Inference with AI-Generated Covariates

Junting Duan and Markus Pelger

University of Hong Kong - Stanford University and NBER



The GenAI Feature Challenge

Goal: Estimate an economic model with a covariate extracted from unstructured data

$$Y_i = X_i^\top \gamma_0 + T_i \beta_0 + \varepsilon_i, \qquad Z_i = (Y_i, X_i, S_i), \quad T_i = f(S_i)$$

Z_i: outcome Y_i, covariates X_i, unstructured input S_i (text/image/audio); T_i: latent feature of S_i (later: general moment conditions $\mathbb{E}[g(Z_i, T_i; \theta_0)] = 0$: regression, IV, MLE, GMM)

New practice: Let a large language model (LLM) read S_i and output a proxy $\tilde{T}_i$

- **Finance:** sentiment and expectations from headlines and earnings calls
- **Health:** diagnostic and severity codes from clinical notes
- **Political economy / innovation:** ideology scores from platforms, patent classes

Why attractive: negligible marginal cost, arbitrary scale, no annotator training

Implicit assumption: the generated feature is an error-free measurement of the latent T_i

$\Rightarrow$ **Can we trust regressions that use LLM-generated covariates?**

LLM Outputs Are Biased and Noisy

Generated features differ from the truth in two distinct ways:

$$\tilde{T}_{i,m,s} = T_i + \underbrace{b_m(Z_i)}_{\text{systematic bias}} + \underbrace{\epsilon_{m,s}(Z_i)}_{\text{seed noise}} + \underbrace{u_{i,m,N}}_{\text{spillovers}} \qquad m : \text{model–prompt pair},\ s : \text{seed}$$

Three economic sources of bias $b_m(Z_i)$:

- Hallucination: content not grounded in the input, input-dependent distortions
- Look-ahead: pretraining corpus contains ex-post information about unit i
- Coverage: undertrained domains, periods, languages (omitted-variable analogue)

Properties of bias: input-dependent, not mean zero, correlated with outcomes and covariates, not sign-identified, different across models and prompts

Two types of noise, beyond the bias:

- Seed noise $\epsilon_{m,s}(Z_i)$: run-to-run randomness; mean zero given input, never fully removed
- Spillovers $u_{i,m,N}$: other sampled units leak in via the corpus; negligible at web scale

$\Rightarrow$ Errors distort the moment conditions, not just the features: bias does not average out

Common Practice and Its Problems

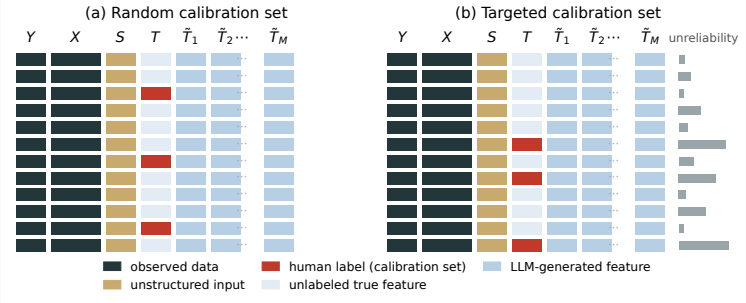
1. **Naive LLM regression:** plug $\tilde{T}_i$ into the moments for all N units
   - Biased, inconsistent, invalid inference, no matter how large N is
   - Conclusions flip across models and prompts: which run do you report?
2. **Calibration-only regression:** hand-label $n \ll N$ units, discard the rest
   - Valid: unbiased under the labeling design
   - Inefficient: wide confidence intervals from a small sample
3. **Diagnostic validation (most common):** report agreement rates on a labeled subset, then run the naive regression anyway
   - Does not change the estimator: the labels never enter the estimation
   - High agreement $\neq$ small bias in the moments: a false sense of security

$\Rightarrow$ **How to combine cheap generated features with few costly human labels, with guarantees?**

This Paper: AI-PI (AI-Powered Inference)

A method-of-moments framework with three components:

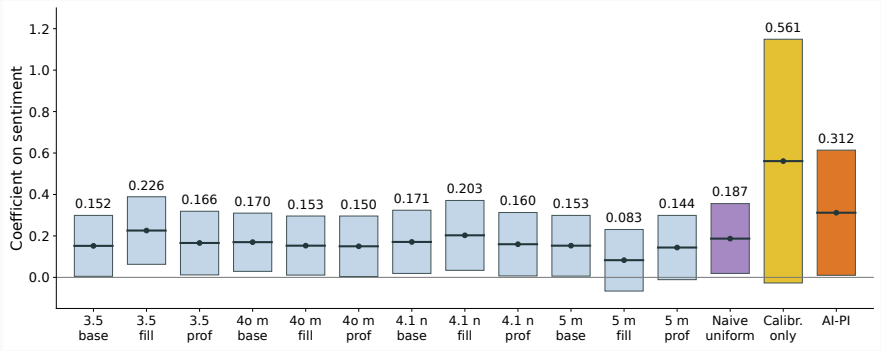
1. Moment-specific bias correction: small human-labeled calibration set removes arbitrary biases
2. Optimal ensembling: adaptive weights over model–prompt pairs downweight unreliable generators
3. Optimal calibration design: choose which units to label to maximize efficiency under a budget



$\Rightarrow$ The three choices interact: AI-PI solves for weights and labeling design jointly

Preview: What AI-PI Delivers on Real Data

News sentiment and next-month stock returns, 107,495 stock-months, 12 model–prompt pairs:



⇒ Naive estimates range 0.083 to 0.226; significance not uniform across GPT model–prompt choices

⇒ Human labels alone (5% of units): valid but the CI is wide (length 1.18)

⇒ AI-PI: 0.312, CI length 0.60, half of calibration-only, anchored to human labels

Key Contributions

1. **Methodology:** debiased moment estimator for LLM-generated covariates, with adaptive model–prompt weights and optimal calibration set design, delivering valid and efficient inference
2. **Statistical theory:**
   (a) Consistency and asymptotic normality under non-random, data-adaptive labeling
   (b) Cross-fitting guarantees when the LLM pipeline is tuned on calibration labels
   (c) Feasible two-stage procedure: asymptotically normal and attains the oracle asymptotic variance of the jointly optimal weights and labeling policy
   (d) Extension: noisy reference labels (model-assisted calibration)
3. **Practice:**
   (a) Simulations: naive inference has 0% coverage; AI-PI is valid, CI length up to -62% vs. calibration-only
   (b) Empirics: 12 fragile sentiment estimates become one valid, stable estimate with half the calibration-only CI, at a 5% labeling budget

$\Rightarrow$ **Validity requires a calibration set; efficiency requires weighting and design**

Related Literature (Incomplete List)

Missing covariates

- Imputation-Powered Inference: Duan and Pelger (2025)

⇒ New here: generated (not missing) covariates, heterogeneous errors, chosen labels

Missing labels and validation-sample frameworks

- Prediction-Powered Inference: Angelopoulos et al. (2023), PPI++, Zrnic–Candès (2024)
- MAR-S: Carlson–Dell (2025): semiparametric missing-data unification, efficient influence-function estimators ▸ Details

⇒ We treat covariates in general moments, weight many generators, design the labels

AI/ML-generated variables

- Ludwig–Mullainathan–Rambachan (2025): fragility; Egami et al. (2024): LLM annotations
- Battaglia et al. (2025): corrections from a model of the error structure ▸ Details
- GPI: Imai–Nakamura (2025): open-model representations deconfound text ▸ Details

⇒ We correct with few human labels under arbitrary errors; closed models fine

Classical roots: survey sampling and AIPW (Neyman 1934; Robins et al. 1994), measurement error (Fuller 2009), DML (Chernozhukov et al. 2018)

An Illustrative Example

Example: A Simple Regression with One Generated Feature and One Generator

Setup: N i.i.d. units (Y_i, X_i, S_i); a human expert could extract $T_i = f(S_i) \in \mathbb{R}$ from the text

$$Y_i = X_i^\top \gamma_0 + T_i \beta_0 + \varepsilon_i, \qquad \mathbb{E}\left[X_i^* \left(X_i^{*\top} \theta_0 - Y_i\right)\right] = 0, \quad X_i^* = (X_i^\top, T_i)^\top$$

