; Guo et al., 2024; Stanton et al., 2021). Here we formalize the analogy between the two domains.

In KD, the training loss is typically a convex combination $\mathscr{L} = \alpha\,\mathrm{KL}(S \,\|\, T) + (1-\alpha)\,\ell_{\mathrm{hard}}(S,y)$, where $\alpha \in [0,1]$ controls fidelity. Empirically, accuracy improves only up to a problem-specific $\alpha^\star$,

Table 1: Mapping between two-stage pipelines and knowledge distillation

Two-stage pipeline	**Knowledge distillation**
Cheap screener S	Student network S
Expert assessor H	Teacher network T
True quantity Q	Ground-truth label y
Inter–stage correlation $\theta = Corr(S,H)$	Fidelity weight α in loss
Critical threshold $\theta^\star = \theta_s/\theta_h$	Optimal $\alpha^\star < 1$ balancing hard vs. soft targets

after which performance degrades—the student overfits the teacher's idiosyncrasies and loses complementary signal from the hard labels (Nagarajan et al., 2023). Our pipeline threshold $\theta^\star$ plays the same role: beyond it, the screener's imitation of the assessor harms final quality.

Both phenomena arise because duplicate information crowds out novel variance. In selection pipelines, when $\theta > \theta^\star$, S rarely contradicts H except on noise, eliminating cases where H alone would add value. In KD, when $\alpha > \alpha^\star$, the student's gradients are dominated by teacher logits, suppressing learning from ground truth and reducing overall performance.

6 Conclusion

The promise of two–stage decision pipelines is to combine the speed and scale of inexpensive screeners with the accuracy of expert assessors. Our analysis reveals a hidden pitfall: *excessive* agreement between stages can invert the screener's value. In the canonical trivariate-Gaussian model, the pipeline is guaranteed to improve expected true quality only while the inter-stage correlation $\theta = Corr(S,H)$ remains below the critical ratio $\theta^\star = \theta_s/\theta_h$, the screener's *normalized predictive power* for the target quantity Q. Once $\theta > \theta^\star$, reversing or bypassing the screener typically increases performance, even when the screener is positively predictive of true quality—a counter-intuitive "good screener gone bad" effect.

Beyond theory, we discussed real-world mechanisms—proxying on H, feature convergence between S and H, and H imprinting on S—by which correlation creep silently pushes θ above θ^*. We proposed several design levers—feature diversification, adversarial decorrelation, random smoothing—to keep $\theta < \theta^\star$ in 2-stage algorithmic decision-making pipelines. Furthermore, drawing an analogy to the fidelity paradox in knowledge distillation, we showed that both domains arrive at the same design principle: optimal collaboration requires partial disagreement.

While the canonical Gaussian model yields clean theoretical results, it also makes many distributional assumptions that invite further study. For example, heavy-tailed score distributions—common in credit or spam detection—may exhibit different critical $\theta^\star$ behavior. Elliptical copulas and rank-based dependence measures (e.g., Kendall's τ) are natural extensions. Another extension would be to consider selection pipelines with more than two stages. Does the information loss due to high correlation in the initial stages get mitigated in later stages?

Nevertheless, the canonical model offers a valuable lesson—that the predictive performance of the screener in predicting Q (θ_S) is an insufficient diagnostic metric for the overall performance of the pipeline. The inter-stage correlation θ is a key diagnostic that should be monitored as carefully as accuracy.

A Proofs

A.1 Proof: Conditional informativeness of the screener, S, given H

Proposition. *For a fixed assessor score H, the screener score S is positively informative about true quality Q if $\theta < \theta_S/\theta_H$. Conversely, it is negatively informative if $\theta > \theta_S/\theta_H$.*

