Generative-AI Powered Inference

Junting Duan* Markus Pelger†

This draft: March, 2026

Abstract

Large language models are increasingly used to generate covariates from unstructured data, yet treating such outputs as observed can induce "hallucination bias" and invalidate structural inference. We propose Generative-AI Powered Inference (Gen-PI), a method-of-moments framework that corrects this bias using a small human-labeled calibration set. Gen-PI accounts for the heterogeneous randomness of generative models, where each model–prompt pair exhibits distinct, input-dependent bias and seed-level noise, through adaptive weighting that downweights unreliable generators and amplifies informative ones. We derive the optimal calibration-set design, a Neyman-type allocation concentrating labels where synthetic features are least reliable, and provide guidance on calibration-set selection. Under cross-fitting for LLM pipeline tuning, we establish that our estimator is consistent and asymptotically normal, which enables the construction of valid confidence intervals. Simulations confirm substantial gains over naive approaches and methods that do not optimally combine generated and labeled data.

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

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

1 Introduction

Modern empirical research increasingly relies on unstructured data—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, all at negligible marginal cost and at 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 a structured covariate panel, 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. Researchers routinely feed text into an LLM, extract a numerical feature, and plug it into a regression 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. Their outputs are subject to *hallucination bias*, a systematic 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.

Treating LLM-generated features as observed data in a structural estimation problem can therefore produce biased point estimates, incorrect standard errors, and misleading hypothesis tests. The

problem is not merely one of attenuation or measurement error in the classical sense. Because the generative model's errors depend on the input content and its training data, the resulting bias can be arbitrarily complex and is generally not sign-identified without additional information. As a result, the naive use of synthetic covariates can invalidate downstream inference.

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. The validation step provides a false sense of security without resolving the underlying inferential problem.

1.1 Our Framework

In this paper, we propose *Generative-AI Powered Inference* (Gen-PI), a general framework for valid estimation and inference when a target covariate or latent feature is produced by a generative model. Our framework proceeds in three steps. First, we generate features from one or more LLM model–prompt pairs for the entire sample, treating each generated feature as a noisy, potentially biased proxy for the true latent quantity. Second, we collect human labels on a small, optimally designed *calibration set* of observations, providing a gold-standard realization of the true feature. Third, we use the calibration set to construct a bias-correction term (LLM rectifier) that removes the systematic distortion introduced by the generative model in the moment conditions relevant to the structural estimand. We then combine this with adaptive weighting across model–prompt pairs to obtain an estimator that is both unbiased and efficient.

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 that the generative model introduces 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.

Beyond debiasing, we address the heterogeneity that arises 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, guaranteeing 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 fixed labeling budget and can yield substantial efficiency 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 Gen-PI estimator. Under standard moment regularity conditions and a mild assumption on 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 the full complexity of modern LLM pipelines. When the prompt, the retrieval-augmented generation (RAG) configuration, or the choice of model is tuned 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 that is natural for LLM pipelines, which requires that the generated features do not change discontinuously as the tuning data vary, cross-fitting restores independence and preserves the asymptotic guarantees.

1.2 Use Cases and Motivating Examples

The Gen-PI framework broadly applies to settings where structured features are extracted from unstructured data using generative models and subsequently used in a downstream statistical analysis. We highlight several concrete use cases that motivate our framework.

Sentiment analysis in empirical finance. A large and growing literature in finance uses LLMs to extract sentiment scores from earnings call transcripts, news articles, and analyst reports, and then studies how these scores predict stock returns, volatility, or corporate actions (Li, Shi, Xia, and Yang, 2025). In this setting, the text S_i is an earnings call transcript, the true feature T_i is a human-assessed sentiment score, and the LLM-generated feature $\tilde{T}_i$ is the model's output. Bias arises when the LLM assigns sentiment based on generic patterns in its training data rather than the specific content of the transcript, or when it introduces spurious details that affect the score. Gen-PI uses a calibration set of human-scored transcripts to correct the bias in the moment conditions of a return-prediction regression.

Clinical feature extraction from medical records. Electronic health records contain large amounts of unstructured clinical text—physician notes, radiology reports, discharge summaries—that encode diagnostic information, treatment decisions, and patient outcomes. LLMs can parse these records into structured features such as disease severity scores, comorbidity indicators, or treatment classifications, enabling large-scale epidemiological studies and health-economics analyses. However, hallucination in the medical domain can be particularly consequential: an LLM may fabricate a diagnosis, misclassify a medication, or conflate two patients' records. Gen-PI provides a principled approach for correcting these errors using a calibration set of expert-reviewed records.

Measuring political and social constructs from text. Political scientists and sociologists increasingly use LLMs to code legislative speeches, court opinions, social media posts, and survey responses for ideological position, policy stance, or affect. These features enter downstream models that study polarization, representation, or policy outcomes. Because political language is nuanced

and context-dependent, LLM-generated codings can exhibit systematic biases, for example by misclassifying irony, regional dialect, or coded language. A calibration set of expert-coded texts allows Gen-PI to correct these biases and produce valid inferences.