Human labels are expensive: an LLM generates $\tilde{T}_i = T_i + b(S_i) + \epsilon_i$ for all N units

- **Naive LLM regression:** replace T_i by $\tilde{T}_i$ everywhere

  Ignores $b(S_i) + \epsilon_i$: biased and inconsistent, no matter how large N is

- **Calibration-only regression:** label $n \ll N$ units (set S_0), discard the generated features

  Unbiased but inefficient: throws away $N - n$ observations

$\Rightarrow$ **Key idea:** Use labeled calibration units (set S_0) to estimate and correct the bias in the moment

$$\text{Bias} \approx \frac{1}{n} \sum_{i \in S_0} \left(\tilde{X}_i (\tilde{X}_i^\top \theta - Y_i) - X_i^* (X_i^{*\top} \theta - Y_i) \right).$$

Example: The Weighted Debiased Estimator

Estimate the moment bias on the calibration set and subtract it; weight the generated part by λ:

$$G_N^\lambda(\theta) = \underbrace{\frac{1}{n}\sum_{i\in S_0} X_i^*(X_i^{*\top}\theta - Y_i)}_{\text{human moment}} + \lambda \cdot \frac{N-n}{N} \underbrace{\left[\frac{1}{N-n}\sum_{i\notin S_0} \tilde{X}_i(\tilde{X}_i^\top\theta - Y_i) - \frac{1}{n}\sum_{i\in S_0} \tilde{X}_i(\tilde{X}_i^\top\theta - Y_i)\right]}_{\text{debiased generated moment: mean zero}}$$

- $\lambda = 0$**:** calibration-only estimator (discard the LLM entirely)
- $\lambda = 1$ **with a perfect LLM (**$\tilde{T}_i = T_i$**):** oracle estimator with all N labels
- The debiasing term has mean zero for every λ: validity is free, λ only governs efficiency

$\Rightarrow$ Optimal λ^* balances the information in $N - n$ extra observations against their noise and bias

$\Rightarrow$ Labels need not be assigned uniformly: targeted labeling is allowed and strongly improves efficiency

Model and Estimation

Problem Setup

Data: N i.i.d. units $Z_i = (Y_i, X_i, S_i)$: outcome, covariates, unstructured input

Goal: Estimate $\theta_0 \in \mathbb{R}^d$ defined by moment conditions with a latent feature vector $T_i = f(S_i)$:

$$\mathbb{E}\big[g(Z_i, T_i; \theta_0)\big] = 0$$

- Covers regression, IV, MLE, GMM with possibly multivariate features
- Structural estimation has no out-of-sample check for $\hat{\theta}$: we need formal guarantees

Generated features: M model–prompt pairs, seed-averaged

$$\tilde{T}_{i,m} = T_i + b_m(Z_i) + \bar{\epsilon}_m(Z_i) + u_{i,m,N}, \qquad m = 1, \ldots, M$$

Calibration set: labels $D_i \in \{0, 1\}$, budget $\mathbb{E}[\pi(W_i)] = n/N =: r_N$ (labeling rate), propensity

$$\mathbb{P}(D_i = 1 \mid W_i) = \pi(W_i) > 0, \qquad W_i : \text{labeling information (e.g., model disagreement)}$$

$\Rightarrow$ Non-random, targeted labeling allowed: the design is chosen and known, not estimated

Weighted Debiased Moment Estimator

AI-PI moment function: for weights $\lambda = (\lambda_1, \ldots, \lambda_M)$,

$$G_{\lambda,\pi}(\theta) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\pi_i}\, g(Z_i, T_i; \theta)}_{\text{IPW human moment}} + \sum_{m=1}^{M} \lambda_m \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} g(Z_i, \tilde{T}_{i,m}; \theta) - \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\pi_i}\, g(Z_i, \tilde{T}_{i,m}; \theta)\right]}_{\text{debiased GenAI moment: } \mathbb{E}[\,\cdot\,]=0}$$

Estimator: $\hat{\theta}_{\lambda,\pi}$ solves $G_{\lambda,\pi}(\theta) = 0$

- **Debiasing:** the calibration set estimates the moment-level discrepancy $\mathbb{E}[g(Z, \tilde{T})] - \mathbb{E}[g(Z, T)]$; the corrected moment is exactly unbiased for any λ and any design π
- **Robustness:** no model of the LLM error: arbitrary, non-classical, input-dependent errors; nonlinear moments handled because the correction acts on moments, not features
- **Re-weighting:** λ_m governs how much to trust each generator
  - $\lambda = 0$: calibration-only estimator (discard the LLM output)
  - $\lambda_m = 1$, perfect generator: oracle estimator with all N labels

$\Rightarrow$ Validity is free for every λ; the weights only govern efficiency: never worse than calibration-only

Inferential Theory and Optimal Design

Asymptotic Normality

Theorem (Asymptotic Normality) ▸ Assumptions

Under overlap (no unit effectively excluded from labeling), standard moment regularity, and negligible spillovers: for fixed λ, as $n, N \to \infty$ with labeling rate $r_N = n/N \to r$,

$$\sqrt{n}\left(\hat{\theta}_{\lambda,\pi} - \theta_0\right) \xrightarrow{d} \mathcal{N}\Big(0,\ J_{\theta_0}^{-1}\,\Omega_{\lambda,\pi}\,J_{\theta_0}^{-\top}\Big), \qquad \Omega_{\lambda,\pi} = \underbrace{r\mathrm{Var}\big(g(Z_i, T_i; \theta_0)\big)}_{\text{labeled sample}} + \underbrace{\mathbb{E}\big[\big(\tfrac{1}{\rho(W_i)} - r\big) R_\lambda(W_i)\big]}_{\text{design-dependent}}$$

with $\rho = \pi/r_N$ and $R_\lambda(W_i) = \mathbb{E}\big[e_{i,\lambda} e_{i,\lambda}^\top \mid W_i\big]$, $e_{i,\lambda} = g(Z_i, T_i; \theta_0) - \sum_m \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)$.

- **Generality:** standard GMM-type regularity plus design conditions; no assumption on the error b_m
- R_λ: discrepancy between weighted GenAI moment and true moment; perfect ensemble: $R_\lambda = 0$, oracle variance
- Valid for vanishing labeling fractions $r = 0$; plug-in variance estimator gives feasible CIs

$\Rightarrow$ Two levers to cut variance: choose λ to shrink R_λ, choose π to label where R_λ is large

Optimal Labeling Policy

For given weight λ, **choose $\pi(\cdot)$ to minimize total asymptotic variance s.t. budget $\mathbb{E}[\pi(W_i)] = r_N$:**

Theorem (Optimal labeling propensity, non-binding case)

Under the assumptions of the asymptotic-normality theorem and mild design assumptions:

$$\pi^*(w \mid \lambda) = c_\lambda \, s_\lambda(w) \;\propto\; s_\lambda(w), \qquad s_\lambda(w) = \operatorname{tr}\big(J_{\theta_0}^{-1} R_\lambda(w) J_{\theta_0}^{-\top}\big)^{1/2},$$

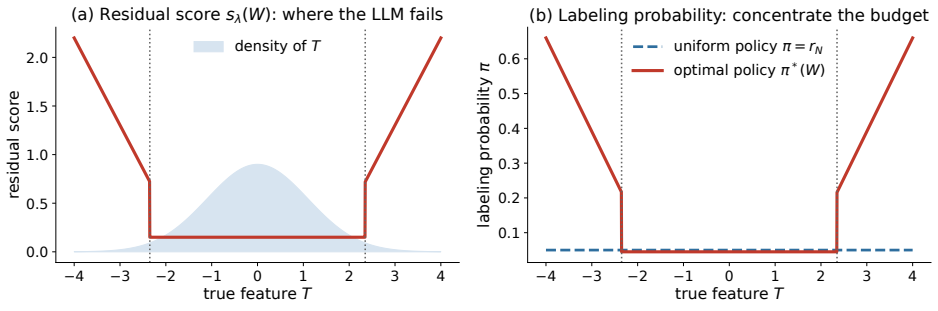
with c_λ set by the budget. (Corner cases: truncated below at $\rho_{\min} r_N$ and above at 1.)