Proof. Positive informativeness is defined by the condition $\frac{\partial}{\partial s}\mathbb{E}[Q|S=s,H=h] > 0$. We derive the conditional expectation $\mathbb{E}[Q|S=s,H=h]$ using the standard formula for multivariate normal distributions. Let $\mathbf{X}_1 = Q$ and $\mathbf{X}_2 = [S,H]^T$. Then,

$$\mathbb{E}[Q|S=s,H=h] = \mathbb{E}[Q] + \mathrm{Cov}(Q,[S,H])\mathrm{Cov}([S,H],[S,H])^{-1}\begin{pmatrix} s-\mathbb{E}[S] \\ h-\mathbb{E}[H] \end{pmatrix}$$

With zero means, this simplifies to:

$$\mathbb{E}[Q|S=s,H=h] = [\theta_S,\theta_H]\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}^{-1}\begin{pmatrix} s \\ h \end{pmatrix}$$

The inverse of the 2×2 covariance matrix is:

$$\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}^{-1} = \frac{1}{1-\theta^2}\begin{pmatrix} 1 & -\theta \\ -\theta & 1 \end{pmatrix}$$

Substituting this back, we get:

$$\begin{aligned}
\mathbb{E}[Q|S=s,H=h] &= \frac{1}{1-\theta^2}[\theta_S,\theta_H]\begin{pmatrix} 1 & -\theta \\ -\theta & 1 \end{pmatrix}\begin{pmatrix} s \\ h \end{pmatrix} \\
&= \frac{1}{1-\theta^2}[\theta_S-\theta\theta_H,\quad \theta_H-\theta\theta_S]\begin{pmatrix} s \\ h \end{pmatrix} \\
&= \underbrace{\left(\frac{\theta_S-\theta\theta_H}{1-\theta^2}\right)}_{\beta_S}s + \underbrace{\left(\frac{\theta_H-\theta\theta_S}{1-\theta^2}\right)}_{\beta_H}h
\end{aligned}$$

The informativeness of S is given by the sign of its coefficient, β_S. Taking the partial derivative with respect to s:

$$\frac{\partial}{\partial s}\mathbb{E}[Q|S=s,H=h] = \frac{\theta_S-\theta\theta_H}{1-\theta^2}$$

Since $0<\theta<1$, the denominator $(1-\theta^2)$ is positive. Thus, the sign is determined by the numerator.

The screener is positively informative if:

$$\theta_S - \theta\theta_H > 0 \iff \theta_S > \theta\theta_H \iff \theta < \frac{\theta_S}{\theta_H}$$

Conversely, the screener is negatively informative if $\theta > \theta_S/\theta_H$. □

A.2 Proof: When the screener is worse than random shortlisting

Proposition. *The screener is detrimental to the average quality of selected candidates, yielding worse performance than random shortlisting, if $\theta > \theta_S/\theta_H$ and the thresholds (τ_S, τ_H) lie within a non-empty region $\mathcal{T} \subset \mathbb{R}^2$. For $(\tau_S, \tau_H) \in \mathcal{T}$, we have:*

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] < \mathbb{E}[Q|H > \tau_H]$$

Proof. The proof is structured in four parts.

1. We first establish an equivalent condition for the screener to be detrimental.
2. We then use this condition to prove that $\theta > \theta_S/\theta_H$ is a necessary condition for a detrimental region to exist.
3. We derive the analytical equation that defines the boundary of the detrimental region $\mathcal{T}$.
4. Finally, we prove that if the condition $\theta > \theta_S/\theta_H$ holds, the region $\mathcal{T}$ is guaranteed to be non-empty for any finite thresholds.

Part 1: An Equivalent Condition for a Detrimental Screener

The benchmark quality, without a screener, is $\mathbb{E}[Q|H > \tau_H]$. We can express this using the Law of Total Expectation by partitioning the sample space $\{H > \tau_H\}$ based on the screener outcome, $S > \tau_S$ or $S \leq \tau_S$.

$$\begin{aligned}\mathbb{E}[Q|H > \tau_H] =& P(S > \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S > \tau_S, H > \tau_H] \\ &+ P(S \leq \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]\end{aligned}$$

