Inference with AI-Generated Covariates

Junting Duan* Markus Pelger†

This draft: July, 2026

Abstract

Empirical researchers increasingly use large language models (LLMs) to extract structured features, such as sentiment scores, classifications, and expectations, from unstructured data and treat these generated features as observed covariates in downstream estimation. This practice can invalidate inference when systematic, input-dependent errors in generated features, such as hallucination and look-ahead bias, distort the downstream moment conditions. Even after these moment-level distortions are corrected, generated features remain noisy proxies for the latent truth, with error profiles that differ across models and prompts. We introduce AI-Powered Inference (AI-PI), a method-of-moments framework that delivers valid and efficient inference by combining three components: (i) a moment-specific bias correction based on a small human-labeled calibration set; (ii) adaptive weights that optimally combine multiple model–prompt pairs, downweighting unreliable generators and exploiting complementary information across generated features; and (iii) an optimal calibration-set design that concentrates costly human labels on the observations where the generated features are least reliable. We establish consistency and asymptotic normality of the AI-PI estimator, allowing for non-random and data-adaptive labeling designs, cross-fitted LLM-pipeline tuning, and overidentified GMM. Simulations confirm that AI-PI delivers substantial gains over naive LLM regressions and over debiasing without optimal weighting or labeling design. In an empirical study of news-based sentiment and stock returns, AI-PI produces stable conclusions in a setting where naive analyses vary substantially across LLM and prompt choices, and its confidence interval is roughly half the length of the interval that uses the human-labeled data alone.

Keywords: Large Language Models, Generated Features, Generative AI, Hallucination Bias, Statistical Inference, Bias Correction, Adaptive Weighting, Method of Moments, Calibration, Unstructured Data

*Stanford University, Department of Management Science & Engineering, Email: duanjt@stanford.edu.
†Stanford University, Department of Management Science & Engineering, and NBER, Email: mpelger@stanford.edu.

1 Introduction

Modern empirical research increasingly relies on unstructured data, such as text, images, audio, and video, as a primary source of information. Earnings calls, clinical notes, satellite imagery, social media posts, and patent filings contain rich economic and scientific content that complements traditional structured covariates. Extracting quantitative features from these sources has historically required expensive manual annotation, limiting both the scale and scope of empirical analysis. The rapid progress of large language models (LLMs) and other generative AI (GenAI) systems has fundamentally altered this landscape: researchers can now prompt an LLM to read an earnings call transcript and output a sentiment score, parse a clinical note into diagnostic codes, or classify a patent into technology categories, at low marginal cost and on a scale that would be infeasible for human annotators.

The appeal of this approach is immediate. A single prompt can transform an entire corpus of unstructured records into structured covariates, enabling regression analysis, causal inference, and policy evaluation on datasets that were previously beyond reach. Consequently, the use of LLM-generated features has proliferated across empirical finance, economics, political science, and the health sciences. For example, in empirical finance, researchers extract sentiment and expectations from news headlines and earnings calls to study return predictability and corporate policies; in health economics, they parse clinical notes into diagnostic and severity codes; and in political economy, they score party platforms and legislative texts on ideological dimensions. In these applications, researchers routinely feed text into an LLM, extract a structured feature, and plug it into a regression, instrumental-variable model, or structural model as if it were an observed covariate.

Despite its practical appeal, this workflow rests on an implicit and often unjustified assumption: that the LLM-generated feature is a reliable proxy for the latent quantity of interest. In practice, generative models are imperfect in ways that matter for downstream inference. First, their outputs are subject to systematic, input-dependent biases. One source is *hallucination bias*, a tendency to produce content that is not fully grounded in the input or underlying data but instead includes fabricated, confabulated, or unsupported information (Ji, Lee, Frieske, Yu, Su, Xu, Ishii, Bang, Madotto, and Fung, 2023). A second important concern is *look-ahead bias*: because LLMs are trained on corpora that may include information from periods after the target measurement date, generated features can incorporate forward-looking content, leading to inflated predictive performance and biased structural estimates (Kaufman, Rosset, Perlich, and Stitelman, 2012; He, Lv, Manela, and

Wu, 2025). Even when an LLM performs well on average, its errors can be systematically correlated with the covariates or outcomes that enter the researcher's model, inducing distortions that do not vanish as the sample size grows. Second, even after the corresponding distortions are corrected, LLM-generated features remain noisy proxies for the latent truth. The magnitude and structure of this residual noise can differ substantially across models and prompts.

Treating LLM-generated features as observed data can therefore produce biased point estimates, incorrect standard errors, and misleading hypothesis tests. The problem is not merely classical measurement error or attenuation: because LLM errors depend on the input content and its training data, their effect on the estimating equations, or moment conditions, can be complex and not sign-identified without additional information.

Many empirical studies attempt to address this concern through a diagnostic validation step. Specifically, the researcher randomly selects a small subset of observations, obtains human labels for the target feature, and reports agreement rates or accuracy measures as evidence that the LLM output is "good enough." While informative, this validation step remains purely diagnostic. It does not change the estimation procedure: the generated features are still treated as observed inputs, and discrepancies between generated and true features are not formally accounted for in the moment conditions that define the estimator. Thus, diagnostic validation alone does not resolve the inferential problem.

In this paper, we propose a general framework for valid estimation and inference when a covariate or latent feature is produced by a generative model. We call our framework AI-PI, for *AI-Powered Inference*; it is distinct from the GenAI-Powered Inference (GPI) framework of Imai and Nakamura (2025), which addresses a complementary but distinct problem, as we discuss in the literature review. AI-PI addresses three linked components of this inferential problem. First, errors in generated features can induce systematic bias in the downstream moment conditions; we correct this bias with a moment-specific correction constructed from a small, human-labeled calibration set. Second, even after the moment correction is applied, generated features remain noisy proxies of the latent truth; we introduce adaptive weights across model–prompt pairs that minimize the asymptotic variance of the downstream estimator. In this sense, AI-PI is "AI-powered": AI-generated covariates provide additional information that improves efficiency beyond what can be achieved from the calibration sample alone, while the calibration sample restores validity by correcting moment-level errors. Third, human labels are costly, so the composition of the calibration set is a design choice; we derive the optimal labeling policy that allocates more labels to observa-

tions where the generated features are least reliable. The three components interact: the value of an additional human label depends on how the generated moments are weighted, and the optimal weights depend on the labeling design. AI-PI solves for the weights and the labeling design jointly.

The core insight of our approach is that the calibration set serves a fundamentally different role from the diagnostic validation that is common in current practice. Rather than merely assessing the quality of the generated features, we use the human-labeled observations to quantify and correct the bias induced by generated features in the specific moment conditions that define the structural parameter. This correction is moment-specific: it targets the estimand of interest rather than the feature reconstruction error, which is generally a different (and often less relevant) objective. Because the correction operates at the level of the moment condition, it accommodates nonlinear models, interaction effects, and any structural specification that can be expressed through moment equations.

A bias correction alone is not sufficient for an efficient estimator. Even after this correction is applied, generated features remain noisy proxies for the latent truth. This residual noise is not eliminated by the moment correction, but its impact on the final estimator can be reduced by appropriately weighting the generated moments. Differences in reliability across generators become central when multiple model–prompt pairs are available. In practice, researchers may experiment with different LLMs (e.g., GPT-4, Claude, Gemini), different prompts (e.g., zero-shot vs. chain-of-thought), and different seeds or temperature settings. Each configuration produces a different proxy for the true feature, with different bias profiles and noise levels. Our framework introduces a weight vector that determines how each model–prompt pair contributes to the final estimator. These weights are chosen to minimize the asymptotic variance of the downstream estimator, automatically downweighting unreliable generators and upweighting informative ones. When all generators are poor, the optimal weights shrink toward zero and the estimator reduces to the calibration-only estimator, providing robustness. When a generator is highly accurate, the weights amplify its contribution, delivering substantial efficiency gains over using the calibration set alone.

In addition to adaptive weighting across model–prompt pairs, we develop an optimal calibration-set design that determines which observations should be selected for human labeling. Rather than labeling observations uniformly at random, our framework allocates the labeling budget to observations where the generated features are least reliable, as measured by the conditional residual between the true and synthetic moments. This targeted labeling strategy minimizes the asymptotic variance of the estimator under a given expected labeling budget and can yield substantial effi-

ciency gains relative to random labeling, particularly when the generator's errors are concentrated in specific regions of the covariate space.

We develop a complete asymptotic theory for the AI-PI estimator. Under standard moment regularity conditions and a missing-at-random condition for the labeling mechanism, we establish consistency and asymptotic normality with an explicit expression for the asymptotic variance. This variance decomposes into two terms: an irreducible component that reflects the variance of calibrated observations, and a design-dependent component that captures the interaction between the labeling policy and the quality of the generative model. The asymptotic theory enables valid confidence intervals and hypothesis testing for the structural parameter, and it provides the objective function for optimizing the weights and the labeling design.

Our framework accommodates an important feature of modern LLM pipelines : the pipeline itself may be tuned using calibration labels. When the prompt, the retrieval-augmented generation (RAG) configuration, or the choice of model is selected using the calibration labels, the generated features are no longer independent of the calibration set, which can invalidate naive inference. We address this through a cross-fitting procedure in which the calibration data are split into folds: one fold is used to tune the LLM pipeline, and the remaining folds are used to construct the debiasing terms. Under a stability condition requiring that the generated features do not change discontinuously as the tuning data vary, cross-fitting restores independence and preserves the asymptotic guarantees.

Our work makes three main contributions. First, we introduce the AI-PI framework, which provides a general approach for conducting valid statistical inference when covariates are produced by generative AI models. The framework treats LLM outputs as imputations within a method-of-moments structure, constructs moment-specific bias corrections using a calibration set, and combines multiple model–prompt pairs through adaptive weighting. Crucially, the framework explicitly accounts for the heterogeneous randomness of generative AI models: each model–prompt pair produces outputs with a distinct, input-dependent systematic bias and seed-level stochastic noise, and these heterogeneous error structures are accommodated through the GenAI feature decomposition. The adaptive weights automatically downweight unreliable model-prompt pairs and place more weight on informative ones.

Second, we develop a complete asymptotic theory for the AI-PI estimator, establishing consistency and asymptotic normality under general conditions. The theory covers non-random labeling designs, multiple model–prompt pairs, cross-fitting with LLM pipeline tuning under a stability condition, and data-driven designs estimated from a pilot sample. We also develop extensions to

overidentified GMM, noisy reference labels, weights that vary with the labeling information, and clustered data. The asymptotic variance admits a transparent decomposition into an irreducible calibration-set component and a design-dependent component that captures the interaction between the labeling policy and the quality of the generative model.

Third, we provide guidance on the construction of the calibration set through the labeling design. We derive the optimal labeling policy, which concentrates human labels on observations where the generative model is least reliable, as measured by a conditional residual score. We also characterize the optimal model–prompt weights, showing that they solve a tractable joint optimization problem. We further develop a feasible two-stage plug-in procedure and prove it attains the oracle asymptotic variance under the jointly optimal weights and labeling policy. These results provide actionable guidance on where to allocate labeling resources and how to combine multiple generative models and prompts efficiently.

We apply AI-PI to a canonical question in empirical finance: whether the sentiment of firm-specific news predicts next-month stock returns. Using approximately 260,000 news headlines matched to U.S. stocks, we construct stock-month sentiment measures from twelve model–prompt pairs and estimate a standard cross-sectional return regression. The application illustrates both the problem and the solution. The twelve naive slope estimates that treat generated sentiment as observed data range from 0.083 to 0.226, a nearly threefold difference, with statistical significance varying across reasonable model and prompt choices. The estimator that uses only the human-labeled calibration sample, about five percent of the units, is valid but imprecise. AI-PI combines both sources of information: it delivers a single bias-corrected estimate of 0.312 whose confidence interval is roughly half as long as the calibration-only interval, and the estimate is stable across alternative control sets. The application thus demonstrates the central promise of the framework: valid and more precise inference at a small labeling cost, with the fragility of naive LLM-based conclusions resolved rather than ignored.

The remainder of the paper is organized as follows. Section 2 develops intuition for our approach through a simple regression example with a single LLM-generated covariate. Section 3 introduces the general estimation framework, including the GenAI feature decomposition, the calibration-set design, the debiased moment estimator, and the cross-fitting procedure. Section 4 establishes consistency and asymptotic normality, derives the form of the asymptotic variance, and proves the consistency of the plug-in variance estimator. Section 5 derives the optimal calibration-set design and model–prompt weights and presents a feasible two-stage algorithm for their joint computation.

Section 6 develops extensions to noisy reference labels and model-assisted calibration, overidentified GMM, and clustered data. Section 7 provides practical guidance. Section 8 presents simulation evidence, and Section 9 applies the framework to news-based sentiment and stock returns. Section 10 concludes. The Appendix contains technical proofs.

1.1 Related Literature

Our paper contributes to several strands of literature.

We build most directly on the imputation-powered inference (IPI) framework of Duan and Pelger (2025), which develops debiased moment estimation with adaptive weighting for downstream inference when structured covariates are missing and imputed by machine-learning methods. AI-PI extends this paradigm to settings where the "imputation" is performed by a generative AI model applied to unstructured data, which introduces several new challenges. The error structure of LLM outputs combines a systematic, input-dependent bias with seed-level stochastic noise, and these error profiles are heterogeneous across model–prompt pairs; the GenAI feature decomposition and the adaptive weighting across generators are designed for exactly this structure. A second distinctive feature is the calibration-set design. In IPI and related imputation frameworks, the set of fully observed units is determined by the missingness pattern and is outside the researcher's control. In the GenAI setting, by contrast, the researcher actively chooses which observations receive costly human labels, subject to a budget constraint. AI-PI formalizes this choice and derives the optimal labeling allocation jointly with the model–prompt weights, a joint design problem with no direct analogue in existing imputation methods.

Our debiasing construction is related to the prediction-powered inference (PPI) framework of Angelopoulos, Bates, Fannjiang, Jordan, and Zrnic (2023) and its extensions (Angelopoulos, Duchi, and Zrnic, 2023; Miao, Miao, Wu, Zhao, and Lu, 2023; Ji, Lei, and Zrnic, 2025; Zrnic and Candès, 2024b), which use a labeled dataset to correct moments computed with machine-learning predictions on a larger unlabeled sample, as well as to the broader semi-supervised literature that exploits unlabeled data (Wasserman and Lafferty, 2007; Chakrabortty and Cai, 2018). An important distinction is that PPI concerns missing *outcome labels*: the prediction enters the loss or estimating equation in place of the missing outcomes. In our setting, the generated object is a covariate or feature, which can be multivariate and can exhibit heterogeneous input-dependent errors across model-prompt pairs. AI-PI addresses this structure through adaptive weighting, derives the optimal labeling design jointly with the weights, and provides cross-fitting guarantees when the LLM

pipeline itself is tuned on the calibration labels. Related work on label-efficient inference includes the active inference approach of Zrnic and Candès (2024a), which allocates labels based on model uncertainty; our optimal labeling policy can be viewed as the generated-feature analogue, derived jointly with the ensemble weights. The residual score that drives our policy differs from uncertainty-based allocation in that it targets the downstream estimator variance: it measures local generator unreliability through the moment Jacobian and the residual covariance of the weighted ensemble, so the labeling policy and the use of the resulting labels are optimized against the same variance criterion.

A classical literature studies estimation with missing covariates (Little, 1986; Abrevaya and Donald, 2017; Freyberger, Höppner, Neuhierl, and Weber, 2024), typically under parametric or semiparametric restrictions on the missingness mechanism and the imputation procedure. Our setting differs in that the latent feature has never been directly observed for most units and must be extracted from unstructured data by a generative model whose error structure is far from classical. Relatedly, the errors-in-variables literature (Fuller, 2009; Carroll, Ruppert, Stefanski, and Crainiceanu, 2006) studies covariates observed with noise, but typically relies on classical (mean-zero, independent) measurement error, repeated measurements, or instruments. The LLM error is generally non-classical: it is input-dependent, can be correlated with outcomes and covariates, and is not sign-identified without additional information. Our calibration set provides direct observations of the true feature, and our correction operates at the level of the moment condition rather than the covariate, which allows for valid inference under flexible error structures.

A growing econometric literature studies inference with AI- and ML-generated variables. Ludwig, Mullainathan, and Rambachan (2024) provide an applied econometric framework for LLMs that distinguishes prediction from estimation problems, documents that downstream regression conclusions can flip across models and prompts, highlights training-leakage concerns, and recommends combining LLM outputs with a small random validation sample. Egami, Hinck, Stewart, and Wei (2024) propose design-based supervised learning for LLM annotations in social science. Gentzkow, Kelly, and Taddy (2019) provide the methodological foundation for text-as-data applications in economics. Imai and Nakamura (2025) introduce GenAI-powered inference, a framework for causal and predictive inference with unstructured data. Their approach uses open-source generative models to regenerate text or images, extract the models' internal representations, and apply machine learning to these representations to adjust for latent confounding encoded in the unstructured data. Their framework and ours address complementary problems: they use these internal representations to

incorporate unstructured treatments or confounders into downstream inference, whereas we correct moment-level errors in AI-generated covariates relative to a human-labeled reference. Because our method requires only the generated outputs, it applies to any AI model, including closed commercial ones.

Most closely related is Battaglia, Christensen, Hansen, and Sacher (2025), who study inference when variables extracted from unstructured data by AI or machine-learning methods are used as regressors in downstream regression models. In an asymptotic regime where the measurement error remains comparable to sampling uncertainty, they show that ordinary least squares keeps the correct asymptotic variance but acquires a first-order bias. As a result, conventional confidence intervals can have the right width but the wrong center. They restore validity either by an analytical bias correction or by joint maximum-likelihood estimation of the regression and the latent variables, using the structure of the first-step error, and they stress that this requires no human-labeled validation sample. Our framework is complementary and differs in four respects. First, rather than modeling the generator's error structure, AI-PI uses a small human-labeled calibration set to correct the discrepancy induced by generated features in the downstream moment conditions. This allows for input-dependent and model-specific LLM errors, including hallucination, look-ahead, and coverage bias, without requiring the error structure to be parametrically specified or sign-identified. Second, our main formulation is a general moment-condition framework, rather than a regression-specific framework, and it allows the generated feature to be multivariate. Third, AI-PI combines multiple model–prompt pairs through adaptive weights, using complementary information across generated features and addressing the model-and-prompt fragility documented by Ludwig, Mullainathan, and Rambachan (2024), rather than relying on a single generated variable. Fourth, we treat the calibration sample as a design choice and derive the labeling policy that minimizes asymptotic variance under a labeling budget, rather than taking a validation sample as given. The two approaches trade off transparently: their structural correction avoids gold labels but relies on the error structure being known or correctly estimated, whereas our calibration set is the price of robustness to flexible generator errors and applies when the feature is a human-replicable annotation. Battaglia, Christensen, Hansen, and Sacher (2025) note that when validation labels are available, one could estimate the target on the labeled subsample alone, so that generated features improve efficiency; this is precisely their role in our framework. In our empirical application, the calibration-only confidence interval is nearly twice as long as the AI-PI interval, while AI-PI uses human labels on about five percent of the sample. When the error structure is itself estimable, the

two approaches connect: our noisy-reference extension in Section 6.1 characterizes, and in special cases corrects, residual attenuation in their spirit.

Our work also connects to the large empirical literature that uses LLMs for data annotation and feature extraction (Li, Shi, Xia, and Yang, 2025; Gilardi, Alizadeh, and Kubli, 2023; Törnberg, 2024). Most of this work treats LLM outputs as observed data and reports diagnostic validation, such as agreement rates with human labels on a small subset. Such validation is informative but does not change the estimator, and high agreement rates do not imply that the bias in the estimated moment conditions is negligible. Our framework instead uses the labeled calibration sample directly in estimation, converting it into a formal moment correction for downstream inference. The concerns we address are well documented: hallucination (Ji, Lee, Frieske, Yu, Su, Xu, Ishii, Bang, Madotto, and Fung, 2023) and look-ahead bias arising from training leakage (Kaufman, Rosset, Perlich, and Stitelman, 2012; He, Lv, Manela, and Wu, 2025).

Finally, our cross-fitting procedure is related to double/debiased machine learning (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018), which uses sample splitting to separate nuisance estimation from inference. In our setting, sample splitting serves to separate LLM-pipeline tuning from moment correction: the prompt, model, or retrieval configuration is selected using one part of the calibration data, while the correction is constructed on held-out calibration observations. Under the stability condition, this separation preserves the asymptotic guarantees.

2 Illustrative Example

We begin with a simple example to illustrate the fundamental challenge of using LLM-generated features in structural estimation and to develop the key ideas underlying our framework, before presenting the general theory.

2.1 Simple Setup

Suppose we observe N i.i.d. units, each consisting of an outcome $Y_i \in \mathbb{R}$, observed covariates $X_i \in \mathbb{R}^p$, and a piece of unstructured text S_i. Let $Z_i = (Y_i, X_i, S_i)$. From the text, a human expert can extract a scalar feature $T_i = f(S_i) \in \mathbb{R}$, for example, a sentiment score, a severity rating, or a topic classification. Our goal is to estimate the linear regression coefficient β_0 in the model

$$Y_i = X_i^\top \gamma_0 + T_i \beta_0 + \varepsilon_i.$$

Let $X_i^* = (X_i^\top, T_i)^\top$ and $\theta_0 = (\gamma_0^\top, \beta_0)^\top$. The target parameter θ_0 is defined as the unique solution to the moment condition

$$\mathbb{E}[X_i^*(X_i^{*\top}\theta_0 - Y_i)] = 0.$$

The challenge is that human labeling is expensive: obtaining T_i for all N units may be infeasible. Instead, the researcher uses an LLM to generate a proxy $\tilde{T}_i$ for each unit by prompting the model with the text S_i. The LLM output may differ from the true feature due to both systematic bias and stochastic variation. Specifically,

$$\tilde{T}_i = T_i + b(S_i) + \epsilon_i,$$

where $b(S_i)$ captures systematic biases of the LLM that may depend on the input S_i, such as hallucination bias and look-ahead bias, and ϵ_i is a mean-zero noise conditional on (Z_i, T_i).

Figure 1 illustrates the resulting data structure. The problem is formally analogous to a missing-covariate problem: the true feature T_i is "missing" for every unit outside the calibration set, and the LLM outputs play the role of imputed values that are available for all units but contaminated by bias and noise. Two features distinguish the GenAI setting from classical missing data. First, several imputations are available for the same unit, one per model–prompt pair, each with its own error profile. Second, the researcher chooses which units enter the calibration set: panel (b) of Figure 1 previews the targeted designs we develop in Section 5, which concentrate the labeling budget on units where the generated features are least reliable.

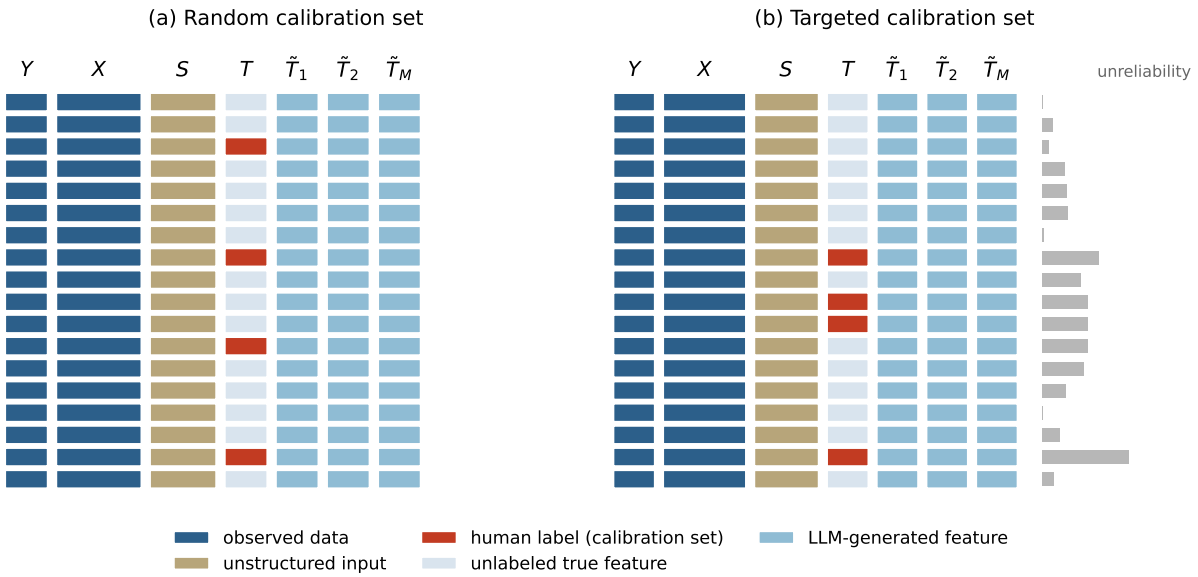
2.2 Naive Approaches and Their Limitations

Naive LLM regression. The most common approach simply substitutes $\tilde{T}_i$ for T_i and runs the regression using all N observations. Specifically, this approach solves

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{X}_i(\tilde{X}_i^\top\theta - Y_i) = 0,$$

where $\tilde{X}_i = (X_i^\top, \tilde{T}_i)^\top$. While this approach is cheap and uses all the data, it ignores the discrepancy $b(S_i) + \epsilon_i$. This generally leads to a biased and inconsistent estimator $\hat{\theta}$, which converges to a limit that differs from θ_0.

Figure 1: Data structure of the AI-PI problem



This figure illustrates the data structure. Each row is a unit i with observed outcome Y_i, observed covariates X_i, and unstructured input S_i. The true feature T_i is observed only for the human-labeled calibration set (red), while LLM-generated features $\tilde{T}_{i,1}, \ldots, \tilde{T}_{i,M}$ are available for all units (light blue). Panel (a) shows a random calibration set. Panel (b) shows a targeted calibration set that concentrates labels on units with high generator unreliability (gray bars).

Calibration-only regression. If a subset S_0 of n units are labeled by human experts, one could discard the LLM-generated features entirely and estimate θ_0 using only the labeled observations:

$$\frac{1}{n} \sum_{i \in S_0} X_i^* (X_i^{*\top} \theta - Y_i) = 0.$$

This estimator is unbiased but inefficient: it discards all information contained in the $N-n$ unlabeled observations and their LLM-generated features. When the labeling budget is small ($n \ll N$), the resulting confidence intervals can be very wide.

A common middle ground approach is to run the naive LLM regression but report agreement rates between $\tilde{T}_i$ and T_i on the labeled subset as evidence of quality. However, this does not change the estimator and therefore does not resolve the bias problem. High agreement rates on the labeled subset do not guarantee that the bias is negligible in the estimated moment conditions.

2.3 Our Solution

Our approach combines the calibration set with the LLM-generated features to construct a bias-corrected estimator that is both unbiased and more efficient than the calibration-only estimator.

The key idea is that the calibration set provides a direct estimate of the bias that generated features induce in the moment condition. Specifically, for each labeled unit $i \in S_0$, we can compute the moment using both the true feature T_i and the LLM-generated feature $\tilde{T}_i$. The difference between these two moments estimates the bias:

$$\begin{aligned}\text{Bias} &= \mathbb{E}\left[\tilde{X}_i(\tilde{X}_i^\top\theta - Y_i) - X_i^*(X_i^{*\top}\theta - Y_i)\right] \\ &\approx \frac{1}{n}\sum_{i\in S_0}\left(\tilde{X}_i(\tilde{X}_i^\top\theta - Y_i) - X_i^*(X_i^{*\top}\theta - Y_i)\right).\end{aligned}$$

Subtracting this estimated bias from the naive LLM moment yields a debiased estimator.

Importantly, the labeled subset S_0 need not be completely-randomly selected. In practice, the accuracy of LLM-generated features varies across observations, and some samples exhibit larger discrepancies between generated and true features. We therefore allow for non-random labeling schemes that allocate human labeling based on observable information. Intuitively, concentrating labels on observations where the generated features are less reliable improves statistical efficiency.

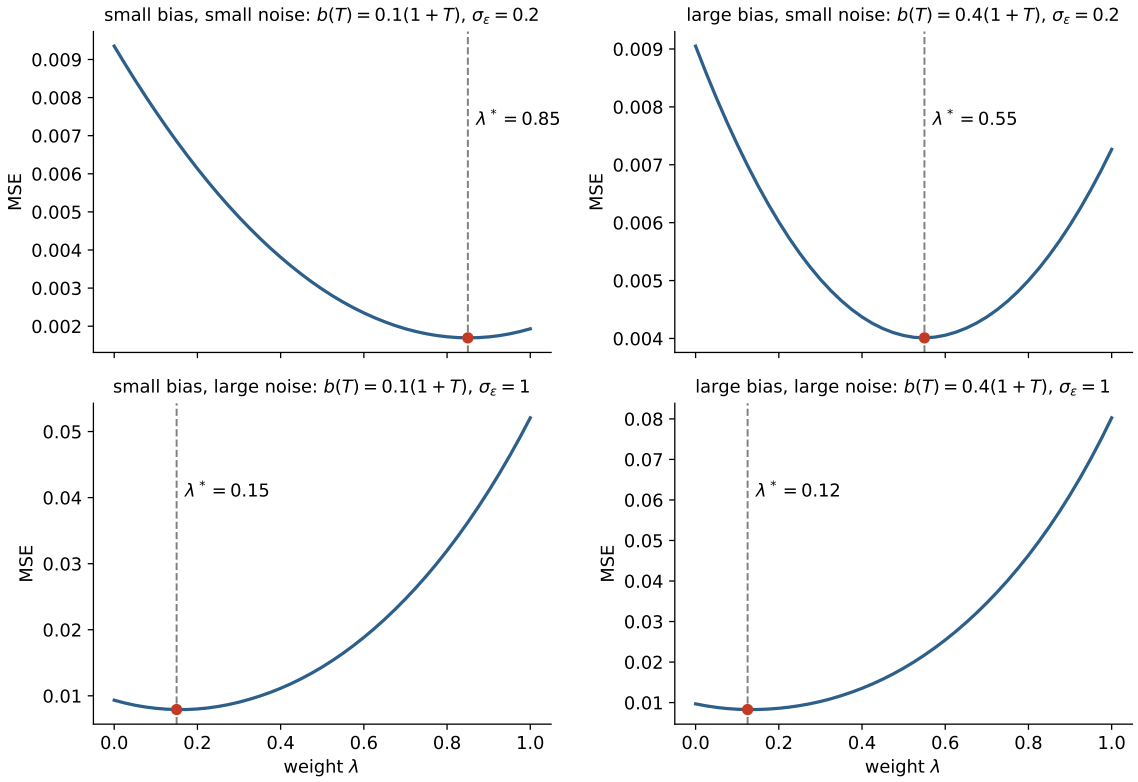
In addition, the LLM-generated features introduce additional noise relative to the true features. We introduce a weight $\lambda \in [0, 1]$ that controls how much the estimator relies on the LLM proxy versus the calibration set. The weighted debiased moment estimator is

$$\begin{aligned}G_N^\lambda(\theta) = &\frac{1}{n}\sum_{i\in S_0} X_i^*(X_i^{*\top}\theta - Y_i) \\ &+ \lambda\cdot\frac{N-n}{N}\left(\frac{1}{N-n}\sum_{i\notin S_0}\tilde{X}_i(\tilde{X}_i^\top\theta - Y_i) - \frac{1}{n}\sum_{i\in S_0}\tilde{X}_i(\tilde{X}_i^\top\theta - Y_i)\right).\end{aligned}$$

When $\lambda = 0$, the estimator reduces to the calibration-only estimator based solely on labeled data. When $\lambda = 1$ and the LLM is perfect ($\tilde{T}_i = T_i$), it recovers the oracle estimator that uses all N labeled features. The optimal weight λ^* balances the efficiency gains from incorporating more data against the additional noise introduced by LLM errors. Accordingly, λ^* is close to one when the LLM is highly accurate and close to zero when the LLM is unreliable.

Figure 2 illustrates this tradeoff in a simulated version of the example. We plot the mean squared error of $\hat{\theta}_\lambda$ as a function of λ for four combinations of systematic bias $b(\cdot)$ and the noise level in

Figure 2: MSE of the weighted debiased estimator as a function of the weight λ

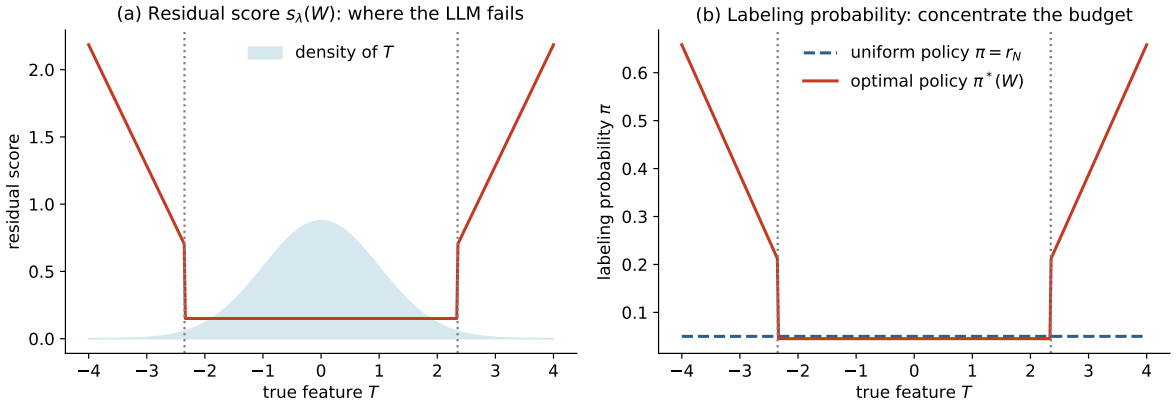


This figure reports the simulated mean squared error of the weighted debiased estimator $\hat{\theta}_\lambda$ in the illustrative example as a function of the weight λ, for four combinations of the generated feature. Data are generated as $Y_i = 0.5 + 1.5\,T_i + \varepsilon_i$ with $T_i \sim \mathcal{N}(0,1)$, $\varepsilon_i \sim \mathcal{N}(0,1)$, $N = 3000$, and a random calibration set of size $n = 200$. The generated feature is $\tilde{T}_i = T_i + b(T_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$; the panels vary the bias function $b(\cdot)$ and the noise level σ_ϵ. The dashed line marks the MSE-minimizing weight λ^*. Results are averaged over 1,500 simulations.

the generated feature. The figure highlights three main findings. First, the debiased estimator is consistent for every λ, so the curves differ only in efficiency, but the efficiency differences are large. Second, the optimal weight λ^* adapts to the quality of the generator: it is close to one when the generated feature is accurate and shrinks toward zero as the bias or the noise grows. Third, the cost of a suboptimal weight is asymmetric: overweighting a poor generator (λ near one in the right panels) is more harmful than underweighting a good one. The weighted estimator is consistently better than the calibration-only estimator at the optimum, since $\lambda = 0$ is always feasible.

The labeling design provides a complementary source of efficiency. Suppose the generator is accurate for typical inputs but fails in a specific region, for example because it hallucinates on rare or ambiguous texts. Figure 3 illustrates this situation. Panel (a) plots a residual score that measures, at each point of the feature space, how much the generated moment deviates from the true

Figure 3: Intuition for the optimal labeling design



This figure illustrates the optimal labeling design in a stylized setting where the generated feature is accurate for typical inputs but unreliable in the tails (vertical dotted lines mark the unreliable region). Panel (a) plots the residual score $s_\lambda(W)$, which measures the local discrepancy between the generated moment and the true moment, together with the density of the true feature. Panel (b) compares the uniform labeling policy $\pi = r_N$ with the optimal policy $\pi^*(W)$ from Theorem 4, which concentrates the labeling budget where the residual score is large while maintaining a strictly positive labeling probability everywhere.

moment; the score spikes in the tails, precisely where the density of observations is low. A uniform labeling policy spends much of the budget on typical observations for which the generated moment is already close to the true moment. The optimal policy in panel (b) instead concentrates the labeling budget on the unreliable region, while maintaining a strictly positive labeling probability everywhere. Section 5 derives this policy formally and shows that the optimal labeling probability is proportional to the residual score.

This illustrative example captures the core mechanism of AI-PI: a moment-specific bias correction makes the estimator valid, the weight λ governs how aggressively the generated features are used, and the labeling design determines where the calibration budget is spent. Section 3 formalizes these ideas and develops the general framework with multiple model–prompt pairs.

3 Main Model and Framework

We now formalize the general AI-PI framework, extending the ideas from Section 2 to accommodate general moment-based estimation, multiple model–prompt pairs, and cross-fitting for pipeline tuning.

3.1 Problem Setup

As in Section 2, suppose we observe N i.i.d. units $Z_i = (Y_i, X_i, S_i)$, where Y_i denotes an outcome variable, X_i is a vector of observed covariates, and S_i is an unstructured input (e.g., text, image, audio). Our goal is to extract a vector of feature representation $T_i = f(S_i) \in \mathbb{R}^p$ from the unstructured input S_i and use it to estimate a parameter vector $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, which is defined as the unique solution to the population moment condition

$$\mathbb{E}\big[g(Z_i, T_i; \theta_0)\big] = 0. \tag{1}$$

Here, g is a user-specified moment function. This formulation encompasses a broad class of structural estimation problems, including linear and nonlinear regression, instrumental variable estimation, maximum likelihood, and generalized method of moments. In the main text, we focus on the just-identified case where $g(\cdot) \in \mathbb{R}^d$; Section 6.2 extends the framework and the inferential theory to overidentified GMM.

Note that our primary focus is on *structural estimation*, where the parameter θ_0 is the object of interest. Unlike predictive modeling, structural estimation generally does not have an analogous out-of-sample evaluation: the goal is to understand how covariates affect outcomes rather than to optimize predictive accuracy. This distinction is important because it means that standard model-validation techniques (e.g., cross-validation on held-out data) cannot directly assess the quality of the structural estimates. Consequently, we require estimation and inference procedures that are formally valid for θ_0.