- **The result:** the labeling probability is proportional to the residual score $s_\lambda(w)$: label where the generators deviate most, since a human label is most informative there
- **Corner cases:** lower truncation preserves overlap ($\pi > 0$); upper is mechanical ($\pi \leq 1$)

$\Rightarrow$ The score is moment-specific: it weighs generator errors by J_{θ_0} and R_λ, not by feature error alone

What Should the Labeling Design Condition On?

- W_i = everything observable before labeling: covariates, generated features, disagreement
- The design may use the data only through W_i: that is what makes labeling MAR by construction
- The score is unobservable ex ante, but models disagree exactly where they are unreliable: cross-model disagreement is the feasible proxy of the score



$\Rightarrow$ Uniform labeling wastes budget where the generator is already right; π^* concentrates labels on the unreliable region, keeping $\pi > 0$ everywhere

Optimal Weights and Joint Design

Optimal weights for given design π: convex quadratic program over $\Lambda = \{\lambda \geq 0, \sum_m \lambda_m \leq 1\}$

$$\min_{\lambda \in \Lambda} \mathbb{E}\left[\left(\frac{r_N}{\pi(W_i)} - r\right)\left(g_i - \tilde{g}_i\lambda\right)^\top J_{\theta_0}^{-\top} J_{\theta_0}^{-1}\left(g_i - \tilde{g}_i\lambda\right)\right]$$

- Downweights noisy generated moments; exploits complementary error patterns across pairs
- The inequality constraint can shrink the entire GenAI contribution:
  if all generators are poor, $\lambda \to 0$ and AI-PI falls back to the calibration-only estimator
- Joint optimum (λ^*, π^*): alternate policy and weight updates (first-order condition at the fixed point); with W-dependent weights the joint problem separates

Feasible two-stage algorithm:

1. **Pilot:** label a random sample n_0; estimate θ_0, J, residual scores, weights and labeling policy
2. **Adaptive stage:** spend the remaining budget $n - n_0$ according to the estimated policy $\hat{\pi}$ (kept > 0 everywhere); solve $G_{\hat{\lambda},\hat{\pi}}(\theta) = 0$

Inference with Feasible Estimated Design

Theorem (Estimated design: validity + oracle asymptotic variance)

Under the assumptions of the asymptotic-normality theorem at optimal design: estimate $(\hat{\lambda}, \hat{\pi})$ on a pilot that grows but keeps an asymptotically negligible share of the labels ($n_0 \to \infty$, $n_0/n \to 0$). Then

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{TS}} - \theta_0\right) \xrightarrow{d} \mathcal{N}\Big(0,\ J_{\theta_0}^{-1}\,\Omega_{\lambda^*,\pi^*}\,J_{\theta_0}^{-\top}\Big) :$$

$\Rightarrow$ asymptotically normal and attains the variance of the infeasible jointly optimal design.

Why this works:

- **Plug-in optimality holds:** $V(\hat{\lambda}, \hat{\pi}) = V(\lambda^*, \pi^*) + o_p(1)$
- **Nothing can break:** the moment is exactly centered for every design: estimating the design can move the variance, never the bias
- **Pilot size:** pooling pilot labels is harmless if the pilot grows more slowly than $\sqrt{n}$

$\Rightarrow$ Confidence intervals from the plug-in variance estimator remain valid for the two-stage procedure

$\Rightarrow$ Adaptivity is free to first order: inference is as if the optimal design had been known all along

Extensions

Noisy reference labels: why it matters ▸ Details

- Human annotators make mistakes; a strong model could serve as the reference on a subsample
- Either way, the calibration label is $T_i^* = T_i + e_i$ rather than the truth T_i

$\Rightarrow$ All results hold for the new reference estimand θ_0^* (solution of $\mathbb{E}[g(Z_i, T_i^*; \theta)] = 0$).

With unbiased label noise (centered given the data):

- Question: How close is θ_0^* to θ_0?
- Linear moments: unbiased $\theta_0^* = \theta_0$, but less efficient
- Non-linear moments: second order attenuation bias: variance $\times$ moment curvature
- Least squares: known attenuation bias, correctable if the noise variance is estimable

Consequence:

- Model-assisted calibration: a strong model debiases cheap models at scale
- Human labels stay the anchor: shared biases (common look-ahead) are the uncorrectable case

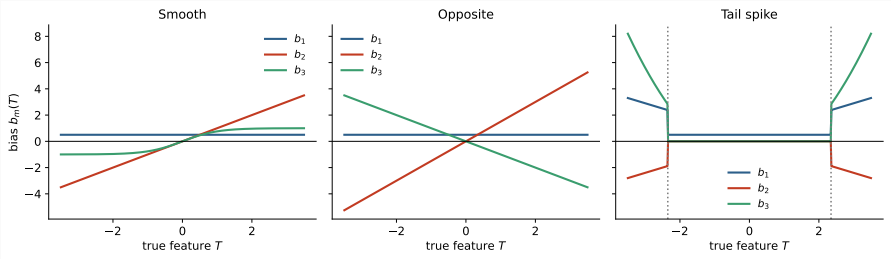
Also in the paper:

- Overidentified GMM: efficient weighting $\hat{A} = \hat{\Omega}_{\lambda,\pi}^{-1}$; the design theory carries over
- Clustered data: cluster = sampling unit; clustered standard errors are automatic

Simulations

Simulation Design

Linear moment, $M = 3$ generators, three bias configurations ($N = 5{,}000$, label budget $r_N = 5\%$):

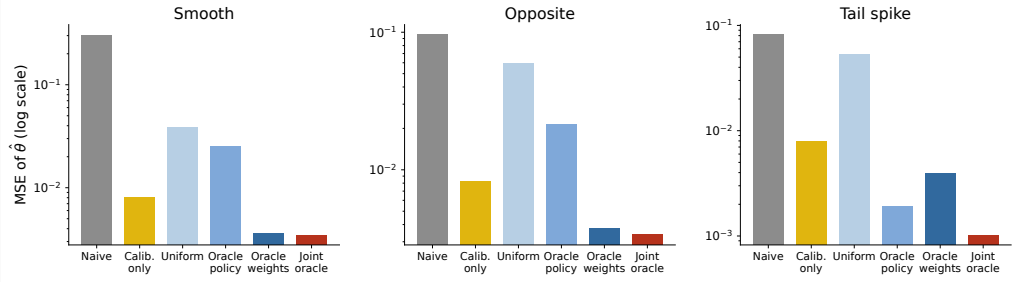


- **Smooth:** all generators informative but distorted over the whole feature space
- **Opposite:** systematic errors in offsetting directions (weights can exploit cancellation)
- **Tail spike:** accurate for typical inputs, hallucinates in the tails (design matters most)

Labeling information W_i: model disagreement $\mathrm{sd}_m(\tilde{T}_{i,m})$

Simulation: Estimator Comparison

Linear moment, $M = 3$ generators; smooth / offsetting / tail-spike bias designs ($N = 5{,}000$, budget $r_N = 5\%$, $W_i =$ model disagreement):



⇒ Naive LLM regression: largest MSE driven by bias

⇒ Debiasing with uniform weights is valid but worse than discarding the LLM data

⇒ Weights drive gains in smooth/offsetting designs; labeling policy drives them in tail-spike design

⇒ The joint design dominates everywhere: the two elements are complements

Simulation: Valid Inference

Coverage of nominal 95% CIs and average CI length ($R = 1{,}000$ replications; W_i = disagreement):