The screener is detrimental if $\mathbb{E}[Q|S > \tau_S, H > \tau_H] < \mathbb{E}[Q|H > \tau_H]$. Substituting the expression for $\mathbb{E}[Q|H > \tau_H]$ into this inequality:

$$\begin{aligned}\mathbb{E}[Q|S > \tau_S, H > \tau_H] <& P(S > \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S > \tau_S, H > \tau_H] \\ &+ P(S \leq \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]\end{aligned}$$

Since $P(S > \tau_S|\cdot) + P(S \leq \tau_S|\cdot) = 1$, we can rewrite the LHS as $1 \cdot \mathbb{E}[Q|S > \tau_S, H > \tau_H]$.

$$(1 - P(S > \tau_S|H > \tau_H)) \cdot \mathbb{E}[Q|S > \tau_S, H > \tau_H] < P(S \leq \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]$$

Using $1 - P(S > \tau_S|\cdot) = P(S \leq \tau_S|\cdot)$, we get:

$$P(S \leq \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S > \tau_S, H > \tau_H] < P(S \leq \tau_S|H > \tau_H) \cdot \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]$$

For any finite τ_S, the probability $P(S \leq \tau_S|H > \tau_H)$ is strictly positive. We can therefore divide by it without changing the inequality's direction. This yields the equivalent condition: the screener is detrimental if and only if

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] < \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]$$

This means the screener is detrimental if and only if the average quality of candidates who pass both tests is strictly lower than the average quality of candidates who pass the assessor but fail the screener.

Part 2: Necessity of the Condition $\theta > \theta_S/\theta_H$

Assume the opposite condition holds: $\theta \leq \theta_S/\theta_H$. From Proposition 1, this implies that the coefficient $\beta_S = \frac{\theta_S - \theta\theta_H}{1-\theta^2}$ is non-negative ($\beta_S \geq 0$). This means that for any fixed $H = h$, the expected quality $\mathbb{E}[Q|S = s, H = h] = \beta_S s + \beta_H h$ is a non-decreasing function of s.

Now consider the difference in expected quality between the "pass" and "fail" groups for a fixed $h > \tau_H$:

$$\mathbb{E}[Q|S > \tau_S, H = h] - \mathbb{E}[Q|S \leq \tau_S, H = h] = \beta_S (\mathbb{E}[S|S > \tau_S, H = h] - \mathbb{E}[S|S \leq \tau_S, H = h])$$

By definition, $\mathbb{E}[S|S > \tau_S, H = h] > \tau_S$ and $\mathbb{E}[S|S \leq \tau_S, H = h] \leq \tau_S$. The term in the parentheses is therefore strictly positive. Since we assumed $\beta_S \geq 0$, the entire expression is non-negative:

$$\mathbb{E}[Q|S > \tau_S, H = h] \geq \mathbb{E}[Q|S \leq \tau_S, H = h] \quad \text{for all } h > \tau_H$$

Now, we average this inequality over the distribution of H in the respective domains:

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] = \int_{\tau_H}^{\infty} \mathbb{E}[Q|S > \tau_S, H = h]\, p(h|S > \tau_S, H > \tau_H)\, dh$$
$$\mathbb{E}[Q|S \leq \tau_S, H > \tau_H] = \int_{\tau_H}^{\infty} \mathbb{E}[Q|S \leq \tau_S, H = h]\, p(h|S \leq \tau_S, H > \tau_H)\, dh$$

Because the inequality holds for every h, and the distributions are over the same domain, it follows that the integrated average for the "pass" group must be greater than or equal to that of the "fail" group. Thus,

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] \geq \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]$$

Based on the equivalent condition from Part 1, this means the screener cannot be detrimental. Therefore, it is a necessary condition that $\theta > \theta_S/\theta_H$ for a detrimental region to exist.

Part 3: Derivation of the Boundary Equation

The boundary of the detrimental region $\mathcal{T}$ is defined by the equality case of the main condition:

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] = \mathbb{E}[Q|H > \tau_H]$$

We derive the formulas for these truncated moments using Owen, 1956 and Tallis, 1961.