3.2 LLM-Generated Features

Suppose we have M model–prompt pairs, indexed by $m \in \{1, \ldots, M\}$, where each pair consists of an LLM model and a specific prompt designed to extract the feature of interest from the unstructured input S_i. For each unit i and model–prompt pair m, we can generate multiple feature realizations using different random seeds $s \in \{1, \ldots, S\}$. The resulting generated features are given by

$$\tilde{T}_{i,m,s} = T_i + b_m(Z_i) + \epsilon_{m,s}(Z_i) + u_{i,m,N}(Z_1, \ldots, Z_N), \tag{2}$$

where $b_m(Z_i)$ denotes a systematic, input-dependent bias specific to model–prompt pair m. The term $\epsilon_{m,s}(Z_i)$ captures seed-specific stochastic variation across different runs of the same model–

prompt pair, and we assume $\mathbb{E}[\epsilon_{m,s}(Z_i) \mid Z_i] = 0$. Finally, $u_{i,m,N} \coloneqq u_{i,m,N}(Z_1, \ldots, Z_N)$ captures the cross-unit dependence induced by the inclusion of the other sampled units $\{Z_j : j \neq i\}$ in the pretraining corpus. We allow this spillover term to be present, but assume that it is asymptotically negligible, in a sense made precise in Assumption 3 below. While own-unit exposure may induce a non-negligible bias in the generated feature, the aggregate contribution of other sampled units to the generated feature for unit i is asymptotically negligible, as the sampled units constitute a negligible fraction of the pretraining corpus.

The systematic bias $b_m(Z_i)$ collects input-dependent deviations in the output of model–prompt pair m. We highlight three economic sources of bias relevant in applications. First, *hallucination bias*: generative models can produce content that is not grounded in the input, including fabricated, confabulated, or unsupported information (Ji, Lee, Frieske, Yu, Su, Xu, Ishii, Bang, Madotto, and Fung, 2023). In feature-extraction tasks this manifests as systematic distortions that depend on the input in complex ways, for example a tendency to assign positive sentiment to headlines containing certain firm names or numeric patterns. Second, *look-ahead bias*: LLMs are trained on corpora that may contain information from periods after the measurement date of unit i, including coverage of the unit itself. A model scoring a 2015 headline about a firm may have been trained on 2020 news about the same firm, so the generated feature can embed ex-post information (Kaufman, Rosset, Perlich, and Stitelman, 2012; He, Lv, Manela, and Wu, 2025). This own-unit exposure is captured by $b_m(Z_i)$, while the exposure to other sampled units is captured by the spillover term $u_{i,m,N}$. Third, *coverage bias*: a model may be undertrained on specific domains, time periods, languages, or types of entities, making its errors systematically correlated with observable characteristics. In all three cases, the bias is input-dependent, generally not mean-zero, and not sign-identified without external information, which is why we anchor the correction in a human-labeled calibration set.

The decomposition (2) also highlights why a moment correction alone does not determine the most efficient estimator. After the correction, generated moments can still differ substantially in their residual variation across models, prompts, and seeds; the adaptive weights in Section 5 choose how much each generated moment should contribute to the final estimator.

For each model–prompt pair m, we construct the seed-averaged feature

$$\tilde{T}_{i,m} \coloneqq \frac{1}{S}\sum_{s=1}^{S} \tilde{T}_{i,m,s} = T_i + b_m(Z_i) + \bar{\epsilon}_m(Z_i) + u_{i,m,N}(Z_1, \ldots, Z_N),$$

where $\bar{\epsilon}_m(Z_i) = S^{-1}\sum_{s=1}^{S} \epsilon_{m,s}(Z_i)$. We compute $\tilde{T}_{i,m}$ for all units and all model–prompt pairs. To

isolate the effect of cross-unit spillovers, we introduce the no-spillover counterpart of the generated feature as

$$\tilde{T}^0_{i,m,s} = T_i + b_m(Z_i) + \epsilon_{m,s}(Z_i),$$

which corresponds to the generated feature when $u_{i,m,N}$ is removed. We further define the corresponding seed-averaged no-spillover feature as $\tilde{T}^0_{i,m} = S^{-1}\sum_{s=1}^{S}\tilde{T}^0_{i,m,s}$, which is used only in the theoretical analysis and is not directly observable. The no-spillover feature removes the cross-unit spillover term from the generated feature and depends only on the data for unit i and seed-level randomness. Therefore, $(Z_i, T_i, \{\tilde{T}^0_{i,m,s}\}_{m,s})$ are i.i.d. across units.

3.3 Calibration Set and Labeling Design

Let $D_i \in \{0,1\}$ indicate whether unit i is labeled by a human expert, where $D_i = 1$ means that the true feature T_i is observed through human labeling, and $D_i = 0$ otherwise. We denote the set of labeled units by $S_0 = \{i : D_i = 1\}$.

We allow the labeling process to be not completely random. Let W_i denote the information available for all units prior to labeling selection, which may include the observation Z_i, LLM-generated features $\tilde{T}_{i,m,s}$, and uncertainty or disagreement measures derived from these features. The labeling decision depends on W_i through the propensity function

$$\mathbb{P}(D_i = 1 \mid W_i) = \pi(W_i) \in (0,1], \tag{3}$$

which governs the probability that unit i is selected for human labeling.

We impose a labeling budget of size n by requiring

$$\mathbb{E}[\pi(W_i)] = \frac{n}{N} =: r_N, \tag{4}$$

which ensures that the expected number of labeled units equals the labeling budget. We consider an asymptotic regime where $n, N \to \infty$ and $r_N \to r \in [0,1]$. Note that this framework allows for vanishing labeling fractions, i.e., $r_N \to 0$ as $N \to \infty$. In such cases, the propensity $\pi(W_i)$ may converge to 0 with N, while remaining strictly positive for each finite N. When $\pi(W_i)$ depends on observable characteristics, it captures targeted labeling strategies that allocate the labeling budget to regions where the LLM is expected to be less reliable. In Section 5, we show how to optimally choose the labeling policy under this budget constraint.

3.4 Weighted Debiased Moment Estimator

We now formalize the weighted debiased moment estimator based on the ideas introduced in Section 2. Let $\lambda = (\lambda_1, \ldots, \lambda_M)$ be a vector of weights assigned to each model–prompt pair. For notational convenience, write $\pi_i \coloneqq \pi(W_i)$. We estimate the moment conditions as

$$\begin{aligned} G_{\lambda,\pi}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\pi_i}\, g(Z_i, T_i; \theta) \\ + \sum_{m=1}^{M} \lambda_m \left[\frac{1}{N}\sum_{i=1}^{N} g(Z_i, \tilde{T}_{i,m}; \theta) - \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\pi_i}\, g(Z_i, \tilde{T}_{i,m}; \theta) \right]. \end{aligned} \tag{5}$$

In this estimator, the first term is an inverse-propensity weighted moment based on human-labeled data, which provides an unbiased estimate of $\mathbb{E}[g(Z_i, T_i; \theta)]$ under the labeling design. The second term aggregates model–prompt specific contributions based on LLM-generated features. Specifically, for each model–prompt pair m, we add the full-sample moment based on LLM-generated features and subtract its inverse-propensity weighted counterpart computed over the calibration set. The difference between the moments with human-labeled and LLM-generated features on the calibration set estimates and corrects the systematic bias. The weights λ then determine how different model–prompt sources are combined in this estimation. Let $\hat{\theta}_{\lambda,\pi}$ denote the estimator defined as the solution to the moment condition $G_{\lambda,\pi}(\theta) = 0$.

To verify unbiasedness, note that under the labeling design (formalized in Assumption 1 below), the labeling indicator satisfies $\mathbb{E}[D_i/\pi_i \mid Z_i, T_i, \tilde{T}_{i,1}, \ldots, \tilde{T}_{i,M}, W_i] = 1$, and therefore

$$\mathbb{E}[G_{\lambda,\pi}(\theta)] = \mathbb{E}[g(Z_i, T_i; \theta)] + \sum_{m=1}^{M} \lambda_m \big(\mathbb{E}[g(Z_i, \tilde{T}_{i,m}; \theta)] - \mathbb{E}[g(Z_i, \tilde{T}_{i,m}; \theta)]\big) = \mathbb{E}[g(Z_i, T_i; \theta)].$$

Thus $G_{\lambda,\pi}(\theta_0) = 0$ in expectation regardless of weights λ and the labeling design π, confirming that the estimator targets the correct parameter.

The centering argument relies on the known labeling design rather than on a model for the LLM errors. Once the generated features are fixed before labeling, the inverse-propensity weights make the labeled units representative of the full sample under Assumption 1. The decomposition (2), the seed-noise structure, and the spillover condition are needed for the asymptotic distribution theory and for the interpretation of the design results, but not for the validity of the moment correction. Because the researcher runs the labeling mechanism, the propensities are known by design, so no propensity-estimation step is needed.

3.5 Cross-Fitting for LLM Pipeline Tuning

In practice, the choice of prompt, the feature construction pipeline, or the selection among candidate models is often tuned using the calibration labels. For example, a researcher might compare several prompts on the labeled subset and select the one with a higher agreement rate, or fine-tune a model using the labeled data as training examples. When such tuning occurs, the LLM-generated features for the calibration set may no longer be drawn from the same distribution as those for the unlabeled data, thereby invalidating the debiasing procedure.

We address this through a cross-fitting procedure. Specifically, we partition the indices $\{1, \ldots, N\}$ into K roughly equal folds $\mathcal{F}_1, \ldots, \mathcal{F}_K$, drawn independently of the data; in practice the partition may be stratified so that labeled observations are spread evenly across folds, while the formal results of Section 4.3 treat a data-independent partition. For each fold k:

1. Use the calibration labels from all folds except k to tune the LLM pipeline.
2. Generate features $\tilde{T}_{i,m}^{(-k)}$ for units in fold k using the tuned pipeline.
3. Compute the fold-specific debiased moment $G_{\lambda,\pi}^{(k)}(\theta)$ using the units in fold k.

The cross-fitted estimator aggregates across K folds as

$$G_{\lambda,\pi}^{\mathrm{CF}}(\theta) = \frac{1}{K} \sum_{k=1}^{K} G_{\lambda,\pi}^{(k)}(\theta), \tag{6}$$

where $G_{\lambda,\pi}^{(k)}(\theta)$ is the debiased moment computed on fold k using the pipeline tuned on the remaining folds. This ensures that, within each fold, the generated features used in moment computation have the same distribution across the uncalibrated and calibration units used for debiasing.

To establish the asymptotic properties of this cross-fitted estimator, the tuning step must be stable. Section 4.3 formalizes this stability condition and shows that, under this condition, the cross-fitted estimator preserves the consistency and asymptotic normality results established for the main estimator in Section 4.

4 Inferential Theory

We now establish the asymptotic properties of the AI-PI estimator. In Section 4.1, we begin by formalizing the assumptions on the labeling mechanism and the moment conditions. Section 4.2

establishes the consistency and asymptotic normality of the proposed estimator, and Section 4.3 extends these results to the cross-fitted estimator.

4.1 Assumptions

Assumption 1 (Human labeling)**.**

1. *Let $\mathcal{Z}_N$ denote the σ-field generated by the full sample $\{(Z_j, T_j, W_j)\}_{j=1}^N$ together with all randomness used in feature generation (the seed noise and the pretraining corpus), so that all generated features $\{\tilde{T}_{j,m,s}\}_{j\le N}$ and their no-spillover counterparts are $\mathcal{Z}_N$-measurable. Conditional on $\mathcal{Z}_N$, the labeling indicators $(D_1, \ldots, D_N)$ are mutually independent with*

$$\mathbb{P}\big(D_i = 1 \,\big|\, \mathcal{Z}_N\big) = \pi(W_i).$$

In particular, conditional on W_i, the labeling indicator is independent of the data and of the generated features,

$$\mathbb{P}\big(D_i = 1 \,\big|\, Z_i,\, T_i,\, \{\tilde{T}_{i,m,s}\}_{m\le M,\, s\le S},\, W_i\big) = \mathbb{P}(D_i = 1 \mid W_i).$$

2. *Define the scaled propensity $\rho(W) = \pi(W)/r_N$. There exists a constant $0 < \rho_{\min} \le 1$ such that for all N, $\rho(W_i) \ge \rho_{\min}$ almost surely.*

Assumption 1.1 requires that the labeling selection depends only on the observable information W_i. This is satisfied when the labeling selection is based on observable characteristics such as the LLM-generated features, model disagreement, or covariates X_i, since any such information can be included in W_i. Note that the conditional independence covers the generated features: the labeling decision may depend on them, but only through W_i. This strengthening relative to a standard missing-at-random condition is useful because the debiasing terms in (5) are evaluated at generated features on the calibration set. Conditioning on the full sample in this assumption accommodates generated features that may contain asymptotically negligible spillovers. The assumption is satisfied by any scheme that, given the realized sample, draws labels independently across units with probabilities $\pi(W_i)$. Assumption 1.2 imposes a uniform overlap condition, ensuring that every unit has a non-negligible probability of being labeled relative to the overall labeling rate r_N.

Assumption 2 (Moment regularity and identification)**.**

1. *For all* $\theta \in \Theta$, $\mathbb{E}\left\|g(Z_i, T_i; \theta)\right\|^2 < \infty$ *and* $\mathbb{E}\left\|g(Z_i, \tilde{T}^0_{i,m}; \theta)\right\|^2 < \infty$ *for each* m.
2. *The population moment* $G(\theta) := \mathbb{E}[g(Z_i, T_i; \theta)]$ *is continuously differentiable in a neighborhood of* θ_0*, with positive definite Jacobian* $J_{\theta_0} := \nabla_\theta G(\theta_0) \in \mathbb{R}^{d\times d}$*, in the sense that* $x^\top J_{\theta_0} x > 0$ *for all* $x \neq 0$ *(*J_{θ_0} *is not required to be symmetric).*
3. *(Lipschitz in* θ*) There exists a neighborhood of* θ_0 *and envelopes* $L_{\theta,H}(Z_i)$*,* $\{L_{\theta,m}(Z_i)\}_{m=1}^M$ *with finite second moments such that for all* θ_1, θ_2 *in the neighborhood,*

$$\left\|g(Z_i, T_i; \theta_1) - g(Z_i, T_i; \theta_2)\right\| \leq L_{\theta,H}(Z_i)\left\|\theta_1 - \theta_2\right\|,$$

$$\left\|g(Z_i, \tilde{T}^0_{i,m}; \theta_1) - g(Z_i, \tilde{T}^0_{i,m}; \theta_2)\right\| \leq L_{\theta,m}(Z_i)\left\|\theta_1 - \theta_2\right\|.$$

4. *(Lipschitz in* T*) There exists a neighborhood of* θ_0 *and an envelope* $L_T(Z_i)$ *with* $\mathbb{E}[L_T(Z_i)^2] < \infty$ *such that for all* θ *in the neighborhood,*

$$\left\|g(Z_i, t_1; \theta) - g(Z_i, t_2; \theta)\right\| \leq L_T(Z_i)\left\|t_1 - t_2\right\|.$$

Assumption 2 collects standard conditions for moment-based estimation. Assumption 2.1 requires the moment functions to have finite second moments, ensuring the applicability of central limit theorems. Assumption 2.2 guarantees local identification of θ_0. Assumption 2.3 imposes a smoothness condition in θ, which is needed for linearization and uniform convergence. Assumption 2.4 imposes a smoothness condition in T, which is used to control the effect of the spillover term by linking perturbations in the generated feature to perturbations in the moment function.

Assumption 3. *For each* $m = 1, \ldots, M$*,*

$$U_{m,N} \coloneqq \max_{1\leq i\leq N} \left\|u_{i,m,N}(Z_1, \ldots, Z_N)\right\| = o_p(n^{-1/2}).$$

Assumption 3 formalizes the negligibility of cross-unit spillovers in the generated features. Specifically, it requires that the maximal spillover contribution vanishes faster than $n^{-1/2}$, where n is the labeling budget. This condition reflects the idea that, although the training corpus of an LLM may include the sampled units, each individual unit contributes only a negligible fraction to the overall training data, so that its influence on the generated features of other units is asymptotically negligible. As a result, the difference between the observed generated feature $\tilde{T}_{i,m}$ and its no-spillover counterpart $\tilde{T}^0_{i,m}$ is asymptotically negligible, which ensures that the spillover effects

do not affect the limiting behavior of the estimator.

The next lemma provides primitive sufficient conditions for Assumption 3 in terms of the influence of individual sampled units on the generated features, which connects the assumption to the scale of the pretraining corpus.

Lemma 1 (Primitive conditions for negligible spillovers)**.** *Fix m and suppose the feature dimension p is fixed. Assumption 3 holds for pair m under either of the following conditions.*

1. *(Bounded influence) $\|u_{i,m,N}\| \le (N-1)\,\bar{\delta}_N$ almost surely for all i, where the maximal per-unit influence satisfies $\bar{\delta}_N = o\big(n^{-1/2}N^{-1}\big)$.*
2. *(Sub-Gaussian aggregation) Each coordinate $u^{(k)}_{i,m,N}$ is sub-Gaussian with parameter σ_N, that is, $\mathbb{P}\big(|u^{(k)}_{i,m,N}| > t\big) \le 2\exp\big(-t^2/(2\sigma_N^2)\big)$ for all $i \le N$, $k \le p$, and $t > 0$, and $\sigma_N\sqrt{\log N} = o\big(n^{-1/2}\big)$.*

Lemma 1 makes the corpus-size intuition precise. Suppose the pretraining corpus contains C_N documents and each document shifts the generated feature of another unit by at most c/C_N. Without cancellation across the $N-1$ sampled peers, condition 1 requires $C_N \gg N\sqrt{n}$; if the peer influences aggregate with random signs, then $\sigma_N \asymp \sqrt{N}/C_N$ and condition 2 requires only $C_N \gg \sqrt{Nn\log N}$. Both are comfortably satisfied for web-scale corpora, whose size exceeds the sample size by many orders of magnitude.

4.2 Consistency and Asymptotic Normality

Theorem 1 (Consistency)**.** *Suppose Assumptions 1-3 hold and $n \to \infty$ as $N \to \infty$. For any fixed $\lambda = (\lambda_1, \ldots, \lambda_M)$, assume $\hat{\theta}_{\lambda,\pi}$ is the unique solution to $G_{\lambda,\pi}(\theta) = 0$ with probability tending to one. If either (i) Θ is compact and Assumptions 2.3 and 2.4 hold globally on Θ, or (ii) Θ is convex and, with probability tending to one, $G_{\lambda,\pi}(\theta)$ is monotone in the sense that $\big(G_{\lambda,\pi}(\theta_1) - G_{\lambda,\pi}(\theta_2)\big)^\top(\theta_1 - \theta_2) \ge 0$, for all $\theta_1, \theta_2 \in \Theta$, then $\hat{\theta}_{\lambda,\pi}$ is consistent, i.e., $\hat{\theta}_{\lambda,\pi} \xrightarrow{p} \theta_0$.*

Theorem 1 shows that our estimator is consistent under the proposed labeling schemes and potentially vanishing labeling rates. Building on this result, we derive its asymptotic distribution.

Theorem 2 (Asymptotic normality)**.** *Let $r_N = n/N$ and $\rho(W_i) = \pi(W_i)/r_N$, where the scaled propensity $\rho(\cdot)$ does not vary with N.*[1] *Fix $\lambda = (\lambda_1, \ldots, \lambda_M)$. Suppose Assumptions 1-3 hold and*

[1] The restriction to a fixed $\rho(\cdot)$ is notational. The proof applies verbatim to a sequence $\rho_N(\cdot) \ge \rho_{\min}$ with $\rho_N(w) \to \rho(w)$ pointwise, with $\Omega_{\lambda,\pi}$ defined at the limit ρ by dominated convergence (using $1/\rho_N \le 1/\rho_{\min}$). This covers the truncated optimal designs of Section 5, whose upper truncation r_N^{-1} varies with N.

that the consistency conditions in Theorem 1 are satisfied. Suppose further that $n \to \infty$ *and* $r_N \to r \in [0,1]$ *as* $N \to \infty$*, that for some* $\delta > 0$*,* $\mathbb{E}\,\|g(Z_i, T_i; \theta_0)\|^{2+\delta} < \infty$ *and* $\mathbb{E}\|g(Z_i, \tilde{T}^0_{i,m}; \theta_0)\|^{2+\delta} < \infty$ *for each* m*, and that the matrix* $\Omega_{\lambda,\pi}$ *defined below is positive definite. Let* $U_N \coloneqq \max_m U_{m,N}$ *and assume that* $\mathbb{E}[L_T(Z_i)^2 U_N^2] = o(1)$*. Then,*

$$\sqrt{n}\,(\hat{\theta}_{\lambda,\pi} - \theta_0) \xrightarrow{d} \mathcal{N}\Big(0,\ J_{\theta_0}^{-1}\Omega_{\lambda,\pi}J_{\theta_0}^{-\top}\Big). \tag{7}$$

The covariance matrix $\Omega_{\lambda,\pi}$ *is*

$$\Omega_{\lambda,\pi} = r\mathrm{Var}(g(Z_i, T_i; \theta_0)) + \mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r\right) R_\lambda(W_i)\right], \tag{8}$$

where $R_\lambda(W_i) = \mathbb{E}\big[e_{i,\lambda}e_{i,\lambda}^\top \mid W_i\big]$ *and* $e_{i,\lambda} = g(Z_i, T_i; \theta_0) - \sum_{m=1}^M \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)$.[2]

The asymptotic variance has a two-part decomposition. The first term, $rJ_{\theta_0}^{-1}\mathrm{Var}(g(Z_i, T_i; \theta_0))J_{\theta_0}^{-\top}$, is the irreducible variance from the labeled sample. This term exists when the labeling rate satisfies $r > 0$. The second term captures the additional design-dependent variance. It depends on the labeling rule through $\rho(W_i)$ and on the quality of the generated-moment ensemble through the conditional residual covariance $R_\lambda(W_i)$, which measures the discrepancy between the true moment and the weighted generated moment.

When the GenAI ensemble perfectly recovers the true moment, in the sense that $g(Z_i, T_i; \theta_0) = \sum_m \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)$ almost surely, we have $R_\lambda(W_i) = 0$ and the second component vanishes. In that case, the asymptotic variance equals the oracle variance from using all N labeled features. Conversely, when the generated features are uninformative, $R_\lambda(W_i)$ is large and the asymptotic variance is dominated by the calibration sample and its allocation across the sample space.

Remark 1 (Variance estimation and confidence intervals)**.** *Theorem 2 yields asymptotically valid confidence intervals once the asymptotic variance is estimated. We use the plug-in estimators*

$$\hat{\Omega}_{\lambda,\pi} = r_N \cdot \frac{1}{N}\sum_{i=1}^N \frac{D_i}{\pi_i}\,\hat{g}_i\hat{g}_i^\top + \frac{1}{N}\sum_{i=1}^N \frac{D_i}{\pi_i}\left(\frac{1}{\rho(W_i)} - r_N\right)\hat{e}_i\hat{e}_i^\top, \qquad \hat{J} = \nabla_\theta G_{\lambda,\pi}(\theta)\big|_{\theta=\hat{\theta}_{\lambda,\pi}},$$

where $\hat{g}_i = g(Z_i, T_i; \hat{\theta}_{\lambda,\pi})$ *and* $\hat{e}_i = \hat{g}_i - \sum_{m=1}^M \lambda_m\, g(Z_i, \tilde{T}_{i,m}; \hat{\theta}_{\lambda,\pi})$*. Both averages are inverse-propensity weighted over the calibration set, so they are computable:* $\hat{g}_i$ *and* $\hat{e}_i$ *require the true feature* T_i*, which is observed for units with* $D_i = 1$*. Proposition 1 below establishes* $\hat{\Omega}_{\lambda,\pi} \xrightarrow{p} \Omega_{\lambda,\pi}$

[2]Here $R_\lambda(W_i)$ is written in terms of the generated features $\tilde{T}_{i,m}$. Since $\tilde{T}_{i,m}$ contains the spillover term, $R_\lambda(W_i)$ may depend on N; the displayed covariance is the limit of the corresponding covariance sequence.

and $\hat{J} \xrightarrow{p} J_{\theta_0}$ *under explicit envelope conditions, so that the intervals are asymptotically exact. For a scalar component* $\theta_{0,j}$*, a* $(1-\alpha)$ *confidence interval is*

$$\hat{\theta}_j \pm z_{\alpha/2} \cdot \hat{\sigma}_j / \sqrt{n}, \qquad \hat{\sigma}_j^2 = \left[\hat{J}^{-1} \hat{\Omega}_{\lambda,\pi} \hat{J}^{-\top}\right]_{jj}.$$

The simulation study in Section 8 confirms that these intervals attain close to nominal coverage in finite samples, while naive confidence intervals that treat the generated features as observed collapse to zero coverage.

Proposition 1 (Consistency of the variance estimator)**.** *Suppose the conditions of Theorem 2 hold and, in addition, the envelopes of Assumption 2 satisfy* $\mathbb{E}[L_{\theta,H}(Z_i)^{2+\delta}] < \infty$*,* $\mathbb{E}[L_{\theta,m}(Z_i)^{2+\delta}] < \infty$ *for each* m*, and* $\mathbb{E}[L_T(Z_i)^{2+\delta}] < \infty$*. Then* $\hat{\Omega}_{\lambda,\pi} \xrightarrow{p} \Omega_{\lambda,\pi}$*. If, moreover,* $g(z,t;\theta)$ *is continuously differentiable in* θ *on a neighborhood* Θ_1 *of* θ_0 *with envelopes* $\sup_{\theta\in\Theta_1} \|\nabla_\theta g(Z_i, T_i;\theta)\| \le M_H(Z_i)$ *and* $\sup_{\theta\in\Theta_1} \left\|\nabla_\theta g(Z_i, \tilde{T}^0_{i,m};\theta)\right\| \le M_m(Z_i)$ *with* $\mathbb{E}[M_H(Z_i)^2] < \infty$ *and* $\mathbb{E}[M_m(Z_i)^2] < \infty$*, and* $\|\nabla_\theta g(z,t_1;\theta) - \nabla_\theta g(z,t_2;\theta)\| \le L_{T,\nabla}(Z_i)\,\|t_1 - t_2\|$ *on* Θ_1 *with* $\mathbb{E}[L_{T,\nabla}(Z_i)^2] < \infty$*, then* $\hat{J} \xrightarrow{p} J_{\theta_0}$*. Consequently,* $\hat{\sigma}_j^2 \xrightarrow{p} \left[J_{\theta_0}^{-1} \Omega_{\lambda,\pi} J_{\theta_0}^{-\top}\right]_{jj}$ *for each* j*, and the confidence intervals of Remark 1 have asymptotic coverage* $1-\alpha$*.*

The proof rests on a weak law of large numbers for inverse-propensity-weighted triangular arrays (Lemma A.1 in the Appendix) that requires only $(1+\delta/2)$ moments of the unweighted summands and is therefore compatible with vanishing labeling rates.

This decomposition also highlights two distinct sources for improving estimation efficiency. First, the weights λ_j can be chosen to reduce the residual covariance $R_\lambda(W_i)$. Second, the labeling policy $\pi(W_i)$ can be chosen to allocate human labels to regions of W_i in which this residual covariance is large. In Section 5, we study the joint choice of λ and $\pi(\cdot)$ that minimizes the resulting asymptotic variance.

4.3 Cross-Fitting with Pipeline Tuning: Formal Guarantees

We now formalize the validity of the cross-fitting procedure of Section 3.5. Recall that the indices are partitioned into K folds $\mathcal{F}_1, \ldots, \mathcal{F}_K$ (we take N/K to be an integer for notational simplicity), and that the features for units in fold k are generated by a pipeline tuned on the data $\mathcal{D}_{-k}$ outside fold k, yielding $\tilde{T}^{(-k)}_{i,m}$ for $i \in \mathcal{F}_k$. The idea of the formal analysis is that, as the tuning data grow, the tuned pipeline settles down: the features it produces converge to those of a *limit pipeline*. When the tuning step selects a prompt from a finite menu, a consistent selection rule picks the

same prompt with probability approaching one, so the tuned features eventually coincide with the limit features; when the pipeline has continuous tuning parameters, such as a retrieval weighting or a temperature, convergence of the tuned parameters translates into convergence of the features whenever the features are continuous in the pipeline parameters.

Assumption 4 (Pipeline stability)**.** *There exist limit features $\tilde{T}^*_{i,m}$, generated unit-wise by a fixed pipeline, such that:*

1. *The limit features admit the decomposition (2) and satisfy Assumptions 2 and 3.*
2. *The labeling information W_i and the propensity $\pi(W_i)$ do not depend on the fold-specific tuned features.*
3. *(Cross-Lipschitz condition) There exist a neighborhood of θ_0 and an envelope $L_{T,\theta}(Z_i)$ with $\mathbb{E}[L_{T,\theta}(Z_i)^2] < \infty$ such that for all θ in the neighborhood and all t_1, t_2,*

$$\left\| \left[g(Z_i, t_1; \theta) - g(Z_i, t_2; \theta) \right] - \left[g(Z_i, t_1; \theta_0) - g(Z_i, t_2; \theta_0) \right] \right\| \leq L_{T,\theta}(Z_i) \, \|t_1 - t_2\| \, \|\theta - \theta_0\| .$$

4. *(Feature stability) With $\bar{L}(Z_i) \coloneqq \max\{L_T(Z_i),\, L_{T,\theta}(Z_i)\}$,*

$$\xi_N \coloneqq \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \mathbb{E}\left[\bar{L}(Z_i)^2 \left\| \tilde{T}^{(-k)}_{i,m} - \tilde{T}^*_{i,m} \right\|^2 \,\middle|\, \mathcal{D}_{-k} \right] = o_p(1),$$

where the conditional expectation is over the distribution of a unit $i \in \mathcal{F}_k$ given the tuning data.

Assumption 4.4 requires only consistency of the tuned features in a weighted L_2 sense, with no rate: cross-fitting converts consistency of the pipeline into validity of the inference. Assumption 4.3 holds, for example, for least-squares moments and other moment functions that are polynomial in the feature, on bounded neighborhoods, with the natural envelopes.

Theorem 3 (Cross-fitted AI-PI)**.** *Suppose Assumptions 1, 2, 3 and the conditions of Theorems 1 and 2 hold with the limit features $\tilde{T}^*_{i,m}$ in place of $\tilde{T}_{i,m}$, the number of folds K is fixed, and Assumption 4 holds. Let $\hat{\theta}^{\mathrm{CF}}_{\lambda,\pi}$ denote the solution of $G^{\mathrm{CF}}_{\lambda,\pi}(\theta) = 0$ in (6), assumed unique with probability tending to one. Then $\hat{\theta}^{\mathrm{CF}}_{\lambda,\pi} \xrightarrow{p} \theta_0$ and*

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

where $\Omega^*_{\lambda,\pi}$ is the covariance matrix of Theorem 2 evaluated at the limit features.

The proof compares the cross-fitted moment to the moment that would be obtained from a fixed stable pipeline. The stability condition makes the difference between these two moments asymptotically negligible, so the cross-fitted estimator inherits the consistency and asymptotic normality of the main estimator.

5 Optimization

In this section, we derive the optimal choices for the labeling policy $\pi(\cdot)$ and the model–prompt weights λ based on the asymptotic variance in Theorem 2. Throughout this section, we let $\Lambda \subset \mathbb{R}^M$ denote a compact convex set of model–prompt weights and assume $0 \in \Lambda$. In applications, a convenient stable choice is $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_m \geq 0,\ \sum_{m=1}^M \lambda_m \leq 1\}$, which allows the generated moments to be shrunk toward zero when they are uninformative.

5.1 Optimal Labeling Selection for Fixed Weights

We first study the problem of choosing which observations to label by selecting the propensity function $\pi(\cdot)$ for a given weight vector $\lambda \in \Lambda$. In principle, one can choose to optimize different criteria based on the asymptotic covariance matrix, for example, the total variance, the variance of a subset of parameters of interest, or other criteria including eigenvalue-based measures. In this paper, we focus on minimizing the total asymptotic variance.

Specifically, the objective is to minimize the total asymptotic variance

$$\begin{aligned} V(\lambda, \pi) &\coloneqq \operatorname{tr}\big(J_{\theta_0}^{-1} \Omega_{\lambda,\pi} J_{\theta_0}^{-\top}\big) \\ &= r \cdot \operatorname{tr}\Big(J_{\theta_0}^{-1} \operatorname{Var}(g(Z_i, T_i; \theta_0)) J_{\theta_0}^{-\top}\Big) + \mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r\right) \operatorname{tr}\big(J_{\theta_0}^{-1} R_\lambda(W_i) J_{\theta_0}^{-\top}\big)\right] \end{aligned}$$

subject to the budget constraint $\mathbb{E}[\pi(W_i)] = r_N$.

Assumption 5 (Design feasibility)**.** *For a given weight vector λ, the residual discrepancy satisfies $\mathbb{P}\big(\operatorname{tr}(J_{\theta_0}^{-1} R_\lambda(W_i) J_{\theta_0}^{-\top}) > 0\big) > r_N$.*

Assumption 5 rules out the degenerate case in which the weighted generated moment already coincides with the true moment on all but a set smaller than the labeling budget. In that case, the labeling allocation is not uniquely determined; otherwise, the assumption ensures that the budget can be meaningfully targeted toward regions with positive residual discrepancy.

Theorem 4 (Optimal labeling propensity). *For a fixed $\lambda \in \Lambda$, under Assumption 5, the problem*

$$\min_{\pi(\cdot)\in(0,1]} V(\lambda, \pi)$$
$$s.t. \quad \mathbb{E}[\pi(W_i)] = r_N, \quad \rho_{\min} r_N \leq \pi(W_i) \leq 1$$

admits an oracle solution of the form

$$\pi^*(w \mid \lambda) = \max\left\{\rho_{\min} r_N,\ \min\left\{1,\ c_\lambda \sqrt{\mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(w) J_{\theta_0}^{-\top})}\right\}\right\}, \tag{9}$$

where $c_\lambda \geq 0$ is chosen to satisfy $\mathbb{E}[\pi^*(W_i \mid \lambda)] = r_N$.

Theorem 4 shows that, when neither truncation binds, the optimal labeling probability $\pi^*(w \mid \lambda)$ is proportional to the "residual score" $s_\lambda(w) = \sqrt{\mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(w) J_{\theta_0}^{-\top})}$, which measures the discrepancy between the weighted generated moment and the true moment at the point w. Regions where the generated moments deviate more from the truth receive higher labeling probability. This is intuitive: the labeling budget should be concentrated where the LLM is least reliable, because that is where human labels provide the most information. The lower bound on the labeling probability aligns with the uniform overlap assumption in Assumption 1.2.

Remark 2 (Targeting a scalar functional). *If the object of interest is a linear functional $a^\top \theta_0$ for a known vector a, for example a single coefficient in a regression with many controls and fixed effects, all results in this section apply with the criterion $a^\top J_{\theta_0}^{-1} \Omega_{\lambda,\pi} J_{\theta_0}^{-\top} a$: Theorem 4 holds with the residual score $s_\lambda(w) = \left(a^\top J_{\theta_0}^{-1} R_\lambda(w) J_{\theta_0}^{-\top} a\right)^{1/2}$, since the objective remains affine in $1/\rho(\cdot)$, and the weight problem in Section 5.2 remains a convex quadratic program. The trace criterion targets the full parameter vector; targeting the coefficient of interest alone tilts the labeling budget toward observations that are informative about that coefficient.*

Remark 3 (Heterogeneous labeling costs). *The budget constraint counts labeled units, but labeling costs may differ across units; in our application, the cost of labeling a stock-month is proportional to its number of headlines. With unit costs $c(W_i)$ and the budget constraint $\mathbb{E}[c(W_i)\pi(W_i)] = B$, the argument of Theorem 4 applies with the pointwise objective $q_\lambda(w)/x + \mu\, c(w)\, x$, and the optimal policy becomes proportional to $s_\lambda(w)/\sqrt{c(w)}$, truncated as before. Costly units must therefore promise proportionally more residual-score reduction to be labeled.*

In Section 5.2, we use the optimal labeling policy characterized in Theorem 4 to study the

resulting optimal ensemble weights.

5.2 Optimal Weights Selection

For a given labeling design π, minimizing $V(\lambda, \pi)$ over $\lambda \in \Lambda$ is equivalent to the weighted quadratic criterion

$$\min_{\lambda \in \Lambda} \mathbb{E}\left[\left(\frac{r_N}{\pi(W)} - r\right)(g_i - \tilde{g}_i\lambda)^\top J_{\theta_0}^{-\top} J_{\theta_0}^{-1}(g_i - \tilde{g}_i\lambda)\right], \tag{10}$$

where $g_i = g(Z_i, T_i; \theta_0)$ and $\tilde{g}_i = [g(Z_i, \tilde{T}_{i,1}; \theta_0), \ldots, g(Z_i, \tilde{T}_{i,M}; \theta_0)]$.

We now characterize the joint optimum by substituting $\pi^*(\cdot \mid \lambda)$ from Theorem 4 into the total asymptotic variance. Define the objective

$$\widetilde{V}(\lambda) := V(\lambda, \pi^*(\cdot \mid \lambda)).$$

The optimal weights balance two effects. First, they make the GenAI-aggregated moment approximate the human moment in the metric induced by $J_{\theta_0}^{-\top} J_{\theta_0}^{-1}$, emphasizing regions with smaller sampling probability $\pi(W)$. Second, they affect the objective indirectly through the dependence of the optimal policy $\pi^*(\cdot \mid \lambda)$ on λ.

Proposition 2 characterizes the first-order condition for optimal weights λ^*.

Proposition 2 (First-order condition for λ^*)**.** *Suppose Assumptions 2.1 and 5 hold. Let $\lambda^* \in \mathrm{int}(\Lambda)$ be a local minimizer of the objective $\widetilde{V}(\lambda)$. Assume that, for all λ in a neighborhood of λ^*, the normalizing constant c_λ is well defined, $\mathbb{P}(c_\lambda s_\lambda(W) = 1) = \mathbb{P}(c_\lambda s_\lambda(W) = \rho_{\min} r_N) = 0$, and differentiation under the expectation is permitted. Then λ^* satisfies*

$$\mathbb{E}\left[\left(\frac{r_N}{\pi^*(W_i \mid \lambda^*)} - r\right)\tilde{g}_i^\top J_{\theta_0}^{-\top} J_{\theta_0}^{-1}\left(g(Z_i, T_i; \theta_0) - \sum_{m=1}^{M} \lambda_m^* g(Z_i, \tilde{T}_{i,m}; \theta_0)\right)\right] = 0, \tag{11}$$

where $\pi^*(\cdot \mid \lambda)$ is defined in Theorem 4.

In Section 5.3, we propose an alternating algorithm to compute (λ^*, π^*). For fixed π, the λ-update in the alternating algorithm solves the sample analogue of the weighted quadratic criterion in (10). When the solution is interior, its population first-order condition is given by (11); for boundary solutions, the corresponding KKT conditions replace the equality.