	Smooth		Opposite		Tail spike	
Estimator	Cover	Length	Cover	Length	Cover	Length
Naive LLM regression	0.0	3.6	0.0	4.8	1.6	7.4
Calibration-only	94.3	24.8	94.5	24.7	95.0	24.6
AI-PI: uniform (λ, π)	92.5	70.4	93.0	91.1	82.5	78.0
AI-PI: joint oracle	95.5	17.4	94.7	17.1	95.9	9.4
AI-PI: feasible two-stage	94.8	17.1	95.6	17.2	87.6	22.2

$\Rightarrow$ Naive inference is entirely misleading: bias dwarfs the tiny standard errors

$\Rightarrow$ AI-PI: nominal coverage, CI length -30% / -31% / -62% vs. calibration-only

$\Rightarrow$ Feasible design matches the oracle except tail spike with a small pilot ($n_0 = 125$): noisy tail score, consistent with the $n_0 \to \infty$ requirement

Empirical Study: News Sentiment and Stock Returns

Empirical Design: A Standard Cross-Sectional Return Regression

Question: Does the sentiment of firm-specific news predict next-month stock returns?

$$R_{i,t+1} = a + b\, T_{i,t} + X'_{i,t}\Gamma + \mu_{j(i)} + \lambda_{t+1} + u_{i,t+1}$$

- $T_{i,t}$: latent human sentiment of stock i's headlines in month t; LLM proxies $\tilde{T}^{(m)}_{i,t}$
- Controls: log size, book-to-market, momentum, same-month return (reversal); industry + month FE; month-clustered SEs
- Sign is an empirical question: drift ($b > 0$, Chan 2003) vs. reversal ($b < 0$, Tetlock 2007)

Data:

- 259,970 headlines: 107,495 stock-months, CRSP-matched, 2009–2020
  (stratified representative sample from 2.2M Kaggle headlines)
- Same headline data as Ludwig–Mullainathan–Rambachan, who document conclusion flips

$\Rightarrow$ Goal: not a definitive premium estimate, but to show that reasonable model/prompt choices change conclusions, and AI-PI resolves this

The Measurement Problem Is Real



12 model–prompt pairs: {GPT-3.5 Turbo, GPT-4o mini, GPT-4.1 nano, GPT-5 mini} $\times$ baseline / fill-in / professional-investor prompts

Headline-level disagreement:

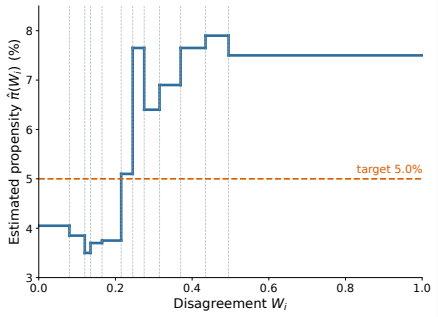
- Labels: positive, neutral, negative
- Average pairwise agreement 83%; lowest pair 76%
- Positive share ranges 26.5% to 42.0% across pairs
- Same model, different prompt: 88% agreement: prompts matter

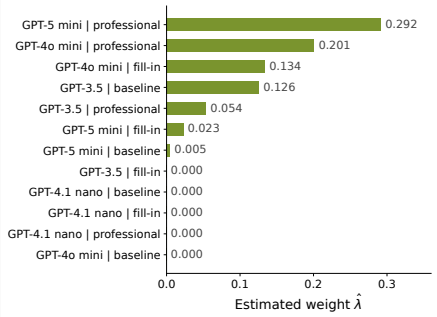
$\Rightarrow$ The level of measured sentiment is model- and prompt-dependent before any regression

$\Rightarrow$ Exactly the setting AI-PI is built for: one valid, stable estimate instead of twelve

Implementing the Two-Stage Design on Real Data

Feasible AI-PI with a 5% labeling budget (pilot: 40% of budget, uniform; W_i = cross-pair disagreement): final calibration set 5,320 stock-months (4.95%), 11,889 headline-level human labels





⇒ Propensities 0.035 to 0.079: targeted, overlap kept; the policy tracks the residual score

⇒ Largest weights: GPT-5 mini and GPT-4o mini, professional prompt; several zero; sum 0.836

Main Result: One Valid, More Precise Estimate

Estimator	Coefficient	SE	95% CI	CI length
12 naive model–prompt estimates	0.083 – 0.226	0.072–0.086	11 of 12 significant at 10%	
Naive uniform ensemble	0.187**	0.086	[0.019, 0.356]	0.337
Calibration-only IPW	0.561*	0.300	[−0.027, 1.149]	1.176
AI-PI	**0.312****	**0.154**	**[0.010, 0.614]**	**0.604**

- Fragility: naive estimates span 0.083–0.226, a factor of almost three
- Validity has a price: using human labels alone is unbiased but the CI is wide (length 1.18)
- AI-PI: anchored to human labels, CI half the calibration-only length

Economic magnitude:

- Uniformly positive vs. neutral news month: +31 bps next-month return ($\approx$ 3.7 pp annualized)
- Modest by design: prices absorb most news within the month (contemporaneous response $\approx$ 3.5, an order of magnitude larger): b is the residual drift
- Attenuation matters economically: the naive scale understates the premium by $\approx$ 40%

Why the Correction Matters: Anatomy of the Measurement Error

Model errors on the 11,889 human-labeled headlines:

	Accuracy	Bias	RMSE
Best pair (GPT-5 mini, professional investor)	86.0%	0.016	0.390
Worst pair (GPT-4.1 nano, fill-in)	75.1%	−0.087	0.516
Range across 12 pairs	75–86%	−0.09–0.08	0.39–0.52

- Mean biases are small, headline-level noise is large: the dominant effect is attenuation, so all 12 naive estimates lie below the human-anchored estimates
- Bias sign flips across pairs: there is no single "safe" model choice
- **The data-driven weights find the reliable pairs:** largest $\hat{\lambda}$ on two of the three most accurate pairs, without using these error metrics as input

$\Rightarrow$ Accuracy of 75–86% sounds reassuring, yet the downstream coefficient moves by a factor of 2.7: diagnostic validation provides a false sense of security

Conclusion

Conclusion

AI-PI: valid and efficient inference with LLM-generated covariates

- Moment-specific bias correction from a small calibration set: validity for arbitrary error structures
- Adaptive weighting of many model–prompt pairs: efficiency
- Optimal labeling design: the budget goes where the generators fail
- Theory: asymptotic normality with adaptive designs, cross-fitting for tuned pipelines, estimated-design oracle variance, GMM, noisy references, clustering

Messages for empirical practice

- Diagnostic validation is not bias correction: convert the validation sample into the estimator
- Debiasing alone is not enough: noise remains, weighting is essential
- In our application, a 5% labeling budget turns 12 fragile estimates into one valid estimate with half the calibration-only CI

$\Rightarrow$ **An automatic, easy-to-use pipeline for credible GenAI-based empirical research**

Appendix

Assumptions

▸ Back ▸ Formal statements

1. **Labeling:** conditional on the full sample and generation randomness, labels are independent Bernoulli with $\mathbb{P}(D_i = 1) = \pi(W_i)$; overlap $\pi(W_i) \geq \rho_{\min} r_N$
2. **Moment regularity:** finite second (and $2 + \delta$) moments at true and generated features; $G(\theta) = \mathbb{E}[g(Z_i, T_i; \theta)]$ continuously differentiable near θ_0 with positive definite Jacobian; Lipschitz in θ and in T with square-integrable envelopes
3. **Spillovers:** $\max_i \|u_{i,m,N}\| = o_p(n^{-1/2})$: any sampled unit is a negligible fraction of the pretraining corpus (primitive conditions: corpus size $\gg$ sample size)
4. **Pipeline stability (cross-fitting):** tuned features converge to a limit pipeline in weighted L_2, design independent of tuned features, cross-Lipschitz condition
5. **Design feasibility:** the weighted generated moment is imperfect on more than an r_N-fraction of the population

Consistency holds under compactness (global Lipschitz) or convexity plus monotone sample moments; vanishing labeling rates $r \to 0$ are allowed.

Assumptions: How Strong Are They?

