Online Resource 3 — Online Appendix to "Disagreement Is Not Debiasing: Human–AI Views in Portfolio Decisions" (*Annals of Operations Research*)

Zhichao Luo[1*]

[1*]Tan Kah Kee College, Xiamen University, Zhangzhou, 363105, Fujian, China.

Corresponding author(s). E-mail(s): zhichao.luo@gmail.com;

Abstract

This online appendix contains material referred to in the main text: the proofs of the propositions and the derivation of the exact block-covariance update (OA.1), the sensitivity analyses and the audit of the estimation windows (OA.2), additional tables and a figure for the main text (OA.3), the mean–variance formulation of the decision problem (OA.4), diagnostics and alternative specifications for the preliminary model (OA.5), the execution layer (OA.6), the decision-layer supplements (OA.7), and additional sensitivities (OA.8). Equation, proposition, section and table numbers without the prefix OA refer to the main text.

Keywords: Online appendix

OA.1 Proofs

Throughout, $\Omega_Q, \Omega_M > 0$, $|C| < \sqrt{\Omega_Q \Omega_M}$ so that Ξ is positive definite, and expectations are over the noise (ε, η) and, where stated, over the prior. The general pick-matrix case is treated at the end.

OA.1.1 Proof of Proposition 1

Write the model (1) as $Q = \theta + b_Q + \varepsilon$, $M = \theta + b_M + \eta$ with $\theta \sim N(\pi, \tau\Sigma)$, where the location π is unknown to the analyst. For any constant vector c, define $\theta' = \theta - c$,

$b'_Q = b_Q + c$, $b'_M = b_M + c$, $\pi' = \pi - c$. Then $Q = \theta' + b'_Q + \varepsilon$ and $M = \theta' + b'_M + \eta$ with the same (ε, η) and $\theta' \sim N(\pi', \tau\Sigma)$, so the joint law of (Q, M) is unchanged: the two parameterizations are observationally equivalent, and no function of (Q, M) can distinguish (b_Q, b_M) from $(b_Q + c, b_M + c)$. The difference $b_Q - b_M$ is invariant to the transformation and is identified by $\mathbb{E}[Q - M]$ from (2).

For the identification conditions: (i) if θ_i is observed for a subset of forecasts, $\mathbb{E}[Q_i - \theta_i] = b_{Q,i}$ and $\mathbb{E}[M_i - \theta_i] = b_{M,i}$ on that subset, which pins c; (ii) if b_M is known (in particular $b_M = 0$), then $b_Q = (b_Q - b_M) + b_M$ is identified from (2); (iii) if π is known, the transformation changes π to $\pi - c$ and is no longer admissible, so $\mathbb{E}[Q] - \pi = b_Q$ identifies b_Q (with the prior serving as the unbiased reference). □

OA.1.2 Proof of Proposition 2

(i) Ranking rules. The ordering of the components of $\hat{\theta} + c\mathbf{1}$ equals that of $\hat{\theta}$, so top-K selection and quantile sorts are unchanged. For the mean–variance problem $\max_h h^\top\hat{\theta} - \frac{\delta}{2}h^\top\Sigma h$ subject to $\mathbf{1}^\top h = 1$ and $0 \le h \le h_{\max}$, the objective under the shifted forecast is $h^\top\hat{\theta} + c\,\mathbf{1}^\top h - \frac{\delta}{2}h^\top\Sigma h = h^\top\hat{\theta} + c - \frac{\delta}{2}h^\top\Sigma h$ on the feasible set, which differs from the original objective by the constant c; the argmax is unchanged. (The statement in the proposition about "a binding cap" describes the low-risk-aversion regime in which the solution places the cap on the highest-$\hat{\theta}$ names; invariance holds for every δ.)

(ii) Level rules. A hurdle rule $\{i : \hat{\theta}_i \ge r\}$ changes whenever c moves some component across r. For the target constraint $h^\top\hat{\theta} \ge$ target and the chance constraint $h^\top\hat{\theta} - z_\alpha\sqrt{h^\top\Sigma h} \ge$ target, the feasible set under $\hat{\theta} + c\mathbf{1}$ equals the feasible set under $\hat{\theta}$ with target replaced by target $- c$ (using $\mathbf{1}^\top h = 1$), so the solution changes with c unless the constraint is slack at both values. For a fully invested portfolio the realized outcome $h^\top\theta$ is unchanged by c while the promised outcome $h^\top(\hat{\theta} + c\mathbf{1}) = h^\top\hat{\theta} + c$ moves by c; the realized shortfall relative to the promise therefore changes by exactly c.

(iii) Heterogeneous bias. Take two assets with $\hat{\theta}_1 > \hat{\theta}_2$ and biases $b_1 > b_2$ with $b_1 - b_2 > \hat{\theta}_1 - \hat{\theta}_2$; then the corrected forecasts satisfy $\hat{\theta}_1 - b_1 < \hat{\theta}_2 - b_2$ and the ranking reverses. □

OA.1.3 Proof of Proposition 3

Drop the asset subscript and the prior ($\tau \to \infty$). With calibrated views $\tilde{Q} = \theta + \varepsilon$, $\tilde{M} = \theta + \eta$ and noise covariance $\Xi = \begin{pmatrix} \Omega_Q & C \\ C & \Omega_M \end{pmatrix}$, the generalized-least-squares combination is $\mu^* = (\mathbf{1}^\top\Xi^{-1}\mathbf{1})^{-1}\mathbf{1}^\top\Xi^{-1}y$. Since $\Xi^{-1} = \frac{1}{\Omega_Q\Omega_M - C^2}\begin{pmatrix} \Omega_M & -C \\ -C & \Omega_Q \end{pmatrix}$,

$$\mathbf{1}^\top\Xi^{-1}\mathbf{1} = \frac{\Omega_Q + \Omega_M - 2C}{\Omega_Q\Omega_M - C^2}, \qquad \mathbf{1}^\top\Xi^{-1}y = \frac{(\Omega_M - C)\tilde{Q} + (\Omega_Q - C)\tilde{M}}{\Omega_Q\Omega_M - C^2},$$

so $\mu^* = (1 - w)\tilde{Q} + w\tilde{M}$ with $w = (\Omega_Q - C)/(\Omega_Q + \Omega_M - 2C)$, as stated, and the error variance is $\mathrm{Var}(\mu^* - \theta) = (\mathbf{1}^\top\Xi^{-1}\mathbf{1})^{-1} = (\Omega_Q\Omega_M - C^2)/(\Omega_Q + \Omega_M - 2C)$.

The denominator is positive because $\Omega_Q + \Omega_M - 2C \geq \Omega_Q + \Omega_M - 2\sqrt{\Omega_Q\Omega_M} = (\sqrt{\Omega_Q} - \sqrt{\Omega_M})^2 \geq 0$, with equality only in the degenerate case excluded by positive definiteness.

(i) Comparing with $\tilde{Q}$ alone,

$$\Omega_Q - \mathrm{Var}(\mu^* - \theta) = \frac{\Omega_Q(\Omega_Q + \Omega_M - 2C) - (\Omega_Q\Omega_M - C^2)}{\Omega_Q + \Omega_M - 2C} = \frac{(\Omega_Q - C)^2}{\Omega_Q + \Omega_M - 2C} \;\geq\; 0,$$

with equality if and only if $C = \Omega_Q$. When $C = \Omega_Q$ the conditional expectation of η given ε is ε itself (the regression coefficient of η on ε is $C/\Omega_Q = 1$), i.e. $\tilde{M}$ adds a copy of the experts' error plus independent noise and carries no information about θ beyond $\tilde{Q}$; conversely any $C \neq \Omega_Q$ makes $\tilde{M} - \tilde{Q} = \eta - \varepsilon$ informative about ε and hence about θ.

(ii) The sign of w is the sign of $\Omega_Q - C$ since the denominator is positive: $w > 0$ for $C < \Omega_Q$, $w = 0$ for $C = \Omega_Q$, and $w < 0$ for $C > \Omega_Q$, the last case being possible only if $\Omega_M > \Omega_Q$ (from $C < \sqrt{\Omega_Q\Omega_M}$). The statement that "the second view always receives positive weight" is the special case $C = 0$.

(iii) Under (4) write $\Omega_Q = \omega_0 + \omega_1 s$, $\Omega_M = \nu_0 + \nu_1 s$ with $s = d^2$. For $C = 0$, $\Delta(s) = \Omega_Q^2/(\Omega_Q + \Omega_M)$ and $\Delta'(s) \propto 2\omega_1\Omega_M + \Omega_Q(\omega_1 - \nu_1)$, which is positive whenever $\Omega_M/\Omega_Q > (\nu_1 - \omega_1)/(2\omega_1)$; with the estimated $\nu_1/\omega_1 \in [1.4, 3]$ this requires $\Omega_M/\Omega_Q > 0.2$–1.0, satisfied in every training window. For $C = \rho\sqrt{\Omega_Q\Omega_M}$ with $\rho > 0$ the conclusion fails: evaluating $\Delta(s)$ on the seven estimated parameter sets $(\omega_0, \omega_1, \nu_1, \rho)$ over $d \in [0, 10\%]$, Δ is non-monotone in every year, typically falling from its value at $d = 0$ to a minimum at $d \approx 2$–6% before rising, and in two years falling throughout; the reason is that C grows with $\sqrt{\Omega_Q\Omega_M}$ and approaches Ω_Q as Ω_M outgrows Ω_Q, driving w toward zero. The proposition therefore claims monotonicity only for $C = 0$. □

OA.1.4 Proof of Lemma 1 and Proposition 4

Lemma. Let h be feasible for (9) with the target constraint binding: $h^\top\mu^* = \text{target} + z_\alpha\hat{\sigma}_h$. Decompose the predictive error in the quarter as $\theta - \mu^* = -\bar{b}\mathbf{1} + u$ with $\bar{b} = -\frac{1}{n}\mathbf{1}^\top(\theta - \mu^*)$ and $\mathbf{1}^\top u = 0$. Then, using $\mathbf{1}^\top h = 1$,

$$h^\top\theta - \text{target} = h^\top\mu^* - \text{target} + h^\top(\theta - \mu^*) = z_\alpha\hat{\sigma}_h - \bar{b} + h^\top u. \quad \square$$

Sign of the selection term under a monotone selection effect. Suppose $\mathbb{E}[u_i \mid \mu_i^*, \Sigma^*] = g(\mu_i^*)$ with g non-increasing, the pairs (u_i, μ_i^*) exchangeable across assets given Σ^*, and $h_i = \phi(\mu_i^*)$ with ϕ non-decreasing (as for (9) when the V_{ii}^{idio} are equal). Then $\mathbb{E}[h^\top u] = \sum_i \mathbb{E}[\phi(\mu_i^*)g(\mu_i^*)]$ and, since $\sum_i u_i = 0$ implies $\sum_i g(\mu_i^*)$ has mean zero, $\mathbb{E}[h^\top u] = \sum_i \mathrm{Cov}\left(\phi(\mu_i^*), g(\mu_i^*)\right) \leq 0$ by Chebyshev's covariance inequality for a non-decreasing and a non-increasing function of the same variable. Without monotonicity of ϕ in μ_i^*, which fails when the V_{ii}^{idio} differ, since (9) trades μ_i^* against V_{ii}^{idio}, no sign is implied, and the term is treated as an empirical quantity in the text. □