The class of estimators can be enlarged by letting the weights depend on the labeling information: replace λ_m by measurable functions $\lambda_m(W_i)$ with $\lambda(W_i) \in \Lambda$ in (5). The next result shows that the asymptotic theory carries over and that, in this larger class, the joint design problem separates.

Proposition 3 (W-dependent weights and separation of the joint design)**.** *Let $\lambda\colon \mathrm{supp}(W) \to \Lambda$ be measurable, and let $\hat{\theta}_{\lambda(\cdot),\pi}$ solve (5) with $\lambda_m(W_i)$ in place of λ_m. Let $q_\lambda(w) = \mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(w) J_{\theta_0}^{-\top})$.*

(a) Under the conditions of Theorems 1 and 2, $\hat{\theta}_{\lambda(\cdot),\pi}$ is consistent and

$$\sqrt{n}\big(\hat{\theta}_{\lambda(\cdot),\pi} - \theta_0\big) \xrightarrow{d} \mathcal{N}\Big(0,\ J_{\theta_0}^{-1}\,\Omega_{\lambda(\cdot),\pi}\,J_{\theta_0}^{-\top}\Big),$$

where

$$\Omega_{\lambda(\cdot),\pi} = r\mathrm{Var}(g(Z_i, T_i; \theta_0)) + \mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r\right) R_{\lambda(W_i)}(W_i)\right].$$

(b) A measurable pointwise minimizer $\lambda^*(w) \in \arg\min_{\lambda\in\Lambda} q_\lambda(w)$ exists. For every admissible design π, $\lambda^*(\cdot)$ minimizes $V(\lambda(\cdot), \pi) = \mathrm{tr}(J_{\theta_0}^{-1}\Omega_{\lambda(\cdot),\pi}J_{\theta_0}^{-\top})$ over all measurable weight functions; in particular, it is weakly more efficient than every constant $\lambda \in \Lambda$.

(c) If $\mathbb{P}\big(q^*(W_i) > 0\big) > r_N$ for the profiled score $q^*(w) = \min_{\lambda\in\Lambda} q_\lambda(w)$, the jointly optimal design is $\big(\lambda^*(\cdot), \pi^*\big)$, where π^* is the policy of Theorem 4 applied to the residual score $\sqrt{q^*(w)}$: the joint problem separates, and no alternation between policy and weight updates is needed.

W-dependent weights are the analogue of the group-specific weights in Duan and Pelger (2025) and quantify the efficiency cost of the constant-weight baseline. The practical price is that $\lambda^*(\cdot)$ must be estimated as a function on the support of W rather than as M scalars, which is demanding exactly when the calibration budget is small; we therefore use constant weights in our baseline implementation and view the profiled-score design as the natural refinement when the pilot sample is large.

Remark 4 (Relation to the semiparametric efficiency bound)**.** *For a fixed labeling design, the efficient estimator in the corresponding missing-data problem augments the inverse-propensity-weighted human moment with the conditional expectation of the true moment given all observables, as in augmented inverse-propensity weighting (Robins, Rotnitzky, and Zhao, 1994) and related semi-supervised estimators (Chakrabortty and Cai, 2018). AI-PI uses a more structured augmentation: it approximates this conditional expectation by a linear combination of generated moments, with constant weights in the baseline and W-dependent weights in Proposition 3. The resulting estimator is optimal within this generated-moment class. This restriction is deliberate: with modest calibration*

samples, estimating a small number of model–prompt weights, or a low-dimensional weight function, is more stable than estimating a fully nonparametric conditional expectation, while preserving the exact centering of the moment correction under flexible generator errors. A full characterization of the semiparametric efficiency bound under the labeling design is left for future work.

5.3 Feasible Joint Optimization

Sections 5.1 and 5.2 show that the optimal labeling policy π^* depends on the weight vector λ through R_λ, and the optimal weight vector λ^* depends on π through the weighting. This gives rise to a joint optimization problem. However, the oracle updates are not directly feasible because they depend on the unobserved true moment $g(Z_i, T_i; \theta_0)$, the Jacobian J_{θ_0}, and the conditional residual covariance $R_\lambda(W_i)$. To address this challenge, we approximate the oracle joint optimum using a feasible two-stage procedure with alternating updates. In the first stage, we use an initial calibration sample to estimate the unobserved components, and in the second stage, we allocate the remaining labeling budget according to the optimized labeling policy. The resulting procedure is summarized in Algorithm 1.

Algorithm 1 (AI-PI Algorithm)**.**

1. ***Stage 1 (Initial calibration):*** *Draw an initial calibration sample with expected size n_0 using a simple design $\pi_i^{(0)}$ (e.g., MCAR), and observe the true features T_i for units with $D_i^{(0)} = 1$.*
2. ***Preliminary estimation:*** *Using the labeled units from Stage 1, estimate the quantities required for policy and weights optimization, including θ_0, the moments $g(Z_i, T_i; \theta_0)$, the Jacobian J_{θ_0}, and the conditional residual covariance $R_\lambda(W_i)$.*
3. ***Alternating updates:*** *Given preliminary estimates, iteratively update the labeling policy and the ensemble weights. For $k = 0, 1, 2, \ldots$,*
   - ***(a) Design update:*** *Estimate $\mathrm{tr}(J_{\theta_0}^{-1} R_{\lambda^{(k)}}(W_i) J_{\theta_0}^{-\top})$ from the current data and set $\pi^{(k+1)}(W_i)$ according to Equation (9).*
   - ***(b) Ensembling update:*** *Given $\pi^{(k+1)}$, update λ by solving the sample analogue of (10).*

   Let $\hat{\pi}(W_i)$ and $\hat{\lambda}$ denote the converged labeling propensity and weights.
4. ***Stage 2 (Adaptive labeling):*** *For units not labeled in Stage 1, draw an additional calibration sample with expected size $n - n_0$ using the optimized labeling policy. Specifically, assign labels*

with conditional probability[3]

$$\hat{a}_i = \frac{\hat{\pi}(W_i) - \pi^{(0)}(W_i)}{1 - \pi^{(0)}(W_i)}.$$

Let $D_i^{(1)}$ *denote this second-stage labeling indicator. The final labeling indicator is*

$$D_i = D_i^{(0)} + (1 - D_i^{(0)})D_i^{(1)}.$$

5. ***Final AI-PI estimation:*** *Compute the final estimator by solving the debiased moment equation* $G_{\hat{\lambda},\hat{\pi}}(\theta) = 0$*, where* $G_{\lambda,\pi}(\theta)$ *is defined in* (5).

The following proposition formalizes the asymptotic optimality of this plug-in procedure.

Algorithm 1 replaces the oracle components by estimates obtained from an initial calibration sample and then computes an estimated weight-policy pair $(\widehat{\lambda}, \widehat{\pi})$ through alternating updates.

The following proposition shows that this feasible plug-in procedure attains, to first order, the oracle asymptotic variance. In other words, provided the estimated variance criterion and the estimated labeling propensities are uniformly consistent, using the data-driven design does not increase the total asymptotic variance relative to the infeasible oracle benchmark.

Proposition 4 (Asymptotic optimality of the feasible joint design)**.** *Let* $r_N = n/N \to r \in [0,1]$*, let* $\lambda^* \in \arg\min_{\lambda\in\Lambda} \widetilde{V}(\lambda)$ *(assumed to exist), and let* $\pi^* = \pi^*(\cdot \mid \lambda^*)$*. For each* $\lambda \in \Lambda$*, define* $\hat{\pi}_\lambda(w) = \hat{\pi}(w \mid \lambda)$*,* $\hat{\rho}_\lambda(w) = \hat{\pi}_\lambda(w)/r_N$*, and suppose* $\hat{\rho}_\lambda(W) \geq \rho_{\min}$ *almost surely for all* $\lambda \in \Lambda$*. Let* $(\widehat{\lambda}, \widehat{\pi})$ *be the weight and policy optimized by Algorithm 1. Suppose that the estimated variance satisfies* $\sup_{\lambda\in\Lambda} |\widehat{V}(\lambda, \widehat{\pi}_\lambda) - V(\lambda, \widehat{\pi}_\lambda)| = o_p(1)$*. Furthermore, assume* $(\widehat{\lambda}, \widehat{\pi})$ *satisfies* $\widehat{V}(\widehat{\lambda}, \widehat{\pi}) \leq \inf_{\lambda\in\Lambda} \widehat{V}(\lambda, \widehat{\pi}_\lambda) + o_p(1)$ *and* $\sup_{\lambda\in\Lambda} \mathbb{E}\left[(\widehat{\rho}_\lambda(W) - \rho^*_\lambda(W))^2\right]^{1/2} = o_p(1)$*. Assume that there exists* $\overline{q}(W)$ *such that* $0 \leq \mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W_i) J_{\theta_0}^{-\top}) \leq \overline{q}(W)$ *for all* $\lambda \in \Lambda$*, and* $\mathbb{E}[\overline{q}(W)^2] < \infty$*. Then,*

$$V(\widehat{\lambda}, \widehat{\pi}) = V(\lambda^*, \pi^*) + o_p(1).$$

Remark 5 (Computation)**.** *For a fixed policy, the weight update is a convex quadratic program, but the profiled objective obtained by substituting the optimal policy need not be convex in* λ*. The alternating updates of Algorithm 1 should therefore be interpreted as a practical procedure for seeking a stationary point, and the near-minimization condition in Proposition 4 is accordingly a condition on the computational output. In practice, we recommend multiple starting values, including uniform*

[3]This conditional probability is well defined provided $\pi^{(0)}(W_i) \leq \hat{\pi}(W_i)$. A sufficient condition is $\pi^{(0)}(W_i) \leq \rho_{\min} r_N$ for all i because $\hat{\pi}(W_i) \geq \rho_{\min} r_N$ by construction.

weights and the best single model–prompt pair. With W-dependent weights, the oracle joint problem separates by Proposition 3(c), and the pointwise weight problem is convex, so the profiled-score design avoids this profiling nonconvexity.

5.4 Inference with the Estimated Design

Proposition 4 shows that the feasible design attains the oracle asymptotic variance. We now establish asymptotic normality for the estimator computed with the estimated weights and adaptively collected labels. For clarity, we first analyze a split version of Algorithm 1: the design $(\hat{\lambda}, \hat{\pi})$ is estimated from the pilot data, while the final moment is computed only on units outside the pilot. We then show in Proposition 5 that the pooled implementation stated in Algorithm 1 is asymptotically equivalent under a stronger pilot-size condition.

Formally, let $D_i^{(0)} \sim \text{Bernoulli}(\pi_N^{(0)})$, i.i.d. and independent of all data, with $\pi_N^{(0)} = n_0/N$, denote the pilot labeling. Let $\mathcal{S}$ be the σ-field generated by the pilot indicators $\{D_i^{(0)}\}_{i=1}^N$ and the pilot units' data, and let the estimated design $(\hat{\lambda}, \hat{\pi})$ be $\mathcal{S}$-measurable with $\hat{\lambda} \in \Lambda$ and $\hat{\pi}(w) \in [\rho_{\min} r_N, 1]$. Units outside the pilot are labeled in the second stage with conditional probability

$$\hat{a}(W_i) \;=\; \frac{\hat{\pi}(W_i) - \pi_N^{(0)}}{1 - \pi_N^{(0)}},$$

independently across units given $(\mathcal{S}, \{W_j\}_{j=1}^N)$; this is well defined and positive for all large N because $\pi_N^{(0)} = o(r_N)$ under condition (i) below. The two-stage estimator $\hat{\theta}^{\text{TS}}$ solves the analogue of moment condition (5) computed on the non-pilot units, with weights $\hat{\lambda}$, labels $D_i^{(1)}$, and propensities $\hat{a}(W_i)$.

Theorem 5 (Asymptotic normality with the estimated design)**.** *Suppose the conditions of Theorem 2 hold at $(\lambda^*, \pi^*(\cdot \mid \lambda^*))$, including the $(2+\delta)$-moment condition and positive definiteness of Ω_{λ^*,π^*}, with $r_N \to r \in [0,1)$. Assume in addition:*

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

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

(iii) $\mathbb{E}_W\big[(\hat{\rho}(W) - \rho^*(W))^2\big] \xrightarrow{p} 0$*, where* $\hat{\rho} = \hat{\pi}/r_N$ *and* $\rho^* = \pi^*(\cdot \mid \lambda^*)/r_N$*, and there exists an envelope* $\bar{q}(W)$ *with* $\mathbb{E}[\bar{q}(W)^2] < \infty$ *such that* $\sup_{\lambda\in\Lambda} \operatorname{tr} R_\lambda(W) \le \bar{q}(W)$ *and* $\mathbb{E}\big[\, \|g(Z_i, T_i; \theta_0)\|^2 + \max_m \big\|g(Z_i, \tilde{T}_{i,m}; \theta_0)\big\|^2 \;\big|\; W\big] \le \bar{q}(W)$.

Then

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

In particular, combined with Proposition 4, the feasible two-stage procedure with a vanishing-fraction pilot is jointly asymptotically normal and *attains the oracle asymptotic variance* $V(\lambda^*, \pi^*)$.

Conditions (ii) and (iii) are exactly the consistency requirements on the estimated design that already underlie Proposition 4; with $n_0 \to \infty$ pilot labels they are attainable by the plug-in estimates of Algorithm 1. The dependence created by adaptive labeling is handled by conditioning: given the pilot data, the second-stage labels are an independent Bernoulli design with a known ($\mathcal{S}$-measurable) propensity function, so the fixed-design theory applies conditionally, and the conditional asymptotic variance converges to the oracle variance.

Remark 6 (Pooling the pilot labels)**.** *In practice it is natural to pool the pilot labels into the final calibration moment, as stated in Algorithm 1: the final labels are* $D_i = D_i^{(0)} + (1 - D_i^{(0)})D_i^{(1)}$ *with propensity* $\hat{\pi}(W_i) = \pi_N^{(0)} + (1 - \pi_N^{(0)})\,\hat{a}(W_i)$ *by construction. The complication is that the pilot units' moments are correlated with the estimated design* $(\hat{\lambda}, \hat{\pi})$*, so exact conditional centering fails for the pooled moment. Proposition 5 shows that this self-influence is negligible when the pilot grows more slowly than* $\sqrt{n}$.

Proposition 5 (Pooled two-stage estimator)**.** *Let* $\hat{\theta}^{\mathrm{P}}$ *solve the moment condition (5) computed on all units with the pooled labels* $D_i = D_i^{(0)} + (1 - D_i^{(0)})D_i^{(1)}$*, propensities* $\hat{\pi}(W_i)$*, and weights* $\hat{\lambda}$*. Suppose the conditions of Theorem 5 hold and, in addition, (iv)* $n_0 = o(\sqrt{n}\,)$*, and (v)* $\hat{\rho}(w) \le \bar{\rho}$ *and* $\rho^*(w) \le \bar{\rho}$ *for a constant* $\bar{\rho} < \infty$*. Then*

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

Condition (iv) covers pilots that grow to infinity but are a vanishing fraction of $\sqrt{n}$ (for example, $n_0 \asymp n^{1/3}$); condition (v) rules out designs that concentrate the entire budget on a vanishing set, and holds, for example, whenever the residual score is bounded on the support of W. For larger pilots, analyzing the pooled estimator would require additional stability conditions controlling the dependence between the pilot observations and the estimated design. The simulations in Section 8 suggest that the pooled implementation can perform well with larger pilot shares in finite samples.

6 Extensions

6.1 Noisy Reference Labels and Model-Assisted Calibration

Our baseline framework treats the human labels T_i on the calibration set as a gold standard. In practice, however, reference labels may themselves be imperfect: human annotators can make mistakes, and researchers may also use a stronger, more expensive generative model to provide auxiliary labels on a subset of units while running cheaper models on the full sample. Both imperfect human labels and model-generated auxiliary labels are covered by the same extension, in which the calibration set reveals a *noisy reference label*

$$T_i^* = T_i + e_i,$$

rather than the true feature, where e_i denotes the reference-label error.

The central question is what the debiased estimator identifies when T_i is replaced by T_i^* in the moment estimator (5). The construction of Section 3.4 removes the discrepancy between the generated moments and the reference moments. The natural target is therefore the reference estimand θ_0^*, defined as a solution of $\mathbb{E}[g(Z_i, T_i^*; \theta)] = 0$. When T_i^* is close to T_i, this estimand is close to the original target θ_0; when the reference shares systematic errors with the cheap generators, such as look-ahead bias from overlapping training corpora, the extension delivers inference for θ_0^* rather than full identification of θ_0. The following assumption regulates the reference errors.

Assumption 6 (Reference labels)**.** *On the calibration set the researcher observes $T_i^* = T_i + e_i$. Let Θ_0 be a closed ball around θ_0 contained in the neighborhood of Assumption 2.2.*

1. *(Conditionally centered errors) $\mathbb{E}[e_i \mid Z_i, T_i] = 0$, and $\mathbb{E}\big[\sup_{\theta\in\Theta_0} \|\partial_t\, g(Z_i, T_i; \theta)\|\ \|e_i\|\big] < \infty$.*
2. *(Smoothness and domination) For every z, the map $t \mapsto g(z, t; \theta)$ is twice continuously differentiable, and there exists $\nu(Z_i, e_i) \geq 0$ such that*

$$\sup_{\theta\in\Theta_0}\ \sup_{u\in[0,1]}\ \big\|\partial_t^2\, g\big(Z_i,\, T_i + u\, e_i;\, \theta\big)\big\| \leq \nu(Z_i, e_i), \qquad \kappa^2 := \mathbb{E}\big[\nu(Z_i, e_i)\ \|e_i\|^2\big] < \infty.$$

Proposition 6 (Noisy reference labels)**.**

(a) (Validity for the reference estimand) Suppose Assumptions 1–3 and the conditions of Theorems 1 and 2 hold with (T_i, θ_0) replaced by (T_i^*, θ_0^*). Then all conclusions of Theorems 1, 2, and 4 hold for θ_0^*, with T_i^* in place of T_i throughout.

(b) Suppose Assumptions 2.2 and 6 hold. There exist constants $c, \bar{\kappa} > 0$, depending only on the population moment G, such that whenever $\kappa^2 \le \bar{\kappa}$, a solution θ_0^* of $\mathbb{E}[g(Z_i, T_i^*; \theta)] = 0$ exists in a neighborhood of θ_0, and every solution in this neighborhood satisfies

$$\|\theta_0^* - \theta_0\| \;\le\; c\,\kappa^2.$$

In particular, if $t \mapsto g(z, t; \theta)$ is affine for all (z, θ) and Assumption 6.1 holds, then $\theta_0^* = \theta_0$.

Proposition 6 formalizes the noisy-reference logic. When the reference labels are noisy but conditionally centered, the gap between the reference estimand and the target estimand θ_0 is second order: by part (b), it scales with the conditional variance of the reference error. By contrast, the calibration step removes the first-order discrepancy between the cheap generators and the reference moments. Thus, if the reference labels are much more accurate than the cheap generators, calibrating to the reference can remove most of the distortion from cheap generated features while leaving only a small gap between θ_0^* and θ_0. This motivates a practical use of strong models: a stronger, more expensive model can be queried on an intermediate calibration sample and used as a noisy reference to debias and weight cheap models on the full sample; by part (a), the resulting estimator delivers valid inference for the reference estimand θ_0^*. What no calibration scheme can deliver is protection against bias that the reference shares with the generators, for example a look-ahead bias common to all models trained on the same corpus. In this precise sense, a gold-standard calibration set remains the legitimacy criterion for full identification of θ_0, while noisy reference labels deliver valid inference on the well-defined reference estimand.

Remark 7. *For the least-squares moment $g(Z_i, t; \theta) = X_i^*(t)\,(Y_i - X_i^*(t)^\top \theta)$ with $X_i^*(t) = (X_i^\top, t)^\top$ and a scalar feature, the second derivative $\partial_t^2 g$ does not depend on t, and the bias admits a closed form. With $\mathbb{E}[e_i \mid Z_i, T_i] = 0$ and $\sigma_e^2 = \mathbb{E}[e_i^2]$,*

$$\mathbb{E}[g(Z_i, T_i^*; \theta)] = \mathbb{E}[g(Z_i, T_i; \theta)] - \sigma_e^2\, S\,\theta, \qquad S = \mathrm{diag}(0, \ldots, 0, 1),$$

so $\theta_0^* = \big(\Sigma_{X^*} + \sigma_e^2 S\big)^{-1} \Sigma_{X^*}\,\theta_0$ with $\Sigma_{X^*} = \mathbb{E}[X_i^* X_i^{*\top}]$: the coefficient on the feature is attenuated by the factor $\sigma_{T|X}^2 / (\sigma_{T|X}^2 + \sigma_e^2)$, where $\sigma_{T|X}^2$ is the residual variance of T_i after projecting on X_i, consistent with the $O(\kappa^2)$ bound of Proposition 6(b). Two practical consequences follow. First, the attenuation is second order when the reference labels are accurate, $\sigma_e^2 \ll \sigma_{T|X}^2$. Second, when σ_e^2 is estimable, for example from duplicate annotations or from the disagreement between two strong

reference models, the attenuation can be corrected exactly by adjusting the second-moment matrix, in the spirit of classical measurement-error corrections (Fuller, 2009) and of the error-structure-based corrections in Battaglia, Christensen, Hansen, and Sacher (2025).

Section 8.5 provides simulation evidence for the noisy-reference extension. The simulations first consider conditionally centered reference errors, where the estimates follow the attenuation formula in Remark 7 and the gap between θ_0^* and θ_0 is second order in the reference-error variance. They also show that, when σ_e^2 is known or estimable, the measurement-error correction restores valid inference for θ_0.

6.2 Overidentified GMM

Our baseline framework considers the just-identified case. We now extend the framework and the inferential theory to overidentified GMM. Let $g(\cdot) \in \mathbb{R}^q$ with $q > d$, and let θ_0 be the unique solution of $G(\theta) = \mathbb{E}[g(Z_i, T_i; \theta)] = 0$ over Θ. The weighted debiased moment vector $G_{\lambda,\pi}(\theta) \in \mathbb{R}^q$ is constructed exactly as in (5); its unbiasedness, $\mathbb{E}[G_{\lambda,\pi}(\theta)] = G(\theta)$, holds coordinate-wise by the same argument as in Section 3.4. The AI-PI GMM estimator is

$$\hat{\theta}^{\mathrm{GMM}}_{\lambda,\pi} \in \arg\min_{\theta\in\Theta}\; G_{\lambda,\pi}(\theta)^\top \hat{A}\, G_{\lambda,\pi}(\theta),$$

for a sequence of symmetric weighting matrices $\hat{A} \xrightarrow{p} A$ with A positive definite.

Theorem 6 (Overidentified GMM)**.** *Suppose Assumptions 1–3 hold with $g(\cdot) \in \mathbb{R}^q$, Θ is compact, Assumptions 2.3 and 2.4 hold globally on Θ, $G(\theta) = 0$ only at $\theta_0 \in \operatorname{int}\Theta$, and G is differentiable at θ_0 with Jacobian $J_{\theta_0} \in \mathbb{R}^{q\times d}$ of full column rank. Suppose further that the moment and rate conditions of Theorem 2 hold, including the $(2+\delta)$-moment condition. Then $\hat{\theta}^{\mathrm{GMM}}_{\lambda,\pi} \xrightarrow{p} \theta_0$ and*

$$\sqrt{n}\,(\hat{\theta}^{\mathrm{GMM}}_{\lambda,\pi} - \theta_0) \xrightarrow{d} \mathcal{N}\Big(0,\; \big(J_{\theta_0}^\top A J_{\theta_0}\big)^{-1} J_{\theta_0}^\top A\, \Omega_{\lambda,\pi}\, A J_{\theta_0} \big(J_{\theta_0}^\top A J_{\theta_0}\big)^{-1}\Big),$$

with $\Omega_{\lambda,\pi}$ as in Theorem 2, now of dimension $q\times q$. The feasible two-step choice $\hat{A} = \hat{\Omega}^{-1}_{\lambda,\pi}$, computed as in Remark 1 at a preliminary consistent estimator, attains the efficient variance $\big(J_{\theta_0}^\top \Omega^{-1}_{\lambda,\pi} J_{\theta_0}\big)^{-1}$ for the given design (λ,π).

Corollary 1 (Design optimization under overidentification)**.** *Fix a weighting matrix A and let $M_A := (J_{\theta_0}^\top A J_{\theta_0})^{-1} J_{\theta_0}^\top A$. The total asymptotic variance equals $\operatorname{tr}\big(M_A \Omega_{\lambda,\pi} M_A^\top\big)$, which is affine in*

$\big(1/\rho(\cdot),\ R_\lambda(\cdot)\big)$ *exactly as in the just-identified case. Consequently, Theorem 4, Proposition 2, and Algorithm 1 apply verbatim with the residual score* $s_\lambda(w) = \sqrt{\mathrm{tr}\big(M_A\, R_\lambda(w)\, M_A^\top\big)}$.

The proof of Corollary 1 is immediate: substituting the sandwich form into the trace criterion reproduces the structure $r \cdot \text{const} + \mathbb{E}[(1/\rho - r)\,\mathrm{tr}(M_A R_\lambda(W) M_A^\top)]$, so the optimization arguments go through with $J_{\theta_0}^{-1}$ replaced by M_A. Under the efficient weighting $A = \Omega_{\lambda,\pi}^{-1}$, the criterion depends on the design through A as well; in practice we iterate the design updates of Algorithm 1 with the weighting-matrix update, re-estimating $\hat{A}$ at each step.

6.3 Dependent Data and Clustering

Our sampling framework treats the units as i.i.d., while many applications, including ours, have clustered or panel structure. The results extend directly when the independent sampling unit is a cluster. Let i index clusters (for example, industries or firms), let $Z_i = O_i = \{(Y_{i,t}, X_{i,t}, S_{i,t})\}_{t \le \tau_i}$ collect all within-cluster observations together with their unstructured inputs, let $T_i = (f(S_{i,t}))_{t \le \tau_i}$ stack the within-cluster latent features and $\tilde{T}_{i,m}$ their generated counterparts, and let $g(Z_i, T_i; \theta) = \sum_{t \le \tau_i} g(Y_{i,t}, X_{i,t}, f(S_{i,t}); \theta)$ be the within-cluster aggregate of an observation-level moment function. Human labeling is assigned at the cluster level with propensity $\pi(W_i)$ for cluster-level information W_i, so that a labeled cluster has all of its features verified; the labeling budget counts labeled clusters, and the optimal design of Section 5 allocates the budget across clusters through the cluster-level residual score.

Proposition 7 (Cluster-level validity)**.** *Suppose the clusters O_i are sampled i.i.d., the population moment condition $\mathbb{E}[g(Z_i, T_i; \theta_0)] = 0$ identifies θ_0 at the cluster level, and Assumptions 1–3, together with the conditions of the relevant result, hold for the cluster-level objects $(Z_i, T_i, \{\tilde{T}_{i,m}\}, W_i, D_i)$ and the cluster-level moment function g. Then the conclusions of Theorems 1–6 and of the propositions of Sections 4–6 hold verbatim, with arbitrary dependence within clusters; all variances are cluster-level (clustered) covariance matrices, and the plug-in estimator of Remark 1 delivers clustered standard errors.*

This extension allows arbitrary dependence within a cluster but still requires independence across clusters. For example, if firms are treated as clusters, common time shocks that induce dependence across firms are outside this framework unless they are absorbed by controls or fixed effects. Partial labeling within clusters, which is the design used in our empirical application, is handled by the following remark.

Remark 8 (Unit-level labeling within clusters)**.** *Suppose labels are assigned to units within clusters: for unit t in cluster i, let $D_{i,t}$ be independent Bernoulli with propensity $\pi(W_{i,t})$, satisfying Assumption 1 at the unit level, and define the cluster-level moment as the within-cluster aggregate*

$$\psi_i(\theta) \;=\; \sum_{t\le\tau_i}\Bigg[\sum_{m=1}^{M}\lambda_m\, g(Y_{i,t},X_{i,t},\tilde{T}_{i,t,m};\theta)+\frac{D_{i,t}}{\pi(W_{i,t})}\Big(g(Y_{i,t},X_{i,t},T_{i,t};\theta)-\sum_{m=1}^{M}\lambda_m\, g(Y_{i,t},X_{i,t},\tilde{T}_{i,t,m};\theta)\Big)\Bigg].$$

Conditional on the full sample, each inverse-propensity-weighted term is exactly centered at its fully labeled counterpart, so $\mathbb{E}[\psi_i(\theta)]$ coincides with the fully labeled cluster moment for every θ, and the aggregates ψ_i remain i.i.d. across clusters. The arguments of Proposition 7 then apply to ψ_i without change, with all variances interpreted as cluster-level covariance matrices of the aggregated moments, and the plug-in estimator of Remark 1, computed on these aggregates, delivers the corresponding clustered standard errors.

7 Practical Guidance

In this section, we discuss practical considerations for implementing AI-PI in empirical research.

7.1 Calibration-Set Design

The calibration set is the cornerstone of AI-PI: its size and composition directly determine the quality of the bias correction and the efficiency of the final estimator. On the budget, Theorem 2 shows that the asymptotic variance scales as $O(1/n)$, so doubling the calibration set halves the variance contribution of the labeled sample; in our simulations, a calibration set of approximately 5–10% of the sample already yields substantial improvements over both the naive LLM approach and the calibration-only estimator. On the composition, Theorem 4 shows that labels should be allocated more heavily to regions where the generated features are least reliable. The oracle residual score $s_\lambda(W)$ is not observable ex ante. In the feasible two-stage design of Algorithm 1, a random pilot labeling stage is used to estimate the residual score functions, and the second-stage labeling probabilities are tilted toward observations with larger estimated scores, while maintaining strictly positive labeling probabilities for every unit to satisfy the overlap condition in Assumption 1. In practice, observable measures such as cross-model disagreement, defined as the standard deviation of generated features across model–prompt pairs or seeds, can serve as useful proxies for generator unreliability and as inputs to the labeling information W. When the data have a panel structure,

stratifying the calibration set across time periods and cross-sectional units guards against temporal drift in LLM performance. Finally, when human annotation is accelerated by pre-labeling with a strong model, we recommend keeping a randomly chosen blind subset that is annotated without seeing the pre-label, in order to audit anchoring effects, and we refer to Section 6.1 for the interpretation of the estimand when reference labels are themselves noisy.

7.2 Prompt Tuning, Model Selection, and RAG

The quality of generated features depends on the prompt, the model, and any retrieval-augmented generation (RAG) configuration. AI-PI accommodates all of these choices in two ways. First, rather than selecting a single best configuration, different prompts, models, and RAG settings can be treated as distinct model–prompt pairs and combined through the adaptive weights, which automatically downweight unreliable configurations and exploit offsetting biases. Second, when the pipeline is explicitly tuned on the calibration labels, for example by selecting the prompt with the highest agreement rate or fine-tuning a retrieval index, the cross-fitting procedure of Section 3.5 preserves validity by ensuring that the features used for debiasing in each fold are generated by a pipeline tuned on the other folds. On the number of configurations: with small pilot samples, adding model–prompt pairs increases the noise in the estimated weights. The constraint set Λ acts as implicit regularization, and a ridge penalty on λ is a natural further safeguard when the number of pairs is large.

7.3 Auditing Generator Drift

LLM outputs may change over time because of model updates, API changes, or shifts in the input distribution. In applications where features are generated repeatedly, periodic calibration labels can be used to track changes in the generator's error profile. Comparing newly estimated moment corrections and residual covariances with historical estimates provides a diagnostic for generator drift. The model-prompt weights can then be re-estimated so that configurations with larger residual covariance receive less weight. A rolling calibration window can also be used to update the correction over time.

8 Simulation Study

In this section, we conduct simulations to evaluate the finite-sample performance of the proposed AI-PI estimator. We vary the designs in both the structure of the LLM bias and the information available for labeling allocation.

8.1 Data Generating Process

We consider a linear moment model with one scalar feature T_i. In all the designs, $T_i \sim \mathcal{N}(0,1)$. The outcome is generated by

$$Y_i = \alpha_0 + \beta_0 T_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. The target parameter $\theta_0 = (\alpha_0, \beta_0)^\top$ is defined by the moment condition $\mathbb{E}[g(Z_i, T_i; \theta_0)] = 0$, where $g(Z_i, T_i; \theta) = X_i^*(X_i^{*\top}\theta - Y_i)$ with $X_i^* = (1, T_i)^\top$.

For each LLM model-prompt pair m, the generated feature is

$$\widetilde{T}_{i,m} = T_i + b_{i,m} + \epsilon_{i,m},$$

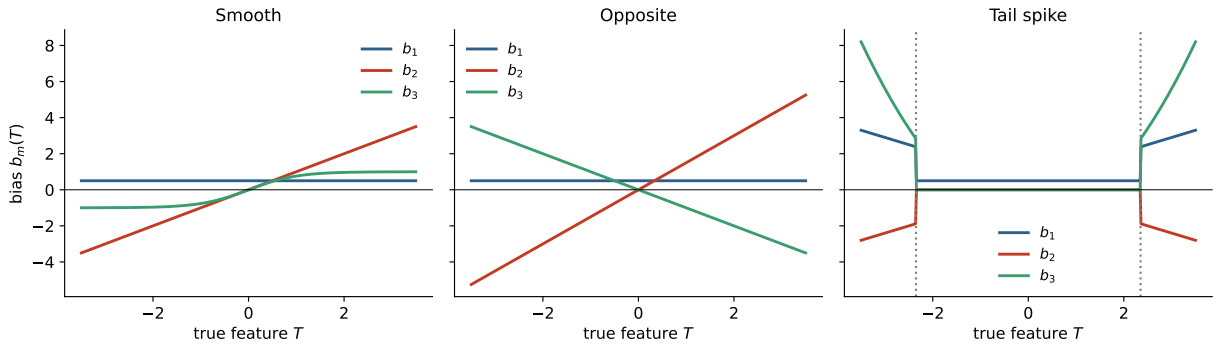
with $\epsilon_{i,m} \sim N(0, \sigma_\epsilon^2)$. We set $M = 3$ model–prompt pairs and, without loss of generality, use one seed for each pair. The bias terms $b_{i,m}$ take one of the three designs.

- **Smooth:** $b_{i,1} = \delta_0, b_{i,2} = \delta T_i, b_{i,3} = \delta \tanh(T_i)$. This design captures hallucination bias that changes smoothly over the feature space.
- **Opposite:** $b_{i,1} = \delta_0, b_{i,2} = 1.5\delta T_i, b_{i,3} = -\delta T_i$. This design represents a setting where different model–prompt pairs have systematic errors in opposite directions.
- **Tail spike:** Let $\mathbf{1}_i^{\text{tail}} = \mathbf{1}\{|T_i| > \text{threshold}\}$. The biases are $b_{i,1} = \delta_0 + \mathbf{1}_i^{\text{tail}} \cdot \delta|T_i|, b_{i,2} = -\mathbf{1}_i^{\text{tail}} \cdot \delta|T_i|, b_{i,3} = \mathbf{1}_i^{\text{tail}} \cdot \delta(T_i^2 - 2)$. This represents a setting where the LLM is accurate for typical inputs but hallucinates badly in the tails.

Figure 4 visualizes the functional form of the bias terms in the three designs. In the smooth design, the three generators have smoothly varying but nonzero biases; in the opposite design, their biases have opposite signs, which the optimal weights can exploit; in the tail-spike design the generators are accurate for typical inputs and fail only in the tails, making the labeling design especially important. The numerical values of all simulation parameters are reported in Figure 5.

Section 8.2 first studies the oracle version of the AI-PI estimator, isolating the gains from optimal weighting and optimal labeling when the population design objects are known. Section 8.3

Figure 4: Functional form of the bias terms $b_m(T)$ in the three simulation designs



This figure plots the bias functions $b_m(T)$ of the three model–prompt pairs ($m = 1, 2, 3$) against the true feature T for the three simulation designs (smooth, opposite, tail spike) with the parameter values used in Figure 5. The vertical dotted lines in the right panel mark the tail-spike threshold ± 2.35.

then examines the feasible two-stage implementation in Algorithm 1, focusing on the pilot-share tradeoff in estimating and exploiting the labeling policy. Section 8.4 presents a unified estimator comparison that includes the feasible two-stage implementation, evaluates the coverage of the confidence intervals of Remark 1, and studies the dependence of the MSE on the weights. Finally, Section 8.5 evaluates the noisy-reference-label extension of Section 6.1, covering both conditionally centered reference errors and systematically biased reference labels.

8.2 Oracle Estimators and Results

We first evaluate the oracle labeling policy and model–prompt weights characterized in Sections 5.1 and 5.2, which use the true data-generating process. They therefore provide an oracle benchmark for the best performance attainable by the AI-PI estimator.

We compare five estimators. Throughout, let $\lambda^{\text{unif}} = (1/M, \ldots, 1/M)^\top$, so that $\lambda_m^{\text{unif}} \geq 0$ for all m and $\sum_{m=1}^M \lambda_m^{\text{unif}} = 1$. For optimized weights, we use the oracle criterion in Section 5.2 over the constrained set $\Lambda = \{\lambda \in \mathbb{R}_+^M : \sum_{m=1}^M \lambda_m \leq 1\}$. Thus, the optimized weights are allowed to shrink the overall contribution of the LLM moments.

- **Naive LLM regression.** This estimator averages the generated features across model–prompt pairs, $\widetilde{T}_i = M^{-1} \sum_{m=1}^M \widetilde{T}_{i,m}$, and then regresses Y_i on $\widetilde{T}_i$, without labeling correction.
- **Uniform debiased estimator.** This estimator uses the debiased moment estimator with uniform labeling probability $\pi(W_i) = r_N$ and uniform weights $\lambda = \lambda^{\text{unif}}$.

- **Oracle-policy debiased estimator.** This estimator uses the debiased moment estimator with uniform weights $\lambda = \lambda^{\text{unif}}$ and oracle optimal labeling rule $\pi^*(\cdot \mid \lambda^{\text{unif}})$ from Section 5.1.
- **Oracle-weight debiased estimator.** This estimator uses the debiased moment estimator with uniform labeling probability $\pi(W_i) = r_N$ and the oracle optimal weights $\lambda^*_\pi \in \arg\min_{\lambda \in \Lambda} V(\lambda, \pi)$ from Section 5.2.
- **Joint-oracle debiased estimator.** This estimator uses the debiased moment estimator with jointly optimized (π^*, λ^*). It alternates between the oracle policy update in Section 5.1 and the oracle weight update in Section 5.2, with each weight update constrained to Λ.