Patent and innovation classification. Researchers studying technological change and innovation routinely classify patents into technology categories, assess novelty, or measure knowledge flows using textual analysis. LLMs can automate this process at scale, but their classifications may reflect the distribution of their training data rather than the nuances of patent law. Gen-PI leverages a calibration set of expert-classified patents to correct these distortions, ensuring that downstream analyses of innovation patterns are not distorted by classification errors.

Generating missing unstructured records. In some applications, the unstructured data themselves are missing for a portion of the sample, and a generative model is used to synthesize plausible records. For example, an LLM might generate synthetic clinical notes for patients who lack documentation, or produce plausible earnings call transcripts for firms that do not hold earnings calls. In this setting, the entire record is synthetic, and the potential for hallucination bias is even greater. Gen-PI treats the features extracted from the synthetic records as imputations and corrects their bias using the calibration set.

1.3 Contributions

Our work makes three main contributions. First, we introduce the Gen-PI framework, which provides a general and rigorous 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 weighting automatically downweights unreliable generators and upweights informative ones; when all generators are poor, the weights shrink toward zero and the estimator reduces to the calibration-only estimator, guaranteeing robustness.

Second, we develop a complete asymptotic theory for the Gen-PI estimator, establishing con-

sistency (under both compactness and monotonicity conditions) and asymptotic normality under general conditions. The theory covers non-random labeling designs, multiple model–prompt pairs, and cross-fitting with LLM pipeline tuning. 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 in terms of both selection and size. 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 efficiency asymptotically. These results provide actionable guidance on where to allocate labeling resources and how many labels are needed to achieve a desired level of precision.

1.4 Related Literature

Our work connects to and builds on several streams of literature.

Prediction-powered inference. Our debiasing approach is most closely related to the prediction-powered inference (PPI) framework of Angelopoulos, Bates, Fannjiang, Jordan, and Zrnic (2023), which uses a labeled dataset to correct the bias in machine-learning predictions for missing outcomes. Recent extensions include Angelopoulos, Duchi, and Zrnic (2023) and Miao, Miao, Wu, Zhao, and Lu (2023), which develop methods that adapt to the precision of ML predictions; Ji, Lei, and Zrnic (2025), which proposes a recalibration approach; and Zrnic and Candès (2024), which introduces cross-prediction-powered inference. A growing body of research in semi-supervised learning studies how to leverage large amounts of data with missing outcomes to improve estimation (Wasserman and Lafferty, 2007; Chakrabortty and Cai, 2018). Our framework generalizes this literature in several directions. We address settings where the generated object is a multivariate covariate rather than a univariate outcome, accommodate multiple model–prompt pairs with heterogeneous bias profiles, introduce optimal calibration-set design with non-random labeling, and provide cross-fitting guarantees for LLM pipeline tuning.

Imputation-powered inference for missing covariates. Our framework builds on the imputation-powered inference (IPI) framework of Duan and Pelger (2025), which develops debiased

estimation and adaptive weighting for downstream inference with missing covariates. While IPI addresses the case where structured covariates are missing and imputed using matrix completion or regression methods, Gen-PI extends the paradigm to settings where the "imputation" is performed by a generative AI model applied to unstructured data. This introduces several new challenges, including the complex bias structure of LLM outputs and the availability of multiple model–prompt pairs with heterogeneous and input-dependent error profiles. A further 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. In contrast, in the GenAI setting the researcher actively chooses which observations receive costly human labels, subject to a budget constraint. Gen-PI formalizes this choice and derives the optimal labeling allocation jointly with the model–prompt weights, providing guidance on both the selection and size of the calibration set that has no analogue in existing imputation methods.

Missing covariates in structural estimation. A classical literature studies estimation with missing covariates (Little, 1986; Abrevaya and Donald, 2017; Freyberger, Höppner, Neuhierl, and Weber, 2024). These methods typically assume that the missing data mechanism and the imputation procedure satisfy specific parametric or semiparametric conditions. Our framework differs in that the "missing" covariate is a latent feature that has never been directly observed for most units; it must be extracted from unstructured data by a generative model. The bias structure of LLM outputs is fundamentally different from that of standard imputation methods, and our framework is specifically designed to accommodate it. Moreover, unlike classical missing-data settings where the observed subset is exogenously determined, Gen-PI allows the researcher to optimally design the calibration set by choosing which observations to label and how many, yielding efficiency gains that are unavailable in traditional approaches.

LLMs for data annotation and feature extraction. A rapidly growing empirical literature uses LLMs for text annotation, sentiment analysis, and feature extraction (Li, Shi, Xia, and Yang, 2025; Gilardi, Alizadeh, and Kubli, 2023; Törnberg, 2024). Gilardi, Alizadeh, and Kubli (2023) demonstrate that ChatGPT can outperform crowd-workers on several text-annotation tasks, while Törnberg (2024) provides best practices for text annotation with LLMs. Most of this work treats LLM outputs as observed data and uses diagnostic validation (e.g., agreement rates with human labels) to argue for their reliability. Our framework provides the missing inferential foundation:

it shows how to move from diagnostic validation to formal bias correction, and it quantifies the conditions under which LLM-generated features can be safely used in structural estimation.