Proposition. The event $\{h^\top\theta \geq \text{target}\}$ equals $\{\bar{b} - h^\top u \leq z_\alpha\hat{\sigma}_h\}$ by the lemma. Conditional on $\mathcal{F}$, $z_\alpha\hat{\sigma}_h$ is a constant, so $\Pr[\bar{b} - h^\top u \leq z_\alpha\hat{\sigma}_h \mid \mathcal{F}] = 1 - \alpha$ if and only

if $z_\alpha \hat{\sigma}_h$ is the conditional $(1-\alpha)$-quantile $q_{1-\alpha}$ (for a continuous conditional law). If the conditional law is $N(m, s^2)$ then $q_{1-\alpha} = m + z_\alpha s$, so coverage is below $1-\alpha$ if and only if $m + z_\alpha s > z_\alpha \hat{\sigma}_h$; the cases $m > 0$ with $s = \hat{\sigma}_h$ (common bias) and $s > \hat{\sigma}_h$ with $m = 0$ (understated dispersion) are the two one-at-a-time instances, and a positive m with $s < \hat{\sigma}_h$, or a negative m with $s > \hat{\sigma}_h$, can fall on either side. For a non-normal conditional law with mean zero and standard deviation $\hat{\sigma}_h$, $q_{1-\alpha}$ may lie on either side of $z_\alpha \hat{\sigma}_h$ (below it for a light-tailed or left-skewed law, above it for a heavy-tailed or right-skewed one at $\alpha = 0.10$), so coverage may exceed or fall short of $1-\alpha$. □

OA.1.5 Proof of Proposition 5

(i) For any feasible h_k, $h_k^\top \mu_k^* = \text{target} + z_\alpha \hat{\sigma}_{h_k} + \text{slack}_k$ by definition of the slack, so Lemma 1's computation gives $S_k = z_\alpha \hat{\sigma}_{h_k} + \text{slack}_k - \bar{b}_k + h_k^\top u_k$ in every quarter; take expectations, difference, and apply the triangle inequality. □

(ii) Under (6), $\hat{\sigma}^2_{h_k} = h_k^\top V_k^{\text{idio}} h_k + v_{g,k}(\mathbf{1}^\top h_k)^2 = h_k^\top V_k^{\text{idio}} h_k + v_{g,k}$ for fully invested h_k. Since V_k^{idio} is diagonal with entries at most $\bar{D}$, $0 \le h_k^\top V_k^{\text{idio}} h_k \le \bar{D}\|h_k\|_2^2$, and $\sqrt{v+x} - \sqrt{v} \le x/(2\sqrt{v})$ for $x \ge 0$ gives the bracket; the difference of two numbers lying in the two brackets is bounded by $|\sqrt{v_{g,2}} - \sqrt{v_{g,1}}| + \max_k \bar{D}\|h_k\|_2^2/(2\sqrt{v_{g,k}})$ realization by realization; take expectations. □

(iii) Conditional calibration $\mathbb{E}[\theta \mid \mathcal{F}_k] = \mu_k^*$ gives $\mathbb{E}[\theta - \mu_k^* \mid \mathcal{F}_k] = 0$ componentwise. Since $\bar{b}_k = -\frac{1}{n}\mathbf{1}^\top(\theta - \mu_k^*)$ and $u_k = (\theta - \mu_k^*) + \bar{b}_k \mathbf{1}$ are linear in $\theta - \mu_k^*$ with coefficients in $\mathcal{F}_k$, $\mathbb{E}[\bar{b}_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[u_k \mid \mathcal{F}_k] = 0$; for h_k measurable with respect to $\mathcal{F}_k$, $\mathbb{E}[h_k^\top u_k \mid \mathcal{F}_k] = h_k^\top \mathbb{E}[u_k \mid \mathcal{F}_k] = 0$. The last sentence is the contrapositive. □

(iv) The CRPS of a normal predictive distribution with variance v and error e is $\sqrt{v}\left(z(2\Phi(z) - 1) + 2\phi(z) - 1/\sqrt{\pi}\right)$ with $z = e/\sqrt{v}$, and the log score is $\frac{1}{2}\log(2\pi v) + e^2/(2v)$. For asset i under specification k, $v_{k,i} = V^{\text{idio}}_{k,ii} + v_{g,k}$ and $e_{k,i} = \theta_i - \mu^*_{k,i}$; the expected difference between specifications is a function of these quantities and involves no portfolio weight. □

OA.1.6 The joint model: the exact block-covariance update, the level–deviation approximation, and the constrained level

Fix a quarter and drop t. The model is $\theta = \pi + g\mathbf{1} + \xi$ with $\pi = \bar{\pi}\mathbf{1}$ (a quarter-level constant), $\xi \sim N(0, \Sigma_\xi)$, $\Sigma_\xi = \text{diag}(1/P_i)$, $P_i = (\tau\sigma_i^2)^{-1}$, g diffuse, and, for the k calibrated views stacked *asset-major* as $y = (y_1; \ldots; y_n) \in \mathbb{R}^{nk}$ with $y_i = (y_{i,1}, \ldots, y_{i,k})^\top$ the views on asset i,

$$y_i = \theta_i \mathbf{1}_k + f + u_i, \quad f = (f_1, \ldots, f_k)^\top \sim N(0, F), \quad u \sim N(0, R), \quad R = \text{diag}_i(\Xi_i),$$

where R is block-diagonal with $k \times k$ blocks $\Xi_i = \Xi_i(d_i)$ that differ across assets. Throughout this appendix the covariance rule $\Xi_i(d_i)$, fitted on history and evaluated at the current disagreement $d_i = |Q_i - M_i|$, is taken as given: what is derived is the exact Gaussian update *conditional on* that rule, and "exact" is relative to the matrix approximation of the next paragraph, not a claim that the disagreement-dependent model has an unconditional generative posterior (Section 3.4).

Exact block-covariance update. Integrating f out of the likelihood, $\mathrm{Cov}(y \mid \theta) = R + WFW^\top$ with $W = \mathbf{1}_n \otimes I_k$ (the $nk \times k$ matrix that assigns the same f to every asset), and by the Woodbury identity

$$(R + WFW^\top)^{-1} = R^{-1} - R^{-1} W\, K\, W^\top R^{-1}, \qquad K = \left(F^{-1} + \textstyle\sum_i \Xi_i^{-1}\right)^{-1},$$

since $W^\top R^{-1} W = \sum_i \Xi_i^{-1}$. With $H = I_n \otimes \mathbf{1}_k$ (the $nk \times n$ matrix that maps θ to $(\theta_i \mathbf{1}_k)_i$) the likelihood precision of θ is $H^\top (R + WFW^\top)^{-1} H = \mathrm{diag}(\mathbf{1}^\top \Xi_i^{-1} \mathbf{1}) - CKC^\top$, where the ith row of $C \in \mathbb{R}^{n \times k}$ is $c_i = \Xi_i^{-1}\mathbf{1}$, and the likelihood score is $H^\top (R + WFW^\top)^{-1} y = (\mathbf{1}^\top \Xi_i^{-1} y_i)_i - CK \sum_i \Xi_i^{-1} y_i$. The prior precision, in the limit of a diffuse g, is $\mathrm{diag}(P_i) - PP^\top / (\mathbf{1}^\top P)$ with $P = (P_i)_i$ (the Schur complement of the diffuse direction), and it annihilates $\pi = \bar\pi \mathbf{1}$, so the prior contributes no location. Hence

$$\begin{gathered}
\Sigma^{\mathrm{exact}} = \left(A - UG_u U^\top\right)^{-1} = A^{-1} + LGL^\top, \qquad L = A^{-1}U, \\
G = (G_u^{-1} - U^\top A^{-1} U)^{-1}, \qquad A = \mathrm{diag}(P_i + \mathbf{1}^\top \Xi_i^{-1} \mathbf{1}), \\
U = [C,\, P], \qquad G_u = \begin{pmatrix} K & 0 \\ 0 & 1/\mathbf{1}^\top P \end{pmatrix},
\end{gathered}$$

and $\mu^{\mathrm{exact}} = \Sigma^{\mathrm{exact}} \cdot \mathrm{score}$. Everything is $n \times (k+1)$ or $(k+1) \times (k+1)$, so the exact update costs $O(nk^2)$ per quarter; the implementation (`exact_fuse`) computes it, checks that G is positive definite in every quarter, and stores (A^{-1}, L, G) so that the portfolio problem uses $h^\top \Sigma^{\mathrm{exact}} h = \sum_i h_i^2 / A_{ii} + (L^\top h)^\top G (L^\top h)$ exactly. The formulas do not depend on the stacking order: a source-major stacking permutes y, R, W and H by the same permutation matrix and leaves A, C, K, U, G and the update unchanged; the code works asset by asset, as written here. The common factors F enter the exact mean through K and G; a diffuse level prior does not integrate them out of the mean, it only removes the shrinkage of the level toward $\bar\pi$.

The level–deviation approximation. Let $J = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $M = I - J$. Applying J and M to each view gives $\bar y_s = \bar\pi + g + \bar\xi + f_s + \bar u_s$ and $My_s = M\xi + Mu_s$. The two sets of observations are *not* independent in general: $\mathrm{Cov}(Ju_s, Mu_{s'}) = JR_{ss'}M$, whose entries are $\frac{1}{n}\big((\Xi_i)_{ss'} - \overline{(\Xi)}_{ss'}\big)$, vanish only when Ξ_i is the same for every asset. Treating them as independent, and dropping the $O(1/n)$ covariance $\Sigma_{\bar\xi\bar u} = \frac{1}{n^2}\sum_i \Xi_i + \frac{1}{n^2}\sum_i P_i^{-1}$ of the mean noises, gives the two-part update of Section 3.4: the deviations are updated asset by asset, $\mathbb{E}[\theta_i - \bar\theta \mid \cdot] = \mathbf{1}^\top \Xi_i^{-1}(y_i - \bar y)/(P_i + \mathbf{1}^\top \Xi_i^{-1}\mathbf{1})$ with variance $1/(P_i + \mathbf{1}^\top \Xi_i^{-1}\mathbf{1})$, which is (8); the level g is the GLS combination of the k means $z_s = \bar y_s - \bar\pi$ with error covariance F alone, $\mathbb{E}[g \mid \cdot] = w_F^\top z$, $w_F = F^{-1}\mathbf{1}/(\mathbf{1}^\top F^{-1}\mathbf{1})$, $\mathrm{Var}(g \mid \cdot) = 1/(\mathbf{1}^\top F^{-1}\mathbf{1}) = v_g$; and $\Sigma^* = V^{\mathrm{idio}} + v_g \mathbf{1}\mathbf{1}^\top$, which is (6). Equations (6)–(8) are therefore an $O(1/n)$ approximation to the exact block-covariance update, not the update itself. The dropped terms are small against the idiosyncratic variances (of order Ξ_i/n with $n \approx 1{,}000$) but they are not small against F when F is near singular: the mean-noise covariance $\bar\Xi/n$ is of the order of 10^{-6} in the units of the yield ($n \approx 1{,}000$, $\bar\Xi \approx 10^{-3}$), while the smallest eigenvalue of the estimated F is at its floor of 10^{-8} in 2019–2020 and 3–7 $\times\, 10^{-6}$ thereafter. In the exact update the

mean-noise term regularizes the level combination; in the approximation nothing does, and the unconstrained w_F extrapolate, to $(-9.5, 10.5)$ in 2020, where $\hat{\rho}_F = 0.999$. Table OA.16 reports the two side by side in every test year: the exact update and the unconstrained approximation differ in the level by 0.01–0.17 points on average in the years in which F is away from its floor and by 6.4 points in 2020, and in the deviations by at most 0.4 points at any asset (median below 0.08).