Before labeling, we observe W_i that can be used to guide the labeling decision. We consider the following three choices:

- **Outcome level:** $W_i = Y_i$. This is a feasible proxy for observations that may be influential for the target moment.
- **Model disagreement:** $W_i = \text{sd}_m(\tilde{T}_{i,m})$. This captures cross-model variation in the generated feature.
- **True feature:** $W_i = T_i$. This is an infeasible oracle information benchmark.

Figure 5 reports the mean squared error (MSE) $R^{-1} \sum_{r=1}^{R} \|\hat{\theta}^{(r)} - \theta_0\|^2$ for $\hat{\theta} = (\hat{\alpha}, \hat{\beta})^\top$ across the three bias designs. We highlight four main findings. First, the naive estimator exhibits substantial error across all designs. This reflects the error induced by directly replacing the true feature with generated features.

Second, debiasing with a small calibration sample substantially improves over the naive regression. This validates the key insight that a small calibration set can correct for the bias in the moment conditions. Even the simplest debiased estimator, with uniform labeling and uniform weights, improves over the naive regression approach in all three settings.

Third, optimizing the model–prompt weights λ provides large additional gains. The optimized weights allow the estimator to downweight the overall GenAI contribution when the generated proxies are too noisy or biased. The improvement is especially visible when different model-prompt pairs have offsetting or heterogeneous biases, because the oracle weights can downweight unreliable generated moments and exploit partial cancellation across model-prompt pairs.

Fourth, optimizing the labeling policy π is most valuable when W_i identifies observations with large residual correction variance. This is especially pronounced in the tail-spike design, where the cross-model disagreement identifies the high-residual region almost as well as the infeasible oracle

signal T_i and much better than the outcome Y_i. The jointly optimized estimator can concentrate labels on the observations where calibration is most valuable. Across the three designs, the jointly optimized estimator delivers the lowest MSE.

To better understand the distinct roles of the labeling policy and the model–prompt weights, we next examine the oracle choices selected in the tail-spike design. The weights λ determine the global contribution of the LLM generated moments, whereas the policy π determines where the fixed labeling budget is spent. These two choices therefore operate through different margins. Weight optimization can shrink or reallocate the contribution of the model–prompt moments, while policy optimization reallocates labels toward observations for which the debiasing correction is most valuable.

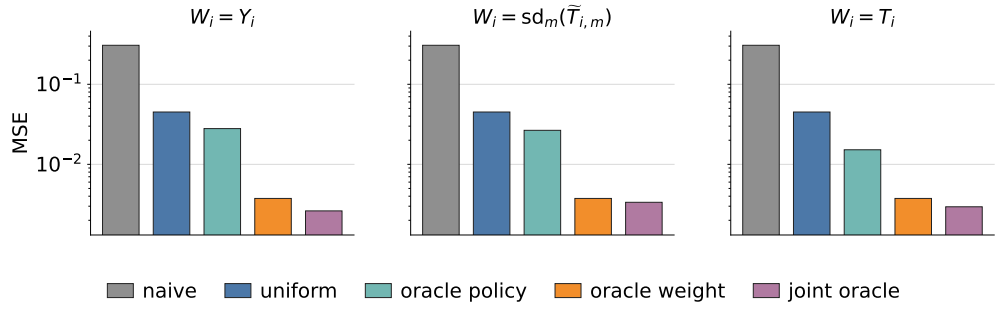
Figure 6 plots the selected labeling probabilities against the true feature T_i in the tail-spike design. The three plots differ in the information W_i used to construct the labeling rule. The vertical dotted lines mark the tail-spike thresholds. The figure confirms that the optimized policies allocate substantially more labels to the tail region, especially when the labeling information is model disagreement or the oracle feature itself. In contrast, the estimators with uniform policy remain flat by construction. In this design, model disagreement is a particularly informative labeling signal. The spike in the generated features occurs precisely in the tail region of the true feature distribution, and this region also generates large cross-model variation. Hence $W_i = \mathrm{sd}_m(\widetilde{T}_{i,m})$ acts as a feasible proxy for the oracle information in T_i: it identifies many of the observations for which the residual correction is largest. This explains why the policy curves and the MSE performance under model disagreement are close to those under the oracle choice $W_i = T_i$, and substantially better than those under the outcome-level proxy $W_i = Y_i$ in the tail-bias design.

8.3 Feasible Implementation

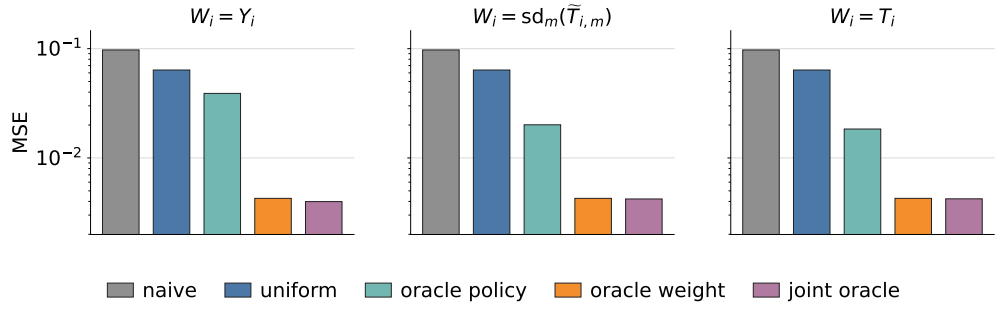
The preceding subsection evaluates the oracle benchmark in which the labeling policy and model–prompt weights are computed from true model parameters. Specifically, the optimal policy depends on the residual score $s_\lambda(W_i)$ and the optimal weight vector depends on the population quadratic criterion. In practice, these population objects are not observed. We therefore evaluate the feasible two-stage implementation described in Section 5.3.

The experiment uses the same data-generating process, the same total labeling rate r_N, and the same weight constraint Λ as in Section 8.2. We focus on the feasible joint debiased estimator proposed in Section 5.3 and compare it to the joint-oracle debiased estimator in Section 8.2. The

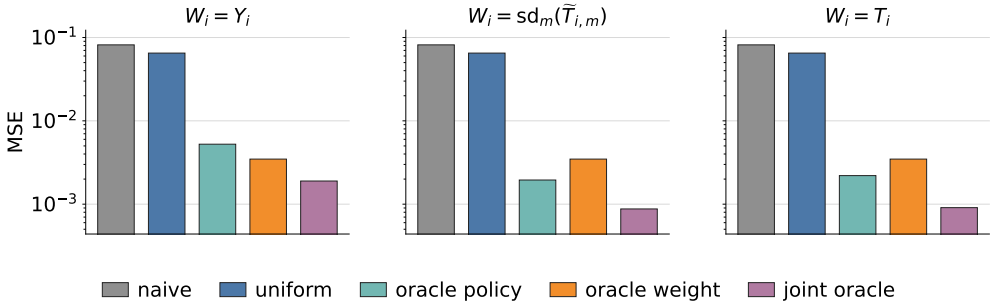
Figure 5: MSE of different estimators across bias designs



(a) Smooth bias



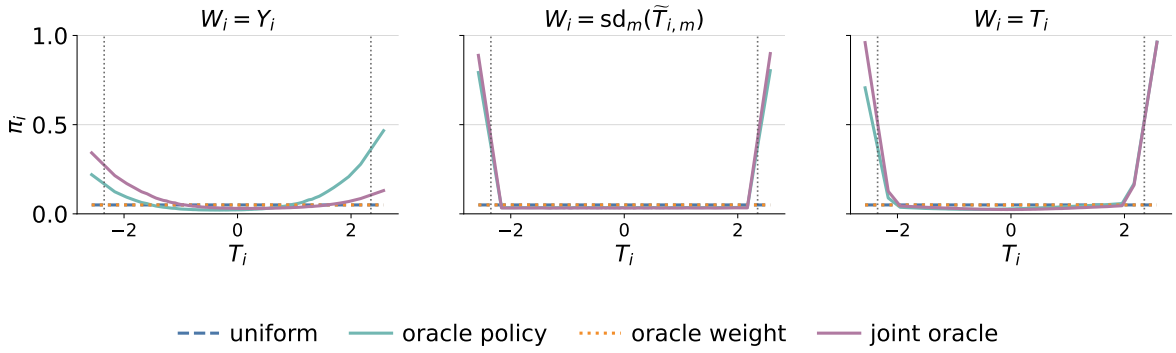
(b) Opposite bias



(c) Tail-spike bias

This figure reports the mean squared error (MSE) $R^{-1}\sum_{r=1}^{R}\|\hat{\theta}^{(r)} - \theta_0\|_2^2$ for the target parameter θ_0. The y-axis is logarithmic. Each subfigure corresponds to a different bias design and has three panels corresponding to $W_i = Y_i, W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$, and $W_i = T_i$. The five bars are, from left to right: naive LLM regression (naive), uniform debiased estimator (uniform), oracle-policy debiased estimator (oracle policy), oracle-weight debiased estimator (oracle weight), and joint-oracle debiased estimator (joint oracle). Oracle policies are computed using the Section 5.1 score, with finite-population conditional expectations approximated by 100 equal-frequency bins of the selected W_i. Oracle weights are computed using the Section 5.2 quadratic criterion subject to $\lambda_m \geq 0$ and $\sum_{m=1}^{M}\lambda_m \leq 1$. The joint estimator starts from uniform λ and alternates the oracle policy and weight updates until $\|\lambda^{(t)} - \lambda^{(t-1)}\|_2 \leq 10^{-10}$ or 100 iterations. In every replication, $T_i \sim \mathcal{N}(0,1)$ and $Y_i = \alpha_0 + \beta_0 T_i + \varepsilon_i$, with $\theta_0 = (\alpha_0, \beta_0)^\top = (0.5, 1.5)^\top$ and $\varepsilon_i \sim N(0,1)$. The generated feature is $\widetilde{T}_{i,m} = T_i + b_{i,m} + \epsilon_{i,m}$, with $\epsilon_{i,m} \sim \mathcal{N}(0, 0.2^2)$, $M = 3$, and one generation seed per model–prompt pair. The bias parameters are as follows: in the smooth design, $(\delta_0, \delta) = (0.5, 1)$; in the opposite design, $(\delta_0, \delta) = (0.5, 1)$; in the tail-spike design, $(\delta_0, \delta, \text{threshold}) = (0.5, 0.8, 2.35)$. We set the sample size $N = 5000$, the labeling rate $r_N = 0.05$, the overlap lower bound $\rho_{\min} = 0.1$, and run $R = 200$ simulations for each setup.

Figure 6: Oracle labeling policies in the tail-spike design



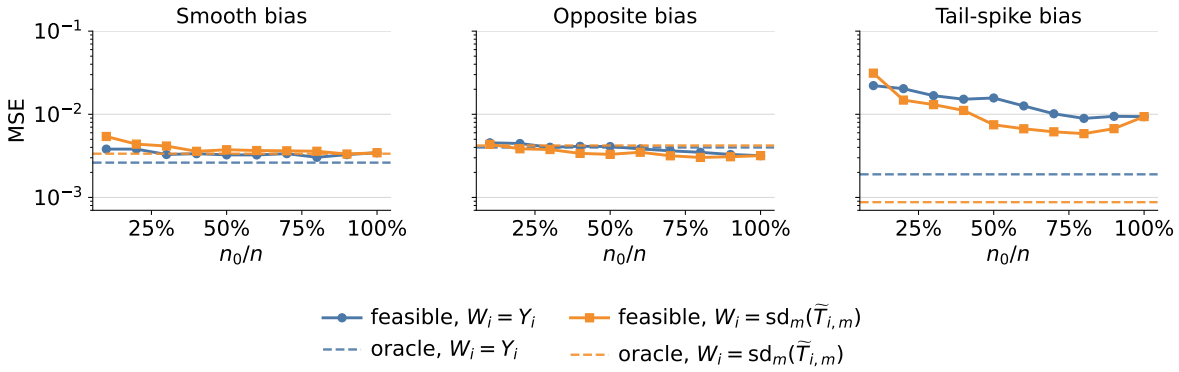
The figure uses the tail-spike design from Figure 5. Each curve reports the Monte Carlo average of $\mathbb{E}[\pi(W_i) \mid T_i]$, computed by binning observations according to the true feature T_i. The three panels correspond to the information used to construct the labeling rule: $W_i = Y_i$, $W_i = \mathrm{sd}_m(\widetilde{T}_{i,m})$, and $W_i = T_i$. The horizontal axis is T_i in all panels. The vertical dotted lines indicate the tail-spike threshold ± 2.35. The four curves are the uniform debiased estimator, the oracle-policy debiased estimator, the oracle-weight debiased estimator, and the joint-oracle debiased estimator. The oracle-weight estimator uses a uniform policy, so its policy curve is flat by construction. For this estimator, the Monte Carlo average weight vector is $(0.177, 0.353, 0.000)$, which is the same across the three choices of W_i. For the joint-oracle estimator, the Monte Carlo average weights are $(0.127, 0.629, 0.022)$ for $W_i = Y_i$, $(0.038, 0.466, 0.442)$ for $W_i = \mathrm{sd}_m(\widetilde{T}_{i,m})$, and $(0.041, 0.461, 0.440)$ for $W_i = T_i$.

feasible estimator splits the total labeling budget into an initial pilot stage and a second adaptive stage. The first stage draws uniform pilot labels with expected size n_0. The pilot labels are then used to construct plug-in estimates of the objects entering the oracle updates, as in Algorithm 1.

Figure 7 reports how the MSE of the feasible joint estimator varies with the pilot share n_0/n of the labeling budget allocated to the pilot stage. The horizontal lines report the corresponding joint-oracle MSEs from Figure 5. The figure illustrates the finite-sample tradeoff in the feasible implementation. When the pilot share is small, the estimated policy and weights are noisy, and the feasible estimator can be far from the oracle benchmark. Increasing the pilot share improves the estimation of the design objects and typically reduces MSE. However, a larger pilot stage also leaves less of the labeling budget for adaptive second-stage labeling. This second force becomes visible in the tail-spike design: the MSE falls at first but rises again when nearly all labels are spent in the pilot stage. At $n_0/n = 1$, the procedure has no remaining adaptive labeling budget, so the estimated policy cannot be used to target the high-residual observations.

The tradeoff is most pronounced in the tail-spike design, where the value of adaptive labeling is largest. In the smooth and opposite designs, the feasible estimator is close to the oracle benchmark for a broad range of pilot shares, because the residual structure is easier to learn from the pilot sample. In contrast, under tail-spike bias, the relevant residual variation is concentrated in the tails. Learning this structure requires a sufficiently large pilot sample, but exploiting it also requires

Figure 7: MSE of feasible joint-optimized estimator



The figure reports the mean squared error (MSE) $R^{-1}\sum_{r=1}^{R}\|\hat{\theta}^{(r)} - \theta_0\|_2^2$ for the feasible joint-optimized weighted debiased estimator described in Section 5.3. The x-axis is the pilot label ratio $n_0/n \in \{0.1, 0.2, \ldots, 1.0\}$. The total labeling rate is fixed at $r_N = 0.05$. In each plot, the two feasible curves correspond to $W_i = Y_i$ and $W_i = \mathrm{sd}_m(\widetilde{T}_{i,m})$. The horizontal dashed lines are the joint-oracle MSEs from Figure 5 for the same bias design and the same choice of W_i. The data-generating process and bias-design parameters are the same as in Figure 5. We set the sample size $N = 5000$ and run $R = 200$ simulations for each setup.

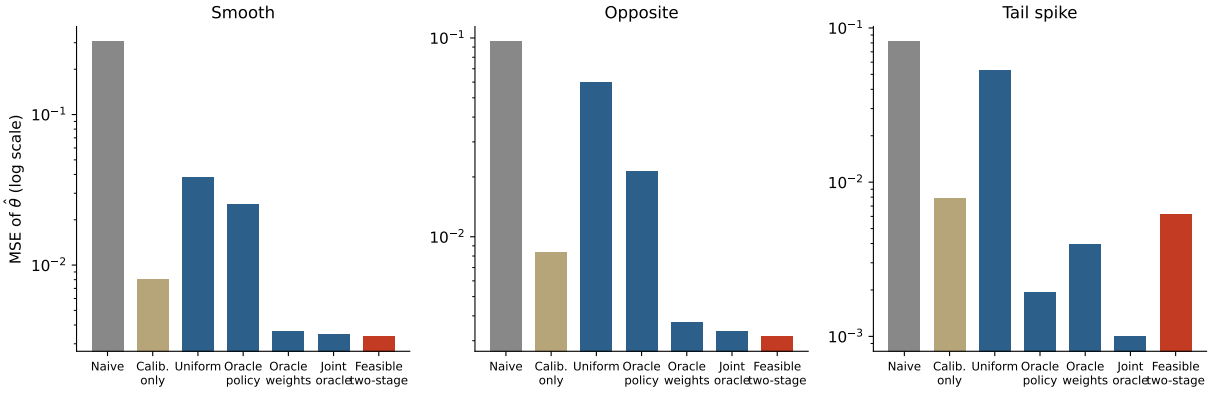
preserving enough labels for the adaptive second stage.

8.4 Estimator Comparison, Coverage, and the Weight Scale

We complement the oracle analysis with a unified comparison that includes the calibration-only benchmark and the feasible two-stage estimator, and with an evaluation of the confidence intervals from Remark 1. Throughout this subsection, the policy-based estimators use the feasible labeling information $W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$ (model disagreement), and the feasible two-stage estimator uses a pilot share of $n_0/n = 0.5$; the data-generating processes and all design parameters are exactly as in Figure 5, with $R = 1{,}000$ replications.

Figure 8 compares the mean squared errors of seven estimators in a single plot per design, and Table 1 reports the corresponding bias, standard deviation, RMSE, empirical coverage of nominal 95% confidence intervals, and average interval length. The results yield five main findings. First, the naive LLM regression is not only biased but produces entirely misleading inference: its empirical coverage is 0.0%, 0.0%, and 1.6% across the three designs, because the bias dominates the sampling uncertainty measured by the standard error from $N = 5,000$ observations. Second, every debiased estimator is approximately unbiased and delivers coverage close to the nominal level, confirming the validity theory of Section 4 and the variance estimator of Remark 1. Third, debiasing alone is not enough: the uniform-design estimator (equal weights, uniform labeling) is less efficient than

Figure 8: Estimator comparison including the feasible two-stage estimator



This figure reports the mean squared error $R^{-1}\sum_{r=1}^{R}\|\hat{\theta}^{(r)}-\theta_0\|_2^2$ (log scale) of seven estimators across the three bias designs: naive LLM regression, calibration-only least squares, and the AI-PI estimator with uniform design, oracle labeling policy, oracle weights, jointly optimized design, and the feasible two-stage implementation. Policy-based estimators use $W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$; the feasible estimator uses a pilot share $n_0/n = 0.5$. The data-generating process and all design parameters are as in Figure 5, with $R = 1{,}000$ replications.

discarding the generated features altogether (RMSE two to three times that of calibration-only), because equally weighting three poor generators adds substantial noise into the correction. The adaptive weights are what convert validity into efficiency. Fourth, the jointly optimized estimator dominates everywhere: relative to calibration-only, it shortens confidence intervals by 30%, 31%, and 62% in the smooth, opposite, and tail-spike designs while maintaining coverage between 94.7% and 95.9%. Fifth, the feasible two-stage estimator essentially attains the joint-oracle performance in the smooth and opposite designs (coverage 94.8% and 95.6%, interval lengths within 2% of the oracle), demonstrating that the oracle gains are achievable with estimated designs. In the tail-spike design, the feasible estimator improves on the uniform design but remains below the oracle estimator and exhibits mild undercoverage (87.6%). With an expected pilot size of $n_0 = 125$, the tail region that drives the residual score contains only a few pilot observations, making the estimated score noisy. This pattern is consistent with the requirement $n_0 \to \infty$ in Theorem 5 and with the pilot-share tradeoff documented in Figure 7.

Finally, Figure 9 examines the role of weight shrinkage by tracing the MSE along the path $\lambda = c\,\lambda^{\mathrm{unif}}$, $c \in [0,1]$, under uniform labeling, together with the population asymptotic variance $\mathrm{tr}(J_{\theta_0}^{-1}\Omega_{\lambda,\pi}J_{\theta_0}^{-\top})/n$ from Theorem 2. The simulated MSE closely follows the population asymptotic variance near the optimum, providing a direct verification of the variance formula that underlies the design optimization. The optimal scale is far below one ($c^* \approx 0.2$ in all three designs), showing

Table 1: Bias, efficiency, and coverage of nominal 95% confidence intervals

Design	Estimator	Bias	SD	RMSE	Coverage	CI length
Smooth	Naive LLM regression	-52.9	0.9	55.4	0.0	3.6
	Calibration-only	0.1	6.4	9.0	94.3	24.8
	AI-PI: uniform (λ, π)	2.9	18.5	19.6	92.5	70.4
	AI-PI: oracle policy	1.4	14.9	15.9	93.8	57.2
	AI-PI: oracle weights	-0.0	4.6	6.0	94.8	18.1
	AI-PI: joint oracle	0.3	4.4	5.9	95.5	17.4
	AI-PI: feasible two-stage	-0.8	4.2	5.8	94.8	17.1
Opposite	Naive LLM regression	-22.6	1.2	31.0	0.0	4.8
	Calibration-only	-0.1	6.6	9.1	94.5	24.7
	AI-PI: uniform (λ, π)	3.4	24.0	24.4	93.0	91.1
	AI-PI: oracle policy	1.1	14.1	14.6	93.7	51.5
	AI-PI: oracle weights	0.2	4.9	6.1	94.5	19.0
	AI-PI: joint oracle	0.2	4.4	5.8	94.7	17.1
	AI-PI: feasible two-stage	0.2	4.3	5.6	95.6	17.2
Tail spike	Naive LLM regression	-7.8	1.9	28.6	1.6	7.4
	Calibration-only	-0.2	6.2	8.9	95.0	24.6
	AI-PI: uniform (λ, π)	2.9	22.7	23.1	82.5	78.0
	AI-PI: oracle policy	-0.0	3.5	4.4	95.2	13.8
	AI-PI: oracle weights	-0.1	5.3	6.3	93.6	20.4
	AI-PI: joint oracle	-0.0	2.4	3.2	95.9	9.4
	AI-PI: feasible two-stage	-1.9	7.2	7.9	87.6	22.2

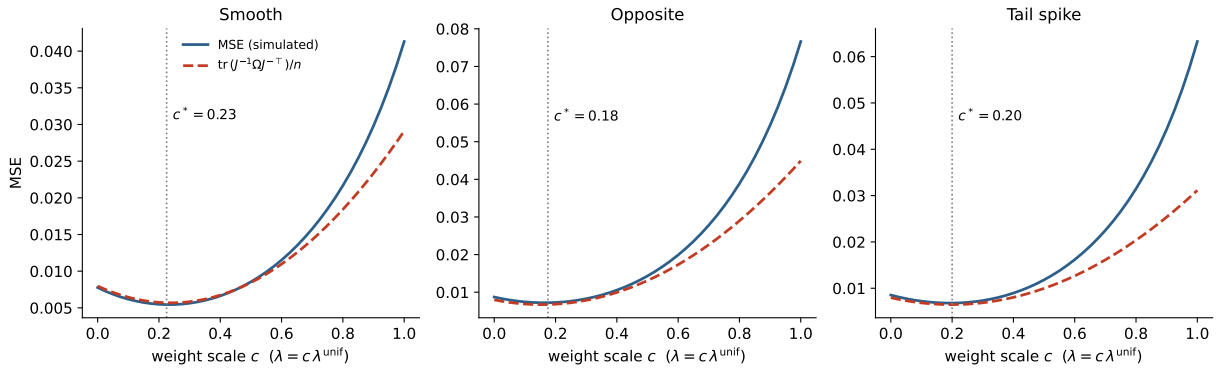
The table reports the bias and standard deviation of $\hat{\beta}$ ($\times 100$), the RMSE of $\hat{\theta}$ ($\times 100$), the empirical coverage of nominal 95% confidence intervals for β_0 (in %), and the average confidence-interval length ($\times 100$), over $R = 1{,}000$ replications of the designs in Figure 5. Confidence intervals for the AI-PI estimators use the plug-in variance estimator of Remark 1; the naive and calibration-only estimators use heteroskedasticity-robust (HC0) standard errors. Policy-based estimators use $W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$; the feasible two-stage estimator uses a pilot share $n_0/n = 0.5$ and pools the pilot labels as in Algorithm 1 (cf. Remark 6).

that the generated moments should be substantially downweighted when they remain noisy after correction. The common practice of implicitly setting $\lambda = 1$, that is, trusting the generated data fully after a bias adjustment, is therefore severely suboptimal.

8.5 Noisy Reference Labels

We close the simulation study by evaluating the noisy-reference extension of Section 6.1. We use the simulations to separate two cases. In the first, reference errors are conditionally centered, so the reference differs from the truth only through noise. In the second, the reference itself contains systematic bias, as may arise when a stronger model shares look-ahead bias with the cheaper generators. The data-generating process is the smooth design of Figure 5. The calibration set now reveals the noisy reference $T_i^* = T_i + e_i$, with $e_i \sim \mathcal{N}(0, \sigma_e^2)$ independent of the data, instead of

Figure 9: MSE and asymptotic variance along the weight path $\lambda = c\,\lambda^{\text{unif}}$



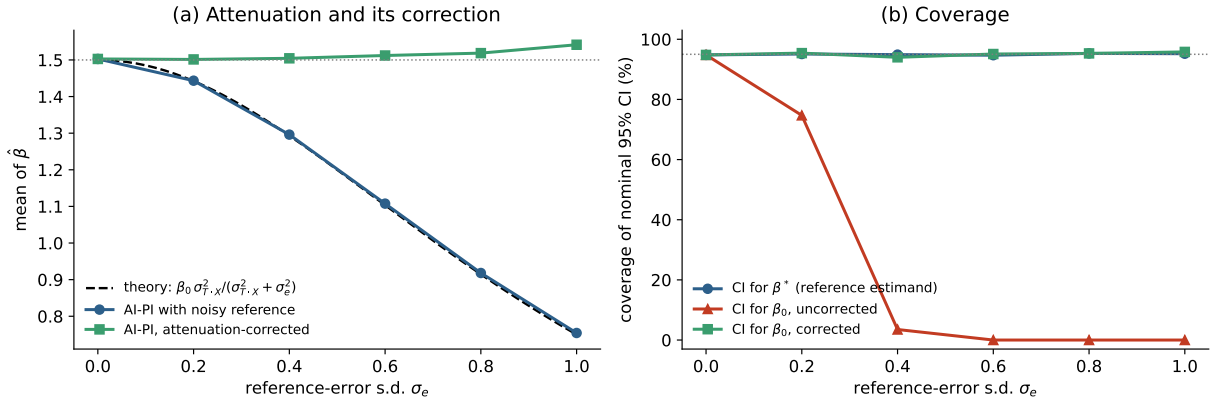
This figure plots the simulated mean squared error of the AI-PI estimator with uniform labeling policy and weights $\lambda = c\,(1/3, 1/3, 1/3)^\top$ as a function of the scale c (solid), together with the population asymptotic variance $\mathrm{tr}(J_{\theta_0}^{-1}\Omega_{\lambda,\pi}J_{\theta_0}^{-\top})/n$ implied by Theorem 2 (dashed), approximated on a sample of 400,000 observations. The vertical dotted line marks the MSE-minimizing scale c^*. The data-generating process and all design parameters are as in Figure 5, with $R = 500$ replications.

the true feature, and the AI-PI estimator (jointly optimized design, labeling information $W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$) treats T_i^* as the truth. Since the residual variance of the feature is $\sigma^2_{T|X} = 1$ in this design, the reference estimand is available in closed form, $\beta^* = \beta_0/(1 + \sigma_e^2)$, which makes the predictions of Proposition 6 and Remark 7 directly testable.

Figure 10 summarizes the centered-reference experiment. Panel (a) shows that the Monte Carlo mean of $\hat{\beta}$ coincides with the attenuation formula of Remark 7 at every error level (for example, 1.296 simulated versus 1.293 predicted at $\sigma_e = 0.4$), and that the corrected estimator, which adjusts the human second-moment matrix by σ_e^2, recovers β_0. Panel (b) shows the inferential counterpart of Proposition 6(a): the confidence intervals cover the reference estimand β^* at rates between 94.7% and 95.3% for every σ_e, while their coverage of β_0 collapses as soon as the attenuation exceeds the interval length (74.7% at $\sigma_e = 0.2$ and 3.5% at $\sigma_e = 0.4$); the corrected intervals restore 94–96% coverage of β_0 throughout. Two caveats apply. The correction requires σ_e^2, in practice an estimate from duplicate annotations or the disagreement of two reference models, and at extreme noise ($\sigma_e = 1$, a reference as noisy as the feature itself) the corrected point estimate drifts upward (1.54 on average against a target of 1.5), the familiar finite-sample instability of measurement-error corrections.

Table 2 reports the model-assisted calibration experiment, in which the reference label is produced by a strong model. The first strong-reference specifications have centered errors with standard

Figure 10: Noisy reference labels: attenuation, correction, and coverage



This figure reports the human-labeling-error experiment in the smooth design of Figure 5. The calibration set reveals $T_i^* = T_i + e_i$ with $e_i \sim \mathcal{N}(0, \sigma_e^2)$; the AI-PI estimator uses the jointly optimized design with $W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$. Panel (a) plots the Monte Carlo mean of $\hat{\beta}$ against σ_e for the uncorrected estimator (circles), the theoretical attenuation $\beta_0\,\sigma^2_{T|X}/(\sigma^2_{T|X}+\sigma^2_e)$ of Remark 7 (dashed), and the attenuation-corrected estimator (squares). Panel (b) reports the empirical coverage of nominal 95% confidence intervals for the reference estimand β^* and for the target β_0, with and without the correction. $R = 1{,}000$ replications.

deviation σ_s; the last row adds the systematic bias $0.2\tanh(T_i)$, illustrating a case in which the reference is itself biased. The results highlight three points. First, the second-order property of Proposition 6(b) is quantitatively accurate: the bias of $\hat{\beta}$ ($\times 100$) is -0.0, -1.3, and -5.6 for $\sigma_s = 0.05, 0.1, 0.2$, closely matching the theoretical attenuation $-\beta_0\sigma_s^2/(1+\sigma_s^2)\times 100 = -0.4, -1.5, -5.8$, while a reference with a first-order systematic bias of comparable magnitude shifts the estimate by -17.2. Second, a strong reference with $\sigma_s \leq 0.1$ essentially attains the gold-label benchmark (RMSE 5.9–6.3 versus 5.8 with gold labels, against 8.7 for the calibration-only estimator), which substantiates the practical recipe of querying an expensive model on the calibration set to debias cheap models on the full sample. Third, the limits are equally visible: at $\sigma_s = 0.2$ the intervals undercover β_0 (74.1%) while still covering β^* (95.4%), and under the shared systematic bias the coverage of β_0 is destroyed (0.3%) although β^* remains covered (94.5%). A gold-standard calibration set therefore remains the criterion for full identification of θ_0, exactly as discussed in Section 6.1.

9 Empirical Study

Whether the tone of news carries information about future stock returns is a long-standing question in asset pricing, and it is economically important because it speaks to how quickly and how completely prices incorporate qualitative, soft information. A large literature documents that

Table 2: Strong models as reference labels

Estimator	Bias	RMSE	Cover β_0	Cover β^*	CI length
Naive cheap-model ensemble	-52.9	55.3	0.0	0.0	3.6
Calibration-only, gold labels	0.7	8.7	94.3	94.3	23.2
AI-PI, gold labels	0.0	5.8	96.1	96.1	17.4
AI-PI, strong ref. ($\sigma_s = 0.05$)	-0.0	5.9	94.6	94.6	17.5
AI-PI, strong ref. ($\sigma_s = 0.1$)	-1.3	6.3	93.0	93.7	17.7
AI-PI, strong ref. ($\sigma_s = 0.2$)	-5.6	8.4	74.1	95.4	18.1
AI-PI, biased ref. ($0.2\tanh(T)$)	-17.2	17.9	0.3	94.5	13.4

The table reports the bias of $\hat{\beta}$ ($\times 100$, relative to β_0), the RMSE of $\hat{\theta}$ ($\times 100$), the empirical coverage of nominal 95% confidence intervals for β_0 and for the reference estimand β^* (in %), and the average interval length ($\times 100$), over $R = 1{,}000$ replications of the smooth design. The cheap generators are as in Figure 5; the reference label is $T_i^* = T_i + \sigma_s e_i$ for the centered designs and $T_i^* = T_i + 0.2\tanh(T_i) + 0.1\, e_i$ for the biased design. AI-PI uses the jointly optimized design with $W_i = \mathrm{sd}_m(\tilde{T}_{i,m})$. For the naive and gold-label rows, $\beta^* = \beta_0$. For the biased reference, β^* is computed from the population least-squares limit.

textual sentiment is related to returns. Tetlock (2007) shows that high media pessimism exerts temporary downward pressure on prices that subsequently reverts, and Tetlock, Saar-Tsechansky, and Macskassy (2008) find that negative words in firm-specific news forecast lower earnings and returns. Subsequent work connects sentiment to the time series and the cross-section of returns using finance-specific dictionaries (Loughran and McDonald, 2011), variation over the business cycle (Garcia, 2013), search-based measures (Da, Engelberg, and Gao, 2015), and broad news flow (Calomiris and Mamaysky, 2019), building on the broader evidence that investor sentiment affects asset prices (Baker and Wurgler, 2006). The direction of the relation is theoretically ambiguous. Under an information or underreaction channel, positive news predicts higher subsequent returns as prices adjust with a delay, producing short-horizon continuation or post-news drift (Chan, 2003; Tetlock, Saar-Tsechansky, and Macskassy, 2008). Under a sentiment or overreaction channel, optimistic news temporarily pushes prices above fundamentals and predicts lower subsequent returns, producing reversal (Tetlock, 2007; Bybee, 2023). Because the two mechanisms have opposite implications, the sign of the predictive relation is an empirical question, and we treat it as such.

The measurement of sentiment from text has evolved from hand-built dictionaries (Loughran and McDonald, 2011) and supervised machine learning (Ke, Kelly, and Xiu, 2019) to large language models, which now extract sentiment and related measures from financial text at scale (Lopez-Lira and Tang, 2023; Bybee, 2023; Jha, Qian, Weber, and Yang, 2024; Schuettler, Audrino, and Sigrist, 2024). This is exactly the practice our framework studies, and it is also the source of the difficulty: LLM-generated sentiment depends on the model and the prompt, so the same headline can be

labeled differently across reasonable choices, and downstream regression conclusions can move with those choices (Ludwig, Mullainathan, and Rambachan, 2024). The relation between news sentiment and stock returns is therefore an ideal laboratory for AI-PI. The economic question is important and well understood, the downstream regression is a standard and transparent object, and the generated feature, the sentiment of a headline, is a human-replicable annotation for which a small human-labeled calibration set both exists by construction and is informative about the error of each model–prompt pair. We use the same news-headline data setting as Ludwig, Mullainathan, and Rambachan (2024), so that our results connect directly to their documentation of model-and-prompt fragility.

Our goal in this application is not to provide a definitive analysis of the effect of news sentiment on the equity premium. It is to demonstrate, in a reasonable and standard research design, that the conclusions a researcher draws from LLM-generated features depend on arbitrary model and prompt choices, and that AI-PI resolves this fragility by combining the generated measures with a small human-labeled calibration set into a single, valid, and more precise estimate.

9.1 Design

We study how news-based sentiment is related to subsequent stock returns. Let the unit of observation be a stock-month pair (i, t), where i is a stock and t is a month. For each unit, define:

- $H_{i,t}$: the set of news headline observations assigned to stock i in month t;
- $T_{i,t}$: the latent human-labeled stock-specific sentiment score based on $H_{i,t}$;
- $\tilde{T}_{i,t}^{(m)}$: the sentiment score generated by LLM model–prompt pair m;
- $Y_{i,t}$: the next-month stock return, $R_{i,t+1}$.

Each headline in the raw data is linked to a stock ticker and a publication date. We aggregate headline-level sentiment to the stock-month level by averaging sentiment scores within each stock-month. This gives one generated sentiment measure $\tilde{T}_{i,t}^{(m)}$ for each model–prompt pair and one latent human-reference sentiment score $T_{i,t}$ for each stock-month unit.

The target regression is

$$R_{i,t+1} = a + bT_{i,t} + X_{i,t}'\Gamma + \mu_{j(i)} + \tau_{t+1} + u_{i,t+1}, \tag{12}$$

where $\mu_{j(i)}$ denotes industry fixed effects for stock i and τ_{t+1} denotes month fixed effects. These fixed effects absorb common shocks to returns in month $t+1$, so the coefficient b is identified from

cross-sectional variation across stocks within the same return month.

The control variables $X_{i,t}$ include log firm size, book-to-market, momentum, and short-term reversal. Specifically, log firm size is the logarithm of monthly market capitalization, book-to-market is book equity divided by monthly market equity, momentum is the cumulative raw return from months $t-12$ through $t-2$, and short-term reversal is the raw return $R_{i,t}$ in the headline month t.[4] [5] The parameter of interest is b.

The sign of b is an empirical question. If $b > 0$, stocks with more positive headline sentiment earn higher next-month returns, consistent with short-horizon continuation or post-news drift. If $b < 0$, stocks with more positive headline sentiment earn lower next-month returns, consistent with sentiment-driven overpricing and reversal.

In practice, $T_{i,t}$ is not observed for the full sample, since obtaining human labels for all headline observations would be costly and time-consuming. We therefore use LLM-generated sentiment measures $\tilde{T}_{i,t}^{(m)}$ as noisy proxies for $T_{i,t}$ and construct a human-labeled calibration set to correct for LLM errors. The next subsections describe the construction of the stock-month headline panel, the generated sentiment measures, the calibration labels, and the AI-PI estimates.

9.2 News Headline Data

Our final cleaned news headline dataset contains 2,195,400 headline observations from August 1, 2009 to May 31, 2020. Each observation is matched to a U.S. common equity and assigned to one of 49 industries. For the stock-month regression analysis, we use a randomly sampled subset of this panel. After merging with CRSP monthly returns and the stock-level controls, the final regression sample contains 259,970 headline observations, corresponding to 107,495 stock-month units.