- **Labeling (A1) is MAR by construction:** the researcher runs the design, so the condition is enforceable by protocol; propensities are known, no estimation step
- **Regularity (A2) is standard:** satisfied by least squares, IV, MLE, smooth GMM; the only nonstandard piece is Lipschitz in T, which converts feature perturbations into moment perturbations
- **Spillovers (A3) is a scale-separation condition:** primitive lemma: bounded influence needs corpus $C_N \gg N\sqrt{n}$; with random signs only $C_N \gg \sqrt{Nn \log N}$: web-scale corpora satisfy both by orders of magnitude
- **What is not assumed:** any structure on the generator error b_m: it may be input-dependent, correlated with outcomes, and sign-unidentified

$\Rightarrow$ Relative to standard GMM, the "covariate observed" assumption is replaced by a labeling design the researcher controls: the identification burden moves from an error model to the design

Formal Assumptions (1/2): Labeling and Moment Regularity

Assumption 1 (Human labeling).

1. Let $\mathcal{Z}_N$ be the σ-field of the full sample $\{(Z_j, T_j, W_j)\}_{j \le N}$ and all generation randomness (seed noise, pretraining corpus). Conditional on $\mathcal{Z}_N$, the labels $(D_1, \ldots, D_N)$ are mutually independent with $\mathbb{P}(D_i = 1 \mid \mathcal{Z}_N) = \pi(W_i)$
2. Overlap: the scaled propensity $\rho(W) = \pi(W)/r_N$ satisfies $\rho(W_i) \ge \rho_{\min} > 0$ almost surely

Assumption 2 (Moment regularity and identification).

1. Finite second moments: $\mathbb{E}\,\|g(Z_i, T_i; \theta)\|^2 < \infty$ and $\mathbb{E}\|g(Z_i, \tilde{T}^0_{i,m}; \theta)\|^2 < \infty$ for each m
2. $G(\theta) = \mathbb{E}[g(Z_i, T_i; \theta)]$ continuously differentiable near θ_0 with positive definite Jacobian J_{θ_0}
3. Lipschitz in θ: envelopes $L_{\theta,H}(Z_i)$, $L_{\theta,m}(Z_i)$ with finite second moments
4. Lipschitz in T: $\|g(Z_i, t_1; \theta) - g(Z_i, t_2; \theta)\| \le L_T(Z_i)\,\|t_1 - t_2\|$, $\mathbb{E}[L_T(Z_i)^2] < \infty$

Formal Assumptions (2/2): Spillovers and Moment Conditions

Assumption 3 (Negligible spillovers). For each m: $\max_{1 \le i \le N} \|u_{i,m,N}\| = o_p(n^{-1/2})$

Lemma (primitive sufficient conditions). Assumption 3 holds under either:

1. Bounded influence: $\|u_{i,m,N}\| \le (N-1)\bar{\delta}_N$ with $\bar{\delta}_N = o(n^{-1/2}N^{-1})$; with corpus size C_N and per-document influence c/C_N, this requires $C_N \gg N\sqrt{n}$
2. Sub-Gaussian aggregation with parameter σ_N and $\sigma_N\sqrt{\log N} = o(n^{-1/2})$; with random signs, $C_N \gg \sqrt{Nn \log N}$ suffices

Web-scale corpora satisfy both by orders of magnitude

Additional conditions for the central limit theorem:

- $(2+\delta)$ moments at θ_0: $\mathbb{E}\,\|g(Z_i, T_i; \theta_0)\|^{2+\delta} < \infty$ and $\mathbb{E}\|g(Z_i, \tilde{T}^0_{i,m}; \theta_0)\|^{2+\delta} < \infty$
- $\Omega_{\lambda,\pi}$ positive definite; $\mathbb{E}[L_T(Z_i)^2\, U_N^2] = o(1)$ with $U_N = \max_m \max_i \|u_{i,m,N}\|$

Formal Theorem: Asymptotic Normality

Theorem 2. Let $r_N = n/N$ and $\rho(W) = \pi(W)/r_N$, with $\rho(\cdot)$ not varying with N. Fix $\lambda = (\lambda_1, \ldots, \lambda_M)$. Suppose Assumptions 1–3 hold, the consistency conditions hold (compact Θ with global Lipschitz conditions, or convex Θ with asymptotically monotone sample moments), $n \to \infty$ and $r_N \to r \in [0, 1]$, the $(2 + \delta)$-moment conditions hold, and $\Omega_{\lambda,\pi}$ is positive definite. Then

$$\sqrt{n}\,(\hat{\theta}_{\lambda,\pi} - \theta_0) \xrightarrow{d} \mathcal{N}\Big(0,\ J_{\theta_0}^{-1}\,\Omega_{\lambda,\pi}\,J_{\theta_0}^{-\top}\Big), \qquad \Omega_{\lambda,\pi} = r\,\mathrm{Var}\big(g(Z_i, T_i; \theta_0)\big) + \mathbb{E}\Big[\Big(\tfrac{1}{\rho(W_i)} - r\Big) R_\lambda(W_i)\Big],$$

where $R_\lambda(W_i) = \mathbb{E}\big[e_{i,\lambda} e_{i,\lambda}^\top \mid W_i\big]$ and $e_{i,\lambda} = g(Z_i, T_i; \theta_0) - \sum_m \lambda_m\, g(Z_i, \tilde{T}_{i,m}; \theta_0)$

- The restriction to fixed $\rho(\cdot)$ is notational: the proof applies to sequences $\rho_N \to \rho$ with $\rho_N \geq \rho_{\min}$, which covers the truncated optimal designs
- Variance estimation: plug-in $\hat{\Omega}_{\lambda,\pi}$ and $\hat{J}$ are consistent under $(2 + \delta)$-moment envelope conditions, so the confidence intervals are asymptotically exact

Formal Theorem: Estimated Design and Pooling

Theorem 5 (split two-stage estimator). Suppose the conditions of Theorem 2 hold at $(\lambda^*, \pi^*(\cdot \mid \lambda^*))$ with $r_N \to r \in [0, 1)$, and:

(i) $n_0 \to \infty$ and $n_0/n \to 0$

(ii) $\hat{\lambda} \xrightarrow{p} \lambda^*$, with Λ compact

(iii) $\mathbb{E}_W\left[(\hat{\rho}(W) - \rho^*(W))^2\right] \xrightarrow{p} 0$, and a square-integrable envelope $\bar{q}(W)$ dominates $\sup_{\lambda \in \Lambda} \operatorname{tr} R_\lambda(W)$ and the conditional second moments of the human and generated moments

Then $\sqrt{n}\left(\hat{\theta}^{\mathrm{TS}} - \theta_0\right) \xrightarrow{d} \mathcal{N}\big(0,\ J_{\theta_0}^{-1}\, \Omega_{\lambda^*, \pi^*}\, J_{\theta_0}^{-\top}\big)$: the oracle variance is attained

Proposition (pooled estimator). If in addition

(iv) $n_0 = o(\sqrt{n})$ (the pilot grows more slowly than $\sqrt{n}$), and

(v) $\hat{\rho}$ and ρ^* are bounded above by a constant $\bar{\rho} < \infty$,

then pooling the pilot labels into the final moment yields the same limit distribution

$\Rightarrow$ Plug-in optimality also holds: $V(\hat{\lambda}, \hat{\pi}) = V(\lambda^*, \pi^*) + o_p(1)$ under uniform consistency of the estimated variance criterion and propensities

Cross-Fitting for LLM Pipeline Tuning

Problem: prompt selection, fine-tuning, RAG configuration are often tuned on the calibration labels

- Features on labeled units are then no longer distributed as on unlabeled units: debiasing breaks