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] = \frac{\theta_S \phi(\tau_S)\Phi\left(\frac{\theta\tau_S - \tau_H}{\sqrt{1-\theta^2}}\right) + \theta_H \phi(\tau_H)\Phi\left(\frac{\theta\tau_H - \tau_S}{\sqrt{1-\theta^2}}\right)}{L(\tau_S, \tau_H; \theta)}$$

$$\mathbb{E}[Q|H > \tau_H] = \mathbb{E}[\theta_H H|H > \tau_H] = \theta_H \mathbb{E}[H|H > \tau_H] = \theta_H \frac{\phi(\tau_H)}{1 - \Phi(\tau_H)}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal PDF and CDF, and $L(\tau_S, \tau_H; \theta) = P(S > \tau_S, H > \tau_H)$ is the bivariate normal survival function. Setting these equal:

$$\frac{\theta_S \phi(\tau_S)\Phi\left(\frac{\theta\tau_S - \tau_H}{\sqrt{1-\theta^2}}\right) + \theta_H \phi(\tau_H)\Phi\left(\frac{\theta\tau_H - \tau_S}{\sqrt{1-\theta^2}}\right)}{L(\tau_S, \tau_H; \theta)} = \frac{\theta_H \phi(\tau_H)}{1 - \Phi(\tau_H)}$$

Multiply by the denominator $L(\tau_S, \tau_H; \theta)$:

$$\theta_S \phi(\tau_S)\Phi\left(\frac{\theta\tau_S - \tau_H}{\sqrt{1-\theta^2}}\right) + \theta_H \phi(\tau_H)\Phi\left(\frac{\theta\tau_H - \tau_S}{\sqrt{1-\theta^2}}\right) = L(\tau_S, \tau_H; \theta)\frac{\theta_H \phi(\tau_H)}{1 - \Phi(\tau_H)}$$

Rearrange to isolate the θ_S term on the left-hand side:

$$\theta_S \phi(\tau_S)\Phi\left(\frac{\theta\tau_S - \tau_H}{\sqrt{1-\theta^2}}\right) = L(\tau_S, \tau_H; \theta)\frac{\theta_H \phi(\tau_H)}{1 - \Phi(\tau_H)} - \theta_H \phi(\tau_H)\Phi\left(\frac{\theta\tau_H - \tau_S}{\sqrt{1-\theta^2}}\right)$$

Factor out the common term $\theta_H \phi(\tau_H)$ on the right-hand side:

$$\theta_S \phi(\tau_S)\Phi\left(\frac{\theta\tau_S - \tau_H}{\sqrt{1-\theta^2}}\right) = \theta_H \phi(\tau_H)\left[\frac{L(\tau_S, \tau_H; \theta)}{1 - \Phi(\tau_H)} - \Phi\left(\frac{\theta\tau_H - \tau_S}{\sqrt{1-\theta^2}}\right)\right] \quad (*)$$

This equation defines the boundary of the region $\mathcal{T}$. The screener is detrimental when the LHS is strictly less than the RHS.

Part 4: Existence of the Detrimental Region $\mathcal{T}$

We now prove that if the necessary condition $\theta > \theta_S/\theta_H$ holds, then the region $\mathcal{T}$ is guaranteed to be non-empty. We use the equivalent condition from Part 1 and analyze its behavior as $\tau_S \to -\infty$. The screener is detrimental if:

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] < \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]$$

We analyze the limits of both sides as $\tau_S \to -\infty$ for a fixed, finite τ_H.

1. **Limit of the "Pass" Group:** As $\tau_S \to -\infty$, the condition $S > \tau_S$ becomes trivial. The population

$\{S > \tau_S, H > \tau_H\}$ converges to the population $\{H > \tau_H\}$.

$$\lim_{\tau_S \to -\infty} \mathbb{E}[Q|S > \tau_S, H > \tau_H] = \mathbb{E}[Q|H > \tau_H] = \theta_H \frac{\phi(\tau_H)}{1 - \Phi(\tau_H)}$$

This limit is a finite constant for any finite τ_H.