Constrained level (main specification). The main specification replaces the unconstrained w_F by $\arg\min_{w\in\Delta} w^\top F w$, which for $k = 2$ is $w_Q = \min\{1, \max\{0, (F_{MM} - F_{QM})/(F_{QQ} + F_{MM} - 2F_{QM})\}\}$, and sets $v_g = w_F^\top F w_F$. This is a constrained GLS pooling of the two views' quarter means, the constrained forecast combination of Bates and Granger (1969), and not a posterior mean of the Gaussian model; $w_F^\top F w_F$ is the error variance of the pooled level under the model, not a posterior variance, and it exceeds $1/(\mathbf{1}^\top F^{-1}\mathbf{1})$ whenever the constraint binds. The predictive distribution of the main specification is thus the empirical-Bayes update of the deviation channel combined with a constrained pooled level; Section 3.4 names it as such, and Remark 1 gives the level channel's information conditions.

An informative level prior. If instead $g \sim N(0, \sigma_g^2)$, the level posterior mean becomes $\bar{\pi} + \sigma_g^2 \mathbf{1}^\top(\sigma_g^2 \mathbf{1}\mathbf{1}^\top + F)^{-1} z$, which shrinks the views' level toward $\bar{\pi}$ by an amount governed by σ_g^2/F; this is identification of the level from the prior location, the route of Proposition 1(iii) that the paper declines for earnings yields, and σ_g^2 itself is identified only from realized outcomes. A construction that computes μ^* asset by asset with the prior included in the level and then attaches $\Lambda F \Lambda^\top$ with the asset-level weights as loadings (the two-stage plug-in construction of Section OA.2) is not the update of this model under any prior, and Section OA.2 reports what it does.

OA.1.7 General pick matrix

With $k < n$ views per source and pick matrix $P \in \mathbb{R}^{k\times n}$, the stacked view vector is $y = H\theta + b + e$ with $H = [P; P]$, $b = (b_Q, b_M)$ and $e = (\varepsilon, \eta) \sim N(0, \Xi)$. Proposition 1 holds with c replaced by any vector in the row space of P (only the projection Pc of a common shift is confounded with $P\theta$); Proposition 3 applies view by view to the 2×2 blocks of Ξ; and the update (7) is the standard two-view Black–Litterman update with H in place of $[I; I]$. The paper uses $P = I$ because every asset in the universe carries both an expert consensus and an AI view.

OA.2 Sensitivity

This appendix records, for every specification, what enters its predictive distribution and what the rolling estimation used at each cutoff; documents the selectivity of the missing prompts; and varies, one at a time, the constraint set of the decision problem, the definition of the expert consensus, the treatment of extreme observations, and the forecast horizon, and reports in each case the two comparisons that carry the paper's conclusion: the calibrated single-source benchmark against standard BL on Q (the value of historical calibration) and the dual-source model against the calibrated benchmark (the value of the second view). Two further sensitivities are in the main

text: the market-impact cost model (Section 7.1) and the split at the LLM's release date (Section 7.2).

Specification audit and estimation windows.

Table OA.19 is generated from the panel of predictive means and variances and states, for each specification, the views entering its predictive mean, the idiosyncratic covariance, the level variance v_g, and the level standard deviation $\sqrt{v_g}$ in each test year; every decision-layer table uses these objects and nothing else. Table OA.1 records the rolling estimation at each cutoff: the firm-quarters in the three-year window, the number dropped because the annual EPS had not been announced by the cutoff (2,200–3,200 per window, a quarter of the window), the counts by season that enter the calibration constants, and the quarters that enter F. Table OA.18 in Section 6.4 gives the year-by-year estimates of F, the level weights before and after the simplex constraint, and the share of weight in the top disagreement quintile.

Table OA.1 Point-in-time audit of every rolling estimation window: the cutoff is the first forecast date of the test year; training firm-quarters whose realized annual EPS was announced on or after the cutoff are dropped before any parameter is estimated; F uses the cross-sectional mean errors of every training quarter before the test year that survives the same rule.

Test year	Firm-quarters in 3-year window	Dropped: EPS announced after cutoff	Used for $\hat{b}$, $\Xi(d)$, τ	By season Q1/Q2/Q3/Q4	Training quarters	Quarters for F	F floor binds	Test firm-quarters
2019	9,210	2,207	7,003	1,364/1,689/1,933/2,017	201603–201812	12	yes	3,686
2020	10,672	3,059	7,613	1,753/1,855/1,990/2,015	201703–201912	16	yes	3,657
2021	10,995	2,683	8,312	2,007/2,060/2,177/2,068	201803–202012	20	no	3,604
2022	10,947	2,688	8,259	1,983/2,055/2,169/2,052	201903–202112	24	no	4,083
2023	11,344	3,153	8,191	1,946/2,018/2,135/2,092	202003–202212	28	no	4,430
2024	12,117	2,997	9,120	2,123/2,224/2,381/2,392	202103–202312	32	no	4,464
2025	12,977	3,199	9,778	2,269/2,407/2,559/2,543	202203–202412	36	no	4,406

The estimation cutoff is 31 March of the test year in every row; the F window runs from 2016Q1 to the last training quarter (expanding). Quarters are labelled by their forecast date (YYYYMM). The seasonal calibration constants use the by-season counts; the variance regressions and τ use all training firm-quarters; the F column counts the quarter means entering the factor covariance.

Constraint set and factor covariance.

Table OA.2 varies the constraint set of the chance-constrained problem (9) and the estimate of the factor covariance F. Because the minimum-variance objective distributes weight across the whole universe, the position cap is slack in every quarter and quarterly turnover stays below the 30% cap; tightening the confidence level to $\alpha = 5\%$ scales the safety margin but leaves both comparisons unchanged. Estimating F on the three-year training window instead of the expanding window, or flooring its eigenvalues at 0.1 or 0.2 pp instead of 0.01, changes the calibration gain by less than 0.01 points and the dual-source model's difference from the benchmark by at most 0.05; the floor binds in 2019 and 2020: in 2019 because twelve training quarters give a negative common variance net of sampling noise, so the 2019 portfolios carry almost no level margin under any of the three floors, and in 2020 because the two views' common errors are correlated at 0.999 and the smaller eigenvalue of F is at the floor. Replacing the normal quantile of the level part of the margin by the empirical 90% quantile of the training-window quarter-mean predictive error (Proposition 4) changes neither comparison. Using the unconstrained GLS level weights of the level–deviation approximation, which leave the unit simplex in 2019–2021 and reach $(-9.5, 10.5)$ in 2020 (Table OA.18), moves the dual-source model's difference to $+0.80$ points ($t = 0.8$); the number is not evidence about the second view but about the approximation, whose level in 2020 is eight points from the exact block-covariance update's (Table OA.16). The exact update itself, in which the mean-noise term regularizes the level, gives $+0.13$ ($t = 0.4$) and $+0.09$, a calibration gain of 0.94 and 0.84, and f10 at -0.41 and -0.50. The last two rows are two-stage plug-in constructions in which the posterior mean is computed asset by asset with the historical prior in the level and a factor covariance is attached afterwards. With asset-specific loadings $\Lambda F \Lambda^\top$ the dual-source model came out 0.38 points behind the benchmark ($t = -1.8$), through the selection term: the heterogeneous loadings gave the optimizer a factor to hedge and the hedge tilted the portfolio toward the high-disagreement names in which the state-dependent bias is largest. With unit loadings it did not ($+0.03$). Neither construction is the predictive distribution of the model of Section 3.4; they are kept as a record of what a covariance that is not the model's own does to a portfolio (Section 6.4).

Forecast horizon.

Splitting the results by fiscal quarter of the forecast date (four horizons, seven test quarters each) leaves both conclusions in place at every horizon (Table OA.24 of the Online Appendix): the calibration gain is 0.82–1.09 points in every quarter type, the second view's proper-score gains are present at every horizon, and its shortfall difference is within ± 0.16 points ($|t| \leq 0.6$). The calibrated benchmark misses its target at the longest horizon (-0.42 points) and exceeds it at the shortest ($+0.79$), so the pooled shortfall averages a miss against an excess; a fixed twelve-month target would be the cleaner design and we have not run it.

Table OA.2 Sensitivity of the decision-layer comparisons to the constraint set and to the factor covariance. Paired shortfall differences in percentage points (positive = first portfolio has the smaller shortfall) with fiscal-year-clustered t_6 in parentheses, under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6).

Constraint set	Target 6%		Target 8%	
	f15 vs calibrated BL	calibrated BL vs standard BL	f15 vs calibrated BL	calibrated BL vs standard BL
Main: α=10%, $h_{\max}$=2%	-0.028 (-0.12)	+0.977 (+4.83)	-0.056 (-0.20)	+0.879 (+4.03)
$h_{\max}$=5%	-0.027 (-0.12)	+0.977 (+4.83)	-0.065 (-0.22)	+0.890 (+3.94)
α=5%	-0.046 (-0.19)	+0.972 (+4.77)	-0.074 (-0.25)	+0.878 (+3.91)
Turnover cap 30%/qtr	-0.026 (-0.11)	+0.972 (+4.87)	-0.026 (-0.10)	+0.853 (+4.13)
F from three-year window	+0.013 (+0.06)	+0.984 (+4.79)	-0.016 (-0.06)	+0.890 (+3.98)
F eigenvalue floor 0.1 pp	-0.041 (-0.18)	+0.977 (+4.83)	-0.071 (-0.25)	+0.879 (+4.03)
F eigenvalue floor 0.2 pp	-0.075 (-0.36)	+0.978 (+4.85)	-0.105 (-0.41)	+0.878 (+4.02)
Empirical quantile margin	-0.070 (-0.29)	+1.006 (+4.81)	-0.109 (-0.36)	+0.917 (+3.89)
Unconstrained level weights w_F	+0.800 (+0.79)	+0.977 (+4.83)	+0.568 (+0.66)	+0.879 (+4.03)
Exact block-covariance update (unconstrained level)	+0.129 (+0.40)	+0.938 (+4.62)	+0.088 (+0.24)	+0.844 (+3.81)
Two-stage plug-in covariance, $\Lambda F \Lambda^\top$	-0.383 (-1.84)	+1.121 (+4.44)	-0.341 (-1.65)	+0.925 (+3.38)
Two-stage plug-in covariance, $\sigma_c^2 \mathbf{1}\mathbf{1}^\top$	+0.030 (+0.19)	+0.732 (+4.28)	-0.012 (-0.06)	+0.611 (+2.76)

The position cap never binds and quarterly turnover stays below the 30% cap in every quarter. Rows 5–7 vary the estimate of the common-error covariance F (three-year window; eigenvalue floors of 0.1 and 0.2 pp instead of 0.01); row 8 replaces the normal quantile of the level part of the margin by the empirical 90% quantile of the training-window quarter-mean predictive error; row 9 uses the unconstrained GLS level weights of the approximation, which leave the unit simplex in 2019–2021; row 10 is the exact block-covariance update of the joint model (Section OA.1.6, Table OA.16); the last two rows are two-stage plug-in covariance constructions that are not predictive distributions of the model of Section 3.4; they are reported as covariance-misspecification diagnostics.