Econometric frameworks for LLMs and text data. Ludwig, Mullainathan, and Rambachan (2024) develop an applied econometric framework for LLMs, addressing estimation and inference challenges that arise when LLM outputs are used in empirical research. Egami, Hinck, Stewart, and Wei (2024) propose design-based supervised learning (DSL) for social science applications, treating LLM annotations as surrogates for downstream inference. Gentzkow, Kelly, and Taddy (2019) provide a comprehensive overview of methods for using text as data in economics, establishing the methodological foundations for quantitative analysis of unstructured text. Our work complements these contributions by providing a unified bias-correction framework specifically designed for the moment-based estimation problems that are central to structural econometrics.

Double/debiased machine learning. Our cross-fitting procedure and debiasing approach share intellectual roots with the double/debiased machine learning (DML) framework of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), which uses sample splitting and Neyman-orthogonal moment conditions to ensure valid inference when nuisance parameters are estimated with machine learning methods. While DML addresses bias from estimating nuisance functions in the moment condition, our framework addresses bias from using LLM-generated proxies for latent features. The cross-fitting structure in Gen-PI serves an analogous purpose: it ensures independence between the LLM pipeline tuning and the debiasing step, preventing overfitting from contaminating the inferential guarantees.

Measurement error and errors-in-variables. The classical errors-in-variables literature (Fuller, 2009; Carroll, Ruppert, Stefanski, and Crainiceanu, 2006) studies estimation when covariates are observed with noise. While there are superficial similarities, our setting differs in important ways. First, the "measurement error" introduced by an LLM is generally not classical (additive, mean-zero, independent of the true value); it can exhibit complex, content-dependent bias patterns. Second, the calibration set in our framework provides direct observations of the true feature, rather than repeated measurements or instrumental variables. Third, our framework operates at the level of moment conditions rather than the covariate level, which allows for more targeted and efficient bias correction.

1.5 Outline

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 problem setup, the debiased moment estimator, and the cross-fitting procedure. Section 4 establishes the asymptotic properties of the estimator, including consistency, asymptotic normality, and the form of the asymptotic variance. Section 5 derives the optimal calibration-set design and model–prompt weights, and presents the alternating optimization algorithm for their joint computation. Section 6 provides practical guidance for calibration-set design, prompt tuning, retrieval-augmented generation, and auditing generator drift. Section 7 demonstrates the performance of our method through simulation studies. Section 8 concludes. The Appendix contains technical proofs.

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 $\tilde{X}_i = (X_i^\top, T_i)^\top$ and $\theta_0 = (\gamma_0^\top, \beta_0)^\top$. The parameter θ_0 is defined as the unique solution to the moment condition

$$\mathbb{E}[\tilde{X}_i(\tilde{X}_i^\top \theta_0 - Y_i)] = 0.$$

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 hallucination bias and

look-ahead bias and stochastic variation:

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

where $b(S_i)$ captures the systematic bias of the LLM and ϵ_i is mean-zero noise conditional on (Z_i, T_i).

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. This treats the LLM output as an observed covariate and solves

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

where $\hat{X}_i = (X_i^\top, \tilde{T}_i)^\top$. While this approach uses all the data, it ignores the bias $b(S_i)$. Because the bias is generally correlated with both T_i and Y_i, the resulting estimator is inconsistent: it converges to a limit that differs from θ_0, and the discrepancy does not vanish as $N \to \infty$.

Calibration-only regression. If a subset 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:

$$\frac{1}{n}\sum_{i \in S_0} \tilde{X}_i(\tilde{X}_i^\top \theta - Y_i) = 0,$$

where S_0 is the set of labeled units. 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.

Diagnostic validation. A common middle ground 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. 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 moment conditions that define θ_0.

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 the LLM introduces in the moment condition. 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:

$$\text{Bias} = \mathbb{E}[g(Z_i, \tilde{T}_i; \theta) - g(Z_i, T_i; \theta)] \approx \frac{1}{n} \sum_{i \in S_0} \big(g(Z_i, \tilde{T}_i; \theta) - g(Z_i, T_i; \theta) \big),$$

where $g(Z_i, T_i; \theta) = \tilde{X}_i(\tilde{X}_i^\top \theta - Y_i)$ is the moment function. Subtracting this estimated bias from the full-sample LLM moment yields a debiased estimator.

Importantly, the labeled subset S_0 need not be 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 relative to uniform random labeling.

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

$$\begin{aligned} G_N^\lambda(\theta) = \frac{1}{n} \sum_{i \in S_0} g(Z_i, T_i; \theta) \\ + \lambda \cdot \frac{N-n}{N} \left(\frac{1}{N-n} \sum_{i \notin S_0} g(Z_i, \tilde{T}_i; \theta) - \frac{1}{n} \sum_{i \in S_0} g(Z_i, \tilde{T}_i; \theta) \right). \end{aligned} \tag{1}$$

When $\lambda = 0$, we recover the calibration-only estimator. When $\lambda = 1$ and the LLM is perfect ($\tilde{T}_i = T_i$), we recover the oracle estimator that uses all N true features. The optimal λ^* balances the efficiency gain from incorporating more data against the additional noise from the LLM's imperfections. If the LLM is highly accurate, λ^* is close to 1 and the estimator achieves efficiency close to the oracle. If the LLM is unreliable, λ^* shrinks toward 0, protecting against hallucination bias and look-ahead bias.