We first clean and deduplicate the headline-level observations from the raw news headline data

[4] The short-term reversal control is measured in the same month as headline sentiment, while momentum follows the standard convention of skipping the most recent month to separate intermediate-horizon momentum from short-term reversal. Book-to-market is winsorized at the 2.5th and 97.5th percentiles, and momentum is winsorized at the 0.5th and 99.5th percentiles. The winsorization follows, for example, Bolton and Kacperczyk (2021); Duan and Pelger (2025). All controls are measured in decimal units, while stock returns are reported in percentage points.

[5] The control variables are the standard determinants of the cross-section of expected returns, and they are the controls used in the empirical literature on news sentiment. Size and book-to-market are the classic return predictors (Banz, 1981; Fama and French, 1992), and momentum captures return continuation at the intermediate horizon (Jegadeesh and Titman, 1993; Carhart, 1997). Firm-level studies of news and returns condition on the same characteristics: Tetlock, Saar-Tsechansky, and Macskassy (2008) control for the return predictors of Fama and French (1992) together with past returns when testing whether the language of firm-specific news forecasts returns, and Chan (2003) measures post-news drift and reversal relative to size and book-to-market benchmarks. The same-month return $R_{i,t}$ is the short-term reversal control (Jegadeesh, 1990; Lehmann, 1990) and is a natural control for price-trend effects in our setting: monthly returns exhibit systematic short-horizon reversal, and headline sentiment is strongly correlated with the same month's return, so omitting the recent price trend would attribute to sentiment the mechanical continuation or reversal of recent price movements rather than the information in the news itself.

from Aenlle (2020), which is publicly available on Kaggle. The raw dataset covers February 14, 2009 to June 11, 2020, and contains headline-level observations with variables including the news headline, publisher, publication date, and the associated stock ticker.[6] We drop observations with missing headline, ticker, or publication date. The unit of observation in the cleaned headline data is a headline-ticker-date observation. We therefore deduplicate the stacked raw data using the key (headline, ticker, publication_date). The same headline associated with different tickers is retained as separate observations, since one news item may be relevant for multiple firms. After cleaning and deduplication, the dataset contains 3,215,142 headline-ticker-date observations, covering 6,619 unique tickers.

We next match the cleaned headline data to CRSP to restrict the sample to U.S. common equities. We use the CRSP CIZ Names file downloaded from WRDS, which provides time-varying ticker records, security classifications, and SIC codes. We match the Aenlle ticker to the CRSP CIZ ticker and require the headline publication date to fall within the corresponding CRSP security date range.[7] We retain U.S. common equities using the CIZ filters SecurityType = EQTY, SecuritySubType = COM, ShareType = NS, and USIncFlg = Y, which correspond to the legacy CRSP common-share restriction of CRSP share codes 10 and 11. This filtering keeps 2,205,984 observations from 3,581 unique headline tickers.

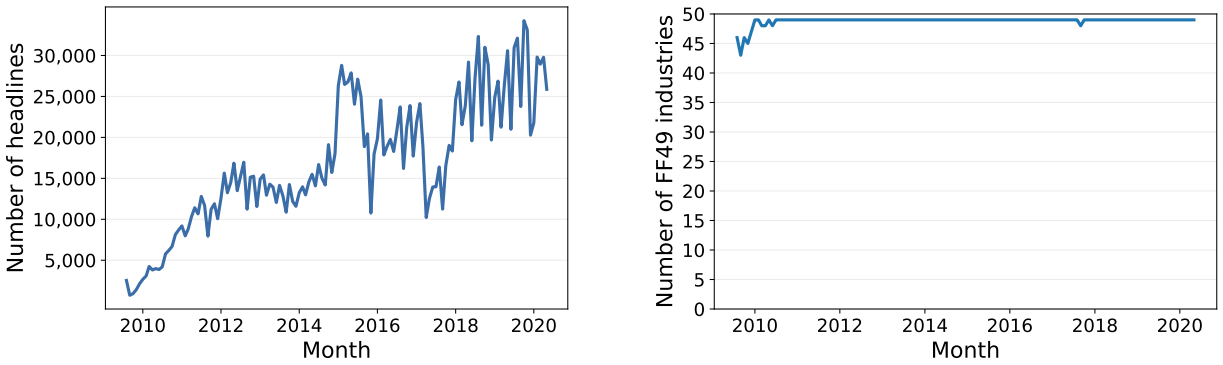
We map each headline observation to an industry using its CRSP SIC code and the 49-industry classification of Fama and French, obtained from Kenneth French's Data Library, which maps four-digit SIC codes into 49 industries, such as Agric, Food, BusSv, or Softw. In the remainder of the paper, we refer to these classifications simply as industries. Out of 2,205,984 CRSP-filtered headline observations, 2,201,048 are successfully assigned to one of the 49 industries; the remaining 4,936 observations have SIC codes that are not covered by the classification and are dropped. We remove the very sparse early coverage before August 2009 and the incomplete endpoint month June 2020. This leaves the final headline panel with 2,195,400 observations, spanning 130 months from August 2009 to May 2020.

Figure 11 describes the time-series coverage of the final news headline dataset. Panel (a) reports the total number of headline observations by month, and Panel (b) reports the number of industries

[6] We use the two raw headline files, `raw_analyst_ratings.csv` and `raw_partner_headlines.csv`. We do not use `analyst_ratings_processed.csv`, since it is a processed version of the analyst-ratings file and may assign different publication dates to otherwise similar headline-ticker observations.

[7] For observations without an exact CRSP ticker-date match, we implement a second-pass match using normalized trading symbols, where punctuation is removed from both the headline ticker and the CRSP trading symbol. This recovers share-class tickers such as BRK.B, which correspond to CRSP trading symbols such as BRKB.

Figure 11: Time-Series Coverage of the News Headline Data



(a) Monthly headline volume

(b) Monthly industry coverage

This figure summarizes the time-series coverage of the final news headline dataset. Panel (a) plots the total number of headline-ticker-date observations in each month. Panel (b) reports the number of industries represented by at least one headline observation in each month. The sample is restricted to U.S. common equities assigned to an industry and covers August 2009 through May 2020.

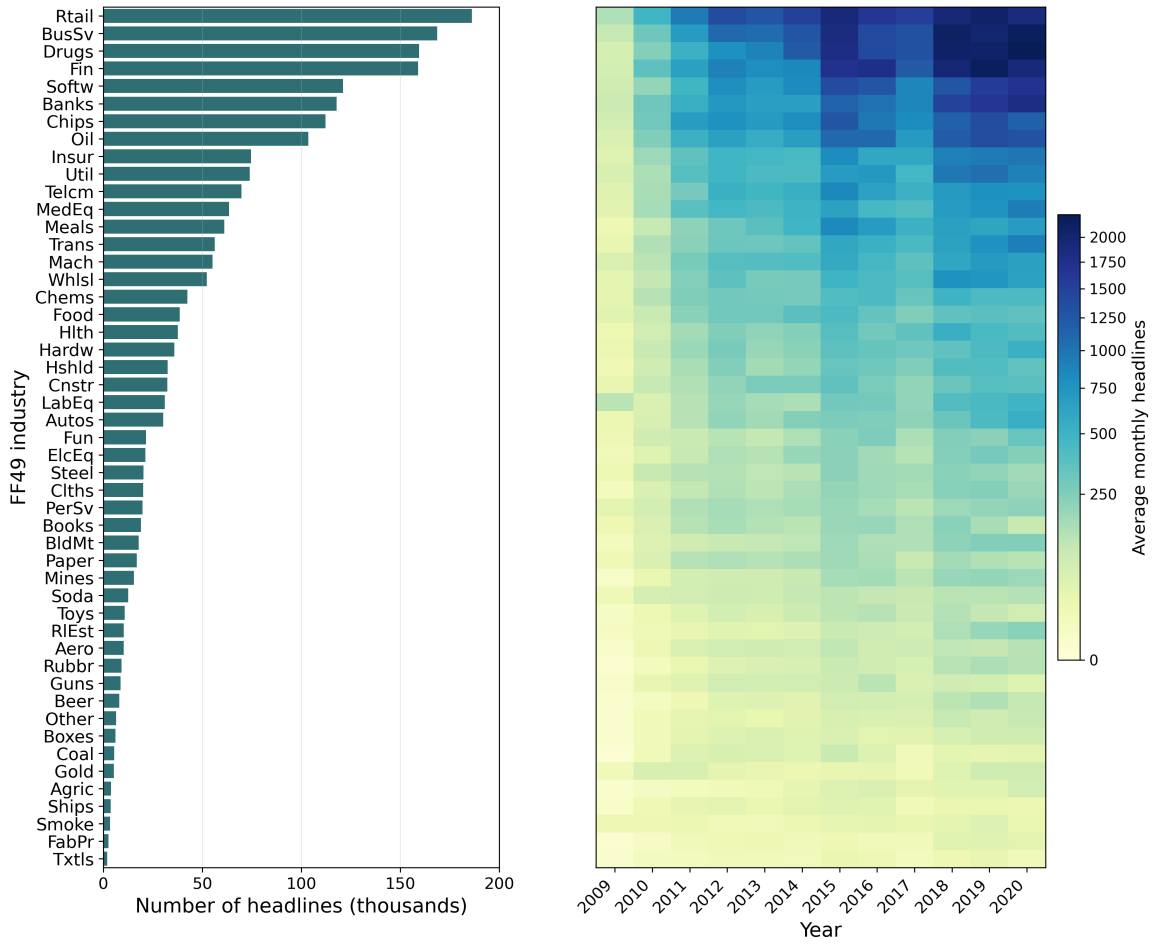
represented in each month. The dataset has broad industry coverage throughout the sample period, with nearly all 49 industries represented in most months. Figure 12 summarizes cross-industry variation in news coverage. The left panel reports total headline counts by industry. Headline volume is concentrated in industries such as Retail, Business Services, Drugs, Finance, Software, Banks, Chips, and Oil. The right panel reports average monthly headline coverage by industry and year, normalizing partial endpoint years by the number of months observed in the sample.

Finally, for this empirical study, we construct a computationally feasible headline sample by retaining up to 50 randomly selected headlines per industry-month. This sampled headline dataset contains 281,759 headline-ticker-date observations. Merging the sampled headlines with stock monthly returns and the control variables yields the final regression sample of 259,970 headline observations and 107,495 stock-month units.

9.3 LLM-Based Sentiment Measures

For each sampled headline, we obtain LLM sentiment labels using multiple model-prompt pairs. Each model-prompt pair classifies whether the headline is positive, neutral, or negative for the associated firm. The main diagnostic analysis uses 12 model-prompt pairs, formed by crossing four models with three prompt designs. The models are GPT-3.5 Turbo, GPT-4o mini, GPT-4.1 nano, and GPT-5 mini. The prompts are the baseline prompt, the fill-in prompt, and the professional-investor prompt.

Figure 12: Cross-Industry Coverage of the News Headline Data



This figure summarizes cross-industry variation in the final news headline dataset. The left panel reports the total number of headline-ticker-date observations assigned to each industry over the final analysis sample, with counts reported in thousands. Industries are sorted by total headline volume. The right panel reports average monthly headline coverage by industry and calendar year. Each industry-year cells equals the total number of headline observations for that industry-year divided by the number of months from that year included in the sample, so that partial endpoint years are comparable to full calendar years.

The baseline prompt asks the model to classify the sentiment of a headline for a specific firm and to return a JSON object with exactly one field, label, whose value must be positive, neutral, or negative. Specifically, the baseline prompt reads:

> *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.*

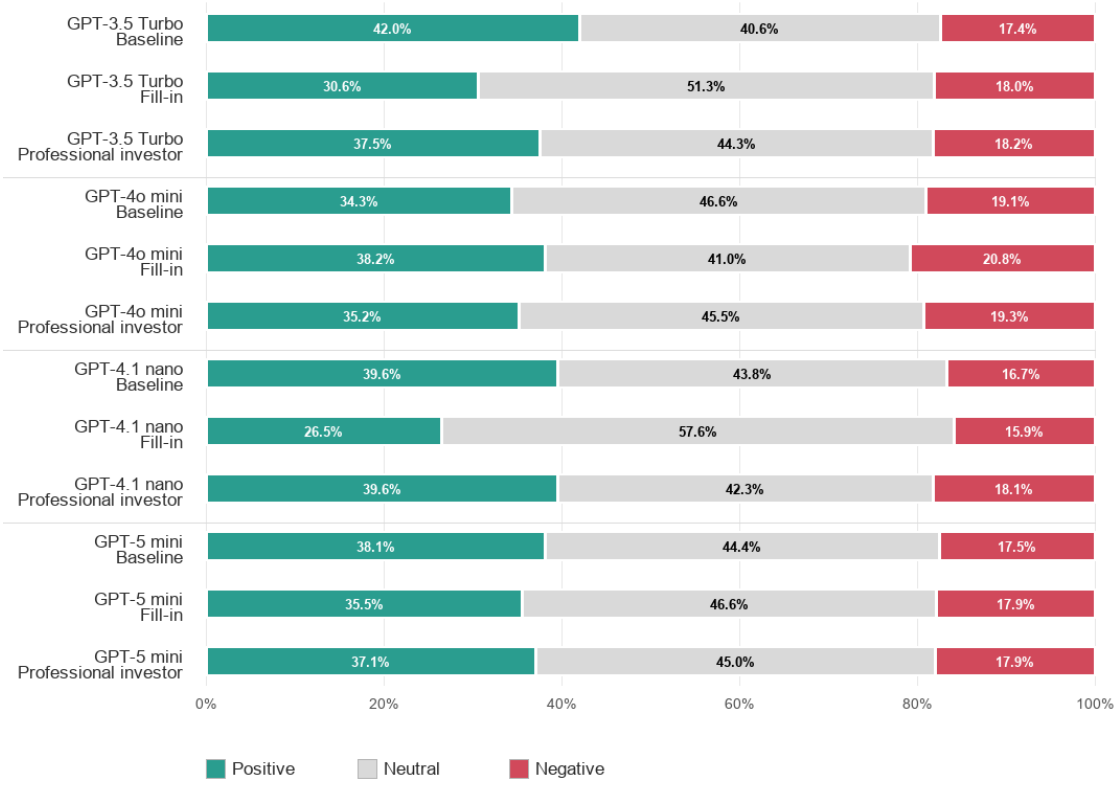
We also use two prompt variants. The fill-in prompt removes the JSON-format instruction and instead asks the model to fill in a blank with one of the three labels. The professional-investor prompt keeps the baseline JSON structure but adds a professional-investor framing before the classification task. The exact prompt templates are reported in Appendix B.1. In all cases, the model receives only the firm name, ticker, and headline text, so the classification is based on information contained in the headline rather than subsequent stock returns or other ex post outcomes. We code positive headlines as +1, neutral headlines as 0, and negative headlines as -1. This produces a headline-level sentiment score for each headline and model-prompt pair.

We first examine how sensitive the headline-level LLM labels are to the choice of model and prompt. Figure 13 reports the distribution of headline-level sentiment labels across the 12 model-prompt pairs in the 281,759-headline sampled dataset. The distribution of labels differs substantially across model-prompt pairs. The positive share ranges from 26.5% for GPT-4.1 nano with the fill-in prompt to 42.0% for GPT-3.5 Turbo with the baseline prompt. The neutral share ranges from 40.6% to 57.6%. Negative labels are more stable, but still range from 15.9% to 20.8%.

Figure 14 reports headline-level pairwise agreement rates. Each off-diagonal entry is the fraction of headlines for which two model-prompt pairs assign the same sentiment label. Agreement is generally high, but far from mechanical. Across all non-identical pairs, the average agreement rate is 83.3%, with the lowest pair agreeing on only 76.3% of headlines. Pairs using the same prompt but different models agree on average 84.2% of the time. Pairs using the same model but different prompts agree more often, with an average agreement rate of 88.3%, although this still leaves meaningful prompt-induced variation.

For each model-prompt pair, we then construct a stock-month sentiment measure by averaging the headline-level sentiment scores across the capped set of headlines assigned to the same stock-month. Specifically, the generated sentiment score for stock i in month t by model-prompt pair m is the average of the headline-level sentiment scores among sampled headlines assigned to stock i

Figure 13: Distribution of LLM sentiment labels across model-prompt pairs



The figure reports the share of sampled headlines classified as positive, neutral, or negative by each model-prompt pair. Positive labels are coded as $+1$, neutral labels as 0, and negative labels as -1.

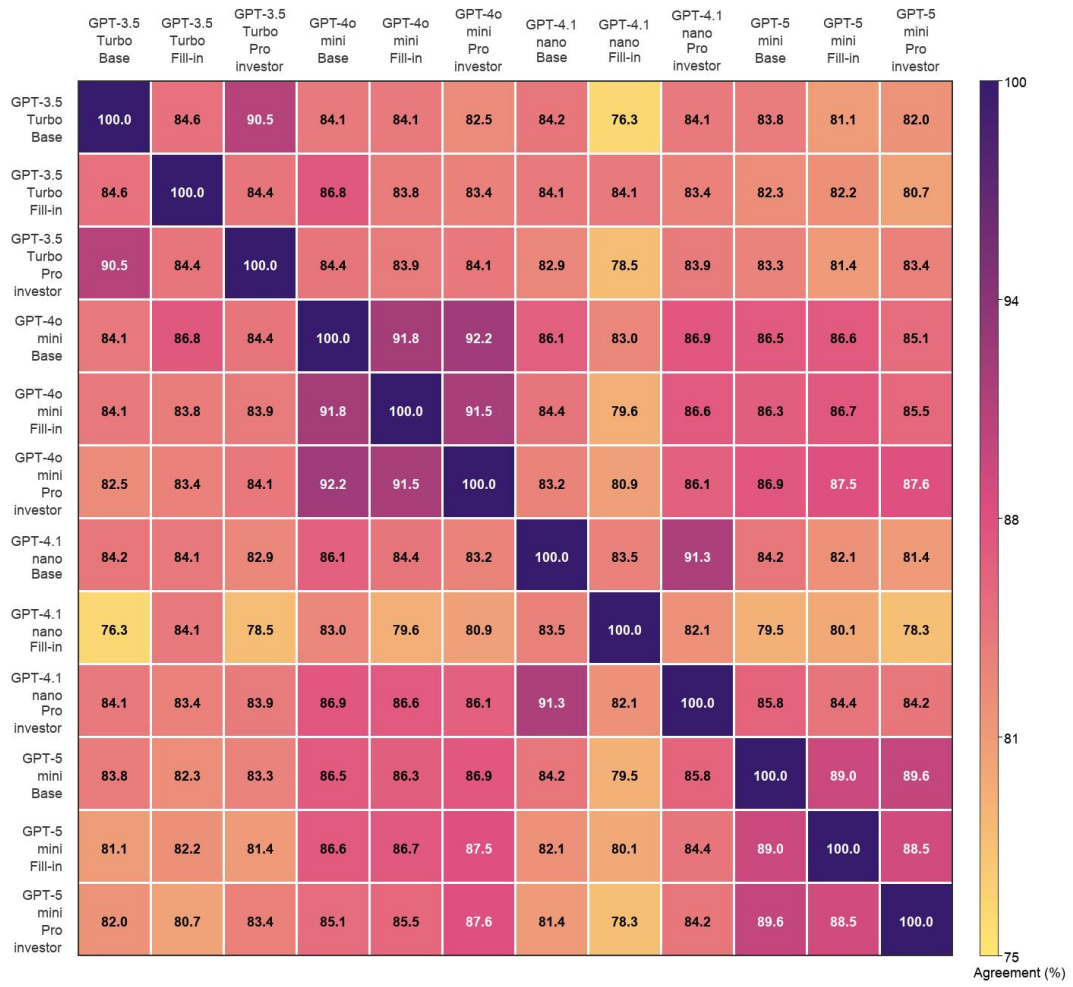
in month t:

$$\tilde{T}_{i,t}^{(m)} = \frac{1}{|H_{i,t}|} \sum_{h \in H_{i,t}} \tilde{s}_h^{(m)},$$

where $\tilde{s}_h^{(m)}$ denotes the headline-level sentiment score for headline h and $H_{i,t}$ denotes the set of headline observations assigned to stock i in month t. A higher value of $\tilde{T}_{i,t}^{(m)}$ indicates more positive news sentiment classified by model-prompt pair m for that stock-month, while a lower value indicates more negative sentiment.

The twelve model–prompt pairs are specified ex ante and are not tuned on the calibration labels, so the cross-fitting procedure of Section 3.5 is not required in this application; it becomes essential when prompts or models are selected using the calibration data.

Figure 14: Headline-level pairwise agreement across model-prompt pairs



Each cell reports the share of sampled headlines for which two model-prompt pairs assign the same sentiment label. Diagonal entries equal one by construction.

9.4 Stock-Level Return Predictability and AI-PI

We now implement AI-PI in the stock-month return regression described in Section 9.1. Specifically, the target parameter is the coefficient on headline sentiment in the regression of next-month stock returns on stock-month sentiment, log firm size, book-to-market, momentum, same-month return, industry fixed effects, and month fixed effects. The estimation sample contains 107,495 stock-month units.

We implement Algorithm 1 with a total human-labeling budget of 5% of stock-month units. The first-stage pilot sample is drawn uniformly at random and accounts for 40% of the total

labeling budget. Using the pilot labels, we estimate both the weighting vector over the generated model-prompt moments and the targeted second-stage labeling policy. In this empirical study, we choose the information variable W_i as the disagreement across the 12 generated sentiment measures, measured as the cross-model-prompt standard deviation of the generated stock-month sentiment. Figure B.1 in Appendix B plots the distribution of W_i across units.

Figure 15 shows the two design objects, labeling policy and weights, learned by Algorithm 1. Figure 15a plots the final labeling probability $\hat{\pi}(W_i)$ as a function of model-prompt disagreement W_i. The mean propensity is 0.050 by construction, with realized propensities ranging from 0.035 to 0.079. The policy is not mechanically monotone in W_i because Algorithm 1 targets the estimated residual moment variability, not disagreement itself. Figure 15b reports the estimated weights $\hat{\lambda}$ on the generated model–prompt moments. The largest weights are placed on GPT-5 mini with the professional-investor prompt and GPT-4o mini with the professional-investor prompt. Several generated measures receive zero weight. The sum of the generated weights is 0.836.

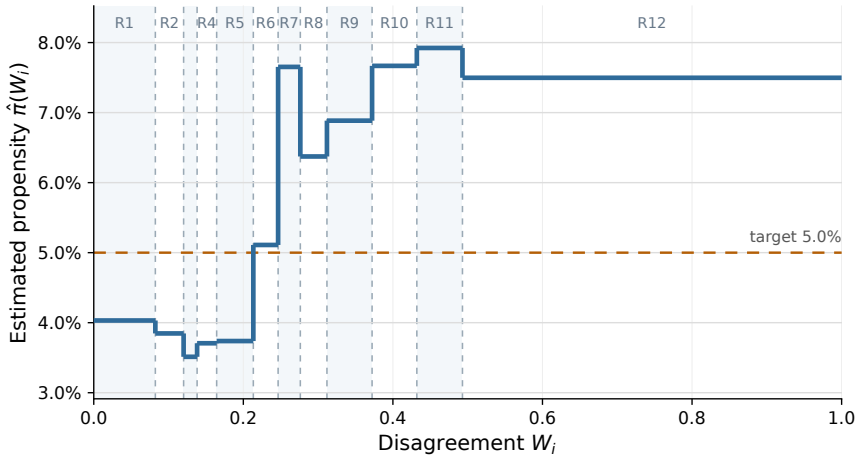
The final labeled sample contains 5,320 stock-month units, corresponding to 4.95% of the estimation sample, and 11,889 headline-level human labels. The headline-level label distribution is stable across the two stages. In the pilot sample, 16.0% of headlines are labeled negative, 50.4% neutral, and 33.6% positive. In the second-stage sample, 15.9% are negative, 49.3% neutral, and 34.8% positive. Labels are coded as -1, 0, and 1 for negative, neutral, and positive sentiment, respectively. Stock-month human-reference sentiment is constructed as the average of headline-level human labels within each labeled stock-month.

The human labels define the reference standard of the analysis, so we document the protocol. All headlines in a selected stock-month were labeled at the headline level by one of the authors, using the same three categories as the LLM classification task. The annotator did not consult any model output during labeling: no pre-labels or model suggestions were used, so the human labels were produced independently of the generated features and are not subject to machine anchoring.[8]

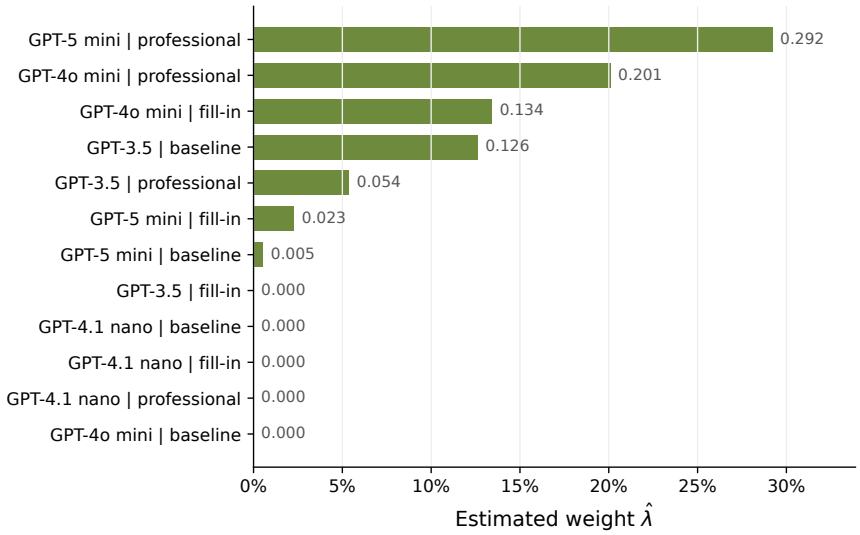
Table 3 reports the main AI-PI estimates. For comparison, the table includes 12 naive generated-sentiment estimators and two benchmark estimators. All specifications use the same controls as in equation (12), and standard errors are clustered by month. The first twelve rows show the estimates a researcher would obtain by choosing a single generated sentiment measure. Specifically,

[8]Because headline sentiment is a judgment rather than an objectively observed quantity, the estimand is protocol-relative: the parameter is defined by this annotation protocol applied to the sampled headline set, and the AI-PI estimate is anchored to this reference standard rather than to an abstract latent truth. Section 6.1 characterizes the consequences of residual imperfections in the reference labels: conditionally centered label noise moves the estimand only at second order in the label-error variance.

Figure 15: AI-PI estimated labeling policy and model-prompt weights



(a) Targeted labeling policy



(b) Model-prompt weights

Panel (a) plots the estimated propensity $\hat{\pi}(W_i)$ used both to draw second-stage labels and in the IPW correction. Panel (b) reports the estimated weights $\hat{\lambda}$ on the generated model-prompt moments.

Table 3: Stock-month sentiment coefficients across estimators

Estimator	Coefficient	SE	95% CI	CI length
GPT-3.5 Turbo – Baseline	0.152**	0.075	[0.005, 0.299]	0.295
GPT-3.5 Turbo – Fill-in	0.226***	0.083	[0.063, 0.389]	0.327
GPT-3.5 Turbo – Professional investor	0.166**	0.078	[0.012, 0.319]	0.307
GPT-4o mini – Baseline	0.170**	0.072	[0.029, 0.310]	0.281
GPT-4o mini – Fill-in	0.153**	0.073	[0.011, 0.296]	0.285
GPT-4o mini – Professional investor	0.150**	0.074	[0.004, 0.296]	0.292
GPT-4.1 nano – Baseline	0.171**	0.078	[0.019, 0.324]	0.305
GPT-4.1 nano – Fill-in	0.203**	0.086	[0.034, 0.371]	0.337
GPT-4.1 nano – Professional investor	0.160**	0.078	[0.007, 0.313]	0.306
GPT-5 mini – Baseline	0.153**	0.075	[0.006, 0.299]	0.293
GPT-5 mini – Fill-in	0.083	0.076	[-0.066, 0.231]	0.297
GPT-5 mini – Professional investor	0.144*	0.079	[-0.011, 0.299]	0.310
Naive uniform estimator	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**

This table reports coefficient estimates on the sentiment measure from the next-month stock return regression (12) across different estimators. The first twelve rows replace the latent human sentiment measure with one generated stock-month sentiment measure. The naive uniform estimator uses the simple average of the 12 generated stock-month sentiment measures and does not use human labels. The calibration-only IPW estimator uses only the human-labeled stock-month sentiment measure and does not use the generated sentiment measures. Both calibration-only IPW and AI-PI estimators use the final 5,320 labeled stock-month units and the recorded propensities from Algorithm 1. All rows include log size, book-to-market, momentum, same-month return $R_{i,t}$, industry fixed effects, and month fixed effects. Standard errors are clustered by month. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

we estimate the stock-month regression in equation (12) using each generated sentiment measure $\tilde{T}_{i,t}^{(m)}$ in place of the latent human sentiment measure $T_{i,t}$. The main pattern is positive short-horizon return predictability, but the magnitude and statistical conclusion depend on the specific model–prompt choice. The coefficients range from 0.083 to 0.226. Eleven of the twelve model–prompt pairs are at least marginally significant at the 10% level, while the GPT-5 mini fill-in prompt remains smaller and statistically insignificant. Thus, the estimated magnitude remains sensitive to the arbitrary model–prompt choice.

The table also includes two benchmark estimators: the naive uniform estimator and calibration-only IPW. The naive uniform estimator uses the simple average of the 12 generated stock-month sentiment measures and does not use human labels. Its coefficient is 0.187, with a standard error of 0.086. The calibration-only IPW estimator uses only the human-labeled stock-month sentiment measure, reweighted by the labeling propensities, and does not use the generated sentiment measures. This estimator is positive but imprecise: its coefficient is 0.561, with a standard error of 0.300

and a 95% confidence interval length of 1.176. The final row reports the AI-PI estimator computed by Algorithm 1. The AI-PI estimate is 0.312, with a standard error of 0.154 and a 95% confidence interval of [0.010, 0.614]. Its confidence interval is roughly half as long as the calibration-only confidence interval, while preserving the human-label correction.[9]

The results in Table 3 show that the naive estimators alone give estimates that vary across model-prompt choices, while AI-PI combines human labels and generated features to obtain a single corrected estimate that is much more precise than the calibration-only estimator, the only other estimator that also uses the human labels. Figure 16 visualizes the point estimates with 95% confidence intervals across different estimators.[10]

In economic terms, the AI-PI estimate implies that a stock whose headlines in a given month are uniformly positive earns a 0.312 percentage point (31 basis point) higher return in the following month than a comparable stock with uniformly neutral news, holding size, book-to-market, momentum, and the same-month return fixed, or roughly 3.7 percentage points annualized. The effect is modest, as expected for next-month predictability once the large contemporaneous price response to news is accounted for (Appendix B.4), but it is measured on a valid scale: the naive uniform estimate of 0.187 is roughly 40 percent smaller, reflecting the attenuation induced by measurement noise in the generated features.

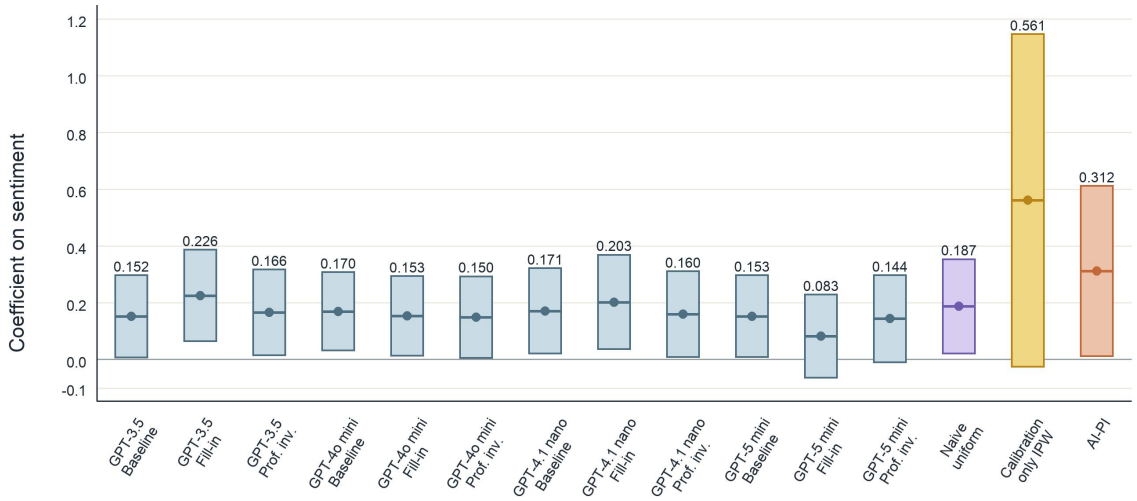
We report additional robustness checks in Appendix B. Appendix B.3 shows that the AI-PI estimate is stable across alternative control sets, ranging from 0.289 to 0.316, so the correction is not driven by the momentum or short-term reversal controls. Appendix B.4 reports the contemporaneous regression, in which all twelve generated measures load strongly on same-month returns, confirming that the sentiment measures capture real information and that the small next-month coefficients reflect the residual short-horizon drift. Appendix B.5 reports a Fama–MacBeth version of the specification, which yields the same qualitative conclusion.

After estimating the return coefficients, we use the human-labeled sample to examine the source of the model-prompt fragility. Table 4 reports headline-level error metrics for each generated model–

[9] The clustered standard errors implement the extension of Section 6.3 with calendar months as the sampling units, of which there are 130, and labeling assigned at the stock-month level within months: the month-level moment is the within-month sum of unit-level moments, with labeled units entering through their inverse-propensity weights as in Remark 8, and the plug-in variance estimator of Remark 1 is applied to these month-level aggregates. Treating months as independent clusters assumes that within-firm dependence across months is absorbed by the controls and fixed effects.

[10] The confidence intervals are computed from the plug-in variance estimator of Remark 1, applied to the month-level aggregates at the realized labeling rate, estimated weights, and estimated propensities. The theory in Section 5.4 justifies this construction when the pilot share vanishes asymptotically, and the simulations in Section 8 show good finite-sample performance of the pooled two-stage implementation with a larger pilot share.

Figure 16: Stock-month sentiment coefficients across estimators



The figure plots coefficient estimates on the sentiment measure and the corresponding 95% confidence intervals from the next-month stock return regression. The first twelve estimates use one generated stock-month sentiment measure at a time. The final three estimates report the naive uniform estimator, the calibration-only IPW estimator, and the AI-PI estimator.

prompt pair on the final human-labeled sample, with rows sorted in descending order of accuracy.

The table shows that accuracy ranges from 75.1% to 86.0% across model–prompt pairs. The best-performing one is GPT-5 mini with the professional-investor prompt, which has accuracy of 86.0%, MAE of 0.144, RMSE of 0.390, and mean bias of 0.016. The strongest models and prompts generally have higher accuracy and lower error, but none is error-free. The average model-minus-human bias ranges from -0.087 to 0.079 across the twelve model–prompt pairs, so the direction and magnitude of the measurement bias depend meaningfully on the generated measure used. This heterogeneity in measurement error helps explain why the naive generated-sentiment estimates vary across model-prompt choices and motivates using the human labels to correct the generated measures rather than selecting a single model-prompt pair.

Two features of Table 4 explain the pattern of the estimates in Table 3. First, the mean biases are small relative to the headline-level noise: the RMSE is between 0.390 and 0.516, several times the largest absolute mean bias. The dominant effect of the measurement error in this application is therefore attenuation: noise in generated sentiment weakens the regression signal and pulls the

Table 4: Model Error Metrics on Human-Labeled Headlines

Model–prompt pair	Accuracy	Bias	MAE	RMSE	Model mean
GPT-5 mini – Professional investor	0.860	0.016	0.144	0.390	0.199
GPT-5 mini – Fill-in	0.857	-0.004	0.147	0.392	0.180
GPT-4o mini – Professional investor	0.852	-0.022	0.150	0.395	0.161
GPT-5 mini – Baseline	0.848	0.031	0.157	0.407	0.214
GPT-4o mini – Fill-in	0.829	-0.003	0.175	0.428	0.180
GPT-4o mini – Baseline	0.818	-0.030	0.185	0.438	0.154
GPT-4.1 nano – Professional investor	0.801	0.044	0.205	0.465	0.227
GPT-3.5 Turbo – Professional investor	0.792	0.022	0.210	0.464	0.205
GPT-3.5 Turbo – Baseline	0.779	0.079	0.224	0.481	0.262
GPT-3.5 Turbo – Fill-in	0.774	-0.061	0.230	0.488	0.122
GPT-4.1 nano – Baseline	0.761	0.058	0.246	0.510	0.241
GPT-4.1 nano – Fill-in	0.751	-0.087	0.255	0.516	0.096
Human mean			0.183		
Headlines			11,889		

The table evaluates the generated headline-level sentiment labels from each model–prompt pair against the human labels in the final labeled sample. Labels are coded as -1, 0, and 1. Accuracy is the share of headlines for which the generated label equals the human label. Bias is the mean model label minus the human label. Mean absolute error (MAE) and root mean squared error (RMSE) are both computed at the headline level.

naive coefficient toward zero. This explains why every naive estimate lies below the human-anchored AI-PI and calibration-only estimates. Second, the estimated weights in Figure 15b concentrate on two of the three most accurate pairs, GPT-5 mini and GPT-4o mini with the professional-investor prompt, so the data-driven weighting recovers the pairs that the ex-post error metrics identify as most reliable, without using this table as an input.

10 Conclusion

This paper proposes AI-Powered Inference (AI-PI), a general framework for valid estimation and inference when covariates are produced by generative AI models. The framework treats LLM outputs as imputations within a method-of-moments structure and addresses the three components of the inferential problem jointly: it corrects the systematic bias induced by errors in generated features using a human-labeled calibration set; it manages residual variation by adaptively weighting multiple model–prompt pairs; and it optimally designs the calibration set by concentrating costly human labels where the generators are least reliable. We establish consistency and asymptotic normality, allowing for non-random and data-adaptive labeling designs, cross-fitting when the LLM pipeline is tuned on the calibration data, and overidentified GMM, and we show that a feasible two-

stage procedure with estimated design is asymptotically normal and attains the oracle asymptotic variance of the jointly optimal weights and labeling policy.