Solution: K-fold cross-fitting

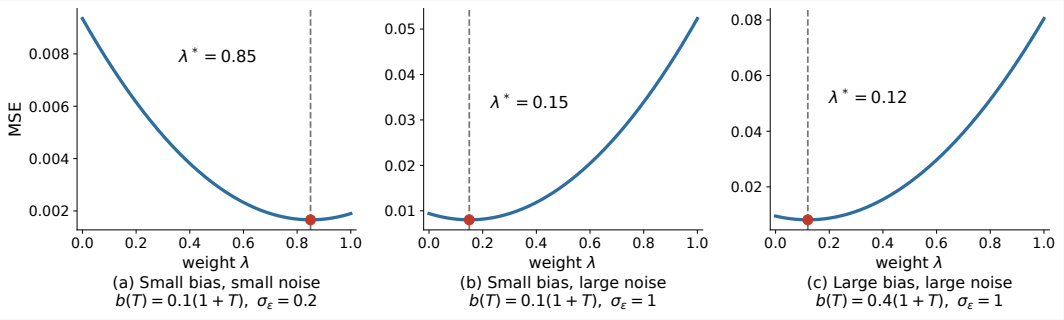
- Partition units into K folds, independently of the data
- Tune the pipeline on the labels outside fold k; generate $\tilde{T}_{i,m}^{(-k)}$ for units in fold k
- Aggregate fold-specific debiased moments: $G_{\lambda,\pi}^{\mathrm{CF}}(\theta) = \frac{1}{K}\sum_k G_{\lambda,\pi}^{(k)}(\theta)$

Guarantee (Theorem 3): under a pipeline-stability condition (tuned features converge to a limit pipeline in weighted L_2, no rate required), the cross-fitted estimator is consistent and asymptotically normal at the limit-pipeline variance

$\Rightarrow$ Debiasing terms are exactly centered conditional on the tuning data, for any realization of the tuned pipeline: tuning affects variance, not validity

Example: How Much to Trust the Generator

Simulated MSE of $\hat{\theta}_\lambda$ as a function of λ ($Y_i = 0.5 + 1.5\,T_i + \varepsilon_i$, $N = 3{,}000$, $n = 200$):



(a) Small bias, small noise
$b(T) = 0.1(1+T),\ \sigma_\varepsilon = 0.2$

(b) Small bias, large noise
$b(T) = 0.1(1+T),\ \sigma_\varepsilon = 1$

(c) Large bias, large noise
$b(T) = 0.4(1+T),\ \sigma_\varepsilon = 1$

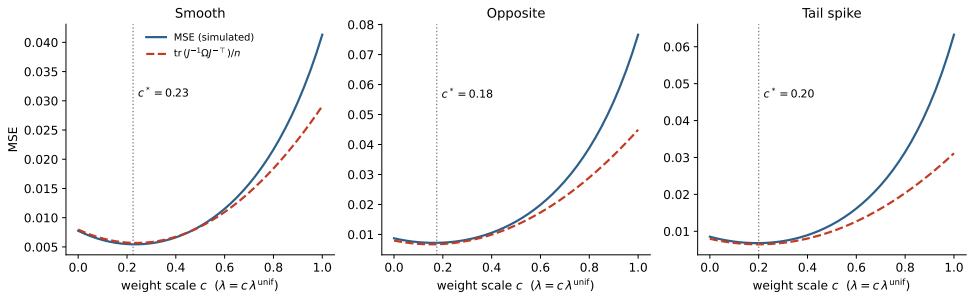
$\Rightarrow$ Debiased estimator is consistent for every λ; efficiency differences are large

$\Rightarrow$ Optimal λ^* adapts to generator quality: near 1 for accurate generators, near 0 for poor ones

$\Rightarrow$ The cost is asymmetric: overweighting a poor generator is worse than underweighting a good one

Simulation: The Weight Scale Matters

MSE along $\lambda = c\,(1/3, 1/3, 1/3)$ and the theoretical asymptotic variance:



⇒ Simulated MSE coincides with the asymptotic variance formula near the optimum

⇒ Optimal scale $c^* \approx 0.2$: substantial shrinkage of the generated moments is essential

⇒ "Debias, then trust the AI" ($\lambda = 1$) is severely suboptimal

Robustness and Measurement Validation

Control sensitivity:

- AI-PI: 0.289 – 0.316 across control sets (drop momentum, drop reversal, size and book-to-market only); naive uniform: 0.162 – 0.193 throughout
- The correction is not driven by a control choice

Contemporaneous returns:

- All 12 generated measures load strongly on same-month returns: 2.99–3.44, all at 1%
- Measures capture real information; the next-month coefficient is the residual drift

Fama–MacBeth:

- AI-PI 0.298 (NW SE 0.151): same conclusion
- Cross-sectional identification confirmed

$\Rightarrow$ Lead message: stability and validity, not a significance star: the CI [0.010, 0.614] just excludes zero

Comparison with Battaglia, Christensen, Hansen, Sacher (2025)

Their result and remedy:

- Naive OLS with an AI/ML-generated regressor: correct asymptotic variance, first-order bias
- Consequence: right CI width, wrong center: nothing looks broken
- Remedy: bias correction or joint MLE from the first-step error structure; no validation sample

Our four differences:

1. Correction at the moment level, identified by gold labels: arbitrary, non-classical errors
2. General method of moments / GMM with multivariate features, not a single regression
3. Many model–prompt pairs combined optimally: resolves the LMR fragility
4. The calibration sample is a design choice: we derive which labels to buy

$\Rightarrow$ The trade-off, honestly: they avoid gold labels but need the error structure; our labels are the price of robustness to unrestricted errors

$\Rightarrow$ "Use the labels alone"? That CI is twice as long: generated features buy first-order efficiency

Comparison with Imai and Nakamura (2025): GenAI-Powered Inference

What GPI does:

- Causal and predictive inference when unstructured data act as treatments or confounders
- Open-source models (re)generate the data, so their internal representations provably capture the content; ML on the representations adjusts for latent confounding
- No fine-tuning, no human labels; three applications and open-source software

How AI-PI differs (complementary problems):

- Different object: they incorporate what the text contains; we correct measurement error in an extracted feature, relative to a human-defined estimand
- Their guarantee concerns deconfounding, not the map from text to the human label: that map is what our calibration set pins down
- GPI needs open-source models; AI-PI needs only outputs: works for closed models
- Combinable: their representations are natural candidates for our labeling information W_i

$\Rightarrow$ Neither nests the other; we renamed our paper and cite GPI to keep the two clearly apart

Comparison with Carlson and Dell (2025): MAR-S

What MAR-S does:

- Unstructured-data inference as missing structured data (Rubin); a fine-tuned network imputes
- Same design assumption as ours: annotation probability known, bounded away from zero
- Efficient influence-function estimators (AIPW, cross-fitting): attain the bound when the single imputation function is consistent; means, OLS, IV, DiD, RDD; EPU and GPR applications

What AI-PI adds (complementary machinery):

- Many black-box generators with variance-optimal weights: model-prompt fragility sits outside a single-imputation framework
- Exact centering for arbitrary moments: no consistency of any generator required
- Joint weight-design optimization, inference under the estimated design (their annotation guidance is the mean-estimation special case)
- Noisy reference labels: excluded by their assumptions, priced in by our theory

$\Rightarrow$ MAR-S: the efficient single-imputation benchmark; AI-PI: the multi-generator estimator with design theory

Where Does AI-PI Sit Relative to the Efficiency Bound?

Question: is AI-PI semiparametrically efficient?