This illustrative example captures the core mechanism of Gen-PI. Section 3 develops the general framework.

3 Weighted Debiased Estimation Framework

We now formalize the general Gen-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

Suppose we observe N i.i.d. samples $Z_i = (Y_i, X_i, S_i)$, where Y_i denotes an outcome variable, X_i is a vector of observed covariates, and S_i is an unstructured input (text, image, audio). Our goal is to extract a 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 $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, defined as the unique solution to the population moment condition

$$\mathbb{E}\big[g(Z_i, T_i; \theta_0)\big] = 0. \tag{2}$$

Here, $g(\cdot)$ is a user-specified moment function, and we focus on the just-identified case where $g(\cdot) \in \mathbb{R}^d$. 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.

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 a generative AI model and a specific prompt designed to extract the feature of interest from the

unstructured input. 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 GenAI features are given by

$$\tilde{T}_{i,m,s} = T_i + b_m(S_i) + \epsilon_{m,s}(S_i), \tag{3}$$

where $b_m(S_i)$ captures the systematic bias specific to model–prompt pair m that depends on the underlying unstructured input S_i, and $\epsilon_{m,s}(S_i)$ represents seed-specific noise. We assume $\mathbb{E}[\epsilon_{m,s}(S_i) \mid Z_i] = 0$, so that, conditional on the data, the GenAI features are centered around the biased target $T_i + b_m(S_i)$.

This decomposition is central to our framework. The systematic bias $b_m(S_i)$ captures hallucination bias, look-ahead bias, and any other reproducible distortion specific to model–prompt pair m. It is generally not mean-zero, can depend on the input in complex ways, and varies across model–prompt pairs. The seed noise $\epsilon_{m,s}(S_i)$ captures the stochastic variation across different runs of the same model–prompt pair and can be reduced by averaging across seeds.

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

$$\tilde{T}_{i,m} \coloneqq \frac{1}{S} \sum_{s=1}^{S} \tilde{T}_{i,m,s},$$

which averages out the seed-level noise while retaining the systematic bias. We compute $\tilde{T}_{i,m}$ for all units and all model–prompt pairs.

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 $S_0 = \{i : D_i = 1\}$ with $|S_0| = n$.

We allow the labeling process to be non-random. Let W_i denote the information available for all units prior to labeling selection, which may include the covariates 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

$$\mathbb{P}(D_i = 1 \mid W_i) = \pi(W_i) \in (0, 1], \tag{4}$$

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

budget of size n by requiring $\mathbb{E}[\pi(W_i)] = n/N =: r_N$, which ensures that the expected number of labeled units equals the labeling budget. We consider an asymptotic regime where $n, N \to \infty$ and $r_N \to r \in [0, 1]$.

This framework encompasses several important special cases. When $\pi(W_i) \equiv r_N$, the calibration set is selected uniformly at random (MCAR). 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.

3.4 Weighted Debiased Moment Estimator

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 (IPW) 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. For each model–prompt pair m, we add the full-sample moment based on LLM-generated features and subtract its IPW counterpart computed over the calibration set. This difference removes the systematic bias introduced by the GenAI features while retaining the efficiency gains from using all N observations. The weights λ then determine how different GenAI sources are combined. Let $\hat{\theta}_{\lambda,\pi}$ denote the estimator defined as the solution to $G_{\lambda,\pi}(\theta) = 0$.

To verify unbiasedness, note that by the labeling assumption $\mathbb{P}(D_i = 1 \mid W_i) = \pi(W_i)$, we have $\mathbb{E}[D_i/\pi_i \mid Z_i, T_i, W_i] = 1$, 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 λ and the labeling design π, confirming that the estimator targets the correct parameter.

3.5 Cross-Fitting for LLM Pipeline Tuning

In practice, the choice of prompt, the retrieval-augmented generation (RAG) configuration, 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 the highest agreement rate, or fine-tune a retrieval system using the labeled data as training examples. When such tuning occurs, the LLM-generated features for the tuning fold are no longer independent of the calibration labels, which can invalidate the debiasing.

We address this through a cross-fitting procedure. Specifically, we partition the indices $\{1, \ldots, N\}$ into K roughly equal folds $\mathcal{F}_1, \ldots, \mathcal{F}_K$, ensuring that each fold contains a comparable number of labeled observations. For each fold k:

1. Use the calibration labels from all folds except k to tune the LLM pipeline (prompt selection, RAG configuration, model choice).
2. Generate features $\tilde{T}_{i,m}^{(-k)}$ for units in fold k using the tuned pipeline.
3. Compute the fold-specific debiased moment using the calibration labels and generated features in fold k.

The cross-fitted estimator aggregates across folds:

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

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 the generated features used in each fold's moment computation are independent of the calibration labels used for debiasing within that fold.

Under a stability condition, which requires that the generated features $\tilde{T}_{i,m}^{(-k)}$ do not change discontinuously as the tuning fold varies, the cross-fitted estimator preserves the asymptotic properties established in Section 4.

4 Inferential Theory

We now establish the asymptotic properties of the Gen-PI estimator. We begin by formalizing the assumptions on the labeling mechanism and the moment conditions.