OA.3 Additional tables and figures for the main text

This section collects the tables and the figure that the main text refers to but does not reproduce: the positioning of the paper relative to the closest LLM-forecasting studies (Table OA.3; Section 2.3), the power budget used to choose the decision problem (Table OA.4; Section 6.1), the paired decision-layer comparisons and non-inferiority tests (Tables OA.5 and OA.6; Section 6.2), the seven fiscal-year means behind the main comparisons (Table OA.7), the realized coverage of the chance constraint with block-bootstrap intervals (Table OA.8; Section 6.3), the forecast errors by decile of signed disagreement (Table OA.9; Section 6.4), the split at the LLM's release date (Table OA.10; Section 7.2), and the probability integral transform of the asset-level predictive distributions (Figure OA.1; Section 5.3).

Table OA.3 Design of the closest predecessors and of this paper. "Sees human output" asks whether the machine forecaster is given the human forecasts (consensus, own history, or ratings) as input; "needs error history" whether the correction requires realized past errors of each expert.

Study	Machine forecaster	Inputs	Sees human output	Needs error history	Decision layer
Chhaochharia et al. (2021)	Neural network (error-correction layer)	Analysts' past errors, size, B/M, horizon	Yes	Yes ($\geq$ 4 lags)	Long–short sorts
Artrip et al. (2026)	LLM personas	Macro releases + lagged SPF consensus	Yes	No	None (accuracy)
An and Lee (2026)	LLM	Firm data + consensus + analyst's own history	Yes	Yes (own history)	None (accuracy)
Chen (2025)	BP, CNN, LSTM, Transformer	Fundamentals; analyst forecasts as features	Yes	No	None (accuracy, win rates)
This paper	LLM on filings (information-set-independent AI view)	PIT financials + annual-report text only	**No** (verified)	**Yes** (both views calibrated)	Chance-constrained target-yield portfolio

B/M: book-to-market. SPF: Survey of Professional Forecasters. PIT: point-in-time.

Table OA.4 Power budget computed before the decision-layer experiment. Proxy portfolios use the constant-weight combination f3 vs. the uncorrected consensus Q with a common Σ; the main specification f10 was not used. MDE $= 2.8\,\sigma_\Delta/\sqrt{T}$ (two-sided 5%, power 80%).

Decision problem	Outcome	Effect f3 vs Q (pp)	MDE (pp)	Effect/MDE
Top-K selection, K=50	realized R/P	+0.018	0.275	0.07
Top-K selection, K=100	realized R/P	+0.017	0.165	0.10
Top-K selection, K=200	realized R/P	+0.109	0.111	0.98
Chance constraint, target 6%	realized shortfall	+0.741	0.145	5.12
Chance constraint, target 8%	realized shortfall	+0.506	0.157	3.22
Chance constraint, target 10%	realized shortfall	+0.192	0.223	0.86
MV with cash leg, δ=3000	realized utility	-0.223	0.176	1.27

Ranking-type problems are insensitive to bias because debiasing adds a (seasonal) constant; level-type problems are sensitive.

Table OA.5 Decision-layer comparisons of the realized shortfall, paired by quarter ($T = 28$), under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6). Positive Δ means the first portfolio has the smaller shortfall. The clustered t uses the seven fiscal-year means (t_6 reference).

Comparison	6%: Δ (pp), i.i.d. t	clustered t	8%: Δ (pp), i.i.d. t	clustered t
Value of historical calibration				
Calibrated single-source BL vs standard BL on Q	+0.977 (+9.32)	+4.83	+0.879 (+7.64)	+4.03
Value of the AI view on top of calibration				
f15 (dual-source calibrated BL) vs calibrated single-source BL	-0.028 (-0.23)	-0.12	-0.056 (-0.37)	-0.20
f3 vs calibrated single-source BL	-0.149 (-1.36)	-0.72	-0.260 (-1.75)	-0.93
f6 vs calibrated single-source BL	+0.013 (+0.14)	+0.08	+0.022 (+0.19)	+0.11
Mechanism counter-example: AI view taken as unbiased				
f10 vs calibrated single-source BL	-0.439 (-2.81)	-1.50	-0.525 (-2.66)	-1.44
f10 vs standard BL on Q	+0.537 (+5.25)	+3.58	+0.354 (+2.85)	+1.82

Non-inferiority of the AI-view specifications relative to the single-source benchmark is tested formally in Table OA.6; the absence of a significant difference here is not evidence of equivalence.

Table OA.6 Non-inferiority and equivalence of the realized shortfall against the calibrated single-source benchmark, paired by quarter with fiscal-year clustering (t_6). Non-inferiority at margin δ (one-sided, 5%) requires the lower bound of the 90% interval to exceed $-\delta$; equivalence (two one-sided tests) requires the whole interval to lie inside $(-\delta, \delta)$. Margins fixed at the design stage (Section 6.1): 0.10 pp of realized yield is a mandate tolerance on the target; 0.30 pp was set relative to the calibration effect seen in the same exploratory analysis and is an internal, descriptive yardstick only.

Target	Comparison	Δ shortfall (pp)	90% CI (clustered)	Non-inferior at $\delta = 0.10$ / 0.30	Equivalent at $\delta = 0.10$ / 0.30
6%	f15 vs calibrated single-source BL	-0.028	[-0.485, +0.430]	no / no	no / no
6%	f3 vs calibrated single-source BL	-0.149	[-0.550, +0.251]	no / no	no / no
6%	f6 vs calibrated single-source BL	+0.013	[-0.306, +0.331]	no / no	no / no
6%	f10 vs calibrated single-source BL	-0.439	[-1.010, +0.132]	no / no	no / no
8%	f15 vs calibrated single-source BL	-0.056	[-0.611, +0.500]	no / no	no / no
8%	f3 vs calibrated single-source BL	-0.260	[-0.806, +0.286]	no / no	no / no
8%	f6 vs calibrated single-source BL	+0.022	[-0.373, +0.416]	no / no	no / no
8%	f10 vs calibrated single-source BL	-0.525	[-1.234, +0.184]	no / no	no / no

At the small margin nothing is established. Read the results at the large margin from the table itself; the text states them.

Table OA.7 Paired differences by fiscal year: the seven annual means of the quarterly differences behind the main comparisons. With seven independent years, these seven numbers are the evidence; the clustered t in the last column is their mean over their standard error.

Comparison	FY19	FY20	FY21	FY22	FY23	FY24	FY25	Mean	t_6
CRPS f15 − calib. (pp)	-0.03	+0.02	-0.07	-0.11	-0.07	-0.04	-0.11	-0.06	-3.5
NLPD f15 − calib.	-0.04	-0.16	-0.10	-0.26	-0.14	-0.13	-0.18	-0.15	-5.8
MSE f15 − calib. ($\times 10^4$)	-0.22	+0.38	-0.16	-0.68	-0.72	+0.94	+0.90	+0.06	+0.2
Shortf. 6% f15 − calib. (pp)	+0.26	+0.99	+0.18	-0.05	-0.07	-0.51	-0.99	-0.03	-0.1
Shortf. 6% calib. − std. (pp)	+0.49	+0.56	+1.08	+0.73	+0.70	+1.27	+2.01	+0.98	+4.8
Shortf. 8% f15 − calib. (pp)	+0.28	+1.21	+0.32	-0.20	-0.31	-0.51	-1.19	-0.06	-0.2
Shortf. 8% calib. − std. (pp)	+0.47	+0.38	+0.93	+0.62	+0.61	+1.10	+2.06	+0.88	+4.0

f15 − calib.: dual-source model minus calibrated single-source BL (negative = f15 better for the forecast metrics); calib. − std.: calibrated minus standard BL on Q (positive = smaller shortfall). Shortfall differences in percentage points at the stated target under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6).

Table OA.8 Realized coverage of the chance constraint $\Pr[h^\top \theta \geq \text{target}] \geq 90\%$ under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6): share of the 28 quarters in which the realized portfolio yield reached the target, with 95% intervals from a fiscal-year block bootstrap (seven blocks, 5,000 draws).

Target	Standard BL on Q	Calibrated single-source BL	f3	f6	f10	f15
6%	21% [7, 43]	57% [32, 82]	64% [39, 89]	64% [36, 86]	32% [7, 61]	61% [36, 89]
8%	29% [11, 46]	61% [39, 82]	50% [21, 79]	68% [46, 89]	29% [4, 61]	54% [29, 82]

No specification approaches the nominal 90%; the shortfall of coverage is decomposed in Section 6.4.

Table OA.9 Mean forecast errors by decile of *signed* disagreement $Q - M$, out of sample ($n = 28{,}330$). Whichever source is higher is the one that is wrong; both are biased upward by about 0.6 points where they agree, and the midpoint is biased upward in both tails.

Decile of $Q - M$	Range (%)	e_Q	e_M	Midpoint	
1	[-38.7, -1.3]	+1.91	+5.10	+3.51	$M \gg Q$
2	[-1.3, -0.5]	+0.87	+1.66	+1.26	
3	[-0.5, -0.1]	+0.64	+0.90	+0.77	
4	[-0.1, 0.1]	+0.56	+0.55	+0.56	
5	[0.1, 0.3]	+0.67	+0.45	+0.56	
6	[0.3, 0.6]	+0.93	+0.47	+0.70	
7	[0.6, 0.9]	+1.17	+0.42	+0.79	
8	[0.9, 1.4]	+1.55	+0.39	+0.97	
9	[1.4, 2.6]	+2.42	+0.49	+1.45	
10	[2.6, 43.5]	+5.30	-0.30	+2.50	$Q \gg M$

Errors in percentage points of price. Firm-clustered regressions: $e_M = 0.014 - 0.594\,(Q - M)$ ($t = -18$); $e_Q = 0.014 + 0.406\,(Q - M)$ ($t = +12$); e_Q on $|Q - M|$: slope 0.63 ($t = 18$); e_M on $|Q - M|$: slope 0.12 ($t = 1.9$).

Table OA.10 Post-release contamination sensitivity: the dual-source model against the calibrated single-source benchmark before and after the LLM's release date (September 2024), and on the four 2025 quarters. Paired quarterly differences (negative = f15 better for forecast metrics; positive = f15 better for the shortfall) with fiscal-year-clustered t in parentheses where at least three fiscal years are available; the post-release rows have one or two fiscal years and are reported without inference.