The benchmark: the most efficient estimator predicts the true moment of each unlabeled unit by its best possible prediction $m_i(\theta)$, and uses the labeled units to correct that prediction:

$$G^{\text{eff}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big[\, m_i(\theta) + \frac{D_i}{\pi_i}\big(g(Z_i, T_i; \theta) - m_i(\theta)\big)\Big], \qquad m_i(\theta) = \mathbb{E}\big[g(Z_i, T_i; \theta) \mid \text{all observables of unit } i\big]$$

AI-PI has this structure, with the ideal prediction replaced by the weighted generated moment:

- $m_i(\theta) \to \sum_m \lambda_m\, g(Z_i, \tilde{T}_{i,m}; \theta)$: the best prediction within the span of the generated moments
- Within this class, (λ^*, π^*) is optimal; W-dependent weights weakly beat any constant λ
- The bound is attained when $m_i(\theta)$ lies in that span; the gap is the span's approximation error
- Estimating $m_i(\theta)$ nonparametrically is what a 5% calibration sample cannot support: the restriction is the price of feasibility

$\Rightarrow$ Full efficient-influence-function comparison: future research

Known vs. Estimated Propensities

Our propensities are known by design:

- The researcher runs the labeling mechanism, so $\pi(W_i)$ is recorded, not estimated (unlike observational missing data, where the propensity must be modeled)

The classic twist (Hirano–Imbens–Ridder 2003):

- Using estimated propensities can be more efficient than using the true ones
- Reason: estimation implicitly projects the moment on the observables: an augmentation effect

Why AI-PI does not need it:

- That augmentation is already explicit: the debiased generated moments supply the projection
- With a rich span (or W-dependent weights) the HIR gain is absorbed; the remaining gap is the span approximation error
- In the two-stage design, $\hat{\pi}$ is known conditional on the pilot: oracle variance still attained

$\Rightarrow$ Validity never requires propensity estimation; the efficiency gain that estimated propensities would deliver is already built in through the debiased generated moments

Refinement: Weights That Depend on the Labeling Information

Enlarge the class: replace constants λ_m by measurable functions $\lambda_m(W_i)$ with $\lambda(W_i) \in \Lambda$

- Asymptotic normality carries over, with $R_{\lambda(W_i)}(W_i)$ in the variance
- The pointwise minimizer $\lambda^*(w) \in \arg\min_{\lambda \in \Lambda} q_\lambda(w)$ is weakly more efficient than every constant λ, for every admissible design
- With W-dependent weights the joint design problem separates: first minimize the score pointwise in w, then apply the optimal-policy theorem to the profiled score $\sqrt{q^*(w)}$: no alternation needed

The practical price:

- $\lambda^*(\cdot)$ must be estimated as a function on the support of W, not as M scalars: demanding exactly when the calibration budget is small

$\Rightarrow$ Baseline implementation: constant weights; the profiled-score design is the natural refinement when the pilot sample is large

How Large Should the Pilot Be?

Theory:

- Split two-stage estimator: pilot $n_0 \to \infty$ with $n_0/n \to 0$
- Pooling the pilot labels into the final moment: $n_0 = o(\sqrt{n})$ (e.g., $n_0 \asymp n^{1/3}$)

Practice, disclosed in the paper:

- Empirics pool with a 40% pilot share ($n_0 \approx 2{,}100$ vs. $\sqrt{n} \approx 73$): outside the rate condition
- A variance approximation, not a bias issue: centering is exact for any pilot share

Evidence and tradeoff:

- Simulations pool with a 50% pilot: coverage 94.8% / 95.6% (smooth / opposite)
- Tail spike 87.6%: traced to the tail score estimated from a small pilot, not to pooling
- Small pilot = noisy design estimates; large pilot = little budget left to target

$\Rightarrow$ The split estimator is conservative benchmark; fixed-pilot-share asymptotics are natural extension

Human Labels: Protocol and Quality

Protocol:

- 11,889 headline labels on 5,320 stock-months; same three categories as the LLMs, aggregated identically
- Labeled by one author from the headline text alone, without seeing any model output: no machine anchoring by construction
- Label representative across pilot and targeted stage (e.g., positive 33.6% vs. 34.8%)

Quality:

- Noisy-reference theory prices in imperfect labels: validity for the reference estimand; the gap to the target is second order
- Remaining risks, stated honestly: single annotator; hindsight (2009–2020 headlines labeled today)
- Possible diagnostics: entity-masked relabeling; chronologically trained reference models

$\Rightarrow$ The human labels define the reference standard; the framework prices in their imperfection

Robustness: Alternative Control Sets

Controls (all with industry + month FE)	Naive uniform	Calibration-only	**AI-PI**
Size, B/M, same-month return $R_{i,t}$	0.193**	0.568*	**0.316****
Size, B/M, momentum	0.162**	0.524*	**0.289****
Size, B/M only	0.168**	0.530*	**0.293****
Main spec (all controls)	0.187**	0.561*	**0.312****

$\Rightarrow$ AI-PI moves only within 0.289 – 0.316: not an artifact of the momentum or reversal controls

$\Rightarrow$ The naive-to-AI-PI gap ($\approx$40%) is stable across specifications: consistent with attenuation

$\Rightarrow$ Same-month return $R_{i,t}$ is the main-spec reversal control: standard when forecasting month $t+1$, and sentiment is most correlated with it

Validation: Contemporaneous Regression and Fama–MacBeth

Contemporaneous regression ($R_{i,t}$ on sentiment) and Fama–MacBeth (next-month):

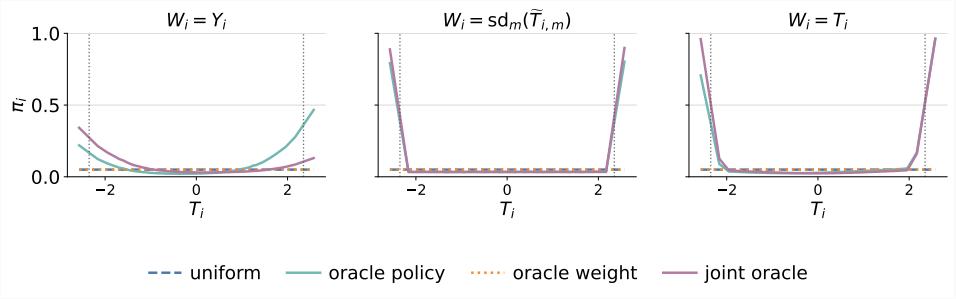
	Contemporaneous		Fama–MacBeth	
	Coef.	SE	Coef.	NW SE
12 naive pairs	2.99 – 3.44***	0.09–0.10	0.102 – 0.234	0.07–0.09
Naive uniform	3.701***	0.107	0.201**	0.084
Calibration-only	3.508***	0.401	0.604*	0.328
AI-PI	3.466***	0.162	0.298**	0.151

⇒ Contemporaneous: all measures load strongly ($t \approx 30$): they capture real news; naive and AI-PI differ by only ≈6%

⇒ The correction matters where the signal is weak (next month), exactly where fragility is dangerous

⇒ Fama–MacBeth: same conclusion; identification is within-month, cross-sectional

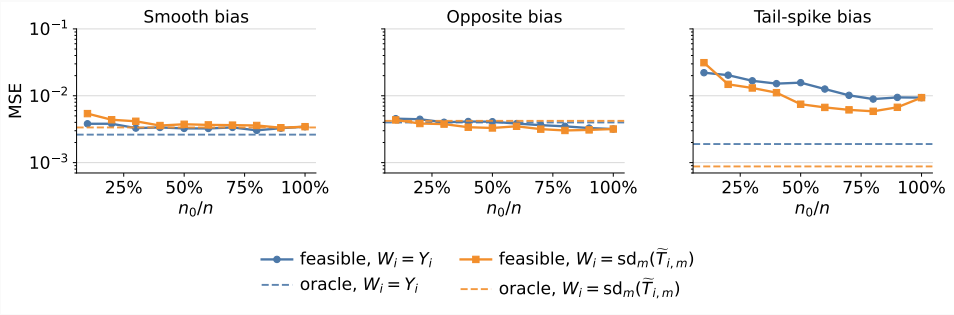
Oracle Labeling Policies in the Tail-Spike Design



$\Rightarrow$ Optimized policies concentrate labels in the tail region, where generators hallucinate

$\Rightarrow$ Model disagreement is nearly as informative as the infeasible oracle $W_i = T_i$