4.1 Assumptions

Assumption 1 (Human labeling)**.**

1. *The labeling selection is conditionally independent of other variables given W_i:*

$$\mathbb{P}(D_i = 1 \mid Z_i, T_i, W_i) = \mathbb{P}(D_i = 1 \mid W_i).$$

2. *Define the scaled propensity $\rho(W) = \pi(W)/r_N$. There exists a constant $\rho_{\min} > 0$ such that for all N, $\rho(W_i) \geq \rho_{\min}$ a.s.*

Assumption 1.1 requires that, conditional on the observable information W_i, the labeling decision does not depend on the unobservable true feature T_i or the structural error. This is satisfied when the labeling selection is based on observable characteristics such as the LLM-generated features, model disagreement, or covariates. 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 $\theta \in \Theta$, $\mathbb{E}\|g(Z_i, T_i; \theta)\|^2 < \infty$ and $\mathbb{E}\|g(Z_i, \tilde{T}_{i,m}; \theta)\|^2 < \infty$ 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 $J_{\theta_0} := \nabla_\theta G(\theta_0) \in \mathbb{R}^{d \times d}$.*
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,*

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

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

Assumption 2 collects standard conditions for moment-based estimation. Assumption 2.1 ensures the applicability of central limit theorems. Assumption 2.2 guarantees local identification of θ_0. Assumptions 2.3 imposes smoothness conditions needed for linearization and uniform convergence.

4.2 Consistency

Theorem 1 (Consistency)**.** *Suppose Assumptions 1–2 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 Assumption 2.3 holds globally on* Θ*, or*

(ii) $G_{\lambda,\pi}(\theta)$ *is monotone in the sense that* $(G_{\lambda,\pi}(\theta_1) - G_{\lambda,\pi}(\theta_2))^\top(\theta_1 - \theta_2) \geq 0$ *for all* $\theta_1, \theta_2 \in \Theta$*,*

then $\hat{\theta}_{\lambda,\pi} \xrightarrow{p} \theta_0$*.*

Theorem 1 shows that our estimator is consistent under the non-random labeling schemes and potentially vanishing labeling rates. The proof proceeds by showing that $G_{\lambda,\pi}(\theta)$ converges uniformly to $G(\theta) = \mathbb{E}[g(Z_i, T_i; \theta)]$ and then applying standard arguments for Z-estimators. Building on this result, we derive its asymptotic distribution in Section 4.3.

4.3 Asymptotic Normality

Theorem 2 (Asymptotic normality)**.** *Let* $r_N = n/N$ *and* $\rho(W_i) = \pi(W_i)/r_N$*. Suppose Assumptions 1–2 hold, and* $n \to \infty$*,* $r_N \to r \in [0, 1]$ *as* $N \to \infty$*. Then, for any fixed* $\lambda = (\lambda_1, \ldots, \lambda_M)$*,*

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

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

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

and

$$R_\lambda(W_i) = \mathbb{E}\left[\left(g(Z_i, T_i; \theta_0) - \sum_{m=1}^{M} \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)\right)\left(g(Z_i, T_i; \theta_0) - \sum_{m=1}^{M} \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0)\right)^\top \,\middle|\, W_i\right].$$

The asymptotic variance $J_{\theta_0}^{-1}\Omega_{\lambda,\pi}J_{\theta_0}^{-\top}$ has a transparent decomposition. The first term, $r \cdot J_{\theta_0}^{-1}\mathrm{Var}(g(Z_i, T_i; \theta_0))J_{\theta_0}^{-\top}$, is the irreducible variance from the labeled sample. When $r > 0$, this component is present regardless of the quality of the GenAI features. The second term is the design- and model-dependent penalty. It depends on the labeling design through $\rho(W_i)$ and on the quality of the GenAI ensemble through the conditional residual covariance $R_\lambda(W_i)$, which measures how well the weighted combination of GenAI moments approximates the true moment, conditional on the observable information W_i.

When the GenAI ensemble perfectly replicates the true moment—that is, $\sum_m \lambda_m g(Z_i, \tilde{T}_{i,m}; \theta_0) = g(Z_i, T_i; \theta_0)$ almost surely—the residual $R_\lambda(W_i) = 0$ and the second term vanishes. In this case, the

asymptotic variance equals the oracle variance from using all N labeled features. Conversely, when the GenAI features are uninformative, $R_\lambda(W_i)$ is large, and the asymptotic variance is dominated by the calibration-set size.

Remark 1 (Confidence intervals)**.** *Theorem 2 directly yields asymptotically valid confidence intervals for the target parameter. For a scalar component $\theta_{0,j}$, a $(1-\alpha)$ confidence interval is*

$$\hat{\theta}_j \pm z_{\alpha/2} \cdot \hat{\sigma}_j / \sqrt{n},$$

where $\hat{\sigma}_j^2$ is the j-th diagonal element of the estimated asymptotic variance $\hat{J}^{-1}\hat{\Omega}_{\lambda,\pi}\hat{J}^{-\top}$, with sample analogues replacing population quantities.