Subsample	T	Δ median (pp)	Δ MSE	Δ CRPS (pp)	Δ NLPD	Δ shortfall 6% (pp)	Coverage f15 / calib.
Before release (2019Q1–2024Q2)	22	+0.023 (+0.3)	-0.17 (-0.7)	-0.050 (-2.8)	-0.143 (-4.9)	+0.229 (+1.4)	50% / 45%
After release (2024Q3–2025Q4)	6	-0.262	+0.92	-0.091	-0.161	-0.970	100% / 100%
2025 only (post-release, all four quarters)	4	-0.276	+0.90	-0.111	-0.181	-0.994	100% / 100%

Forecast dates after the release cannot be contaminated by information the model absorbed after the forecast date; earlier dates may be. The split is a post-release contamination sensitivity, not a model-development holdout: the 2025 quarters share one realized annual EPS and post-date every modelling decision except the last.

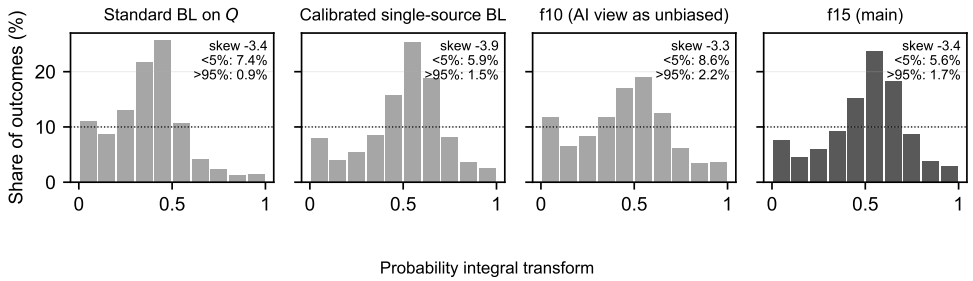

Fig. OA.1 Probability integral transform of the realized earnings yield under each specification's asset-level predictive distribution, out of sample (n = 28,330). A calibrated predictive distribution puts 10% of outcomes in each decile (dotted line). Insets: skewness of the standardized error and the shares of outcomes below the 5% and above the 95% predictive quantiles (nominal 5% each).

OA.4 Mean–variance formulation

Section 6.1 argued, on the basis of a power budget, that a fully invested mean–variance portfolio is close to a ranking problem and therefore a poor test bed for a bias correction. Table OA.11 reports the mean–variance portfolios themselves, $\max_h h^\top \mu - (\delta/2) h^\top \Sigma h$ subject to $\mathbf{1}^\top h = 1$ and $0 \le h \le h_{\max}$, for three risk-aversion levels and two position caps, and uses them as a direct check of Proposition 2(i).

Table OA.11 Mean–variance formulation $\max_h h^\top \mu - (\delta/2) h^\top \Sigma h$, $\mathbf{1}^\top h = 1$, $0 \le h \le h_{\max}$: realized portfolio earnings yield by specification. With a common Σ (the predictive variance of standard BL on Q) the historical calibration that is worth 0.98 points at the level decision moves the realized yield of the ranking problem by at most 0.2 points and in either direction, and the second view adds at most 0.1.

δ	$h_{\max}$	Names held	Realized R/P (%), common Σ					Paired Δ (pp), t	
		BL(Q)/calib.f15	BL(Q)	Calib. BL	f15	f3	calib. − BL(Q)	f15 − calib.	f15 − calib., own Σ^*
1000	2%	74/71/70	12.81	12.64	12.76	12.67	-0.17 (-2.7)	+0.11 (+1.4)	+0.24 (+3.1)
1000	5%	45/41/42	15.28	15.23	15.22	15.10	-0.05 (-0.6)	-0.01 (-0.0)	+0.21 (+0.8)
3000	2%	128/120/114	12.43	12.41	12.44	12.40	-0.02 (-0.7)	+0.03 (+0.6)	+0.24 (+2.0)
3000	5%	114/102/97	13.46	13.65	13.57	13.64	+0.19 (+2.0)	-0.08 (-1.9)	-0.24 (-1.5)
10000	2%	336/318/290	10.39	10.48	10.55	10.57	+0.09 (+2.0)	+0.07 (+1.1)	+0.00 (+0.0)
10000	5%	335/314/288	10.46	10.55	10.63	10.65	+0.09 (+2.1)	+0.07 (+1.2)	-0.20 (-0.8)

Quarterly averages over 28 quarters; fiscal-year-clustered t_6. Names held are counted under each specification's own predictive variance. The last column compares the dual-source model and the calibrated benchmark each under its own Σ^*.

With a common Σ (the posterior variance of standard BL on Q, used for every μ), the historical calibration that is worth 0.98 points of shortfall at the level decision of Section 6 moves the realized yield of the mean–variance portfolio by between -0.18 and $+0.20$ points, with the sign depending on δ and the cap, and with fiscal-year-clustered $|t| \leq 2.7$. The calibration is not exactly a common shift (each asset's predictive mean moves by the seasonal constant times its own weight on the view, which varies with the prior precision), so the invariance of Proposition 2(i) holds only approximately, and what survives is a small, sign-indefinite effect on the ranking. The second view adds between -0.06 and $+0.10$ points on top of the calibrated benchmark under the common Σ, in no cell significant under clustering, and is indistinguishable from the free-intercept combination f3. The ranking problem therefore cannot see the level correction that the target-yield problem pays for in full, which is why the paper's decision layer uses the latter.

With its own Σ^*, each specification holds a different number of names (335 for standard BL, about 315 for the calibrated benchmark and 290 for the dual-source model at $\delta = 10{,}000$), because a tighter predictive variance concentrates the solution at a given δ. The last column compares the dual-source model and the calibrated benchmark each under its own predictive variance; the differences range from -0.24 to $+0.24$ points and are a comparison of concentration as much as of μ. A specification with a smaller Σ^* collects the higher realized yield of a more concentrated bet on high-μ names and pays for it with a larger promised-minus-realized gap, the selection effect of Proposition 4. Comparisons across specifications with different Σ scales at a fixed δ are therefore reported under the common Σ. The chance-constrained formulation of the main text does not have this problem, because the target constraint rather than δ pins the risk taken, and the optimized portfolios all hold about 1,000 names regardless of Σ.

OA.5 Diagnostics and alternative specifications for the preliminary model

OA.5.1 Diagnostics for the preliminary closed-form specification

The specification f7 in Table 3 implemented a preliminary version of the model with two features later removed: the weak-prior limit $\tau \to \infty$, and a debiasing function with a signed disagreement term, $\tilde{Q} = Q - a_q - \rho\,(Q - M)$. To attribute its failure we replaced one ingredient at a time, re-estimating on the same rolling windows, and recorded the median absolute error in the highest disagreement quintile (Q5) and the low-disagreement quintile Q1 (percentage points, averaged over the seven test years with quintile edges from each training window; on the same basis the constant-weight benchmark f3 has 2.01 in Q5 and 0.55 in Q1).

Counterfactual	What changes	Q5	Q1
f7 as estimated	—	2.37	0.50
Weights replaced by empirical quintile optimum	w only	2.38	0.50
Weights replaced by those implied by f6	w only	2.34	0.54
$w = 0$: signed debiasing only, no anchor weight	w only	2.08	0.85
Seasonal debiasing, endogenous Ω, constant prior shrinkage ("f9")	signed term removed, prior added	1.91	0.57
Seasonal debiasing, endogenous Ω, endogenous prior weight (f10)	both	1.99	0.42

Source: `code/diag/diag_f7.py`, `diag_f7b.py`, `diag_f7c.py`, `diag_f7d.py`.

Replacing the precision weights by any alternative leaves Q5 essentially unchanged, so the functional form of Ω is not the cause. Setting the AI view weight to zero recovers the benchmark in Q5, because the signed debiasing term $\rho(Q - M)$ already puts weight ρ on M, and the precision weight w then adds to it; the AI view's total weight $(1 - w)\rho + w$ reached 0.6–0.8 against a least-squares optimum of 0.3–0.5, in the quintile where the AI view's variance is largest ($\nu_1/\omega_1 \approx 1.4$–3). Removing the signed term and adding a constant prior weight fixes Q5 (indeed improves on f3) but loses Q1, because a constant shrinkage over-corrects where both sources are precise; making the prior weight endogenous through (8) recovers Q1 while keeping Q5. The two modelling choices thus account for separate parts of the failure, and the corrected specification is f10, which keeps the AI view as an uncalibrated anchor; Section 6 shows what that costs at the decision layer. The general point, that the calibration covariates must not carry signed information about the other view, is Assumption 2.

OA.5.2 Alternative specifications evaluated after the decision-layer analysis

The first decision-layer experiment compared f10 with the free-intercept combination f3 and found a shortfall 0.35 points worse ($t = -12$); Section 6.4 traces this to the AI view's own bias, which f10 does not remove. Two alternative specifications were evaluated at the time, before the calibrated single-source benchmark and the dual-source model of Section 3.4 existed; Table OA.12 reports their forecast-layer statistics and Table OA.13 their decision-layer outcome. They are reported as the history of the model, not as candidates.

Table OA.12 Post-hoc modifications of the closed-form predictive mean, introduced after the first decision-layer test of the earlier design had failed (Section 6.4 of the main text). Same rolling design and kill-test statistics as Table 3 of the main text.

Specification	Median $\lvert e\rvert$	MSE	Bias	Q1 median	Q5 median	Q1–Q2 vs Q (t)	Q5 vs f3 (t)
f10 (AI view taken as unbiased)	0.841	12.62	+0.74	0.474	2.179	+0.022 (+0.57)	+0.047 (+1.29)
f12: f10 + even debias term $\kappa_Q\lvert Q-M\rvert$ (post hoc)	0.751	11.28	+0.57	0.363	1.965	-0.054 (-2.50)	-0.110 (-3.03)
f12b: f12 + seasonal debias of M (post hoc)	0.843	11.11	+0.24	0.573	2.013	+0.115 (+3.09)	-0.071 (-1.50)

These specifications were formulated after the decision-layer result of Section 6.4 of the main text and are reported as history; neither is used in the main text.

Table OA.13 Decision-layer outcome of the two post-hoc repairs of the closed form under the chance-constrained problem (9), paired against the free-intercept combination f3 ($T = 28$).

Specification	Target 6%		Target 8%	
	Shortfall (pp)	vs f3: Δ (t)	Shortfall (pp)	vs f3: Δ (t)
f3 (reference)	-0.12	—	-0.25	—
f10 (AI view taken as unbiased)	-0.47	-0.349 (-12.42)	-0.64	-0.393 (-10.45)
f11: correlated view noise ($\hat{\rho} \approx 0.8$)	-0.44	-0.319 (-12.82)	-0.54	-0.287 (-7.36)
f11b: f11 + seasonal debias of M	-0.16	-0.041 (-1.90)	-0.29	-0.041 (-1.04)
f12: f10 + even term $\kappa_Q \lvert Q - M \rvert$	-0.55	-0.422 (-12.28)	-0.75	-0.497 (-10.57)
f12b: f12 + seasonal debias of M	-0.26	-0.138 (-5.70)	-0.49	-0.241 (-6.31)

Both specifications were formulated after the decision-layer comparison of f10 with f3; neither is used in the main text. Shortfall levels use the diagonal predictive covariance of the preliminary experiment and are not comparable with Table 7 of the main text, which uses the factor covariance; the paired differences are the informative column.