Pilot-Share Tradeoff for the Feasible Design

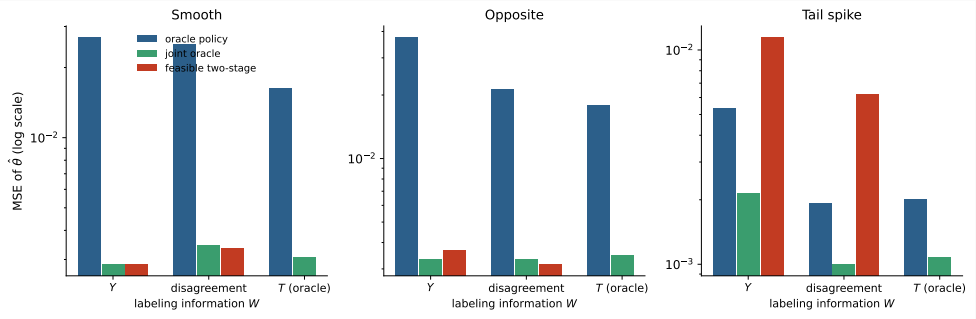


$\Rightarrow$ Small pilot: noisy design estimates; large pilot: no budget left for targeting

$\Rightarrow$ Tradeoff most pronounced in the tail-spike design, where adaptive labeling is most valuable

$\Rightarrow$ Empirical implementation uses a 40% pilot share of the 5% budget

The Role of the Labeling Information W_i



Labeling information W_i: outcome Y_i, model disagreement $\mathrm{sd}_m(\tilde{T}_{i,m})$, or oracle T_i

$\Rightarrow$ Cross-model disagreement is a powerful feasible labeling signal: close to oracle information in the tail-spike design, far better than the outcome level

Noisy Reference Labels

Back

Assumption (reference labels): calibration reveals $T_i^* = T_i + e_i$ with conditionally centered errors, $\mathbb{E}[e_i \mid Z_i, T_i] = 0$, and finite curvature-weighted error $\kappa^2 = \mathbb{E}[\nu(Z_i, e_i)\,\|e_i\|^2]$; here $\nu(Z_i, e_i)$ bounds the second derivative of the moment in the feature between T_i and T_i^*

Results (Taylor expansion around T_i; conclusions, not assumptions):

(a) All AI-PI guarantees hold for the reference estimand θ_0^* (solving $\mathbb{E}[g(Z_i, T_i^*; \theta)] = 0$)

(b) $\|\theta_0^* - \theta_0\| \le c\,\kappa^2$: the first-order term vanishes because the errors are centered; what remains is the second-order curvature term

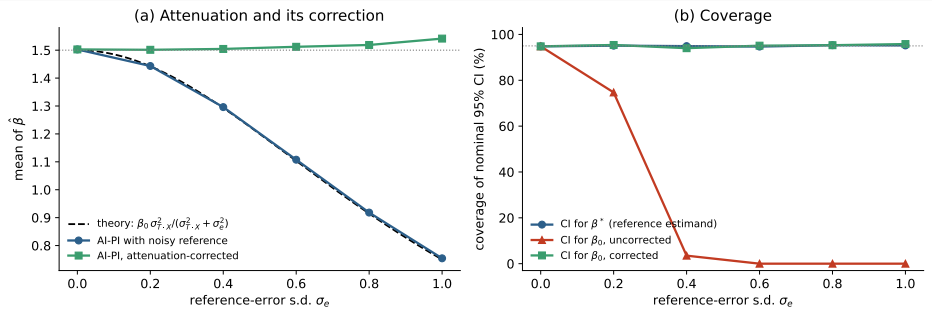
(c) Affine moments have zero curvature: $\theta_0^* = \theta_0$ exactly

Least squares: exact attenuation of the feature coefficient by $\sigma^2_{T|X}/(\sigma^2_{T|X} + \sigma^2_e)$; correctable when σ^2_e is estimable (duplicate annotations, two reference models)

Model-assisted calibration: strong model on an intermediate sample debiases cheap models on the full sample; human labels remain the anchor

Limit: no calibration scheme protects against bias shared by reference and generators (e.g., common look-ahead)

Simulation: Noisy Reference Labels

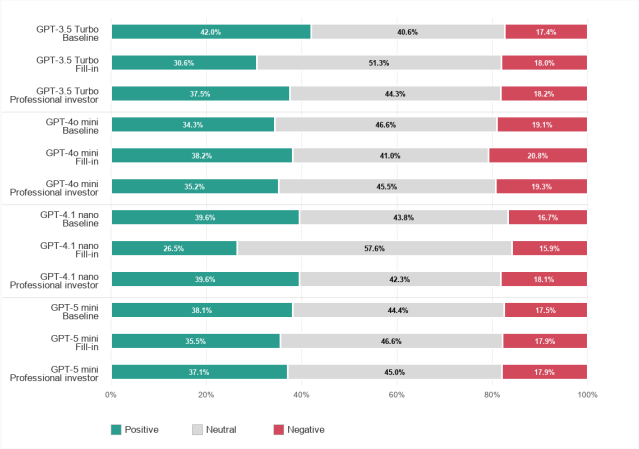


⇒ Simulated attenuation matches $\beta_0\,\sigma^2_{T|X}/(\sigma^2_{T|X}+\sigma^2_e)$ exactly; known-σ^2_e correction restores β_0

⇒ CIs cover the reference estimand β^* at ≈95% for every σ_e: validity, as the theory predicts

⇒ Strong-model reference with $\sigma_s \leq 0.1$ attains the gold-label benchmark (RMSE 5.9 vs. 5.8); a shared systematic bias destroys coverage of θ_0 (0.3%) while θ_0^* stays covered

LLM Sentiment Measures: Label Distributions



⇒ Positive share ranges from 26.5% to 42.0% across the 12 model–prompt pairs: the level of measured sentiment is itself model- and prompt-dependent

Model Error Metrics on the Human-Labeled Sample

Model–prompt pair	Accuracy	Bias	MAE	RMSE
GPT-5 mini – Professional investor	0.860	0.016	0.144	0.390
GPT-5 mini – Fill-in	0.857	−0.004	0.147	0.392
GPT-4o mini – Professional investor	0.852	−0.022	0.150	0.395
GPT-5 mini – Baseline	0.848	0.031	0.157	0.407
GPT-4o mini – Fill-in	0.829	−0.003	0.175	0.428
GPT-4o mini – Baseline	0.818	−0.030	0.185	0.438
GPT-4.1 nano – Professional investor	0.801	0.044	0.205	0.465
GPT-3.5 Turbo – Professional investor	0.792	0.022	0.210	0.464
GPT-3.5 Turbo – Baseline	0.779	0.079	0.224	0.481
GPT-3.5 Turbo – Fill-in	0.774	−0.061	0.230	0.488
GPT-4.1 nano – Baseline	0.761	0.058	0.246	0.510
GPT-4.1 nano – Fill-in	0.751	−0.087	0.255	0.516

$\Rightarrow$ 11,889 human-labeled headlines; labels coded $-1/0/1$; human mean 0.183

LLM Prompt Templates (1/2)

In each prompt, [firm name], [ticker], and [headline] are replaced by the values for the headline observation; the model sees only these inputs (no returns, no ex-post information)

Baseline prompt:

Here is a news headline about [firm name] ([ticker]): [headline]
Is this headline positive, negative, or neutral about the company?
Return a JSON object with exactly one field, label. The value of label must be exactly one of: positive, negative, or neutral.

Fill-in prompt (removes the JSON structure; intentionally less constrained):

Here is a news headline about [firm name] ([ticker]): [headline]
Is this headline positive, negative, or neutral about the company?
Write your answer as: ____ (fill in with one of positive/negative/neutral)

LLM Prompt Templates (2/2)

Professional-investor prompt (baseline JSON structure plus an investor framing):

Suppose you are a professional investor evaluating the implications of firm-specific news. Here is a news headline about [firm name] ([ticker]): [headline]
Is this headline positive, negative, or neutral about the company?
Return a JSON object with exactly one field, label. The value of label must be exactly one of: positive, negative, or neutral.

Design notes:

- Baseline and professional-investor prompts constrain the response to the three labels; the fill-in prompt is deliberately less structured
- The three formats separate model sensitivity from prompt format and framing sensitivity; headline labels coded $+1/0/-1$, averaged to stock-months

$\Rightarrow$ Same model, different prompt: only 88% agreement: prompts are part of the measurement