Our results provide a clear message for empirical practice. Diagnostic validation of LLM outputs is not a substitute for bias correction: valid inference requires a calibration set, and the calibration set should be used to correct the moment conditions rather than merely to report agreement rates. The adaptive weighting scheme guarantees graceful degradation: when the generators are poor, the estimator reduces to the calibration-only estimator; when they are informative, it delivers large efficiency gains. In our simulations, even a modest calibration rate of 5–10% generates substantial improvements, and the labels are more valuable when allocated by the optimal design. The empirical application confirms this in practice: with human labels on five percent of the sample, AI-PI turns twelve fragile naive estimates into a single valid estimate whose confidence interval is roughly half the length of the interval based on the human labels alone. When reference labels are noisy, our extension characterizes the reference estimand and clarifies when inference remains anchored to the original target. Stronger models can provide useful noisy auxiliary references for debiasing cheaper models, but human-labeled calibration remains the benchmark when shared model biases are a concern.

References

Abrevaya, J., and S. G. Donald (2017): "A GMM approach for dealing with missing data on regressors," *Review of Economics and Statistics*, 99(4), 657–662.

Aenlle, M. (2020): "Daily Financial News for 6000+ Stocks," https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests, Accessed May 18, 2026.

Aliprantis, C. D., and K. C. Border (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, 3rd edn.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," *arXiv preprint arXiv:2301.09633*.

Angelopoulos, A. N., J. C. Duchi, and T. Zrnic (2023): "Ppi++: Efficient prediction-powered inference," *arXiv preprint arXiv:2311.01453*.

Baker, M., and J. Wurgler (2006): "Investor sentiment and the cross-section of stock returns," *The Journal of Finance*, 61(4), 1645–1680.

Banz, R. W. (1981): "The relationship between return and market value of common stocks," *Journal of financial economics*, 9(1), 3–18.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2025): "Inference for Regression with Variables Generated by AI or Machine Learning," *arXiv preprint arXiv:2402.15585*.

Bolton, P., and M. Kacperczyk (2021): "Do investors care about carbon risk?," *Journal of financial economics*, 142(2), 517–549.

Bybee, J. L. (2023): "The ghost in the machine: Generating beliefs with large language models," *arXiv preprint arXiv:2305.02823*.

Calomiris, C. W., and H. Mamaysky (2019): "How news and its context drive risk and returns around the world," *Journal of Financial Economics*, 133(2), 299–336.

Carhart, M. M. (1997): "On persistence in mutual fund performance," *The Journal of Finance*, 52(1), 57–82.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006): *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC.

Chakrabortty, A., and T. Cai (2018): "Efficient and adaptive linear regression in semi-supervised settings," *The Annals of Statistics*, 46(4), 1541–1572.

Chan, W. S. (2003): "Stock price reaction to news and no-news: Drift and reversal after headlines," *Journal of Financial Economics*, 70(2), 223–260.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

Da, Z., J. Engelberg, and P. Gao (2015): "The sum of all FEARS investor sentiment and asset prices," *The Review of Financial Studies*, 28(1), 1–32.

Duan, J., and M. Pelger (2025): "Imputation-Powered Inference for Missing Covariates," *Working Paper, Stanford University*.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2024): "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models," in *Advances in Neural Information Processing Systems*, vol. 36.

Fama, E. F., and K. R. French (1992): "The cross-section of expected stock returns," *the Journal of Finance*, 47(2), 427–465.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber (2024): "Missing data in asset pricing panels," *The Review of Financial Studies*, p. hhae003.

Fuller, W. A. (2009): *Measurement Error Models*. John Wiley & Sons.

Garcia, D. (2013): "Sentiment during recessions," *The Journal of Finance*, 68(3), 1267–1300.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as data," *Journal of Economic Literature*, 57(3), 535–574.

Gilardi, F., M. Alizadeh, and M. Kubli (2023): "ChatGPT outperforms crowd-workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, 120(30), e2305016120.

He, S., L. Lv, A. Manela, and J. Wu (2025): "Chronologically consistent large language models," *arXiv preprint arXiv:2502.21206*.

Imai, K., and K. Nakamura (2025): "GenAI-Powered Inference," *arXiv preprint arXiv:2507.03897*.

Jegadeesh, N. (1990): "Evidence of predictable behavior of security returns," *The Journal of Finance*, 45(3), 881–898.

Jegadeesh, N., and S. Titman (1993): "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of Finance*, 48(1), 65–91.

Jha, M., J. Qian, M. Weber, and B. Yang (2024): "Generative AI, managerial expectations, and economic activity," *arXiv preprint arXiv:2410.03897*.

Ji, W., L. Lei, and T. Zrnic (2025): "Predictions as surrogates: Revisiting surrogate outcomes in the age of ai," *arXiv preprint arXiv:2501.09731*.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung (2023): "Survey of hallucination in natural language generation," *ACM computing surveys*, 55(12), 1–38.

Kaufman, S., S. Rosset, C. Perlich, and O. Stitelman (2012): "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4), 1–21.

Ke, Z. T., B. T. Kelly, and D. Xiu (2019): "Predicting returns with text data," Working Paper 26186, National Bureau of Economic Research.

Lehmann, B. N. (1990): "Fads, martingales, and market efficiency," *The Quarterly Journal of Economics*, 105(1), 1–28.

Li, S., Z. Shi, Y. Xia, and B. Yang (2025): "The Value of Stock Analysis in News," *Available at SSRN 5754262*.

Little, R. J. (1986): "Survey nonresponse adjustments for estimates of means," *International Statistical Review/Revue Internationale de Statistique*, pp. 139–157.

Lopez-Lira, A., and Y. Tang (2023): "Can ChatGPT forecast stock price movements? Return predictability and large language models," Working paper, University of Florida.

Loughran, T., and B. McDonald (2011): "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, 66(1), 35–65.

Ludwig, J., S. Mullainathan, and A. Rambachan (2024): "Large Language Models: An Applied Econometric Framework," *NBER Working Paper No. 33344*.

Miao, J., X. Miao, Y. Wu, J. Zhao, and Q. Lu (2023): "Assumption-lean and data-adaptive post-prediction inference," *arXiv preprint arXiv:2311.14220*.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, vol. 4, pp. 2111–2245. Elsevier.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866.

Schuettler, J., F. Audrino, and F. Sigrist (2024): "Does sentiment help in asset pricing? A novel approach using large language models and market-based labels," Research Paper 24-69, Swiss Finance Institute.

Tetlock, P. C. (2007): "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, 62(3), 1139–1168.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008): "More than words: Quantifying language to measure firms' fundamentals," *The Journal of Finance*, 63(3), 1437–1467.

Törnberg, P. (2024): "Best practices for text annotation with large language models," *Sociologica*, 18(2), 67–85.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3. Cambridge university press.

Wasserman, L., and J. Lafferty (2007): "Statistical analysis of semi-supervised regression," *Advances in Neural Information Processing Systems*, 20.

Zrnic, T., and E. J. Candès (2024a): "Active Statistical Inference," *arXiv preprint arXiv:2403.03208*.

——— (2024b): "Cross-prediction-powered inference," *Proceedings of the National Academy of Sciences*, 121(15), e2322083121.

A Proofs

A.1 Proof of Theorem 1

Proof. Let $\tilde{T}^0_{i,m,s} \coloneqq T_i + b_m(Z_i) + \epsilon_{m,s}(Z_i)$ denote the no-spillover counterpart of $\tilde{T}_{i,m,s}$, obtained by setting $u_{i,m,N} = 0$. Let $\tilde{T}^0_{i,m}$ be the seed average of $\tilde{T}^0_{i,m,s}$. Let $G^0_{\lambda,\pi}(\theta)$ denote the corresponding no-spillover version of $G_{\lambda,\pi}(\theta)$, obtained by replacing $\tilde{T}_{i,m}$ with $\tilde{T}^0_{i,m}$. Define

$$G_\Delta(\theta) \coloneqq G_{\lambda,\pi}(\theta) - G^0_{\lambda,\pi}(\theta).$$

By construction,

$$G_\Delta(\theta) = \sum_{m=1}^{M} \lambda_m \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{D_i}{\pi_i}\right) \left(g(Z_i, \tilde{T}_{i,m}; \theta) - g(Z_i, \tilde{T}^0_{i,m}; \theta)\right).$$

We first analyze the asymptotic behavior of $G^0_{\lambda,\pi}(\theta)$. For notational simplicity, define

$$g_i(\theta) \coloneqq g(Z_i, T_i; \theta), \qquad b_i(\theta) \coloneqq \sum_{m=1}^{M} \lambda_m g(Z_i, \tilde{T}^0_{i,m}; \theta).$$

Then, $G^0_{\lambda,\pi}(\theta) = N^{-1} \sum_{i=1}^{N} \psi_i(\theta)$, where

$$\psi_i(\theta) \coloneqq b_i(\theta) + \frac{D_i}{\pi_i}\big(g_i(\theta) - b_i(\theta)\big).$$

By Assumption 1.1, $\mathbb{E}[D_i/\pi_i \mid \mathcal{Z}_N] = 1$, and hence, by iterated expectations over any sub-σ-field of $\mathcal{Z}_N$,

$$\mathbb{E}\left[\frac{D_i}{\pi_i} \,\middle|\, Z_i, T_i, W_i\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{D_i}{\pi_i} \,\middle|\, \mathcal{Z}_N\right] \middle| Z_i, T_i, W_i\right] = 1.$$

Therefore, for any fixed θ,

$$\mathbb{E}\left[\frac{D_i}{\pi_i} g_i(\theta)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{D_i}{\pi_i} g_i(\theta) \,\middle|\, Z_i, T_i, W_i\right]\right] = \mathbb{E}[g_i(\theta)] = G(\theta),$$

and, since $\tilde{T}^0_{i,m}$ is $\mathcal{Z}_N$-measurable, for each model–prompt pair m,

$$\mathbb{E}\left[\frac{D_i}{\pi_i} g(Z_i, \tilde{T}^0_{i,m}; \theta)\right] = \mathbb{E}[g(Z_i, \tilde{T}^0_{i,m}; \theta)].$$

Therefore, $\mathbb{E}[\psi_i(\theta)] = G(\theta)$, and hence $\mathbb{E}[G^0_{\lambda,\pi}(\theta)] = G(\theta)$ for every $\theta \in \Theta$.

We next prove the pointwise convergence of $G^0_{\lambda,\pi}(\theta)$ for any fixed $\theta \in \Theta$. Since M and λ are fixed, Assumption 2.1 implies $\mathbb{E}\|b_i(\theta)\|^2 < \infty$. Moreover,

$$\|\psi_i(\theta)\|^2 \le 2\|b_i(\theta)\|^2 + 2\Big(\frac{D_i}{\pi_i}\Big)^2 \|g_i(\theta) - b_i(\theta)\|^2.$$

By Assumption 1.2, $\pi_i \ge \rho_{\min} r_N$ almost surely,

$$\mathbb{E}\left[\Big(\frac{D_i}{\pi_i}\Big)^2 \,\middle|\, W_i\right] = \frac{1}{\pi_i} \le \frac{1}{\rho_{\min} r_N}.$$

Thus, $\mathbb{E}\|\psi_i(\theta)\|^2 = O(r_N^{-1})$. Because the unit-level variables are i.i.d., the no-spillover features are unit-wise functions of them, and, by Assumption 1.1, the labels are conditionally independent Bernoulli draws given $\mathcal{Z}_N$ with parameters $\pi(W_i)$, the vectors $(Z_i, T_i, \tilde{T}^0_{i,\cdot}, W_i, D_i)$ are i.i.d. across i. Hence,

$$\mathrm{Var}(G^0_{\lambda,\pi}(\theta)) = \frac{1}{N}\mathrm{Var}(\psi_i(\theta)) = O\left(\frac{1}{N r_N}\right) = O\left(\frac{1}{n}\right) = o(1).$$

Since $\mathbb{E}[G^0_{\lambda,\pi}(\theta)] = G(\theta)$, Chebyshev's inequality yields $G^0_{\lambda,\pi}(\theta) \xrightarrow{p} G(\theta)$ for any $\theta \in \Theta$.

(i) Consistency of the estimator $\hat{\theta}_{\lambda,\pi}$ under compactness.

We now establish uniform convergence of $G^0_{\lambda,\pi}(\theta)$ on compact sets. Let $K \subset \Theta$ be any compact set contained in the neighborhood of θ_0 on which Assumptions 2.3 and 2.4 hold. For any $\theta_1, \theta_2 \in K$,

$$\|g_i(\theta_1) - g_i(\theta_2)\| \le L_{\theta,H}(Z_i)\|\theta_1 - \theta_2\|,$$

and

$$\|g(Z_i, \tilde{T}^0_{i,m}; \theta_1) - g(Z_i, \tilde{T}^0_{i,m}; \theta_2)\| \le L_{\theta,m}(Z_i)\|\theta_1 - \theta_2\|.$$

Therefore

$$\|\psi_i(\theta_1) - \psi_i(\theta_2)\| \le L_i\|\theta_1 - \theta_2\|,$$

where

$$L_i = \frac{D_i}{\pi_i} L_{\theta,H}(Z_i) + \sum_{m=1}^{M} |\lambda_m| \Big(1 + \frac{D_i}{\pi_i}\Big) L_{\theta,m}(Z_i).$$

By Assumption 1.1,

$$\mathbb{E}\left[\frac{D_i}{\pi_i} L_{\theta,H}(Z_i)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{D_i}{\pi_i} \,\middle|\, Z_i, T_i, W_i\right] L_{\theta,H}(Z_i)\right] = \mathbb{E}[L_{\theta,H}(Z_i)] < \infty,$$

and similarly,

$$\mathbb{E}\left[\frac{D_i}{\pi_i} L_{\theta,m}(Z_i)\right] = \mathbb{E}[L_{\theta,m}(Z_i)] < \infty.$$

Thus $\mathbb{E}[L_i] < \infty$ uniformly in N. Moreover, since the L_i form a row-wise i.i.d. triangular array with $\mathbb{E}[L_i^2] \leq C/(\rho_{\min} r_N)$ by Assumptions 1.2 and 2.3, Chebyshev's inequality gives $N^{-1}\sum_{i=1}^N L_i = \mathbb{E}[L_i] + O_p\big((N r_N)^{-1/2}\big) = O_p(1)$.

Fix $\delta > 0$. Since K is compact, there exists a finite δ-net $\{\theta_1, \ldots, \theta_J\} \subset K$. For any $\theta \in K$, choose θ_j such that $\|\theta - \theta_j\| \leq \delta$. Then

$$\begin{aligned}\|G^0_{\lambda,\pi}(\theta) - G(\theta)\| &\leq \|G^0_{\lambda,\pi}(\theta_j) - G(\theta_j)\| + \|G^0_{\lambda,\pi}(\theta) - G^0_{\lambda,\pi}(\theta_j)\| + \|G(\theta) - G(\theta_j)\| \\ &\leq \max_{1\leq j\leq J} \|G^0_{\lambda,\pi}(\theta_j) - G(\theta_j)\| + \delta \cdot \frac{1}{N}\sum_{i=1}^N L_i + \delta \cdot \mathbb{E}[L_i].\end{aligned}$$

The first term is $o_p(1)$ because J is finite and pointwise convergence holds at each θ_j. The second term is $O_p(\delta)$, and the third term is $O(\delta)$. Since $\delta > 0$ is arbitrary, we obtain

$$\sup_{\theta\in K} \|G^0_{\lambda,\pi}(\theta) - G(\theta)\| = o_p(1).$$

We now show that the spillover term is negligible. By Assumption 2.4,

$$\|g(Z_i, \tilde{T}_{i,m}; \theta) - g(Z_i, \tilde{T}^0_{i,m}; \theta)\| \leq L_T(Z_i)\|u_{i,m,N}\|$$

uniformly over $\theta \in K$. Hence,

$$\sup_{\theta\in K} \|G_\Delta(\theta)\| \leq \sum_{m=1}^M |\lambda_m|\, U_{m,N} \cdot \underbrace{\frac{1}{N}\sum_{i=1}^N \left|1 - \frac{D_i}{\pi_i}\right| \cdot L_T(Z_i)}_{O_p(1)},$$

where $U_{m,N} = \max_i \|u_{i,m,N}\|$. The indicated average is $O_p(1)$ by Markov's inequality, because $\mathbb{E}[|1 - D_i/\pi_i| L_T(Z_i)] \leq \mathbb{E}[L_T(Z_i)] + \mathbb{E}[(D_i/\pi_i)L_T(Z_i)] = 2\,\mathbb{E}[L_T(Z_i)] < \infty$, using Assumption 1.1. By Assumption 3, $U_{m,N} = o_p(n^{-1/2})$, and therefore we have $\sup_{\theta\in K} \|G_\Delta(\theta)\| = o_p(n^{-1/2})$. Combining the no-spillover uniform convergence and the spillover bound gives

$$\sup_{\theta\in K} \|G_{\lambda,\pi}(\theta) - G(\theta)\| \leq \sup_{\theta\in K} \|G^0_{\lambda,\pi}(\theta) - G(\theta)\| + \sup_{\theta\in K} \|G_\Delta(\theta)\| = o_p(1).$$

In particular, if Θ is compact and Assumptions 2.3 and 2.4 hold globally on Θ, taking $K = \Theta$ gives

$$\sup_{\theta \in \Theta} \|G_{\lambda,\pi}(\theta) - G(\theta)\| = o_p(1).$$

Since $G(\theta) = 0$ has the unique solution θ_0 and G is continuous, for every $\varepsilon > 0$,

$$\inf_{\theta \in \Theta:\, \|\theta - \theta_0\| \geq \varepsilon} \|G(\theta)\| > 0.$$

Hence,

$$\inf_{\theta \in \Theta:\, \|\theta - \theta_0\| \geq \varepsilon} \|G_{\lambda,\pi}(\theta)\| \geq \inf_{\theta \in \Theta:\, \|\theta - \theta_0\| \geq \varepsilon} \|G(\theta)\| - \sup_{\theta \in \Theta} \|G_{\lambda,\pi}(\theta) - G(\theta)\|.$$

The second term is $o_p(1)$, so with probability tending to one there is no zero of $G_{\lambda,\pi}$ outside the ε-ball around θ_0. Since $\hat{\theta}_{\lambda,\pi}$ is the unique solution to $G_{\lambda,\pi}(\theta) = 0$ with probability tending to one, it follows that $\hat{\theta}_{\lambda,\pi} \xrightarrow{p} \theta_0$.

(ii) Consistency of the estimator $\hat{\theta}_{\lambda,\pi}$ under the monotonicity condition.

Choose $\varepsilon > 0$ small enough so that the closed ball

$$B_{2\varepsilon}(\theta_0) := \{\theta \in \Theta : \|\theta - \theta_0\| \leq 2\varepsilon\}$$

lies inside the neighborhood where Assumptions 2.3 and 2.4 hold. By (i),

$$\sup_{\theta \in B_{2\varepsilon}(\theta_0)} \|G_{\lambda,\pi}(\theta) - G(\theta)\| = o_p(1).$$

Since $G(\theta_0) = 0$, G is differentiable at θ_0, and J_{θ_0} is positive definite, there exist constants $c > 0$ and $\varepsilon > 0$ such that

$$G(\tilde{\theta})^\top (\tilde{\theta} - \theta_0) \geq c\varepsilon^2 \qquad \text{for all } \tilde{\theta} \in \partial B_\varepsilon(\theta_0),$$

where $\partial B_\varepsilon(\theta_0) = \{\theta \in \Theta : \|\theta - \theta_0\| = \varepsilon\}$. Hence, with probability tending to one,

$$G_{\lambda,\pi}(\tilde{\theta})^\top (\tilde{\theta} - \theta_0) \geq \frac{1}{2} c\varepsilon^2 \qquad \text{for all } \tilde{\theta} \in \partial B_\varepsilon(\theta_0).$$

Now fix any $\theta \in \Theta$ such that $\|\theta - \theta_0\| \geq 2\varepsilon$, and define

$$\tilde{\theta} := \theta_0 + \varepsilon \frac{\theta - \theta_0}{\|\theta - \theta_0\|} \in \partial B_\varepsilon(\theta_0).$$

By monotonicity,

$$\big(G_{\lambda,\pi}(\theta) - G_{\lambda,\pi}(\tilde{\theta})\big)^\top(\theta - \tilde{\theta}) \geq 0,$$

which is equivalent to

$$G_{\lambda,\pi}(\theta)^\top(\tilde{\theta} - \theta) \leq G_{\lambda,\pi}(\tilde{\theta})^\top(\tilde{\theta} - \theta).$$

Since

$$\tilde{\theta} - \theta = -\frac{\|\theta - \tilde{\theta}\|}{\varepsilon}(\tilde{\theta} - \theta_0),$$

we obtain

$$G_{\lambda,\pi}(\theta)^\top(\tilde{\theta} - \theta) \leq -\frac{\|\theta - \tilde{\theta}\|}{\varepsilon} G_{\lambda,\pi}(\tilde{\theta})^\top(\tilde{\theta} - \theta_0) \leq -\frac{1}{2}c\varepsilon\|\theta - \tilde{\theta}\| < 0.$$

Therefore $G_{\lambda,\pi}(\theta) \neq 0$ for every θ with $\|\theta - \theta_0\| \geq 2\varepsilon$, with probability tending to one. Since the corresponding estimator is the unique solution to $G_{\lambda,\pi}(\theta) = 0$ with probability tending to one, it must lie inside $B_{2\varepsilon}(\theta_0)$ with probability tending to one. Hence $\hat{\theta} \xrightarrow{p} \theta_0$. □

A.2 Proof of Theorem 2

Proof. As in the proof of Theorem 1, define

$$g_i(\theta) := g(Z_i, T_i; \theta), \qquad b_i(\theta) := \sum_{m=1}^{M} \lambda_m g(Z_i, \tilde{T}^0_{i,m}; \theta),$$

and

$$\psi_i(\theta) := b_i(\theta) + \frac{D_i}{\pi_i}\big(g_i(\theta) - b_i(\theta)\big).$$

Let $G^0_{\lambda,\pi}(\theta) = N^{-1}\sum_{i=1}^{N}\psi_i(\theta)$. As shown in the proof of Theorem 1, Assumption 1 implies that $\mathbb{E}[\psi_i(\theta)] = G(\theta)$, and therefore, $\mathbb{E}[\psi_i(\theta_0)] = 0$. The AI-PI moment $G_{\lambda,\pi}$ differs from the no-spillover moment $G^0_{\lambda,\pi}$ by $G_\Delta(\theta) = G_{\lambda,\pi}(\theta) - G^0_{\lambda,\pi}(\theta)$. By Assumptions 2.4 and 3, as established in the proof of Theorem 1, $\sup_{\theta\in K}\|G_\Delta(\theta)\| = o_p(n^{-1/2})$ uniformly for θ in a neighborhood K of θ_0, so that $\sqrt{n}\sup_{\theta\in K}\|G_\Delta(\theta)\| = o_p(1)$.

By Theorem 1, $\hat{\theta}_{\lambda,\pi} \xrightarrow{p} \theta_0$. Since $G_{\lambda,\pi}(\hat{\theta}_{\lambda,\pi}) = 0$ and $G(\theta_0) = 0$,

$$0 = G^0_{\lambda,\pi}(\theta_0) + \big(G(\hat{\theta}_{\lambda,\pi}) - G(\theta_0)\big) + \Big[(G^0_{\lambda,\pi} - G)(\hat{\theta}_{\lambda,\pi}) - (G^0_{\lambda,\pi} - G)(\theta_0)\Big] + G_\Delta(\hat{\theta}_{\lambda,\pi}).$$

By Assumption 2.2, G is differentiable at θ_0, so

$$G(\hat{\theta}_{\lambda,\pi}) - G(\theta_0) = J_{\theta_0}(\hat{\theta}_{\lambda,\pi} - \theta_0) + o(\|\hat{\theta}_{\lambda,\pi} - \theta_0\|).$$

Next, Assumption 2.3 implies that, in a neighborhood of θ_0,

$$\|\psi_i(\theta) - \psi_i(\theta_0)\| \le L_i \|\theta - \theta_0\|,$$

where

$$L_i := \frac{D_i}{\pi_i} L_{\theta,H}(Z_i) + \sum_{m=1}^{M} |\lambda_m| \Big(1 + \frac{D_i}{\pi_i}\Big) L_{\theta,m}(Z_i).$$

Using Assumption 1.2, we have $\mathbb{E}[r_N L_i^2] \lesssim \mathbb{E}[L_{\theta,H}(Z_i)^2] + \sum_{m=1}^{M} \mathbb{E}[L_{\theta,m}(Z_i)^2] < \infty$. Hence, for every $\delta > 0$,

$$\mathbb{E}\left[\sup_{\|\theta-\theta_0\|\le\delta} r_N \|\psi_i(\theta) - \psi_i(\theta_0)\|^2\right] \le C\delta^2.$$

Since $\hat{\theta}_{\lambda,\pi} \xrightarrow{p} \theta_0$, there exists a deterministic sequence $\delta_N \downarrow 0$ such that

$$P(\|\hat{\theta}_{\lambda,\pi} - \theta_0\| > \delta_N) \to 0.$$

For each N, the class $\mathcal{F}_N = \{\sqrt{r_N}(\psi_i(\theta) - \psi_i(\theta_0)) : \|\theta - \theta_0\| \le \delta\}$ is indexed by a bounded subset of $\mathbb{R}^d$ and is Lipschitz in θ with square-integrable envelope $F_N = \sqrt{r_N}\, L_i\, \delta$, where $\mathbb{E}[F_N^2] \le C\delta^2$ uniformly in N by the display above. Hence its L_2-bracketing numbers satisfy $N_{[]}(\varepsilon \|F_N\|_{P,2}, \mathcal{F}_N, L_2(P)) \le C(1/\varepsilon)^d$ with constants that do not depend on N (Van der Vaart, 2000, Example 19.7), and the bracketing maximal inequality (Van der Vaart, 2000, Lemma 19.34) applies for each N with constants depending only on the bracketing numbers and the envelope norm, hence uniformly along the triangular array. Therefore, we obtain

$$\sup_{\|\theta-\theta_0\|\le\delta_N} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big[\sqrt{r_N}\big(\psi_i(\theta) - \psi_i(\theta_0)\big) - \mathbb{E}\big[\sqrt{r_N}\big(\psi_i(\theta) - \psi_i(\theta_0)\big)\big]\big] \right\| = o_p(1).$$

Equivalently,

$$\sqrt{n}\Big[(G^0_{\lambda,\pi} - G)(\hat{\theta}_{\lambda,\pi}) - (G^0_{\lambda,\pi} - G)(\theta_0)\Big] = o_p(1).$$

Combining the previous steps gives

$$J_{\theta_0}\sqrt{n}\,(\hat{\theta}_{\lambda,\pi}-\theta_0) = -\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0) - o\Big(\sqrt{n}\,\|\hat{\theta}_{\lambda,\pi}-\theta_0\|\Big) + o_p(1).$$

We now show that $\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0) = O_p(1)$. By independence across i and $\mathbb{E}[\psi_i(\theta_0)] = 0$,

$$\begin{aligned}\mathbb{E}\left\|\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0)\right\|^2 &= \mathbb{E}\,\|\sqrt{r_N}\,\psi_i(\theta_0)\|^2 \\ &\le 2r_N\mathbb{E}\|b_i(\theta_0)\|^2 + 2r_N\mathbb{E}\left[\left(\frac{D_i}{\pi_i}\right)^2\|g_i(\theta_0)-b_i(\theta_0)\|^2\right].\end{aligned}$$

Since $\mathbb{E}\left[(D_i/\pi_i)^2\,\middle|\,W_i\right] = \pi_i^{-1}$, we have

$$\begin{aligned}r_N\mathbb{E}\left[\left(\frac{D_i}{\pi_i}\right)^2\|g_i(\theta_0)-b_i(\theta_0)\|^2\right] &= \mathbb{E}\left[\frac{r_N}{\pi_i}\|g_i(\theta_0)-b_i(\theta_0)\|^2\right] \\ &\le \frac{1}{\rho_{\min}}\mathbb{E}\|g_i(\theta_0)-b_i(\theta_0)\|^2 < \infty.\end{aligned}$$

Hence $\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0) = O_p(1)$. Since J_{θ_0} is positive definite, it is invertible, and therefore

$$\sqrt{n}\,\|\hat{\theta}_{\lambda,\pi}-\theta_0\| \le \|J_{\theta_0}^{-1}\|\left(\|\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0)\| + o_p(1) + o\Big(\sqrt{n}\,\|\hat{\theta}_{\lambda,\pi}-\theta_0\|\Big)\right).$$

This implies $\sqrt{n}\,(\hat{\theta}_{\lambda,\pi}-\theta_0) = O_p(1)$, and thus,

$$\sqrt{n}\,(\hat{\theta}_{\lambda,\pi}-\theta_0) = -J_{\theta_0}^{-1}\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0) + o_p(1).$$

Finally, we apply a central limit theorem to $\sqrt{n}\,G^0_{\lambda,\pi}(\theta_0)$. The summands $\sqrt{r_N}\,\psi_i(\theta_0)$ form a row-wise i.i.d. triangular array of mean-zero random vectors (the array structure arises because $\pi_i = r_N\rho(W_i)$ varies with N), so we verify the Lyapunov condition. Using $\|\psi_i(\theta_0)\|^{2+\delta} \le C\big(\|b_i(\theta_0)\|^{2+\delta} + (D_i/\pi_i)^{2+\delta}\|g_i(\theta_0)-b_i(\theta_0)\|^{2+\delta}\big)$, $\mathbb{E}[(D_i/\pi_i)^{2+\delta} \mid W_i] = \pi_i^{-(1+\delta)} \le (\rho_{\min}r_N)^{-(1+\delta)}$, and the $(2+\delta)$-moment condition of the theorem,

$$\mathbb{E}\big\|\sqrt{r_N}\,\psi_i(\theta_0)\big\|^{2+\delta} \le C\,r_N^{1+\delta/2}\Big(\mathbb{E}\|b_i(\theta_0)\|^{2+\delta} + (\rho_{\min}r_N)^{-(1+\delta)}\,\mathbb{E}\|g_i(\theta_0)-b_i(\theta_0)\|^{2+\delta}\Big) = O\big(r_N^{-\delta/2}\big).$$

Hence the Lyapunov ratio satisfies $N^{-\delta/2}\,\mathbb{E}\|\sqrt{r_N}\psi_i(\theta_0)\|^{2+\delta} = O\big((Nr_N)^{-\delta/2}\big) = O(n^{-\delta/2}) \to 0$. Since $\Omega^0_{\lambda,\pi}$ defined below is positive definite by assumption, the Lindeberg–Feller central limit

theorem (applied to $a^\top \sqrt{r_N}\psi_i(\theta_0)$ for each $a \in \mathbb{R}^d$, together with the Cramér–Wold device) yields

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sqrt{r_N}\, \psi_i(\theta_0) \;\xrightarrow{d}\; \mathcal{N}(0, \Omega^0_{\lambda,\pi}),$$

where $\Omega^0_{\lambda,\pi} = \lim_{N\to\infty} \mathbb{E}[r_N \psi_i(\theta_0)\psi_i(\theta_0)^\top]$, and the limit exists because $\rho(\cdot)$ is fixed and $r_N \to r$.

To simplify notation, define $\psi_i \coloneqq \psi_i(\theta_0), g_i \coloneqq g_i(\theta_0)$ and $b_i \coloneqq b_i(\theta_0)$. Observe that

$$\begin{aligned}
\mathbb{E}[\psi_i \psi_i^\top] &= \mathbb{E}\left[\left(1 - 2\frac{D_i}{\pi_i} + \frac{D_i}{\pi_i^2}\right) b_i b_i^\top + \left(\frac{D_i}{\pi_i} - \frac{D_i}{\pi_i^2}\right)(b_i g_i^\top + g_i b_i^\top) + \frac{D_i}{\pi_i^2} g_i g_i^\top\right] \\
&= \mathbb{E}\left[g_i g_i^\top + \frac{1-\pi_i}{\pi_i}(g_i - b_i)(g_i - b_i)^\top\right].
\end{aligned}$$

Define $R^0_\lambda(W_i) = \mathbb{E}[(g_i - b_i)(g_i - b_i)^\top \mid W_i]$. Multiplying by r_N gives

$$\mathbb{E}[r_N \psi_i \psi_i^\top] = r_N \mathbb{E}[g_i g_i^\top] + \mathbb{E}\left[r_N \left(\frac{1}{\pi_i} - 1\right) R^0_\lambda(W_i)\right].$$

Since $\pi_i = r_N \rho(W_i)$,

$$r_N \left(\frac{1}{\pi_i} - 1\right) = \frac{1}{\rho(W_i)} - r_N.$$

Also $\mathbb{E}[g_i] = G(\theta_0) = 0$, so $\mathbb{E}[g_i g_i^\top] = \mathrm{Var}(g_i)$. Therefore

$$\mathbb{E}[r_N \psi_i \psi_i^\top] = r_N \mathrm{Var}(g_i) + \mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r_N\right) R^0_\lambda(W_i)\right],$$

and

$$\Omega^0_{\lambda,\pi} = r\mathrm{Var}(g_i) + E\left[\left(\frac{1}{\rho(W_i)} - r\right) R^0_\lambda(W_i)\right].$$

We now show that the same limiting covariance can be written using the observed generated features $\tilde{T}_{i,m}$. Define

$$e^0_i(\lambda) = g(Z_i, T_i; \theta_0) - \sum_{m=1}^{M} \lambda_m g(Z_i, \tilde{T}^0_{i,m}; \theta_0),$$

and

$$e_i(\lambda) = g(Z_i, T_i; \theta_0) - \sum_{m=1}^{M} \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0).$$

By Assumption 2.4,

$$\|e_i(\lambda) - e^0_i(\lambda)\| \le L_T(Z_i) \sum_{m=1}^{M} |\lambda_m|\, \|u_{i,m,N}\| \le C \cdot L_T(Z_i) U_N,$$

where $U_N = \max_m U_{m,N}$. By Assumption 1.2, $|1/\rho(W_i) - r_N|$ is uniformly bounded. Hence,

$$\begin{aligned}
&\left\|\mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r_N\right)\left(R_\lambda(W_i) - R^0_\lambda(W_i)\right)\right]\right\| \\
&= \left\|\mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r_N\right)\mathbb{E}\left[e_i(\lambda)e_i(\lambda)^\top - e^0_i(\lambda)e^0_i(\lambda)^\top \mid W_i\right]\right]\right\| \\
&\lesssim \mathbb{E}\left\|e_i(\lambda)e_i(\lambda)^\top - e^0_i(\lambda)e^0_i(\lambda)^\top\right\| \\
&\lesssim \mathbb{E}\big[L_T(Z_i)U_N\|e^0_i(\lambda)\|\big] + \mathbb{E}\big[L_T(Z_i)^2U_N^2\big],
\end{aligned}$$

where the last inequality is derived using $\|aa^\top - bb^\top\| \le 2\|a-b\|\,\|b\| + \|a-b\|^2$ with $a = e_i(\lambda)$ and $b = e^0_i(\lambda)$. By Cauchy-Schwarz inequality,

$$\mathbb{E}\big[L_T(Z_i)U_N\|e^0_i(\lambda)\|\big] \le \left(\mathbb{E}[L_T(Z_i)^2U_N^2]\cdot\mathbb{E}[\|e^0_i(\lambda)\|^2]\right)^{1/2}.$$

By Assumption 2.1 and fixed λ, $\mathbb{E}[\|e^0_i(\lambda)\|^2] < \infty$. Therefore, under the additional assumption $\mathbb{E}\big[L_T(Z_i)^2U_N^2\big] = o(1)$, we have

$$\mathbb{E}\big[L_T(Z_i)U_N\|e^0_i(\lambda)\|\big] = o(1),$$

and hence

$$\left\|\mathbb{E}\left[\left(\frac{1}{\rho(W_i)} - r_N\right)\left(R_\lambda(W_i) - R^0_\lambda(W_i)\right)\right]\right\| = o(1).$$

Thus the covariance term computed with $R^0_\lambda(W_i)$ is asymptotically equivalent to the corresponding term computed with $R_\lambda(W_i)$. Combining this with the central limit theorem gives

$$\sqrt{n}\,(\hat\theta_{\lambda,\pi} - \theta_0) \xrightarrow{d} \mathcal{N}\Big(0,\, J^{-1}_{\theta_0}\Omega_{\lambda,\pi}J^{-\top}_{\theta_0}\Big).$$

This proves the theorem. □

A.3 Proof of Lemma 1

Proof. Part 1 is immediate: $U_{m,N} \le (N-1)\bar\delta_N \le N\bar\delta_N = o(n^{-1/2})$ deterministically.

For part 2, a union bound over $i \le N$ and $k \le p$ gives, for any $t > 0$,

$$\mathbb{P}\big(U_{m,N} > t\sqrt{p}\,\big) \;\le\; \mathbb{P}\Big(\max_{i\le N,\,k\le p}\big|u^{(k)}_{i,m,N}\big| > t\Big) \;\le\; 2Np\,\exp\big(-t^2/(2\sigma_N^2)\big),$$

using $\|u\| \le \sqrt{p}\max_k |u^{(k)}|$. Taking $t_N = \sigma_N\sqrt{2\log(Np) + 2x_N}$ with $x_N = \log\log N \to \infty$ shows $\mathbb{P}(U_{m,N} > t_N\sqrt{p}) \le 2e^{-x_N} \to 0$, so $U_{m,N} = O_p\big(\sigma_N\sqrt{\log N}\big)$ for fixed p. Hence $\sigma_N\sqrt{\log N} = o(n^{-1/2})$ implies $U_{m,N} = o_p(n^{-1/2})$. □

A.4 A Weak Law for Inverse-Propensity-Weighted Arrays

The following auxiliary lemma is used in the proof of Proposition 1. It requires only $(1+\varepsilon)$ moments of the unweighted summands and accommodates vanishing labeling rates.

Lemma A.1 (IPW triangular-array weak law)**.** *Let $h_i = h(Z_i, T_i, \tilde{T}^0_{i,\cdot}) \ge 0$ be a scalar function of the unit-level variables with $\mathbb{E}[h_i^{1+\varepsilon}] \le C_h$ for some $\varepsilon \in (0,1]$, and let $a_N(\cdot)$ be functions of W_i with $\sup_N \sup_w |a_N(w)| \le C_a$. Suppose Assumption 1 holds and $n \to \infty$. Then*

$$\frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\pi_i}\, a_N(W_i)\, h_i \;-\; \mathbb{E}\big[a_N(W_i)\, h_i\big] \;\xrightarrow{p}\; 0.$$

Proof. Fix $\tau_N \to \infty$ with $\tau_N^{1-\varepsilon}/n \to 0$; for $\varepsilon < 1$ take $\tau_N = n^{1/(2(1-\varepsilon))}$, and for $\varepsilon = 1$ any $\tau_N \to \infty$. Decompose $h_i = h_i\mathbf{1}\{h_i \le \tau_N\} + h_i\mathbf{1}\{h_i > \tau_N\} =: h_i^{\le} + h_i^{>}$.