5 Optimization

The asymptotic variance in Theorem 2 depends on two design choices: the labeling policy $\pi(\cdot)$ and the model–prompt weights λ. In this section, we derive the optimal choices for both.

5.1 Optimal Calibration-Set Design

We first study the problem of choosing which observations to label by selecting the propensity function $\pi(\cdot)$ for a given weight vector λ. 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

$$V(\lambda, \pi) \coloneqq \operatorname{tr}\big(J_{\theta_0}^{-1}\Omega_{\lambda,\pi}J_{\theta_0}^{-\top}\big)$$

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

Assumption 3 (Design feasibility)**.** *For any fixed λ, $\mathbb{P}\big(\operatorname{tr}(J_{\theta_0}^{-1}R_\lambda(W_i)J_{\theta_0}^{-\top}) > 0\big) > r_N$.*

Assumption 3 assumes that the GenAI-aggregated moment is imperfect on at least a fraction r_N of the population. This ensures that the labeling budget is economically meaningful.

Theorem 3 (Optimal labeling propensity)**.** *Under Assumption 3, the problem*

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

admits an oracle solution of the form

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

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

Theorem 3 shows that, when neither truncation binds, the optimal labeling probability 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 GenAI-aggregated moment and the true moment at the point w. Regions where the generated features 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.

5.2 Optimal Model–Prompt Weights

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

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

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 3 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. In applications, one may restrict to a convex ensemble

$\lambda \in \{\lambda \geq 0,\ \sum_m \lambda_m = 1\}$ to stabilize finite-sample behavior. Additional regularization, such as ridge or entropy penalties, can further prevent extreme weights.

Proposition 1 (First-order condition for λ^*)**.** *Assume: (i) $\lambda \mapsto \mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W) J_{\theta_0}^{-\top})$ is differentiable for a.e. W and $\mathbb{E}[\sup_{\lambda\in\Lambda} \|\nabla_\lambda \mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W_i) J_{\theta_0}^{-\top})\|] < \infty$ for a neighborhood Λ of λ^*; (ii) for each $\lambda \in \Lambda$, the inner problem admits an optimizer $\pi^*(\cdot \mid \lambda)$ satisfying the KKT conditions; (iii) $\mathbb{P}(c(\lambda)\mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W_i) J_{\theta_0}^{-\top}) = 1) = 0$ for λ in a neighborhood of λ^*. Then any (local) minimizer λ^* of $\widetilde{V}(\lambda)$ 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 $\tilde{g}_i \coloneqq [g(Z_i, \tilde{T}_{i,1}; \theta_0), \ldots, g(Z_i, \tilde{T}_{i,M}; \theta_0)]$ and c_{λ^*} is determined by the budget $\mathbb{E}[\pi^*(W_i \mid \lambda^*)] = r_N$.

In Section 5.3, we propose an alternating algorithm to compute (λ^*, π^*). For fixed π, the condition (11) coincides with the equations defining the λ-update in this algorithm. Therefore, the alternating optimization algorithm can be interpreted as seeking a solution of this first-order condition.

5.3 Joint Optimization via Alternating Algorithm

Sections 5.1 and 5.2 show that the optimal π depends on λ (through R_λ) and the optimal λ depends on π (through the weighting). We solve this joint optimization through the following alternating procedure:

1. **Initialize:** Set $\lambda^{(0)}$ to uniform weights (e.g., $\lambda_m^{(0)} = 1/M$) and run a pilot audit with MCAR labeling.
2. 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 the quadratic program (10).

In large samples, each step decreases (or leaves unchanged) the empirical approximation to the joint criterion. In finite samples, we recommend cross-fitting: estimate λ and $\mathrm{tr}(J_{\theta_0}^{-1} R_{\lambda^{(k)}}(W_i) J_{\theta_0}^{-\top})$ on one fold of the calibration data and plug them into the estimator on another fold.

5.4 Feasible Adaptive Calibration Design

The conditional residual $R_\lambda(W)$ involves the unobserved true moment $g(Z_i, T_i; \theta_0)$. We propose a feasible two-stage procedure:

1. **Pilot stage:** Draw an initial audit sample of size n_0 using a simple design (e.g., MCAR). Observe the true features T_i for audited units.
2. **Score estimation:** Using the pilot labels and covariates W_i, estimate the conditional residual score $s_\lambda(W) = (\mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W) J_{\theta_0}^{-\top}))^{1/2}$ via regression or machine learning. Denote the fitted score by $\hat{s}_\lambda(W)$.
3. **Adaptive labeling:** Set the propensity

$$\hat{\pi}(W) \coloneqq \max\{\rho_{\min} r_N, \min\{1,\ \hat{c}\, \hat{s}_\lambda(W)\}\},$$

   with $\hat{c}$ chosen to satisfy the remaining budget constraint. Sample additional audits using $\hat{\pi}(\cdot)$.
4. **Estimation:** Compute $\hat{\theta}_{\lambda,\pi}$ using the known design propensities π_i.