Correlated view noise (f11). The training-window residual correlation between the two views is 0.68–0.86. Replacing the diagonal Ω in (7) by the full 2×2 view covariance with $c = \hat{\rho}\sqrt{\Omega_Q \Omega_M}$ changes the predictive-mean weights by only a few points, because with $\Omega_Q \approx \Omega_M$ the optimal weight $(\Omega_M - c)/(\Omega_Q + \Omega_M - 2c)$ is insensitive to c; the decision-layer shortfall moves from 0.47 to 0.44 and the test still fails ($t = -12.8$). Noise correlation is not the mechanism.

Seasonal debiasing of the AI view (f11b, f12b). Treating the two sources symmetrically, by subtracting the AI view's training-window seasonal mean error as is done for the experts, closes most of the decision-layer gap: the shortfall becomes 0.17 against f3's 0.12 ($t = -1.9$ at 6%, $t = -1.0$ at 8%). But it breaks the forecast layer. The AI view's bias is +0.5 points where the two sources agree and +5 points where the AI view is the optimist (Table OA.9); its seasonal average of +1.0 is therefore too much for the centre of the distribution, where errors are smallest and a level shift of half a point is fatal to the median. The low-disagreement median deteriorates from 0.422 to 0.617, significantly worse than the uncorrected consensus ($t = +3.7$), and the two-sided screening criterion of Section 5.1 fails in its other direction.

The even disagreement term (f12) improves the forecast layer throughout but leaves the AI view's tail bias untouched and so does not help the decision layer (shortfall 0.55, $t = -12.3$).

Two rounds of diagnostics were sufficient to reveal the structure of the problem, and the main text resolves it by changing the question rather than the estimator. A seasonal calibration of both views, which is what f12b does and what the dual-source model of Section 3.4 also does, must either leave the AI view's tail bias in (and lose the level decision) or remove it with a constant (and lose the centre of the median-error distribution). The calibrated single-source benchmark pays the same centre cost (low-disagreement median 0.624 against 0.379 for the raw consensus, Table 3), because any constant correction of a bias that is small where the two sources agree and large where they disagree is too much for the centre. Proposition 2 says which of the two losses matters: a level decision is charged the full common bias, a median is charged a half-point level shift where errors are smallest. The main text therefore evaluates the calibrated specifications on the decision they are built for and on the predictive

distribution as a whole (Section 5.3), where the dual-source model improves on the single-source benchmark, rather than on a median that rewards leaving the bias in. A calibration that varies with the signed disagreement would remove the centre cost, but Assumption 2 rules it out for the reason given in Section 3.4: it collapses the posterior to a regression of the outcome on the two views.

OA.6 Execution layer: paired net returns

Table OA.14 reports the realized performance of the target-6% portfolios after the execution frictions described in Section 7.1 of the main text, and Table OA.15 pairs their quarterly net returns (constant-cost and square-root market-impact models) with fiscal-year-clustered t-statistics. None of the comparisons was predicted and none is part of the paper's contribution.

Table OA.14 Realized performance of the target-6% portfolios after execution frictions, 2019-04 to 2026-04 (1,725 trading days), with weights from the factor covariance of Section 6. Trades at the close of the first trading day after each quarter-end; buy orders blocked at the upper price limit and sell orders at the lower limit (retried for three days, then filled at a 5% penalty); suspended names carried, delistings written to zero; commission 3 bp and slippage 10 bp per side, stamp duty 5 bp on sales; weights drift between rebalances. The last column replaces the constant slippage with a square-root market-impact model at an AUM of ¥10 bn.

Portfolio (target 6%)	Gross (%)	Net (%)	Vol. (%)	Sharpe	Max DD (%)	Turnover/qtr (%)	Blocked days	Net, AC cost (%)
Equal weight	11.89	11.67	21.7	0.64	-35.0	34	61	11.79
Value weight	8.98	8.87	18.9	0.56	-31.7	17	59	8.93
Standard BL on Q	11.80	11.49	20.5	0.65	-29.5	45	62	11.65
Standard BL on M	11.90	11.56	20.4	0.66	-29.9	50	62	11.73
Calibrated single-source BL	9.69	9.38	19.9	0.57	-29.3	47	63	9.54
f3	9.51	9.22	20.0	0.56	-30.6	44	63	9.37
f6	9.66	9.36	19.9	0.57	-30.1	45	63	9.51
f10 (AI view as unbiased)	11.47	11.14	20.2	0.64	-29.2	48	61	11.30
f15 (main)	10.56	10.22	19.9	0.61	-28.6	50	63	10.39
CSI 300 (ref.)	3.22	—	18.7	0.27	-45.6	—	—	—
CSI 500 (ref.)	5.81	—	21.8	0.38	-41.8	—	—	—

Annualized. Turnover is two-sided per quarter. Blocked days: number of days on which at least one order was blocked by a price limit or suspension, out of 28 rebalances. Index rows are price indices on a different universe, without dividends or costs, shown for orientation only.

Table OA.15 Paired differences in quarterly net returns ($T = 28$; fiscal-year-clustered t_6 in parentheses). Forward earnings yields have no return predictability in the sample, and none was expected.

Cost model	Comparison	6%: Δ, t	8%: Δ, t
Constant cost	f15 vs calibrated single-source BL	+0.207 (+1.65)	+0.180 (+1.51)
Constant cost	f15 vs equal weight	-0.460 (-0.99)	-0.874 (-1.15)
Constant cost	calibrated single-source BL vs standard BL on Q	-0.560 (-2.89)	-0.432 (-2.82)
Constant cost	f10 vs calibrated single-source BL	+0.442 (+2.52)	+0.354 (+2.19)
Constant cost	equal weight vs CSI 300 (reference)	+2.224 (+2.77)	+2.224 (+2.77)
AC impact, ¥10 bn	f15 vs calibrated single-source BL	+0.209 (+1.67)	+0.182 (+1.53)
AC impact, ¥10 bn	f15 vs equal weight	-0.449 (-0.96)	-0.859 (-1.13)
AC impact, ¥10 bn	calibrated single-source BL vs standard BL on Q	-0.560 (-2.88)	-0.434 (-2.82)
AC impact, ¥10 bn	f10 vs calibrated single-source BL	+0.443 (+2.53)	+0.355 (+2.19)
AC impact, ¥10 bn	equal weight vs CSI 300 (reference)	+2.250 (+2.80)	+2.250 (+2.80)

Positive Δ favours the first portfolio.

OA.7 Decision-layer supplements

This section holds four tables that the main text's Sections 3 and 6 refer to: the comparison of the exact block-covariance update of the joint model (conditional on $\Xi_i(d_i)$) with its level–deviation approximation and with the main specification's constrained level pooling (Table OA.16; Section 3.4 of the main text and Section OA.1.6 above), the coverage–margin trade-off as the nominal violation probability of the chance constraint is varied (Table OA.17), the year-by-year stability of the objects behind the comparison of the dual-source model with the calibrated benchmark (Table OA.18), and the machine-generated specification audit (Table OA.19), which lists for every specification the objects that enter its predictive mean, its idiosyncratic covariance and its level variance, with the level standard deviation by test year.

Table OA.16 The exact block-covariance update of the joint model of Section 3.4 (full block covariance, Woodbury, conditional on the fitted $\Xi_i(d_i)$), its level–deviation $O(1/n)$ approximation with unconstrained GLS level weights, and the main specification (the same approximation with the level weights constrained to the simplex), on the same panel with the same $\Xi_i(d)$, F and τ grid.

Panel A: f15 predictive means by test year (pp): level = quarter-mean difference, deviation = largest absolute demeaned difference

Test year	Exact − approximation (unconstrained)			Exact − main (constrained pooling)			sd ratio exact / main
	level mean	max |level|	max |dev.|	level mean	max |level|	max |dev.|	
2019	+0.105	0.243	0.173	+0.056	0.166	0.173	1.000
2020	+6.439	7.960	0.409	-0.762	1.375	0.409	0.997
2021	-0.137	0.265	0.226	-0.079	0.240	0.226	1.000
2022	-0.167	0.215	0.195	-0.167	0.215	0.195	1.000
2023	-0.114	0.168	0.087	-0.114	0.168	0.087	1.000
2024	+0.029	0.083	0.213	+0.029	0.083	0.213	1.000
2025	+0.009	0.037	0.114	+0.009	0.037	0.114	1.000

Panel B: forecast layer, pooled out of sample (n = 28,330)

Construction	Spec.	Bias (pp)	Median |e|	CRPS (pp)	NLPD	90% cov.	Δ f15 − calib. CRPS / NLPD (t_6)
Exact block update	f15	+0.23	1.021	1.554	-1.977	92.5%	-0.010 (-0.2) / -0.119 (-3.2)
	calib. BL	+0.31	1.004	1.571	-1.853	92.7%	
Approx., unconstr. w_F	f15	-0.56	1.069	2.193	-1.410	82.2%	+0.691 (+0.9) / +0.502 (+0.8)
	calib. BL	+0.31	1.004	1.571	-1.853	92.7%	
Main (constr. pooling)	f15	+0.37	0.954	1.509	-2.002	92.7%	-0.059 (-3.5) / -0.147 (-5.8)
	calib. BL	+0.31	1.004	1.571	-1.853	92.7%	

Panel C: decision layer (pp; T = 28; fiscal-year-clustered t_6)

Construction	Target	Shortfall f15 / coverage	Shortfall calib. / coverage	Margin f15 / calib.	Δ f15 − calib. (t_6)	Δ calib. − std. BL (t_6)	Δ f10 − calib. (t_6)
Exact block update	6%	+0.38 / 64%	+0.25 / 57%	0.52 / 0.58	+0.129 (+0.4)	+0.938 (+4.6)	-0.412 (-1.4)
	8%	+0.22 / 54%	+0.13 / 61%	0.54 / 0.60	+0.088 (+0.2)	+0.844 (+3.8)	-0.501 (-1.4)
Approx., unconstr. w_F	6%	+1.05 / 61%	+0.25 / 57%	0.38 / 0.58	+0.800 (+0.8)	+0.977 (+4.8)	-0.634 (-2.0)
	8%	+0.69 / 54%	+0.13 / 61%	0.21 / 0.60	+0.568 (+0.7)	+0.879 (+4.0)	-0.746 (-2.2)
Main (constr. pooling)	6%	+0.22 / 61%	+0.25 / 57%	0.53 / 0.58	-0.028 (-0.1)	+0.977 (+4.8)	-0.439 (-1.5)
	8%	+0.07 / 54%	+0.13 / 61%	0.55 / 0.60	-0.056 (-0.2)	+0.879 (+4.0)	-0.525 (-1.4)

Panel A: the level difference is the difference of the quarter means of the predictive mean (identical for all assets in a quarter); the deviation difference is the largest absolute difference after removing the quarter means; the sd ratio is the median across assets of the ratio of predictive standard deviations. Panel C: shortfall is the signed shortfall at the target; margin is the promised yield minus the target. The exact update and the unconstrained approximation differ only through the $O(1/n)$ terms of Appendix OA.1.6; the main specification differs from the approximation only through the simplex constraint on w_F.