Tail part. By Assumption 1 and Markov's inequality, $\mathbb{E}\big[\frac{1}{N}\sum_i \frac{D_i}{\pi_i}|a_N(W_i)|\, h_i^{>}\big] = \mathbb{E}\big[|a_N(W_i)|\, h_i^{>}\big] \le C_a\,\mathbb{E}\big[h_i\mathbf{1}\{h_i > \tau_N\}\big] \le C_aC_h\,\tau_N^{-\varepsilon} \to 0$, so the inverse-propensity-weighted tail average converges to zero in L_1, and $\mathbb{E}[a_N(W_i)h_i^{>}] \to 0$ likewise.

Truncated part. The summands $(D_i/\pi_i)\, a_N(W_i)\, h_i^{\le}$ are i.i.d. across i with mean $\mathbb{E}[a_N(W_i)h_i^{\le}]$, and, using $\mathbb{E}[D_i/\pi_i^2 \mid \mathcal{Z}_N] = 1/\pi_i \le (\rho_{\min}r_N)^{-1}$ and $\mathbb{E}[(h_i^{\le})^2] \le \tau_N^{1-\varepsilon}\mathbb{E}[h_i^{1+\varepsilon}]$ for $\varepsilon \le 1$,

$$\mathrm{Var}\Big(\frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\pi_i}\, a_N(W_i)\, h_i^{\le}\Big) \;\le\; \frac{C_a^2}{N\rho_{\min}r_N}\,\mathbb{E}\big[(h_i^{\le})^2\big] \;\le\; \frac{C_a^2\, C_h}{\rho_{\min}} \cdot \frac{\tau_N^{1-\varepsilon}}{n} \;\to\; 0.$$

Chebyshev's inequality gives convergence of the truncated average to its mean, and combining the two parts completes the proof. □

A.5 Proof of Proposition 1

Proof. Throughout, C denotes a finite constant that may change between displays, $\hat{\theta} = \hat{\theta}_{\lambda,\pi}$, and K is a fixed compact neighborhood of θ_0 contained in the neighborhoods of Assumption 2 and Θ_1; by consistency, $\hat{\theta} \in K$ with probability tending to one, and we argue on this event. Write $g_i(\theta) = g(Z_i, T_i; \theta)$, $\tilde{g}_{i,m}(\theta) = g(Z_i, \tilde{T}_{i,m}; \theta)$, $\tilde{g}^0_{i,m}(\theta) = g(Z_i, \tilde{T}^0_{i,m}; \theta)$, $e_i(\theta) = g_i(\theta) - \sum_m \lambda_m \tilde{g}_{i,m}(\theta)$,

and let $e_i^0(\theta)$ denote the no-spillover analogue. Whenever Lemma A.1 is applied to a matrix- or vector-valued average, it is applied to the positive and negative parts of each entry, each dominated by the stated scalar h_i.

Step 1 (convergence at θ_0 with no-spillover features). For the first term of $\hat{\Omega}_{\lambda,\pi}$, apply Lemma A.1 with $h_i = \|g_i(\theta_0)\|^2$, $a_N \equiv r_N \leq 1$, and $\varepsilon = \delta/2$: the moment condition of Theorem 2 gives $\mathbb{E}[h_i^{1+\delta/2}] = \mathbb{E}\,\|g_i(\theta_0)\|^{2+\delta} < \infty$, and the mean $r_N\mathbb{E}[g_i g_i^\top]$ converges to $r\mathrm{Var}(g_i)$ since $\mathbb{E}[g_i(\theta_0)] = 0$. For the second term with $\tilde{T}^0$ in place of $\tilde{T}$, apply Lemma A.1 with $h_i = \left\|e_i^0(\theta_0)\right\|^2$ and $a_N(W) = 1/\rho(W) - r_N$, $|a_N| \leq \rho_{\min}^{-1} + 1$: by the c_r inequality, $\mathbb{E}[h_i^{1+\delta/2}] \leq C\big(\mathbb{E}\,\|g_i(\theta_0)\|^{2+\delta} + \sum_m \mathbb{E}\|\tilde{g}_{i,m}^0(\theta_0)\|^{2+\delta}\big) < \infty$, and the mean converges to $\mathbb{E}[(1/\rho - r)R_\lambda^0(W)]$ by dominated convergence. By the spillover-replacement step in the proof of Theorem 2, the resulting limit equals $\Omega_{\lambda,\pi}$.

Step 2 (replacing $\tilde{T}^0$ by $\tilde{T}$). By Assumption 2.4, $\sup_{\theta\in K}\left\|e_i(\theta) - e_i^0(\theta)\right\| \leq CL_T(Z_i)U_N$. Using $\left\|aa^\top - bb^\top\right\| \leq \|a-b\|\,(\|a\| + \|b\|)$,

$$\begin{aligned}
&\sup_{\theta\in K}\Big\|\frac{1}{N}\sum_i \frac{D_i}{\pi_i}\,a_N(W_i)\big(e_ie_i^\top - e_i^0e_i^{0\top}\big)(\theta)\Big\| \\
&\leq C\,U_N\,\frac{1}{N}\sum_i \frac{D_i}{\pi_i}L_T(Z_i)\sup_{\theta\in K}\left\|e_i^0(\theta)\right\| + C\,U_N^2\,\frac{1}{N}\sum_i \frac{D_i}{\pi_i}L_T(Z_i)^2.
\end{aligned}$$

The two IPW averages are $O_p(1)$ by Lemma A.1 with $\varepsilon = \delta/2$: $\sup_K\left\|e_i^0(\theta)\right\| \leq \left\|e_i^0(\theta_0)\right\| + L_i\,\mathrm{diam}(K)$ with $L_i = L_{\theta,H} + \sum_m |\lambda_m| L_{\theta,m}$, and the Cauchy–Schwarz inequality together with $\mathbb{E}[L_T^{2+\delta}] < \infty$, $\mathbb{E}[L_{\theta,H}^{2+\delta}] < \infty$, $\mathbb{E}[L_{\theta,m}^{2+\delta}] < \infty$ bounds all required $(1+\delta/2)$ moments. Since $U_N = o_p(1)$ by Assumption 3 and $n \to \infty$, the replacement error is $o_p(1)$.

Step 3 (replacing θ_0 by $\hat{\theta}$). On K, $\left\|\hat{g}_i\hat{g}_i^\top - g_i(\theta_0)g_i(\theta_0)^\top\right\| \leq L_{\theta,H}(Z_i)\left\|\hat{\theta} - \theta_0\right\|\left(2\,\|g_i(\theta_0)\| + L_{\theta,H}(Z_i)\left\|\hat{\theta} - \theta_0\right\|\right)$, and analogously for $\hat{e}_i^0\hat{e}_i^{0\top}$ with the envelope L_i. The IPW averages of $L_{\theta,H}\,\|g_i(\theta_0)\|$, $L_{\theta,H}^2$, $L_i\left\|e_i^0(\theta_0)\right\|$, and L_i^2 are $O_p(1)$ by Lemma A.1 exactly as in Step 2. Since $\left\|\hat{\theta} - \theta_0\right\| = o_p(1)$, the replacement error is $o_p(1)$. Combining Steps 1–3 yields $\hat{\Omega}_{\lambda,\pi} \xrightarrow{p} \Omega_{\lambda,\pi}$.

Step 4 (Jacobian). Write

$$\hat{J} = \frac{1}{N}\sum_{i=1}^N\Big[\frac{D_i}{\pi_i}\nabla_\theta g_i(\hat{\theta}) + \sum_m \lambda_m\Big(1 - \frac{D_i}{\pi_i}\Big)\nabla_\theta\tilde{g}_{i,m}(\hat{\theta})\Big].$$

First, replacing $\tilde{T}_{i,m}$ by $\tilde{T}_{i,m}^0$ incurs an error bounded by $CU_N\cdot\frac{1}{N}\sum_i(1+\frac{D_i}{\pi_i})L_{T,\nabla}(Z_i) = o_p(1)\cdot O_p(1)$, by the gradient Lipschitz condition, the plain law of large numbers, and Lemma A.1 with $\varepsilon = 1$. Second, with $\omega_i(\varrho) \coloneqq \sup_{\|\theta-\theta_0\|\leq\varrho}\big(\,\|\nabla_\theta g_i(\theta) - \nabla_\theta g_i(\theta_0)\| + \sum_m |\lambda_m|\,\|\nabla_\theta\tilde{g}_{i,m}^0(\theta) - \nabla_\theta\tilde{g}_{i,m}^0(\theta_0)\|\big)$, con-

tinuity of $\nabla_\theta g$ in θ and dominated convergence (envelope $2(M_H + \sum_m |\lambda_m| M_m)$, square-integrable) give $\mathbb{E}[\omega_i(\varrho)] \to 0$ as $\varrho \downarrow 0$; choosing $\varrho_N \downarrow 0$ with $\mathbb{P}(\left\|\hat{\theta} - \theta_0\right\| > \varrho_N) \to 0$ and applying the plain law of large numbers and Lemma A.1 ($\varepsilon = 1$) to $\omega_i(\varrho_N)$ shows that replacing $\hat{\theta}$ by θ_0 incurs an $o_p(1)$ error. Third, at $(\theta_0, \tilde{T}^0)$: $\frac{1}{N}\sum_i \frac{D_i}{\pi_i} \nabla_\theta g_i(\theta_0) \to_p \mathbb{E}[\nabla_\theta g_i(\theta_0)]$ by Lemma A.1 ($\varepsilon = 1$, envelope M_H), while the debiasing part converges to zero because the unweighted and the inverse-propensity-weighted averages of $\nabla_\theta \tilde{g}^0_{i,m}(\theta_0)$ converge to the same limit $\mathbb{E}[\nabla_\theta \tilde{g}^0_{i,m}(\theta_0)]$ (plain law of large numbers and Lemma A.1, envelope M_m). Finally, $\mathbb{E}[\nabla_\theta g_i(\theta_0)] = \nabla_\theta \mathbb{E}[g_i(\theta)]\big|_{\theta_0} = J_{\theta_0}$, where the interchange is justified by the envelope M_H and dominated convergence.

The convergence $\hat{\sigma}_j^2 \to_p [J_{\theta_0}^{-1} \Omega_{\lambda,\pi} J_{\theta_0}^{-\top}]_{jj}$ follows from the continuous mapping theorem, since J_{θ_0} is invertible, and the exact asymptotic coverage follows from Theorem 2 and Slutsky's theorem, using positive definiteness of $\Omega_{\lambda,\pi}$. □

A.6 Proof of Theorem 4

Proof. For notational simplicity, define

$$q_\lambda(W_i) = \operatorname{tr}(J_{\theta_0}^{-1} R_\lambda(W_i) J_{\theta_0}^{-\top}) = s_\lambda(W_i)^2,$$

where $s_\lambda(w) = \sqrt{q_\lambda(w)} \geq 0$ since $R_\lambda(w)$ is positive semidefinite. For fixed λ, minimizing $V(\lambda, \pi)$ over admissible propensities is equivalent to minimizing

$$\mathcal{J}(\rho) \coloneqq \mathbb{E}\left[\frac{q_\lambda(W_i)}{\rho(W_i)}\right]$$

subject to $\mathbb{E}[\rho(W_i)] = 1$ and $\rho(W_i) \in [\rho_{\min}, r_N^{-1}]$.

Let

$$\mathcal{R}_N \coloneqq \left\{\rho : \ \rho \text{ measurable}, \ \rho_{\min} \leq \rho(W_i) \leq r_N^{-1}, \ \mathbb{E}[\rho(W_i)] = 1\right\}.$$

This set is nonempty because the constant rule $\rho(W_i) \equiv 1$ belongs to $\mathcal{R}_N$. We have

$$\mathcal{J}(\rho) \leq \frac{1}{\rho_{\min}} \mathbb{E}[q_\lambda(W_i)].$$

By Assumption 2.1 and fixed λ, $\mathbb{E}[q_\lambda(W_i)] < \infty$. Therefore, $\mathcal{J}(\rho)$ is finite for every $\rho \in \mathcal{R}_N$.

For any $\mu > 0$ and w, consider the problem

$$\ell_\mu(x; w) \coloneqq \frac{q_\lambda(w)}{x} + \mu x, \qquad x \in [\rho_{\min}, r_N^{-1}]$$

Since

$$\frac{\partial}{\partial x}\ell_\mu(x;w) = -\frac{q_\lambda(w)}{x^2} + \mu, \qquad \frac{\partial^2}{\partial x^2}\ell_\mu(x;w) = \frac{2q_\lambda(w)}{x^3} \geq 0,$$

the function $\ell_\mu(\cdot;w)$ is convex on $[\rho_{\min}, r_N^{-1}]$. Its unconstrained minimizer is $x = \sqrt{q_\lambda(w)/\mu}$. After imposing the constraints, the unique minimizer becomes

$$x_\mu(w) = \max\left\{\rho_{\min},\ \min\left\{r_N^{-1},\ \sqrt{\frac{q_\lambda(w)}{\mu}}\right\}\right\}.$$

Writing $a = \mu^{-1/2}$ and using $q_\lambda(w) = s_\lambda(w)^2$, we obtain

$$x_\mu(w) = \max\left\{\rho_{\min},\ \min\left\{r_N^{-1},\ a\, s_\lambda(w)\right\}\right\}.$$

We next show that one can choose a to satisfy the budget constraint. Define

$$h(a) \coloneqq \mathbb{E}\left[\max\left\{\rho_{\min},\ \min\left\{r_N^{-1},\ a\, s_\lambda(W_i)\right\}\right\}\right], \qquad a \geq 0.$$

We have $h(a)$ is continuous and nondecreasing in a. Moreover, $h(0) = \rho_{\min} \leq 1$. If $\rho_{\min} = 1$, then any admissible $\rho \in \mathcal{R}_N$ must satisfy $\rho(W_i) = 1$. Consider the case $\rho_{\min} < 1$. As $a \to \infty$,

$$\max\left\{\rho_{\min},\ \min\left\{r_N^{-1},\ a\, s_\lambda(W_i)\right\}\right\} \to r_N^{-1}\mathbf{1}\{s_\lambda(W_i) > 0\} + \rho_{\min}\mathbf{1}\{s_\lambda(W_i) = 0\}.$$

Therefore, $\lim_{a\to\infty} h(a) = r_N^{-1}\mathbb{P}(s_\lambda(W_i) > 0) + \rho_{\min}\mathbb{P}(s_\lambda(W_i) = 0)$. By Assumption 5,

$$\mathbb{P}(s_\lambda(W_i) > 0) = \mathbb{P}(q_\lambda(W_i) > 0) > r_N,$$

and thus $\lim_{a\to\infty} h(a) > 1$. Since $h(0) \leq 1$ and h is continuous, there exists at least one $a_\lambda \geq 0$ such that $h(a_\lambda) = 1$. Define

$$\rho^*(w \mid \lambda) = \max\left\{\rho_{\min},\ \min\left\{r_N^{-1},\ a_\lambda\, s_\lambda(w)\right\}\right\}.$$

Then $\rho^*(\cdot \mid \lambda) \in \mathcal{R}_N$.

It remains to prove optimality. If $a_\lambda = 0$, then $h(0) = \rho_{\min} = 1$, in which case $\mathcal{R}_N = \{\rho \equiv 1\}$ and optimality is trivial; assume henceforth $a_\lambda > 0$ and let $\mu_\lambda = a_\lambda^{-2}$. For any $\rho \in \mathcal{R}_N$, pointwise

optimality of $x_{\mu_\lambda}(w) = \rho^*(w \mid \lambda)$ implies

$$\frac{q_\lambda(W_i)}{\rho(W_i)} + \mu_\lambda \rho(W_i) \geq \frac{q_\lambda(W_i)}{\rho^*(W_i \mid \lambda)} + \mu_\lambda \rho^*(W_i \mid \lambda) \qquad \text{a.s.}$$

Taking expectations and using

$$\mathbb{E}[\rho(W_i)] = \mathbb{E}[\rho^*(W_i \mid \lambda)] = 1,$$

we obtain $\mathcal{J}(\rho) \geq \mathcal{J}(\rho^*)$. Therefore $\rho^*(\cdot \mid \lambda)$ is a global minimizer.

Finally, since $\pi(W_i) = r_N \rho(W_i)$, multiplying $\rho^*(w \mid \lambda)$ by r_N yields

$$\pi^*(w \mid \lambda) = \max\left\{\rho_{\min} r_N,\ \min\left\{1,\ c_\lambda s_\lambda(w)\right\}\right\}, \qquad c_\lambda = r_N a_\lambda.$$

This completes the proof. □

A.7 Proof of Proposition 2

Proof. Denote $s_\lambda(w) = \sqrt{\mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W) J_{\theta_0}^{-\top})}$. The part of $V(\lambda, \pi)$ that depends on (λ, π) is

$$V_1(\lambda, \pi) := r_N \, \mathbb{E}\left[\frac{s_\lambda(W)^2}{\pi(W)}\right] - r \, \mathbb{E}[s_\lambda(W)^2].$$

For a fixed λ, the optimal rule $\pi^*(\cdot \mid \lambda)$ solves

$$\begin{aligned} &\min_{\pi(\cdot) \in (0,1]} \mathbb{E}\left[\frac{s_\lambda(W)^2}{\pi(W)}\right] \\ &\text{s.t.} \quad \mathbb{E}[\pi(W_i)] = r_N, \quad \rho_{\min} r_N \leq \pi(W_i) \leq 1. \end{aligned}$$

The Lagrangian for this inner problem can be written as

$$\begin{aligned} \mathcal{L}(\pi, \mu, \eta_1, \eta_2; \lambda) = \mathbb{E}\left[\frac{s_\lambda(W)^2}{\pi(W)}\right] &+ \mu(\mathbb{E}[\pi(W_i)] - r_N) \\ &+ \eta_1(W_i)(\pi(W_i) - 1) + \eta_2(W_i)(\rho_{\min} r_N - \pi(W_i)), \end{aligned}$$

where $\eta_1(W_i), \eta_2(W_i) \geq 0$. The pointwise stationarity condition is therefore

$$-\frac{s_\lambda(W)^2}{\pi(W)^2} + \mu + \eta_1(W) - \eta_2(W) = 0.$$

together with the budget constraint and complementary slackness.

By the constrained envelope theorem, since the constraints do not depend on λ, the derivative of the inner value $\phi(\lambda) \coloneqq \mathbb{E}[s_\lambda(W)^2/\pi^*(W \mid \lambda)]$ is

$$\nabla_\lambda \phi(\lambda) = \mathbb{E}\left[\frac{1}{\pi^*(W \mid \lambda)} \nabla_\lambda s_\lambda(W)^2\right].$$

Define $\widetilde{V}_1(\lambda) = V_1(\lambda, \pi^*(\cdot \mid \lambda)) = r_N\, \phi(\lambda) - r\, \mathbb{E}[s_\lambda(W)^2]$. Then we have

$$\nabla_\lambda \widetilde{V}_1(\lambda) = \mathbb{E}\left[\left(\frac{r_N}{\pi^*(W \mid \lambda)} - r\right) \nabla_\lambda s_\lambda(W)^2\right].$$

Let $e_i(\lambda) = g(Z_i, T_i; \theta_0) - \sum_{m=1}^M \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)$. We have

$$\nabla_\lambda s_\lambda(W)^2 = -2\, \mathbb{E}\Big[\tilde{g}_i^\top J_{\theta_0}^{-\top} J_{\theta_0}^{-1} e_i(\lambda) \mid W_i = W\Big].$$

Substituting this into $\nabla_\lambda \widetilde{V}_1(\lambda)$ and using the law of iterated expectations yields

$$\nabla_\lambda \widetilde{V}_1(\lambda) = -2\mathbb{E}\left[\left(\frac{r_N}{\pi^*(W \mid \lambda)} - r\right) \tilde{g}_i^\top J_{\theta_0}^{-\top} J_{\theta_0}^{-1} e_i(\lambda)\right].$$

Since λ^* is an interior local minimizer, setting $\nabla_\lambda \widetilde{V}_1(\lambda^*) = 0$ gives the stated first-order condition.

□

A.8 Proof of Proposition 3

Proof. (a) Write $b_i(\theta) = \sum_m \lambda_m(W_i)\, g(Z_i, \tilde{T}_{i,m}^0; \theta)$ and $\psi_i(\theta) = b_i(\theta) + (D_i/\pi_i)\big(g_i(\theta) - b_i(\theta)\big)$ for the no-spillover moment, as in the proofs of Theorems 1 and 2. We verify that every step of those proofs remains valid; the weights enter in exactly three places.

(i) *Centering and independence.* $\lambda(W_i)$ is $\mathcal{Z}_N$-measurable, so by Assumption 1.1, $\mathbb{E}\big[(1-D_i/\pi_i)\, b_i(\theta)\big] = \mathbb{E}\big[\big(1 - \mathbb{E}[D_i/\pi_i \mid \mathcal{Z}_N]\big)\, b_i(\theta)\big] = 0$, hence $\mathbb{E}[\psi_i(\theta)] = G(\theta)$ for every θ, exactly as before. Since $\lambda(W_i)$ is a unit-wise function of W_i, the vectors $(Z_i, T_i, \tilde{T}_{i,\cdot}^0, W_i, D_i)$ and hence the summands $\psi_i(\theta)$ remain i.i.d. across i.

(ii) *Envelopes.* Since Λ is compact, $B_\Lambda \coloneqq \sup_{\lambda \in \Lambda} \sum_{m=1}^M |\lambda_m| < \infty$. Thus

$$\|b_i(\theta)\| \le B_\Lambda \max_{m \le M} \|g(Z_i, \tilde{T}_{i,m}^0; \theta)\| \le B_\Lambda \sum_{m=1}^M \|g(Z_i, \tilde{T}_{i,m}^0; \theta)\|.$$

Similarly, the Lipschitz bound becomes

$$\|\psi_i(\theta_1) - \psi_i(\theta_2)\| \leq \left[\frac{D_i}{\pi_i} L_{\theta,H}(Z_i) + B_\Lambda \left(1 + \frac{D_i}{\pi_i}\right) \sum_{m=1}^{M} L_{\theta,m}(Z_i)\right] \|\theta_1 - \theta_2\|,$$

and the spillover contribution is bounded by

$$\sup_\theta \|G_\Delta(\theta)\| \leq B_\Lambda U_N \cdot \frac{1}{N} \sum_i \left|1 - \frac{D_i}{\pi_i}\right| L_T(Z_i).$$

These envelopes are finite and do not depend on the particular measurable function $\lambda(\cdot)$. Hence the pointwise and uniform laws of large numbers, the identification arguments, the stochastic-equicontinuity step, and the Lyapunov computation apply without change.

(iii) *Variance.* The covariance computation in the proof of Theorem 2 uses only $D_i^2 = D_i$ and $\mathbb{E}[D_i/\pi_i \mid \mathcal{Z}_N] = 1$, giving $\mathbb{E}[\psi_i \psi_i^\top] = \mathbb{E}\big[g_i g_i^\top + (1/\pi_i - 1)(g_i - b_i)(g_i - b_i)^\top\big]$; conditioning on W_i, on which $\lambda(W_i)$ is fixed, identifies the conditional residual covariance as $R_{\lambda(W_i)}(W_i)$. The rescaling by r_N, the passage to the limit, and the spillover replacement are unchanged, which gives the stated $\Omega_{\lambda(\cdot),\pi}$ and completes part (a).

(b) For each w, $\lambda \mapsto q_\lambda(w)$ is a convex quadratic, hence continuous, and $w \mapsto q_\lambda(w)$ is measurable for each λ, so q is a Carathéodory function on $\mathrm{supp}(W) \times \Lambda$ with Λ compact. By the measurable maximum theorem (Aliprantis and Border, 2006, Theorem 18.19), $q^*(w) = \min_{\lambda \in \Lambda} q_\lambda(w)$ is measurable and the argmin correspondence admits a measurable selection $\lambda^*(\cdot)$. Every admissible design satisfies $\rho(w) \leq 1/r_N$ for all N, hence $\rho(w) \leq \lim_N 1/r_N = 1/r$ when $r > 0$, and $1/\rho(w) - r \geq 0$ almost surely (trivially so when $r = 0$). Therefore, for any measurable $\lambda(\cdot)$,

$$V(\lambda(\cdot), \pi) - V(\lambda^*(\cdot), \pi) = \mathbb{E}\Big[\Big(\frac{1}{\rho(W)} - r\Big)\big(q_{\lambda(W)}(W) - q^*(W)\big)\Big] \;\geq\; 0,$$

as an expectation of a product of nonnegative terms. Constant weights are the special case $\lambda(\cdot) \equiv \lambda$.

(c) For any pair $(\lambda(\cdot), \pi)$, part (b) gives $V(\lambda(\cdot), \pi) \geq V(\lambda^*(\cdot), \pi)$. Given $\lambda^*(\cdot)$, the design-dependent part of the criterion is $\mathbb{E}[(1/\rho - r)\, q^*(W)]$, and the proof of Theorem 4 applies verbatim with $q^*(\cdot)$ in place of $q_\lambda(\cdot)$: q^* is nonnegative, measurable, and integrable, since $q^*(w) \leq q_0(w)$, the value at $\lambda = 0$, whose integral is finite by Assumption 2.1; the existence of the budget constant uses $\mathbb{P}(q^*(W) > 0) > r_N$ in place of Assumption 5. The resulting policy minimizes $V(\lambda^*(\cdot), \cdot)$ over admissible designs, so $(\lambda^*(\cdot), \pi^*)$ attains the joint minimum, and the two minimizations can be performed in this order without alternation. □

A.9 Proof of Proposition 4

Proof. Let

$$\Delta_N := \sup_{\lambda \in \Lambda} \left| V(\lambda, \widehat{\pi}_\lambda) - \widetilde{V}(\lambda) \right|.$$

Since $\widehat{\pi}_\lambda = r_N \widehat{\rho}_\lambda$ and $\pi^*(\cdot \mid \lambda) = r_N \rho^*_\lambda(\cdot)$, the parts of $V(\lambda, \widehat{\pi}_\lambda)$ and $V(\lambda, \pi^*(\cdot \mid \lambda))$ that do not depend on the labeling rule cancel. Therefore, for each $\lambda \in \Lambda$,

$$\begin{aligned}
\left| V(\lambda, \widehat{\pi}_\lambda) - \widetilde{V}(\lambda) \right| &= \left| \mathbb{E}\left[\left\{ \frac{1}{\widehat{\rho}_\lambda(W)} - \frac{1}{\rho^*_\lambda(W)} \right\} \operatorname{tr}\left(J_{\theta_0}^{-1} R_\lambda(W) J_{\theta_0}^{-\top} \right) \right] \right| \\
&\leq \frac{1}{\rho^2_{\min}} \mathbb{E}\left[|\widehat{\rho}_\lambda(W) - \rho^*_\lambda(W)| \operatorname{tr}\left(J_{\theta_0}^{-1} R_\lambda(W) J_{\theta_0}^{-\top} \right) \right] \\
&\lesssim \mathbb{E}\left[(\widehat{\rho}_\lambda(W) - \rho^*_\lambda(W))^2 \right]^{1/2},
\end{aligned}$$

where the first inequality uses the uniform lower bounds $\widehat{\rho}_\lambda(W) \geq \rho_{\min}$ and $\rho^*_\lambda(W) \geq \rho_{\min}$. Taking the supremum over $\lambda \in \Lambda$ gives

$$\Delta_N = o_p(1).$$

Let $\widehat{\widetilde{V}}(\lambda) := \widehat{V}(\lambda, \widehat{\pi}_\lambda)$ and

$$\Gamma_N := \sup_{\lambda \in \Lambda} \left| \widehat{\widetilde{V}}(\lambda) - \widetilde{V}(\lambda) \right|.$$

By the triangle inequality,

$$\Gamma_N \leq \sup_{\lambda \in \Lambda} \left| \widehat{\widetilde{V}}(\lambda) - V(\lambda, \widehat{\pi}_\lambda) \right| + \sup_{\lambda \in \Lambda} \left| V(\lambda, \widehat{\pi}_\lambda) - \widetilde{V}(\lambda) \right| = o_p(1),$$

since the first term is $o_p(1)$ by assumption and the second term is $\Delta_N = o_p(1)$.

We have

$$\begin{aligned}
\widetilde{V}(\widehat{\lambda}) &\leq \widehat{\widetilde{V}}(\widehat{\lambda}) + \Gamma_N \leq \inf_{\lambda \in \Lambda} \widehat{\widetilde{V}}(\lambda) + \Gamma_N + o_p(1) \\
&\leq \widehat{\widetilde{V}}(\lambda^*) + \Gamma_N + o_p(1) \\
&\leq \widetilde{V}(\lambda^*) + 2\Gamma_N + o_p(1).
\end{aligned}$$

Since λ^* minimizes $\widetilde{V}(\lambda)$ over Λ, we also have

$$\widetilde{V}(\lambda^*) \leq \widetilde{V}(\widehat{\lambda}).$$

Combining these two inequalities yields

$$\widetilde{V}(\widehat{\lambda}) = \widetilde{V}(\lambda^*) + o_p(1).$$

Finally, since $\widehat{\pi} = \widehat{\pi}_{\widehat{\lambda}}$,

$$\left|V(\widehat{\lambda}, \widehat{\pi}) - \widetilde{V}(\widehat{\lambda})\right| \leq \sup_{\lambda \in \Lambda} \left|V(\lambda, \widehat{\pi}_\lambda) - \widetilde{V}(\lambda)\right| = \Delta_N = o_p(1).$$

Therefore,

$$V(\widehat{\lambda}, \widehat{\pi}) = \widetilde{V}(\widehat{\lambda}) + o_p(1) = \widetilde{V}(\lambda^*) + o_p(1) = V(\lambda^*, \pi^*) + o_p(1),$$

which proves the claim. □

A.10 Proof of Theorem 3

Proof. Since the limit features are generated unit-wise by a fixed pipeline, the cross-fitted moment evaluated at the limit features equals the fixed-pipeline moment of Section 3.4: writing $G^*_{\lambda,\pi}(\theta)$ for the moment (5) computed with $\tilde{T}^*_{i,m}$, we have the decomposition

$$G^{\mathrm{CF}}_{\lambda,\pi}(\theta) = G^*_{\lambda,\pi}(\theta) + \Delta(\theta), \qquad \Delta(\theta) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{F}_k} f^{(k)}_{\theta,i},$$

where

$$f^{(k)}_{\theta,i} \coloneqq \left(1 - \frac{D_i}{\pi_i}\right) \sum_{m=1}^{M} \lambda_m \left[g\big(Z_i, \tilde{T}^{(-k)}_{i,m}; \theta\big) - g\big(Z_i, \tilde{T}^*_{i,m}; \theta\big)\right].$$

Step 1 (conditional centering). Fix a fold k and condition on the tuning data $\mathcal{D}_{-k}$. Because the tuning data exclude the units in fold k, the variables $\{(Z_i, T_i, W_i) \colon i \in \mathcal{F}_k\}$, the associated seed noise, and the labeling indicators $\{D_i \colon i \in \mathcal{F}_k\}$ are jointly independent of $\mathcal{D}_{-k}$ and i.i.d. across $i \in \mathcal{F}_k$, and by Assumption 4.2 the labeling structure of Assumption 1 is preserved conditionally. Hence, for every fixed θ and every $i \in \mathcal{F}_k$, with $\Delta g_{i,m}(\theta) \coloneqq g(Z_i, \tilde{T}^{(-k)}_{i,m}; \theta) - g(Z_i, \tilde{T}^*_{i,m}; \theta)$,

$$\mathbb{E}\big[f^{(k)}_{\theta,i} \,\big|\, \mathcal{D}_{-k}\big] = \mathbb{E}\left[\left(1 - \mathbb{E}\Big[\frac{D_i}{\pi_i} \,\Big|\, W_i, Z_i, T_i, \tilde{T}\text{-features}, \mathcal{D}_{-k}\Big]\right) \sum_{m=1}^{M} \lambda_m \Delta g_{i,m}(\theta) \,\middle|\, \mathcal{D}_{-k}\right] = 0.$$

The debiasing terms are therefore exactly centered for any realization of the tuned pipeline.

Step 2 (uniform negligibility and consistency). Let $c_i \coloneqq \sum_{m=1}^{M} |\lambda_m| \, \|\tilde{T}^{(-k(i))}_{i,m} - \tilde{T}^*_{i,m}\|$, where $k(i)$

is the fold of unit i. By Assumption 2.4, uniformly over θ in the relevant neighborhood (or over Θ under the global conditions),

$$\|f_{\theta,i}^{(k)}\| \le \Big|1 - \frac{D_i}{\pi_i}\Big| \, L_T(Z_i)\, c_i.$$

By the argument in the proof of Theorem 1, $\mathbb{E}[(D_i/\pi_i)L_T(Z_i)c_i \mid \mathcal{D}_{-k}] = \mathbb{E}[L_T(Z_i)c_i \mid \mathcal{D}_{-k}]$, and by the Cauchy–Schwarz inequality and Assumption 4.4,

$$\mathbb{E}\Big[\Big|1 - \frac{D_i}{\pi_i}\Big|L_T(Z_i)c_i \,\Big|\, \mathcal{D}_{-k}\Big] \le 2\sum_{m=1}^{M} |\lambda_m| \left(\mathbb{E}\big[\bar{L}(Z_i)^2 \big\|\tilde{T}_{i,m}^{(-k)} - \tilde{T}_{i,m}^{*}\big\|^2 \,\big|\, \mathcal{D}_{-k}\big]\right)^{1/2} \le C_\lambda \sqrt{\xi_N},$$

where $C_\lambda := 2\sum_{m=1}^{M} |\lambda_m|$. Splitting $\sup_\theta \|\Delta(\theta)\|$ by folds and applying the conditional Markov inequality given $\mathcal{D}_{-k}$ to each fold average, followed by dominated convergence to remove the conditioning (the conditional bounds are bounded by one and $o_p(1)$), then gives $\sup_\theta \|\Delta(\theta)\| = O_p(\sqrt{\xi_N}) = o_p(1)$. Since $G_{\lambda,\pi}^{*}$ satisfies the uniform law of large numbers established in the proof of Theorem 1 (the limit features satisfy all maintained assumptions), the identification argument there applies to $G^{\mathrm{CF}} = G^{*} + \Delta$ and yields $\hat{\theta}_{\lambda,\pi}^{\mathrm{CF}} \xrightarrow{p} \theta_0$.

Step 3 ($\sqrt{n}$-negligibility of the tuning perturbation). (a) *At θ_0:* conditional on $\mathcal{D}_{-k}$, the summands $\{f_{\theta_0,i}^{(k)} : i \in \mathcal{F}_k\}$ are i.i.d. with zero mean by Step 1, and using $\mathbb{E}[(1 - D_i/\pi_i)^2 \mid W_i, \cdot\,] = 1/\pi_i - 1 \le (\rho_{\min} r_N)^{-1}$ together with $c_i^2 \le M\sum_m \lambda_m^2 \|\tilde{T}_{i,m}^{(-k)} - \tilde{T}_{i,m}^{*}\|^2$,

$$\mathbb{E}\big[\|f_{\theta_0,i}^{(k)}\|^2 \,\big|\, \mathcal{D}_{-k}\big] \le \frac{1}{\rho_{\min} r_N}\, \mathbb{E}\big[L_T(Z_i)^2 c_i^2 \,\big|\, \mathcal{D}_{-k}\big] \le \frac{C\,\xi_N}{\rho_{\min} r_N}.$$

Hence

$$\mathrm{Var}\Big(\sqrt{n}\,\frac{1}{N}\sum_{i\in\mathcal{F}_k} f_{\theta_0,i}^{(k)} \,\Big|\, \mathcal{D}_{-k}\Big) \le \frac{n}{N^2}\cdot\frac{N}{K}\cdot\frac{C\xi_N}{\rho_{\min} r_N} = \frac{C\,\xi_N}{K\rho_{\min}} = o_p(1),$$

and conditional Chebyshev followed by dominated convergence gives $\sqrt{n}\,\Delta(\theta_0) = o_p(1)$ after summing over the K folds. (b) *Increments:* by Assumption 4.3, for θ in the neighborhood of θ_0,

$$\big\|f_{\theta,i}^{(k)} - f_{\theta_0,i}^{(k)}\big\| \le \Big|1 - \frac{D_i}{\pi_i}\Big| \, L_{T,\theta}(Z_i)\, c_i \;\|\theta - \theta_0\|\,,$$

so that $\|\Delta(\theta) - \Delta(\theta_0)\| \le H_N \|\theta - \theta_0\|$ with $H_N := N^{-1}\sum_i |1 - D_i/\pi_i|\, L_{T,\theta}(Z_i)\, c_i$. Decomposing H_N by folds and applying, fold by fold, the conditional bound $\mathbb{E}\big[\,|1 - D_i/\pi_i|\, L_{T,\theta}(Z_i)\, c_i \,\big|\, \mathcal{D}_{-k}\big] \le C_\lambda\sqrt{\xi_N}$ together with the conditional Markov inequality and dominated convergence, as in Step 2, gives $H_N = o_p(1)$.

Step 4 (conclusion). Write

$$0 = G^{\mathrm{CF}}(\hat{\theta}^{\mathrm{CF}}) = G^{*}(\hat{\theta}^{\mathrm{CF}}) + \Delta(\theta_0) + \big[\Delta(\hat{\theta}^{\mathrm{CF}}) - \Delta(\theta_0)\big]$$

and expand $G^{*}(\hat{\theta}^{\mathrm{CF}})$ exactly as in the proof of Theorem 2; all the conditions hold for the limit features. Collecting terms,

$$J_{\theta_0}\sqrt{n}\,(\hat{\theta}^{\mathrm{CF}} - \theta_0) = -\sqrt{n}\,G^{*,0}_{\lambda,\pi}(\theta_0) + o_p(1) + o\big(\sqrt{n}\,\|\hat{\theta}^{\mathrm{CF}} - \theta_0\|\big) + H_N \cdot \sqrt{n}\,\|\hat{\theta}^{\mathrm{CF}} - \theta_0\|,$$

where $G^{*,0}_{\lambda,\pi}$ is the no-spillover version of the limit-feature moment. Since $H_N = o_p(1)$, the same fixed-point argument as in the proof of Theorem 2 yields $\sqrt{n}\,\|\hat{\theta}^{\mathrm{CF}} - \theta_0\| = O_p(1)$, after which the last two terms are $o_p(1)$ and

$$\sqrt{n}\,(\hat{\theta}^{\mathrm{CF}} - \theta_0) = -J_{\theta_0}^{-1}\sqrt{n}\,G^{*,0}_{\lambda,\pi}(\theta_0) + o_p(1) \xrightarrow{d} \mathcal{N}\big(0,\ J_{\theta_0}^{-1}\Omega^{*}_{\lambda,\pi}J_{\theta_0}^{-\top}\big)$$

by the central limit theorem step in the proof of Theorem 2. □

A.11 Proof of Theorem 5

Proof. All conditional statements are given $\mathcal{S}$, which fixes the pilot indicators, the pilot data, and hence $(\hat{\lambda}, \hat{\pi}, \hat{a})$. We write $X_N = o_{p|\mathcal{S}}(1)$ if $\mathbb{P}(\|X_N\| > \varepsilon \mid \mathcal{S}) \xrightarrow{p} 0$ for every $\varepsilon > 0$; by dominated convergence (the conditional probabilities are bounded by one), $X_N = o_{p|\mathcal{S}}(1)$ implies $X_N = o_p(1)$.