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

Proposition 2 (Asymptotic optimality of the plug-in design)**.** *Suppose: (i) there exists a measurable oracle score $s_\lambda(W)$ satisfying Assumption 3; (ii) the fitted score satisfies $\mathbb{E}[|\hat{s}_\lambda(W) - s_\lambda(W)|] \to 0$ in probability; (iii) $\hat{c} \to c$ in probability where c solves the budget constraint; and (iv) the design uses $\hat{\pi}(W) = \max\{\rho_{\min} r_N, \min\{1, \hat{c}\, \hat{s}_\lambda(W)\}\}$. Then the design-dependent penalty satisfies*

$$\mathbb{E}\left[\frac{1 - \hat{\pi}(W)}{\hat{\pi}(W)}\, s_\lambda(W)^2\right] \to \mathbb{E}\left[\frac{1 - \pi^*(W)}{\pi^*(W)}\, s_\lambda(W)^2\right],$$

so the plug-in design attains the oracle trace criterion asymptotically.

The proof follows by dominated convergence and consistency of $\hat{c}$ and $\hat{s}_\lambda(w)$. If $\hat{s}_\lambda(\cdot)$ is estimated flexibly (e.g., via random forests or neural networks), cross-fitting can be used to avoid overfitting bias in the score estimation step.

6 Practical Guidance

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

6.1 Calibration-Set Design

The calibration set is the cornerstone of Gen-PI. Its size and composition directly affect the quality of the bias correction and the efficiency of the final estimator. We offer the following guidance.

Budget allocation. The labeling budget n should be chosen to balance cost against statistical precision. Theorem 2 shows that the asymptotic variance scales as $O(1/n)$, so doubling the calibration set halves the variance contribution from the labeled sample. However, the efficiency gain from incorporating GenAI features also depends on n through the debiasing terms. In practice, even a modest calibration set (5–10% of the total sample) can yield substantial improvements over the naive LLM approach.

Targeted labeling. Theorem 3 shows that the optimal labeling design concentrates labels on observations where the GenAI features are least reliable. In practice, researchers can use model disagreement (the standard deviation of outputs across model–prompt pairs or seeds) as a proxy for the residual score $s_\lambda(W)$. Units with high disagreement should be prioritized for human labeling.

Temporal and cross-sectional stratification. When the data have a panel structure, it is often useful to stratify the calibration set across time periods and cross-sectional units to ensure representativeness. If the LLM's performance varies over time (e.g., due to changes in the input distribution), temporal stratification ensures that the bias correction captures these dynamics.

6.2 Prompt Tuning and Model Selection

The quality of the GenAI features depends critically on the prompt and model choice. Gen-PI accommodates prompt tuning and model selection through the cross-fitting procedure described in Section 3.5. Researchers can:

- Compare multiple prompts (e.g., zero-shot, few-shot, chain-of-thought) on the calibration set and select the one that minimizes the estimated asymptotic variance.
- Use the calibration set to fine-tune a retrieval-augmented generation pipeline, where relevant documents are retrieved and included in the prompt context.
- Combine outputs from multiple models and prompts through the adaptive weighting scheme, rather than selecting a single best configuration.

The cross-fitting procedure ensures that the resulting inference is valid even when the pipeline is tuned on the calibration data.

6.3 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) can improve the quality of GenAI features by providing the model with relevant context at inference time. For example, when extracting features from earnings calls, RAG can retrieve prior earnings reports, analyst forecasts, or industry benchmarks and include them in the prompt. This additional context can reduce hallucination and improve feature accuracy.

Within Gen-PI, RAG is accommodated as part of the model–prompt pair. Different RAG configurations (e.g., different retrieval strategies, different numbers of retrieved documents) can be treated as different model–prompt pairs and combined through adaptive weighting. If the RAG configuration is tuned using the calibration set, cross-fitting ensures validity.

6.4 Auditing Generator Drift

In practice, LLM outputs can change over time due to model updates, API changes, or shifts in the input distribution. Gen-PI provides a natural framework for detecting and accommodating such drift. By periodically collecting new calibration labels and comparing the estimated bias with historical values, researchers can detect when the generator's performance has degraded. The adaptive weighting scheme automatically accommodates drift by downweighting model–prompt pairs whose bias has increased.

Formally, if the bias function $b_m(S_i)$ changes over time, the cross-fitted estimator with a rolling calibration window adapts to the current bias level. We recommend maintaining a rolling calibration set that is refreshed at regular intervals to ensure that the bias correction remains current.

7 Simulation Study

We conduct extensive simulations to evaluate the finite-sample performance of Gen-PI under different bias designs and choices of labeling information.

7.1 Data Generating Process

Let $U_i \sim \mathcal{N}(0,1)$ denote a latent semantic factor extracted from text. We generate the true feature as $T_i = U_i$ and the outcome according to the linear model

$$Y_i = \alpha_0 + \beta_0 T_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,1),$$

with $(\alpha_0, \beta_0) = (0.5, 1.5)$. 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 $g(Z_i, T_i; \theta) = X_i(Y_i - X_i^\top \theta)$ with $X_i = (1, T_i)^\top$.

GenAI feature generation. We simulate $M = 3$ model–prompt pairs. For each model m and seed $s = 1, \ldots, S$,

$$\tilde{T}_{i,m,s} = T_i + b_m(U_i) + \epsilon_{i,m,s}, \qquad \epsilon_{i,m,s} \sim \mathcal{N}(0, \sigma^2_{\text{seed}}),$$

and we use the seed average $\tilde{T}_{i,m} = S^{-1} \sum_{s=1}^{S} \tilde{T}_{i,m,s}$.