Table OA.17 Coverage–margin trade-off at the 6% target: realized coverage (share of 28 quarters at or above target), the average model margin (promised yield minus target, pp) and the realized shortfall, as the nominal violation probability α of (9) is varied.

Nominal α	Calibrated single-source BL			f15 (main)			f6		
	Coverage	Margin	Shortfall	Coverage	Margin	Shortfall	Coverage	Margin	Shortfall
5%	61%	0.74	+0.41	64%	0.68	+0.36	68%	0.74	+0.42
10%	57%	0.58	+0.25	61%	0.53	+0.22	64%	0.57	+0.26
20%	54%	0.38	+0.06	54%	0.34	+0.06	57%	0.38	+0.08
30%	50%	0.23	-0.07	46%	0.21	-0.06	54%	0.23	-0.06

Nominal coverage is $1 - \alpha$; realized coverage stays 20–35 points below it at every α.

Table OA.18 Year-by-year stability of the objects behind the decision-layer comparison of f15 with the calibrated single-source benchmark: the estimated common-error covariance F (standard deviations in pp and correlation), the level weights before and after the simplex constraint, whether the eigenvalue floor binds, the level standard deviation of each specification, the share of weight in the top disagreement quintile at the 6% target, and the fiscal-year mean of the selection-term difference (pp).

Test year	sd f_Q	sd f_M	ρ_F	w_F unconstrained	w_F simplex	Floor binds	$\sqrt{v_g}$ f15 / calib.	Top-quintile wt. f15 / calib.	Δselection f15 − calib.
2019	0.10	0.02	0.85	(-0.17, +1.17)	(0.00, 1.00)	yes	0.02 / 0.10	18% / 24%	+0.08
2020	0.46	0.42	1.00	(-9.48, +10.48)	(0.00, 1.00)	yes	0.42 / 0.46	24% / 31%	+0.45
2021	0.50	0.61	0.91	(+1.45, -0.45)	(1.00, 0.00)	no	0.50 / 0.50	29% / 34%	+0.27
2022	0.48	0.53	0.89	(+0.87, +0.13)	(0.87, 0.13)	no	0.48 / 0.48	16% / 26%	+0.02
2023	0.40	0.44	0.76	(+0.70, +0.30)	(0.70, 0.30)	no	0.39 / 0.40	15% / 25%	-0.15
2024	0.52	0.45	0.74	(+0.23, +0.77)	(0.23, 0.77)	no	0.44 / 0.52	17% / 27%	+0.08
2025	0.54	0.48	0.73	(+0.29, +0.71)	(0.29, 0.71)	no	0.46 / 0.54	21% / 31%	-0.31

Unconstrained weights outside $[0, 1]$ extrapolate between the two views; the simplex constraint replaces them by a boundary solution in those years.

Table OA.19 Specification audit, generated from the panel of predictive means and variances (`v2_posterior.parquet`, fusion = joint): what enters each specification's predictive mean, its idiosyncratic covariance, and its level variance under (6), with the level standard deviation by test year. Every decision-layer table uses $\hat{\sigma}_h^2 = h^\top V^{\text{idio}} h + v_g$ with these objects.

Specification	Posterior mean	$\Xi_i(d)$ for V^{idio}	v_g	$\sqrt{v_g}$ by test year (pp)						
				2019	2020	2021	2022	2023	2024	2025
Standard BL on Q	raw Q; level $= \bar{Q}_t$	$\Omega_Q^{\text{raw}}(d)$	F_{QQ}	0.10	0.46	0.50	0.48	0.40	0.52	0.54
Standard BL on M	raw M; level $= \bar{M}_t$	$\Omega_M(d)$	F_{MM}	0.02	0.42	0.61	0.53	0.44	0.45	0.48
Calibrated single-source BL	$\tilde{Q}$; level $= \bar{\tilde{Q}}_t$	constant Ω_Q	F_{QQ}	0.10	0.46	0.50	0.48	0.40	0.52	0.54
f10 (AI view as unbiased)	$\tilde{Q}$, raw M; level $= w_F^\top(\bar{\tilde{Q}}_t, \bar{M}_t)$	$\Omega_Q(d), \Omega_M(d)$, $C = 0$	$w_F^\top F w_F$	0.02	0.42	0.50	0.48	0.39	0.44	0.46
f15 (main)	$\tilde{Q}, \tilde{M}$; level $= w_F^\top(\bar{\tilde{Q}}_t, \bar{\tilde{M}}_t)$	$\Omega_Q(d), \Omega_M(d)$, $C = \rho\sqrt{\Omega_Q \Omega_M}$	$w_F^\top F w_F$	0.02	0.42	0.50	0.48	0.39	0.44	0.46
f3	OLS on (Q, M)	borrowed from B6	borrowed from B6	0.10	0.46	0.50	0.48	0.40	0.52	0.54
f6	OLS with $\lvert Q - M \rvert$ interactions	borrowed from B6	borrowed from B6	0.10	0.46	0.50	0.48	0.40	0.52	0.54

Level weights w_F are the simplex-constrained GLS weights of Section 3.4; for a single source the level is that source's quarter mean. The point combinations f3 and f6 have no covariance of their own and borrow the calibrated single-source BL's.

OA.8 Additional sensitivities

This section collects the sensitivities that the main text summarizes in one sentence each: the definition of the expert consensus, the selectivity of the firm-quarters without a prompt, the treatment of extreme observations, the disagreement-quintile table and the sensitivity of its edges, the forecast horizon, and the descriptive table for the second machine view of Section 7.2 of the main text.

Definition of the consensus.

The main specification aggregates each institution's most recent forecast in a 180-day window by the median. Table OA.20 rebuilds Q with a 90-day window and with the mean, holding the AI view, prices and outcomes fixed.

Table OA.20 Sensitivity to the definition of the expert consensus. The AI view, prices and outcomes are unchanged; only Q is rebuilt. With a 90-day active window fewer firm-quarters have three institutions, and the consensus is more accurate.

Consensus	n	Q med.	Calib. BL med.	f15 med.	Δ median (t)	Δ CRPS (t)	Δ NLPD (t)	Δ shortfall f15 vs calib. (t)	Δ shortfall calib. vs std. (t)
Main: 180 d, median	28,330	0.903	1.004	0.954	-0.038 (-0.5)	-0.059 (-3.5)	-0.147 (-5.8)	-0.028 (-0.1)	+0.977 (+4.8)
90 d, median	23,189	0.737	0.823	0.847	+0.011 (+0.2)	-0.010 (-0.6)	-0.094 (-2.9)	-0.038 (-0.2)	+0.904 (+4.7)
180 d, mean	28,328	0.929	1.004	0.958	-0.034 (-0.4)	-0.052 (-2.9)	-0.137 (-4.9)	-0.045 (-0.2)	+1.013 (+5.3)

Median $|e|$ in percentage points of price. Δ columns are paired quarterly differences ($T = 28$, fiscal-year-clustered t_6): forecast metrics f15 minus calibrated single-source BL (negative = f15 better); shortfall differences at the 6% target under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6) (positive = first portfolio has the smaller shortfall).

Table OA.21 Firm-quarters for which a prompt could not be assembled, compared with those that enter the panel. A prompt requires a usable annual-report passage in the processed corpus and a point-in-time financial record before the forecast date.

Characteristic	With prompt (n = 37,596)		Without prompt (n = 2,576)		Difference (t)
	Mean	Median	Mean	Median	
Consensus forecast yield Q/P (%)	5.29	4.41	3.61	3.14	-1.68 (-13.7)
Realized yield R/P (%)	3.72	3.29	2.36	2.03	-1.36 (-9.5)
Realized loss (share, %)	6.62	—	5.36	—	-1.26 (-1.5)
Active institutions (count)	11.63	9.00	6.66	5.00	-4.97 (-19.0)
Forecast dispersion / price (%)	0.69	0.36	0.52	0.30	-0.17 (-7.7)
Market capitalization (¥bn)	43.53	19.22	14.95	8.19	-28.58 (-9.7)
Years since listing	12.94	11.64	6.21	5.02	-6.73 (-24.5)
Main board (share, %)	78.42	—	50.54	—	-27.87 (-8.9)

Difference = without minus with; t from a regression on the missingness indicator with firm-clustered standard errors. Market capitalization at the forecast date. Every firm in the universe was still listed in 2026 (see the text).

With the mean consensus the estimates are essentially unchanged. The 90-day window changes the sample (one firm-quarter in five no longer has three active institutions) and produces a more accurate consensus (median error 0.74 against 0.90 points). Both conclusions survive: the calibration gain is 0.90 points (clustered $t = 4.7$), and the second view improves the log score (-0.09, $t = -2.9$) but not CRPS (-0.01 points, $t = -0.7$), with a shortfall difference of -0.04 points ($t = -0.2$). The second view's proper-score gain is between one sixth (CRPS) and two thirds (NLPD) of its size in the main sample, which is what Proposition 3 predicts when the expert view carries less error to begin with: the conditional information value of the second view falls with the error it could remove.

Missing prompts.

Table OA.21 compares the 2,576 firm-quarters of the universe for which no prompt could be assembled with the 37,596 that enter the panel. The missing firm-quarters belong to smaller (median market capitalization ¥8 bn against ¥19 bn), younger (five years since listing against twelve) firms with about half as many active analysts, more often listed outside the main boards, and with lower consensus and realized earnings yields. All differences but the loss share are significant with firm-clustered standard errors. The missingness is thus a property of firms whose annual reports are less completely represented in the processed corpus, not of the outcome; it concentrates in 2016, a training-only year. We report it as a description of the population to which the results apply rather than attempt to reweight, since no result in the paper is a population average.

Quintile edges.

The disagreement quintiles of Table 3 of the main text and Table OA.22 below are formed with pooled out-of-sample quantiles of $|Q - M|$, which uses the future distribution of disagreement to label observations (though not to estimate anything). Recomputing the edges from each training window (`code/diag/robust_bins.py`; the

edges drift from $[0.22, 0.48, 0.87, 1.68]$ to $[0.29, 0.63, 1.14, 2.24]$ points between 2019 and 2025, and 22% of test observations then fall in the top quintile) changes no conclusion: the low-disagreement penalty of the calibrated specifications relative to Q is $+0.21$ ($t = 5.7$) for the single-source benchmark and $+0.15$ ($t = 3.9$) for f15, against $+0.20$ and $+0.14$ with pooled edges; the top-quintile comparisons with f3 remain insignificant for the calibrated specifications ($t = -0.9$, -1.1) and f6's advantage there grows (-0.09, $t = -2.7$ against -2.0).

Table OA.22 Constant-weight combination vs. the uncorrected consensus by quintile of human–AI disagreement $|Q - M|$ (out of sample). Constant weights hurt where the two sources agree and help where they disagree.

Disagreement quintile	$\lvert Q - M \rvert$ range (%)	Median $\lvert e \rvert$ (%)			MSE ($\times 10^4$)		
		Q	f3	Δ	Q	f3	Δ
Q1	[0.00, 0.24]	0.379	0.555	+46%	3.36	2.92	-13%
Q2	[0.24, 0.55]	0.567	0.671	+18%	5.70	4.83	-15%
Q3	[0.55, 1.01]	0.841	0.808	-4%	7.72	6.38	-17%
Q4	[1.01, 2.00]	1.334	1.117	-16%	12.73	9.74	-23%
Q5	[2.00, 43.54]	2.688	2.102	-22%	50.00	34.87	-30%

Quintile edges from the pooled out-of-sample distribution of $|Q - M|$ scaled by price.