2. **Limit of the "Fail" Group:** The condition $S \leq \tau_S$ describes an infinitesimally thin slice of the distribution's left tail as $\tau_S \to -\infty$. The expected quality in this slice can be approximated by the conditional expectation at the boundary, $S = \tau_S$.

$$\lim_{\tau_S \to -\infty} \mathbb{E}[Q|S \leq \tau_S, H > \tau_H] = \lim_{s \to -\infty} \mathbb{E}[Q|S = s, H > \tau_H]$$

Using the Law of Total Expectation and the result from Proposition 1, $\mathbb{E}[Q|S = s, H = h] = As + Bh$:

$$\mathbb{E}[Q|S = s, H > \tau_H] = \mathbb{E}[As + Bh|H > \tau_H] = As + B \cdot \mathbb{E}[H|S = s, H > \tau_H]$$

The conditional distribution of H given $S = s$ is normal: $H|S = s \sim \mathcal{N}(\theta s, 1 - \theta^2)$. The term $\mathbb{E}[H|S = s, H > \tau_H]$ is the mean of this normal distribution truncated from below at τ_H. Its value is:

$$\mathbb{E}[H|S = s, H > \tau_H] = \theta s + \sqrt{1 - \theta^2} \cdot \lambda\left(\frac{\tau_H - \theta s}{\sqrt{1 - \theta^2}}\right)$$

where $\lambda(z) = \phi(z)/(1 - \Phi(z))$ is the inverse Mills ratio. A key property of $\lambda(z)$ is that $\lim_{z \to \infty} \lambda(z) = z$. As $s \to -\infty$, the argument $z = (\tau_H - \theta s)/\sqrt{1 - \theta^2} \to +\infty$ (since $\theta > 0$). Therefore,

$$\lim_{s \to -\infty} \mathbb{E}[H|S = s, H > \tau_H] = \lim_{s \to -\infty} \left[\theta s + \sqrt{1 - \theta^2}\left(\frac{\tau_H - \theta s}{\sqrt{1 - \theta^2}}\right)\right] = \lim_{s \to -\infty} [\theta s + \tau_H - \theta s] = \tau_H$$

Substituting this back into the limit for the "fail" group's quality:

$$\lim_{s \to -\infty} \mathbb{E}[Q|S = s, H > \tau_H] = \lim_{s \to -\infty} (\beta_S s + \beta_H \tau_H)$$

Under our pre-condition for detrimental screening, $\theta > \theta_S/\theta_H$, we have $\beta_S < 0$. Therefore:

$$\lim_{s \to -\infty} (\beta_S s + \beta_H \tau_H) = +\infty$$

We have shown that as $\tau_S \to -\infty$, the expected quality of the "pass" group converges to a finite constant, while the expected quality of the "fail" group converges to positive infinity. Therefore, there must exist some threshold τ_S^* such that for all $\tau_S < \tau_S^*$, the inequality

$$\mathbb{E}[Q|S > \tau_S, H > \tau_H] < \mathbb{E}[Q|S \leq \tau_S, H > \tau_H]$$

holds. This proves that the detrimental region $\mathcal{T}$ is non-empty for any fixed τ_H when the condition $\theta > \theta_S/\theta_H$ is met. □

References

Guo, Chenqi et al. (2024). "Why Does Knowledge Distillation Work? Rethink its Attention and Fidelity Mechanism". In: *arXiv preprint arXiv:2405.00739*.

Nagarajan, Vaishnavh et al. (2023). "On Student–Teacher Deviations in Distillation: Does It Pay to Disobey?" In: *arXiv preprint arXiv:2301.12923*.

Owen, Donald B. (1956). "Tables for Computing Bivariate Normal Probabilities". In: *The Annals of Mathematical Statistics* 27.4, pp. 1075–1090.

Pearl, Judea (1988). "Embracing causality in default reasoning". In: *Artificial Intelligence* 35.2, pp. 259–271.

Stanton, Samuel et al. (2021). "Does Knowledge Distillation Really Work?" In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv: 2106.05945 [cs.LG].

Tallis, G. M. (1961). "The moment generating function of the truncated multi-normal distribution". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 23.1, pp. 223–229.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (2018). "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340.