Step 0 (conditional design regularity). Because the pilot selection is independent of the data, conditional on $\mathcal{S}$ the non-pilot units are i.i.d. draws from the population, with $N' \coloneqq \#\{i\colon D_i^{(0)} = 0\}$ fixed given $\mathcal{S}$ and $N'/N \xrightarrow{p} 1$, since $\pi_N^{(0)} = (n_0/n)\,r_N \to 0$ by condition (i). The second-stage labels are conditionally independent Bernoulli draws with propensity $\hat{a}(W_i)$, an $\mathcal{S}$-measurable function of W_i alone; thus Assumption 1 holds conditionally for the non-pilot subsample. Let $\tilde{r} \coloneqq \mathbb{E}[\hat{a}(W) \mid \mathcal{S}]$ denote the conditional labeling rate and $\tilde{n} \coloneqq N'\tilde{r}$. By condition (iii) and the Cauchy–Schwarz inequality, $\mathbb{E}[\hat{\rho}(W) \mid \mathcal{S}] - 1 = \mathbb{E}[\hat{\rho} - \rho^{*} \mid \mathcal{S}] = o_p(1)$ (recall $\mathbb{E}[\rho^{*}(W)] = 1$), so $\mathbb{E}[\hat{\pi}(W) \mid \mathcal{S}] = r_N(1 + o_p(1))$ and

$$\tilde{r} = \frac{\mathbb{E}[\hat{\pi}(W) \mid \mathcal{S}] - \pi_N^{(0)}}{1 - \pi_N^{(0)}} = r_N\,(1 + o_p(1)), \qquad \tilde{n} = n\,(1 + o_p(1)).$$

Moreover, $\hat{a}(W) \geq (\rho_{\min} r_N - \pi_N^{(0)})/(1 - \pi_N^{(0)})$, and since $\pi_N^{(0)}/r_N \to 0$, with probability tending to one

$\hat{a}(W) \geq \frac{1}{2}\rho_{\min} r_N \geq \frac{1}{4}\rho_{\min}\tilde{r}$ for all w: the overlap condition of Assumption 1.2 holds conditionally with constant $\rho_{\min}/4$.

Step 1 (conditional linearization). Conditional on $\mathcal{S}$, the estimator $\hat{\theta}^{\mathrm{TS}}$ is the fixed-design AI-PI estimator applied to the N' non-pilot units with design $(\hat{\lambda}, \hat{a})$. Every bound in the proofs of Theorems 1 and 2 depends on the design only through (a) the overlap constant, bounded below by $\rho_{\min}/4$ by Step 0; (b) the weights, which lie in the compact set Λ and enter the envelopes linearly; and (c) the moment envelopes of Assumption 2, which are design-free. Hence the linearization holds conditionally, uniformly over realizations of the design:

$$\sqrt{\tilde{n}}\left(\hat{\theta}^{\mathrm{TS}} - \theta_0\right) = -J_{\theta_0}^{-1} \frac{1}{\sqrt{N'}} \sum_{i:\, D_i^{(0)}=0} \sqrt{\tilde{r}}\, \psi_i\big(\theta_0; \hat{\lambda}, \hat{a}\big) + o_{p|\mathcal{S}}(1),$$

where $\psi_i(\theta_0; \lambda, a) = b_i(\lambda; \theta_0) + \big(D_i^{(1)}/a(W_i)\big)\big(g_i(\theta_0) - b_i(\lambda; \theta_0)\big)$ and $b_i(\lambda; \theta_0) = \sum_m \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)$.

Step 2 (convergence of the conditional variance). Conditional on $\mathcal{S}$, the summands are i.i.d. with zero mean for any design satisfying the overlap condition, with conditional covariance

$$\tilde{\Omega} := \mathbb{E}\big[\tilde{r}\,\psi_i\psi_i^\top \mid \mathcal{S}\big] = \tilde{r}\,\mathrm{Var}\big(g_i(\theta_0)\big) + \mathbb{E}\Big[\Big(\frac{\tilde{r}}{\hat{a}(W)} - \tilde{r}\Big) R_{\hat{\lambda}}(W) \,\Big|\, \mathcal{S}\Big],$$

by the covariance computation in the proof of Theorem 2. We claim $\tilde{\Omega} \xrightarrow{p} \Omega_{\lambda^*,\pi^*}$. First, $\tilde{r} \to_p r$ by Step 0. Second, since $\hat{a} = \hat{\pi}\,(1 - \pi_N^{(0)}/\hat{\pi})/(1 - \pi_N^{(0)})$ and $\pi_N^{(0)}/\hat{\pi} \leq \pi_N^{(0)}/(\rho_{\min} r_N) \to 0$, we have $\tilde{r}/\hat{a}(W) = (1/\hat{\rho}(W))\,(1 + o_p(1))$ uniformly in W. Third, decompose

$$\mathbb{E}\Big[\frac{1}{\hat{\rho}}\, R_{\hat{\lambda}}\Big] - \mathbb{E}\Big[\frac{1}{\rho^*}\, R_{\lambda^*}\Big] = \mathbb{E}\Big[\Big(\frac{1}{\hat{\rho}} - \frac{1}{\rho^*}\Big) R_{\hat{\lambda}}\Big] + \mathbb{E}\Big[\frac{1}{\rho^*}\big(R_{\hat{\lambda}} - R_{\lambda^*}\big)\Big].$$

For the first term, $|1/\hat{\rho} - 1/\rho^*| \leq \rho_{\min}^{-2}\,|\hat{\rho} - \rho^*|$ and $\|R_{\hat{\lambda}}(W)\| \leq \bar{q}(W)$, so by the Cauchy–Schwarz inequality and condition (iii) it is bounded in norm by $\rho_{\min}^{-2}\,\mathbb{E}[(\hat{\rho} - \rho^*)^2]^{1/2}\,\mathbb{E}[\bar{q}^2]^{1/2} \to_p 0$. For the second term, $e_i(\hat{\lambda}) - e_i(\lambda^*) = -\sum_m (\hat{\lambda}_m - \lambda_m^*)\, g(Z_i, \tilde{T}_{i,m}; \theta_0)$ implies $\|R_{\hat{\lambda}}(W) - R_{\lambda^*}(W)\| \leq C\,\|\hat{\lambda} - \lambda^*\|\,\bar{q}(W)$, so it is bounded in norm by $C\rho_{\min}^{-1}\,\|\hat{\lambda} - \lambda^*\|\,\mathbb{E}[\bar{q}] \to_p 0$ by condition (ii). The terms multiplying $\tilde{r}$ and r converge analogously. Combining, $\tilde{\Omega} \xrightarrow{p} \Omega_{\lambda^*,\pi^*}$.

Step 3 (conditional CLT and conclusion). The conditional Lyapunov computation in the proof of Theorem 2 applies with overlap constant $\rho_{\min}/4$ and the $(2+\delta)$ moments, with bounds uniform over the design class, so the conditional Lyapunov ratio is $O_p(\tilde{n}^{-\delta/2}) = o_p(1)$. Take any subsequence; extract a further subsequence along which $\tilde{\Omega} \to \Omega_{\lambda^*,\pi^*}$ a.s., the overlap and rate statements hold

a.s., and the conditional remainder terms vanish a.s. Along this subsequence, the Lindeberg–Feller theorem applies conditionally for almost every realization, and the Cramér–Wold device yields, for all $a \in \mathbb{R}^d$ and $t \in \mathbb{R}$,

$$\mathbb{P}\Big(a^\top \sqrt{\tilde{n}}\,(\hat{\theta}^{\mathrm{TS}} - \theta_0) \le t \,\Big|\, \mathcal{S}\Big) \to \Phi\big(t/\sigma_a\big) \quad \text{a.s.}, \qquad \sigma_a^2 = a^\top J_{\theta_0}^{-1} \Omega_{\lambda^*,\pi^*} J_{\theta_0}^{-\top} a.$$

Dominated convergence removes the conditioning; since the limit does not depend on the subsequence, the convergence holds along the full sequence. Finally, $\tilde{n}/n \to_p 1$ and Slutsky's theorem give the same limit with the $\sqrt{n}$ scaling. □

A.12 Proof of Proposition 5

Proof. Let K be a compact neighborhood of θ_0 on which the envelopes of Assumption 2 apply (under the global conditions of Theorem 1(i), take $K = \Theta$); all bounds below are uniform over $\theta \in K$. Define the unit-level envelope

$$H_i := \|g_i(\theta_0)\| + \sum_{m=1}^{M} \big\|\tilde{g}^0_{i,m}(\theta_0)\big\| + \Big(L_{\theta,H}(Z_i) + \sum_{m=1}^{M} L_{\theta,m}(Z_i)\Big)\,\mathrm{diam}(K),$$

so that, using $\hat{\lambda} \in \Lambda$ and Assumption 2.4, $\sup_{\theta\in K}\big(\,\|g_i(\theta)\| + \|b_i(\theta)\|\,\big) \le H_i + M L_T(Z_i) U_N$ with $\mathbb{E}[H_i] < \infty$ and $\mathbb{E}[L_T(Z_i)] < \infty$, where $b_i(\theta) = \sum_m \hat{\lambda}_m \tilde{g}_{i,m}(\theta)$. Write $\pi_N^{(0)} = n_0/N = (n_0/n)\, r_N$, $N_0 = \#\{i\colon D_i^{(0)} = 1\}$, $N' = N - N_0$, and recall from Step 0 of the proof of Theorem 5 that, with probability tending to one, $\hat{a}(W) \ge \rho_{\min} r_N/2$ and $\hat{\pi}(W) \ge \rho_{\min} r_N$; we argue on this event. Let $G^{\mathrm{P}}(\theta) = \frac{1}{N}\sum_i \psi_i^{\mathrm{P}}(\theta)$ with $\psi_i^{\mathrm{P}}(\theta) = b_i(\theta) + \big(D_i/\hat{\pi}(W_i)\big)\big(g_i(\theta) - b_i(\theta)\big)$, and let $G^{\mathrm{S}}(\theta) = \frac{1}{N'}\sum_{i\colon D_i^{(0)}=0} \psi_i^{\mathrm{S}}(\theta)$ with $\psi_i^{\mathrm{S}}(\theta) = b_i(\theta) + \big(D_i^{(1)}/\hat{a}(W_i)\big)\big(g_i(\theta) - b_i(\theta)\big)$ denote the split moment of Theorem 5.

Step 1 (pilot units). For pilot units, $D_i = 1$ and $\sup_{\theta\in K}\big\|\psi_i^{\mathrm{P}}(\theta)\big\| \le \big(1 + 2/(\rho_{\min} r_N)\big)\big(H_i + M L_T(Z_i) U_N\big)$. The pilot indicators are i.i.d. Bernoulli$(\pi_N^{(0)})$ and independent of the data. As a result, $\mathbb{E}\big[\frac{1}{N}\sum_{i\colon D_i^{(0)}=1} H_i\big] = \pi_N^{(0)}\mathbb{E}[H_i]$, and Markov's inequality gives $\frac{1}{N}\sum_{\mathrm{pilot}} H_i = O_p(n_0/N)$, and likewise for L_T. With $U_N = o_p(1)$,

$$\sup_{\theta\in K}\Big\|\frac{1}{N}\sum_{i\colon D_i^{(0)}=1} \psi_i^{\mathrm{P}}(\theta)\Big\| = O_p\Big(\frac{n_0}{N r_N}\Big) = O_p\Big(\frac{n_0}{n}\Big).$$

Step 2 (propensity discrepancy on non-pilot units). For non-pilot units, $\psi_i^{\mathrm{P}}(\theta) - \psi_i^{\mathrm{S}}(\theta) = D_i^{(1)}\big(1/\hat{\pi}(W_i) -$

$1/\hat{a}(W_i)\big)\big(g_i(\theta) - b_i(\theta)\big)$. Since $\hat{a} \le \hat{\pi}$ and $\hat{\pi} - \hat{a} = \pi_N^{(0)}(1-\hat{\pi})/(1-\pi_N^{(0)}) \le 2\pi_N^{(0)}$ for all large N,

$$\Big|\frac{1}{\hat{\pi}(W_i)} - \frac{1}{\hat{a}(W_i)}\Big| \le \frac{2\pi_N^{(0)}}{\hat{\pi}(W_i)\,\hat{a}(W_i)} \le \frac{4\,\pi_N^{(0)}}{\rho_{\min}^2 r_N^2} = \frac{4}{\rho_{\min}^2} \cdot \frac{n_0}{n} \cdot \frac{1}{r_N}.$$

By condition (v), $\hat{a}(W) \le \hat{\pi}(W) = r_N\hat{\rho}(W) \le r_N\bar{\rho}$, so, conditionally on $\mathcal{S}$, $\mathbb{E}\big[\frac{1}{N}\sum_{i:\, D_i^{(0)}=0} D_i^{(1)} H_i \mid \mathcal{S}\big] \le r_N\bar{\rho}\,\mathbb{E}[H_i]$ and hence $\frac{1}{N}\sum_{\text{non-pilot}} D_i^{(1)} H_i = O_p(r_N)$; the same bound holds with $L_T(Z_i)$ in place of H_i, and the U_N contribution factors out as $o_p(1)$. Therefore

$$\sup_{\theta\in K}\Big\|\frac{1}{N}\sum_{i:\, D_i^{(0)}=0}\big(\psi_i^{\mathrm{P}} - \psi_i^{\mathrm{S}}\big)(\theta)\Big\| \;=\; O_p\Big(\frac{n_0}{n}\cdot\frac{1}{r_N}\Big)\cdot O_p(r_N) \;=\; O_p\Big(\frac{n_0}{n}\Big).$$

Step 3 (normalization). $\frac{1}{N}\sum_{i:\, D_i^{(0)}=0}\psi_i^{\mathrm{S}}(\theta) = \frac{N'}{N}\,G^{\mathrm{S}}(\theta)$, and $1 - N'/N = N_0/N = O_p(n_0/N)$. Moreover, $\mathbb{E}\big[\sup_{\theta\in K}\big\|\psi_i^{\mathrm{S}}(\theta)\big\| \mid \mathcal{S}\big] \le \mathbb{E}[H_i] + \mathbb{E}\big[(D_i^{(1)}/\hat{a}(W_i))\,(H_i + ML_TU_N) \mid \mathcal{S}\big] = O_p(1)$ by the conditional unbiasedness of $D_i^{(1)}/\hat{a}(W_i)$, so $\sup_{\theta\in K}\big\|G^{\mathrm{S}}(\theta)\big\| = O_p(1)$ and the normalization error is $O_p(n_0/N)$, dominated by $O_p(n_0/n)$. Combining Steps 1–3,

$$\sup_{\theta\in K}\big\|G^{\mathrm{P}}(\theta) - G^{\mathrm{S}}(\theta)\big\| = O_p\Big(\frac{n_0}{n}\Big).$$

Step 4 (conclusion). Since $n_0/n \to 0$, G^{P} inherits the uniform convergence of G^{S} to G established (conditionally on $\mathcal{S}$) in Step 1 of the proof of Theorem 5, so the identification argument of Theorem 1 yields $\hat{\theta}^{\mathrm{P}} \xrightarrow{p} \theta_0$. Moreover,

$$\big\|G^{\mathrm{S}}(\hat{\theta}^{\mathrm{P}})\big\| \le \big\|G^{\mathrm{P}}(\hat{\theta}^{\mathrm{P}})\big\| + O_p(n_0/n) = O_p(n_0/n) = o_p(n^{-1/2})$$

by condition (iv), so $\hat{\theta}^{\mathrm{P}}$ is an $o_p(n^{-1/2})$-approximate zero of G^{S}. The conditional linearization in Step 1 of the proof of Theorem 5 uses only that the estimator is consistent and makes the split moment $o_p(n^{-1/2})$ small, so it applies to $\hat{\theta}^{\mathrm{P}}$ and yields the same expansion as for $\hat{\theta}^{\mathrm{TS}}$, up to an additional $o_p(1)$ term after scaling by $\sqrt{\tilde{n}}$. The conditional central limit theorem and the de-conditioning argument of Steps 2 and 3 of that proof, together with $\tilde{n}/n \to_p 1$ and Slutsky's theorem, give the stated limit. □

A.13 Proof of Proposition 6

Proof. (a) The estimator, the assumptions, and every argument in the proofs of Theorems 1, 2, and 4 use only that the calibration set reveals, for units with $D_i = 1$, a feature value at which the moment function is evaluated, and that the estimand is identified by the corresponding population moment. Replacing (T_i, θ_0) by (T_i^*, θ_0^*) throughout leaves all arguments unchanged.

(b) Write $G^*(\theta) := \mathbb{E}[g(Z_i, T_i^*; \theta)]$ and $h(\theta) := G^*(\theta) - G(\theta)$. A second-order Taylor expansion with integral remainder gives, for $\theta \in \Theta_0$ and componentwise in g,

$$h(\theta) = \mathbb{E}\big[\partial_t\, g(Z_i, T_i; \theta)\, e_i\big] + \mathbb{E}\Big[\int_0^1 (1-u)\, e_i^\top\, \partial_t^2\, g\big(Z_i, T_i + u\, e_i; \theta\big)\, e_i\, du\Big].$$

The first term vanishes: by Assumption 6.1 and Fubini's theorem, justified by the domination in 1,

$$\mathbb{E}\big[\partial_t\, g(Z_i, T_i; \theta)\, e_i\big] = \mathbb{E}\big[\partial_t\, g(Z_i, T_i; \theta)\, \mathbb{E}(e_i \mid Z_i, T_i)\big] = 0.$$

By Assumption 6.2, the second term is bounded in norm by $\frac{1}{2}\mathbb{E}[\nu(Z_i, e_i)\|e_i\|^2]$, uniformly over $\theta \in \Theta_0$:

$$\sup_{\theta \in \Theta_0} \|G^*(\theta) - G(\theta)\| \;\le\; \tfrac{1}{2}\,\kappa^2.$$

By Assumption 2.2, G is continuously differentiable near θ_0 with J_{θ_0} positive definite, so by the inverse function theorem there exist $r_0, c_0, \eta > 0$ such that G is a homeomorphism from $\bar{B}_{r_0}(\theta_0)$ onto its image, $\|G(\theta)\| \ge c_0\, \|\theta - \theta_0\|$ on $\bar{B}_{r_0}(\theta_0)$, and $G(\bar{B}_{r_0}(\theta_0)) \supseteq \bar{B}_\eta(0)$.

Existence. Set $\bar{\kappa} := 2\min(\eta,\, c_0 r_0)$ and suppose $\kappa^2 \le \bar{\kappa}$. The map $\phi(\theta) := G^{-1}\big(-h(\theta)\big)$ is well defined and continuous on $\bar{B}_{r_0}(\theta_0)$, since $\|h(\theta)\| \le \kappa^2/2 \le \eta$ and h is continuous by dominated convergence. By the lower bound on G, ϕ maps $\bar{B}_{r_0}(\theta_0)$ into the closed convex set $\bar{B}_{\kappa^2/(2c_0)}(\theta_0) \subseteq \bar{B}_{r_0}(\theta_0)$; in particular ϕ maps $\bar{B}_{\kappa^2/(2c_0)}(\theta_0)$ into itself, and Brouwer's fixed-point theorem yields θ_0^* with $\phi(\theta_0^*) = \theta_0^*$, that is, $G(\theta_0^*) = -h(\theta_0^*)$ and hence $G^*(\theta_0^*) = 0$.

Bound. Any solution $\theta_0^* \in \bar{B}_{r_0}(\theta_0)$ of $G^*(\theta) = 0$ satisfies $\|G(\theta_0^*)\| = \|h(\theta_0^*)\| \le \kappa^2/2$, hence

$$\|\theta_0^* - \theta_0\| \;\le\; \frac{\kappa^2}{2c_0} \;=:\; c\,\kappa^2.$$

(c) If $t \mapsto g(z, t; \theta)$ is affine, the second-order term vanishes identically, so $h \equiv 0$ by part (b)'s first-term computation, $G^* = G$, and $\theta_0^* = \theta_0$. □

A.14 Proof of Theorem 6

Proof. Write $G_N(\theta) \coloneqq G_{\lambda,\pi}(\theta)$, $Q_N(\theta) \coloneqq G_N(\theta)^\top \hat{A}\, G_N(\theta)$, and $Q(\theta) \coloneqq G(\theta)^\top A\, G(\theta)$.

Consistency. The proof of Theorem 1 establishes $\sup_{\theta\in\Theta} \|G_N(\theta) - G(\theta)\| = o_p(1)$ under the global Lipschitz conditions on the compact set Θ; the argument is coordinate-wise and nowhere uses $q = d$. Since $\sup_{\theta\in\Theta} \|G(\theta)\| < \infty$ by Assumption 2.1 and Jensen's inequality, and $\hat{A} \to_p A$,

$$\sup_{\theta\in\Theta} \left|Q_N(\theta) - Q(\theta)\right| \le \|\hat{A}\| \sup_{\theta} \Big(\|G_N - G\| \, \big(\|G_N\| + \|G\| \big) \Big) + \|\hat{A} - A\| \sup_{\theta} \big(\|G\|^2 \big) = o_p(1).$$

Q is continuous, nonnegative, and zero only at θ_0 by identification and positive definiteness of A; Θ is compact. The standard argmin argument (e.g., Newey and McFadden, 1994, Theorem 2.1) gives $\hat{\theta}^{\mathrm{GMM}}_{\lambda,\pi} \to_p \theta_0$.

Asymptotic normality. We verify the conditions of Theorem 3.3 of Pakes and Pollard (1989) for the criterion $\theta \mapsto \left\|A^{1/2} G_N(\theta)\right\|$. The substitution of $\hat{A}^{1/2}$ for $A^{1/2}$ is innocuous: $\hat{A} \to_p A$ with A positive definite implies that the minimizer of $\left\|\hat{A}^{1/2} G_N(\cdot)\right\|$ minimizes $\left\|A^{1/2} G_N(\cdot)\right\|$ up to a multiplicative factor $1+o_p(1)$, and hence up to an additive $o_p(n^{-1/2})$ term, since $\inf_\theta \left\|A^{1/2} G_N(\theta)\right\| \le \left\|A^{1/2} G_N(\theta_0)\right\| = O_p(n^{-1/2})$ by condition (iv) below. The conditions are: (i) $\left\|A^{1/2} G_N(\hat{\theta}^{\mathrm{GMM}})\right\| \le \inf_{\theta\in\Theta} \left\|A^{1/2} G_N(\theta)\right\| + o_p(n^{-1/2})$, which holds by the definition of the estimator and the preceding remark; (ii) G is differentiable at θ_0 with derivative J_{θ_0} of full column rank, as assumed; (iii) stochastic equicontinuity: for every sequence $\delta_N \downarrow 0$,

$$\sup_{\|\theta-\theta_0\|\le\delta_N} \sqrt{n}\, \|(G_N - G)(\theta) - (G_N - G)(\theta_0)\| = o_p(1),$$

which is established in the proof of Theorem 2: the empirical-process argument applies to each of the q components, and the spillover contribution is $o_p(1)$ after scaling by $\sqrt{n}$ by Assumption 3; (iv) $\sqrt{n}\, G_N(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega_{\lambda,\pi})$, established in the proof of Theorem 2 (the central limit theorem step is dimension-free given the $(2+\delta)$ moments); (v) $\hat{\theta}^{\mathrm{GMM}}_{\lambda,\pi} \to_p \theta_0$, established above. The conclusion of Theorem 3.3 of Pakes and Pollard (1989) is

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{GMM}}_{\lambda,\pi} - \theta_0\right) = -\left(J_{\theta_0}^\top A J_{\theta_0}\right)^{-1} J_{\theta_0}^\top A\, \sqrt{n}\, G_N(\theta_0) + o_p(1),$$

which yields the stated limit distribution. For the efficient choice, a preliminary consistent estimator (e.g., with $\hat{A} = I_q$) delivers $\hat{\Omega}_{\lambda,\pi} \to_p \Omega_{\lambda,\pi}$ by Remark 1, so $\hat{A} = \hat{\Omega}^{-1}_{\lambda,\pi} \to_p \Omega^{-1}_{\lambda,\pi}$, and substituting $A = \Omega^{-1}_{\lambda,\pi}$ in the sandwich gives $(J_{\theta_0}^\top \Omega^{-1}_{\lambda,\pi} J_{\theta_0})^{-1}$. □

A.15 Proof of Proposition 7

Proof. Every assumption and every step of the proofs in this paper constrains only (i) the joint distribution of the i.i.d. unit-level vector $(Z_i, T_i, \{\tilde{T}_{i,m,s}\}, W_i, D_i)$, (ii) the moment function g through the moment, identification, Lipschitz, and envelope conditions of Assumptions 1–3 (and, where invoked, Assumptions 4–6 and the design conditions of Section 5), and (iii) the labeling mechanism through Assumption 1. No step restricts the internal composition of Z_i, the dimension of T_i, which may differ across units (formally, T_i takes values in the disjoint union $\sqcup_\tau \mathbb{R}^{p\tau}$, and norms and Lipschitz conditions are imposed on the relevant component), or the dependence among components within a unit. The clustered model defines exactly such a unit-level sequence: clusters are i.i.d. by assumption, g is the within-cluster aggregate, the population moment condition holds at the cluster level, and the labeling is a cluster-level design satisfying Assumption 1. Hence every proof applies verbatim after this substitution.

The quantities $\mathrm{Var}(g(Z_i, T_i; \theta_0))$ and $R_\lambda(W_i)$ are covariance matrices of within-cluster sums, that is, clustered covariance matrices, and the plug-in estimator of Remark 1 estimates exactly these objects, so the resulting standard errors are clustered. Finally, the cluster-level envelope conditions hold under primitive observation-level conditions: if g is Lipschitz in the feature with observation-level envelope $\ell_{i,t}$, the Cauchy–Schwarz inequality gives, for cluster-level features t_1, t_2,

$$\|g(Z_i, t_1; \theta) - g(Z_i, t_2; \theta)\| \le \sum_{t \le \tau_i} \ell_{i,t} \, \|t_{1,t} - t_{2,t}\| \le \Big(\sum_{t \le \tau_i} \ell_{i,t}^2 \Big)^{1/2} \|t_1 - t_2\| \,,$$

so $L_T(Z_i) = (\sum_{t \le \tau_i} \ell_{i,t}^2)^{1/2}$, which has finite second moments, for example, when cluster sizes are bounded and the observation-level envelopes are square-integrable. □

B Supplementary Material for the Empirical Study

This appendix collects supplementary material for the empirical study of Section 9: the prompt templates, the distribution of the labeling information variable, the sensitivity of the estimates to the momentum and short-term reversal controls, the contemporaneous regression, and the Fama–MacBeth robustness tests.

B.1 LLM Prompt Designs

This appendix reports the prompt templates used to construct the LLM-based headline sentiment labels. In each prompt, [firm name], [ticker], and [headline] are replaced with the corresponding values for a headline observation. The model is asked to classify whether the headline is positive, neutral, or negative for the associated firm.

The prompt templates below are shown in italics.

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: *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)

Professional-investor prompt: *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?

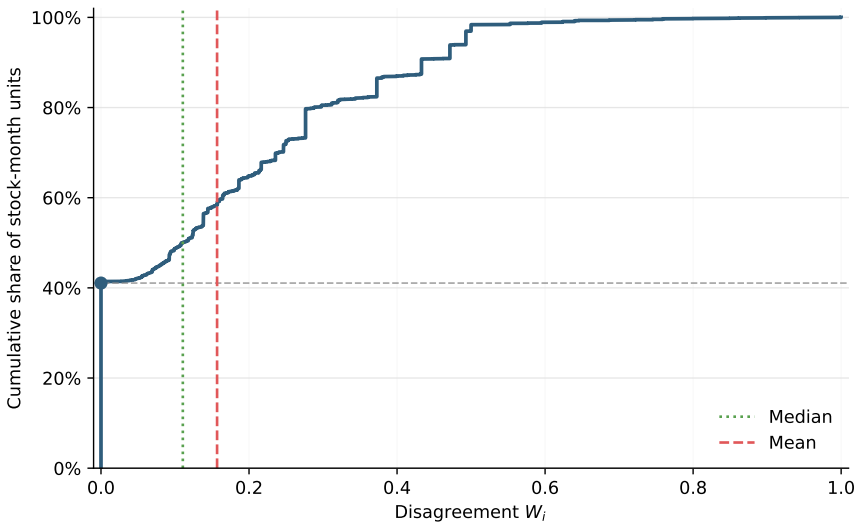
Return a JSON object with exactly one field, label. The value of label must be exactly one of: positive, negative, or neutral.

For the baseline and professional-investor prompts, the API response is constrained to return a label field taking one of the three allowed values. The fill-in prompt is intentionally less structured: it removes the JSON-format requirement and asks the model to directly fill in one of the three labels. This design allows us to examine whether the LLM sentiment measures are sensitive not only to model choice but also to prompt format and framing.

B.2 Distribution of the Labeling Information Variable

This appendix describes the distribution of the information variable used in the targeted labeling design. In the empirical implementation of Algorithm 1, we set W_i equal to the cross-model-prompt disagreement in generated sentiment for stock-month unit i. Specifically, W_i is the standard deviation of the 12 generated stock-month sentiment measures across model–prompt pairs. Units with

Figure B.1: Distribution of model-prompt disagreement



The figure plots the empirical cumulative distribution function of the information variable W_i, defined as the cross-model-prompt standard deviation of the 12 generated stock-month sentiment measures. The jump at zero corresponds to stock-month units for which all 12 model-prompt pairs assign the same generated sentiment score.

larger W_i are those for which the generated sentiment measures disagree more strongly, and are therefore natural candidates for targeted human labeling.

Figure B.1 plots the empirical cumulative distribution function of W_i across the stock-month estimation sample. The distribution is discrete as the headline-level sentiment labels take values in $\{-1, 0, 1\}$ and stock-month sentiment is constructed by averaging a finite number of sampled headline labels. A substantial mass occurs at zero, corresponding to stock-months for which all 12 model–prompt pairs assign the same generated sentiment score. Away from zero, the distribution summarizes the degree of disagreement available for the targeted second-stage labeling policy.

B.3 Sensitivity to Momentum and Short-Term Reversal Controls

This appendix examines whether the next-month sentiment coefficient is driven by the momentum and short-term reversal controls. Table B.1 reports three alternative specifications. Panel A drops momentum but keeps the same-month return $R_{i,t}$. Panel B keeps momentum but drops the same-month return. Panel C keeps only log size and book-to-market.

The results are stable across the three specifications. The AI-PI estimate is 0.316 when controlling for log size, book-to-market, and same-month return; 0.289 when controlling for log size, book-to-market, and momentum; and 0.293 when controlling only for log size and book-to-market. The naive uniform estimator is smaller in all three specifications, ranging from 0.162 to 0.193.

Table B.1: AI-PI Estimates Under Alternative Control Sets

Estimator	Coefficient	SE	95% CI	CI length
Panel A. Controls: log size, book-to-market, same-month return $R_{i,t}$				
Naive uniform estimator	0.193**	0.087	[0.022, 0.365]	0.343
Calibration-only IPW	0.568*	0.300	[-0.019, 1.155]	1.174
AI-PI	0.316**	0.155	[0.012, 0.620]	0.607
Panel B. Controls: log size, book-to-market, momentum				
Naive uniform estimator	0.162**	0.080	[0.005, 0.319]	0.314
Calibration-only IPW	0.524*	0.294	[-0.052, 1.099]	1.152
AI-PI	0.289**	0.146	[0.003, 0.575]	0.572
Panel C. Controls: log size and book-to-market only				
Naive uniform estimator	0.168**	0.081	[0.009, 0.326]	0.317
Calibration-only IPW	0.530*	0.294	[-0.046, 1.105]	1.151
AI-PI	0.293**	0.146	[0.006, 0.580]	0.574

This table compares coefficient estimates on the sentiment measure under alternative control sets. The dependent variable is next-month stock return. All rows include industry fixed effects and month fixed effects. Momentum is the cumulative CRSP raw return from months $t-12$ through $t-2$. Same-month return $R_{i,t}$ is measured in the same month as headline sentiment. The AI-PI estimators reuse the realized human labels, labeling propensities, and Algorithm 1 weights from the main specification. Standard errors are clustered by publication month. The sample contains 107,495 stock-month observations and 5,320 human-labeled stock-month units. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Thus, the correction from the naive generated estimator to AI-PI does not appear to be driven by a particular choice of momentum or short-term reversal control.

B.4 Contemporaneous Stock-Month Regression

In Table B.2, we replace next-month stock returns with same-month stock returns in the stock-month specification. Since the dependent variable is now the same-month return $R_{i,t}$, the short-term return control is the previous-month return $R_{i,t-1}$. The regressions also include log size, book-to-market, momentum, industry fixed effects, and month fixed effects.

Table B.2 shows that the coefficients are positive and statistically significant at the 1% level for all 12 model–prompt pairs, ranging from 2.993 to 3.436. The naive uniform estimator is 3.701, and the AI-PI estimate is 3.466. This pattern indicates that the generated headline sentiment measures strongly capture contemporaneous stock-month news. The much smaller next-month coefficients in the main specification therefore reflect only the remaining short-horizon drift after most of the information is included within the publication month.

Table B.2: Contemporaneous Stock-Month Return Regressions

Estimator	Coefficient	SE
GPT-3.5 Turbo – Baseline	3.088***	0.092
GPT-3.5 Turbo – Fill-in	3.279***	0.103
GPT-3.5 Turbo – Professional investor	2.993***	0.089
GPT-4o mini – Baseline	3.212***	0.102
GPT-4o mini – Fill-in	3.070***	0.094
GPT-4o mini – Professional investor	3.143***	0.097
GPT-4.1 nano – Baseline	3.077***	0.098
GPT-4.1 nano – Fill-in	3.436***	0.103
GPT-4.1 nano – Professional investor	3.009***	0.093
GPT-5 mini – Baseline	3.226***	0.100
GPT-5 mini – Fill-in	3.287***	0.099
GPT-5 mini – Professional investor	3.196***	0.100
Naive uniform estimator	3.701***	0.107
Calibration-only IPW	3.508***	0.401
AI-PI	3.466***	0.162

This table reports coefficient estimates on the sentiment measure from the contemporaneous stock-month return regressions in which the dependent variable is the same-month stock return $R_{i,t}$. Each model–prompt row replaces the latent sentiment measure with one generated stock-month sentiment measure. The naive uniform estimator uses the simple average of the 12 generated stock-month sentiment measures. The calibration-only IPW estimator uses only the human-labeled stock-month sentiment measure. Both calibration-only IPW and AI-PI estimators use the final 5,320 labeled stock-month units and the propensities from Algorithm 1. All rows include log size, book-to-market, momentum, previous-month return $R_{i,t-1}$, industry fixed effects, and month fixed effects. Standard errors are clustered by month. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B.5 Fama–MacBeth Robustness Tests

As an additional robustness check, we estimate a Fama–MacBeth version of the stock-month regression. We use the same outcome, sentiment measures, controls, and industry fixed effects as in the main specification. For comparability with the AI-PI estimates, we first residualize the outcome and sentiment variables with respect to the controls and industry fixed effects. We then estimate the residualized sentiment-return relation separately within each publication month and report the time-series average of the monthly coefficients. Standard errors are Newey–West standard errors for the time-series mean of the monthly coefficients, using three lags.

Table B.3 reports the results. The Fama–MacBeth estimates are similar to the main month-clustered estimates. The naive generated sentiment coefficients remain positive but vary across model–prompt pairs, ranging from 0.102 to 0.234. The naive uniform estimator is 0.201, while the AI-PI estimate is 0.298 with a Newey–West standard error of 0.151. The calibration-only IPW estimator remains much less precise. Thus, the Fama–MacBeth exercise gives the same qualitative

Table B.3: Fama–MacBeth Robustness for Stock-Month Sentiment Coefficients

Estimator	Coefficient	SE
GPT-3.5 Turbo – Baseline	0.162**	0.072
GPT-3.5 Turbo – Fill-in	0.234***	0.079
GPT-3.5 Turbo – Professional investor	0.168**	0.079
GPT-4o mini – Baseline	0.176**	0.069
GPT-4o mini – Fill-in	0.169**	0.071
GPT-4o mini – Professional investor	0.158**	0.076
GPT-4.1 nano – Baseline	0.175**	0.078
GPT-4.1 nano – Fill-in	0.204**	0.086
GPT-4.1 nano – Professional investor	0.175**	0.079
GPT-5 mini – Baseline	0.162**	0.076
GPT-5 mini – Fill-in	0.102	0.074
GPT-5 mini – Professional investor	0.153*	0.079
Naive uniform estimator	0.201**	0.084
Calibration-only IPW	0.604*	0.328
AI-PI	0.298**	0.151

This table reports Fama–MacBeth robustness estimates for stock-month sentiment coefficients. The dependent variable is next-month stock return. The outcome and sentiment variables are first residualized with respect to the controls and industry fixed effects, using the same controls as the main specification: log size, book-to-market, momentum, and same-month return $R_{i,t}$. The residualized sentiment-return relation is then estimated separately within each publication month, and the reported coefficient is the time-series average of monthly coefficients. Standard errors are Newey–West standard errors for the time-series mean with three lags. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

conclusion as the main specification: generated sentiment measures imply positive short-horizon continuation, and AI-PI gives a larger corrected estimate than the naive uniform estimator while remaining substantially more precise than calibration-only IPW.