Extreme observations.

The main analysis drops firm-quarters whose realized forecast error exceeds half the share price for either view (56 firm-quarters over the whole panel, 48 of them out of sample). Because this rule uses the outcome, Table OA.23 repeats the main comparisons keeping every observation, and with the alternative of clipping the three scaled quantities Q/P, M/P and R/P to a fixed ex-ante range of $[-50\%, 100\%]$ of price, which uses no outcome information.

Table OA.23 Sensitivity to the treatment of extreme observations. The main analysis drops firm-quarters whose realized forecast error exceeds half the share price for either view, a rule that uses the outcome. The alternatives keep every observation, or clip the three scaled quantities to a fixed ex-ante range. Paired differences with fiscal-year-clustered t; shortfall in percentage points at the 6% target under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6).

Treatment of extreme observations	n (OOS)	Q: median / MSE	Calib. BL median	f15 median	Δ CRPS (t_{cl})	Δ NLPD (t_{cl})	Δ shortfall 6%: f15 vs calib. (t_{cl})	Δ shortfall 6%: f15 vs BL(Q) (t_{cl})
Drop $\lvert e\rvert \geq 0.5$ (main; realized-error rule)	28,330	0.903 / 15.9	1.004	0.954	-0.059 (-3.5)	-0.147 (-5.8)	-0.028 (-0.1)	+0.949 (+7.3)
Keep all observations	28,378	0.906 / 32.6	1.069	1.143	-0.088 (-0.9)	-0.754 (-1.1)	+0.077 (+0.2)	+0.920 (+4.0)
Clip Q/P, M/P, R/P to $[-0.5, 1]$ (ex-ante rule)	28,378	0.906 / 20.4	1.029	1.090	-0.055 (-1.5)	-0.224 (-8.4)	+0.104 (+0.3)	+0.991 (+5.0)
Clip (ex-ante rule) × exact block-covariance update	28,378	0.906 / 20.4	1.029	1.089	-0.020 (-0.3)	-0.202 (-4.4)	+0.202 (+0.5)	+1.046 (+4.1)

Keeping the extremes doubles the MSE of every specification (a handful of firms with earnings collapses exceeding half the share price) and raises the median errors of the calibrated models, because the extremes enter the training-window calibration; the ranking of specifications and the decision-layer conclusions do not change. The last row combines the ex-ante clipping rule with the exact block-covariance update of Appendix OA.1.6 in place of the constrained level pooling.

Keeping the extreme observations roughly doubles the mean squared error of every specification, since a handful of observations with errors of several times the share price dominate a squared loss, and raises the medians of the calibrated specifications by 0.06–0.2 points; the proper-score differences grow in the second view's favour but become noisy (NLPD -0.75, clustered $t = -1.1$), because the same handful of observations dominates the log score of whichever predictive variance is smaller at those points. At the decision layer the dual-source model's difference from the benchmark is $+0.08$ points ($t = 0.2$) and its advantage over standard BL on Q is 0.92 points ($t = 4.0$). Clipping, the ex-ante rule, leaves every conclusion in place: CRPS -0.06 points ($t = -1.5$) and log score -0.22 ($t = -8.4$) for the second view, a shortfall difference of $+0.10$ points ($t = 0.3$), and 0.99 points ($t = 5.0$) over standard BL on Q. The last row combines the ex-ante clipping with the exact block-covariance update in place of the constrained level pooling: the NLPD gain is -0.20 ($t = -4.4$), the CRPS difference -0.02 ($t = -0.3$), and the decision-layer comparisons are $+0.20$ ($t = 0.5$) against the calibrated benchmark and $+1.05$ ($t = 4.1$) against standard BL on Q. The exclusion rule is a convenience for the squared-error columns, not a driver of the results; the CRPS significance is the one statistic that depends on it.

Forecast horizon.

The four fiscal quarters of the forecast date are four horizons: a Q1 forecast is made with almost the whole fiscal year ahead, a Q4 forecast with almost none of it. Table OA.24 splits the main results by fiscal quarter, with seven test quarters of each type.

Table OA.24 Results by forecast horizon: fiscal quarter of the forecast date (Q1 forecasts are made with almost the whole fiscal year ahead, Q4 forecasts with almost none of it). Biases and medians in percentage points of price; Δ columns are paired differences over the seven quarters of each type, one per fiscal year (t_6): forecast metrics f15 minus calibrated single-source BL (negative = f15 better); shortfall at the 6% target under the factor covariance $\Sigma^* = V^{\text{idio}} + v_g \mathbf{1}\mathbf{1}^\top$ of (6) (positive = first portfolio has the smaller shortfall).

Fiscal quarter of forecast date	n	Bias of Q	Bias, calib. BL	Median $\lvert e \rvert$, calib. BL	Δ median (t)	Δ CRPS (t)	Δ NLPD (t)	Shortfall, calib. BL	Δ calib. vs std. (t)	Δ f15 vs calib. (t)	Coverage f15 / calib.
Q1	6,672	+2.19	+0.53	1.296	-0.074 (-1.6)	-0.041 (-2.8)	-0.094 (-2.5)	-0.42	+1.035 (+5.3)	+0.043 (+0.4)	43% / 29%
Q2	6,958	+1.83	+0.33	1.179	-0.073 (-0.8)	-0.062 (-2.9)	-0.172 (-10.1)	+0.12	+1.084 (+4.1)	+0.070 (+0.2)	43% / 43%
Q3	7,450	+1.43	+0.23	0.972	-0.016 (-0.1)	-0.059 (-2.0)	-0.163 (-5.2)	+0.52	+0.972 (+4.0)	-0.066 (-0.2)	57% / 71%
Q4	7,250	+1.01	+0.15	0.680	+0.011 (+0.1)	-0.073 (-3.7)	-0.157 (-4.9)	+0.79	+0.816 (+5.2)	-0.158 (-0.6)	100% / 86%

Coverage is the share of the seven quarters in which the realized yield reached the target.

The experts' bias falls from 2.2 points at the longest horizon to 1.0 at the shortest, and the seasonal calibration removes most of it at every horizon, leaving 0.15–0.5 points; the calibration gain at the decision layer is 0.82–1.09 points in every quarter type (t from 4.0 to 5.3 on seven quarters, one per fiscal year). The residual bias is largest at the longest horizon and it is there that the calibrated benchmark misses its target (-0.42 points, coverage 29%), while at the shortest horizon it exceeds it ($+0.79$, coverage 86%): the pooled shortfall of Table 7 averages a miss at long horizons against an excess at short ones, and a practitioner running the problem at one horizon should recalibrate the target to it. The second view's proper-score gains are present at every horizon (ΔNLPD from -0.09 to -0.17, t from -2.5 to -10.0), largest in the second and third fiscal quarters where the two sources have the most to disagree about, while its shortfall difference is within ± 0.16 points at every horizon ($|t| \leq 0.6$). Neither conclusion is an artefact of pooling horizons. A fixed twelve-month target would be the cleaner design and we have not run it.

Second machine view.

Table OA.25 supports the paragraph of Section 7.2 on the Llama-3.1-70B view: descriptive errors and correlations for the 8,857 firm-quarters of 2024–2025 that carry all three views, the regressions of each machine's error on the machine–machine disagreement, and a constant-covariance fusion calibrated on 2024 and evaluated on 2025. The 37 firm-quarters excluded here are those with a Llama error beyond half the share price, by the main sample's rule.

Table OA.25 A second machine view. Llama-3.1-70B-Instruct (4-bit AWQ, same prompts, same decoding) forecasts the same 8,894 firm-quarters of 2024–2025 as the main AI view. Panel A: errors and their correlations with the experts' and with the main AI view's. Panel B: regressions of each machine's error on the machine–machine disagreement. Panel C: predictive means from a constant-covariance fusion of the calibrated views, with idiosyncratic covariance and seasonal calibration estimated on 2024 and τ as in the main specification.

Panel A: forecast errors, 2024Q1–2025Q4 ($n = 8{,}857$ firm-quarters with all three views)

View	Bias (pp)	Median $\lvert e\rvert$	MSE	Share > 0	Corr. with e_Q (asset / quarter mean)	Corr. with e_{M_1} (asset / quarter mean)
Expert consensus Q	+1.74	0.962	14.2	82%	1.00 / 1.00	0.68 / 0.48
M_1: Qwen2.5-72B (main AI view)	+1.54	1.251	17.4	73%	0.68 / 0.48	1.00 / 1.00
M_2: Llama-3.1-70B (second machine view)	+3.16	1.966	37.3	84%	0.68 / 0.66	0.74 / 0.90

Panel B: what the machine–machine disagreement $M_1 - M_2$ identifies (firm-clustered regressions, errors in pp)

$e_{M_1} = +1.53\,(+18.4)\; - 0.002\,(-0.1) \times (M_1 - M_2)$; $\quad e_{M_2} = +1.53\,(+18.4)\; - 1.002\,(-24.9) \times (M_1 - M_2)$; $\quad \mathbb{E}[M_1 - M_2] = -1.63$, midpoint bias +2.35

Panel C: constant-covariance fusion, calibrated on the four 2024 quarters, evaluated on the four 2025 quarters (asset level; no inference)

Views in the predictive mean	Median $\lvert e\rvert$	Bias (pp)	MSE	CRPS (pp)	NLPD	Median weights (Q, M_1, M_2)
Q only (calibrated)	1.170	-0.36	9.65	1.415	-2.053	+0.77
Q + M1 (Qwen)	1.138	-0.34	8.93	1.378	-2.091	+0.62, +0.15
Q + M2 (Llama)	1.154	-0.35	9.69	1.413	-2.051	+0.82, -0.05
Q + M1 + M2	1.056	-0.31	8.56	1.346	-2.112	+0.67, +0.25, -0.14

Errors in percentage points of price; MSE $\times 10^4$. Panel C is descriptive: four training and four test quarters, and the common-factor variance estimated on four quarters is at its floor, so the CRPS and NLPD use the idiosyncratic variance only.

References

An H, Lee S (2026) Simulating analyst forecast behavior using LLMs. Working Paper WP26-14, KDI School of Public Policy and Management, https://doi.org/10.2139/ssrn.6940898

Artrip A, Horton JJ, Kazinnik S, et al (2026) Simulating the Survey of Professional Forecasters. SSRN Working Paper No 5066286 https://doi.org/10.2139/ssrn.5066286

Bates JM, Granger CWJ (1969) The combination of forecasts. Journal of the Operational Research Society 20(4):451–468. https://doi.org/10.1057/jors.1969.103

Chen Y (2025) AI analysts and analysts' earnings forecasts: differences and collaboration (in Chinese). Master's thesis, Shandong University of Finance and Economics, https://doi.org/10.27274/d.cnki.gsdjc.2025.000945

Chhaochharia V, Kumar A, Murali S, et al (2021) Aggregating artificially intelligent earnings forecasts. SSRN Working Paper No 3781108 https://doi.org/10.2139/ssrn.3781108