We consider three bias designs that capture different patterns of LLM hallucination:

- **Smooth:** $b_1(U) = \delta_0$, $b_2(U) = \delta U$, $b_3(U) = \delta \tanh(U)$. This represents a setting where hallucination bias varies smoothly across the feature space, with some models exhibiting constant bias and others exhibiting bias proportional to the true feature.
- **Opposite:** $b_1(U) = \delta_0$, $b_2(U) = 1.5\delta U$, $b_3(U) = -\delta U$. This represents a setting where different model–prompt pairs exhibit biases in opposite directions, so that adaptive weighting can partially cancel the biases.
- **Tail spike:** With $M(U) = \mathbf{1}\{|U| > 2.5\}$: $b_1(U) = \delta_0 + \delta M(U)|U|$, $b_2(U) = \delta M(U)|U|$, $b_3(U) = \delta M(U)(U^2 - 2)$. This represents a setting where the LLM is accurate for typical inputs but hallucinates badly in the tails of the distribution—a pattern commonly observed in practice.

We set $\delta_0 = 0.5$ and $\delta = 1$.

7.2 Estimators and Labeling Designs

Before labeling, we observe a scalar summary W_i that can be used to guide the labeling decision. We consider:

- **Outcome level:** $W_i = Y_i$. A feasible proxy available in practice.
- **Model disagreement:** $W_i = \text{std}(\tilde{T}_{i,m})$. Measures cross-model variation.

- **Oracle index:** $W_i = U_i$. The ideal (infeasible) information set.

We compare the following five estimators:

1. **Naive estimator:** Uses $\tilde{T}_{i,m}$ with no labeling correction.
2. **Uniform π and λ:** Debiased estimator with MCAR labeling and equal weights.
3. **Optimal π, uniform λ:** Optimizes the labeling design with equal model weights.
4. **Uniform π, optimal λ:** MCAR labeling with optimized model weights.
5. **Jointly optimized (λ^*, π^*):** Full optimization over both weights and design.

Our primary metric is the Monte Carlo mean squared error (MSE) of the slope estimator, i.e., $\text{MSE}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta_0)^2]$. We set $N = 3000$, $r_N = 0.10$, and run 200 Monte Carlo replications.

7.3 Results

Figures 1–3 present the results across the three bias designs. We highlight four main findings.

First, the naive estimator exhibits substantial bias across all designs. Its MSE can exceed the debiased estimators by an order of magnitude, particularly under the smooth and tail-spike designs where the bias is large relative to the noise.

Second, even the simplest debiased estimator (uniform π and λ) dramatically improves over the naive approach, demonstrating that the core debiasing mechanism is effective. This validates the key insight that a small calibration set can correct for the bias in the moment conditions.

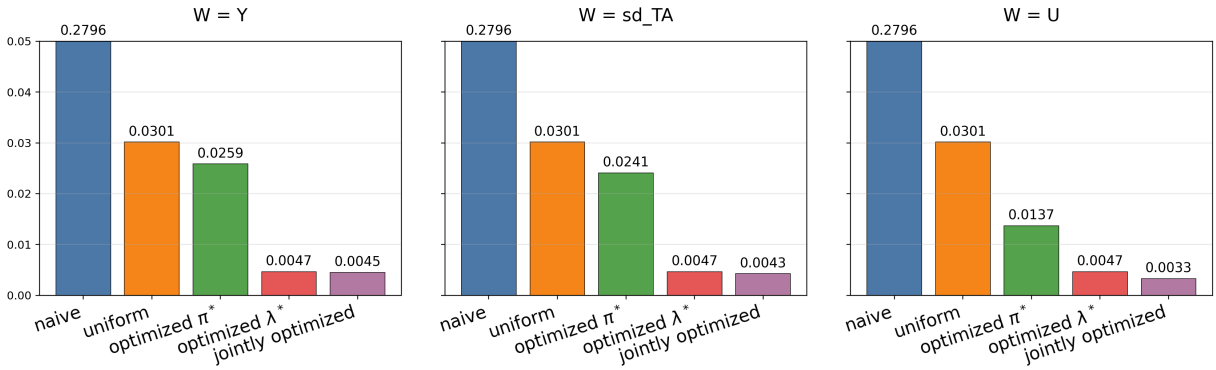
Third, optimizing λ (the model–prompt weights) typically provides larger gains than optimizing π (the labeling design) in our setting. This is because the weights directly control how the generated ensemble approximates the true moment, whereas the labeling design affects only how the debiasing terms are estimated. Under the "opposite" bias design, weight optimization is particularly valuable because it can exploit the opposing bias directions of different models.

Fourth, the jointly optimized estimator consistently achieves the lowest MSE across all designs and labeling information sets. Under the tail-spike design with oracle information $W = U$, the joint optimization targets labels precisely in the tail region where the LLM hallucinates, producing efficiency gains of 40–60% relative to the uniform debiased estimator.

7.4 Distinct Roles of π and λ: Single-Model Interpretation

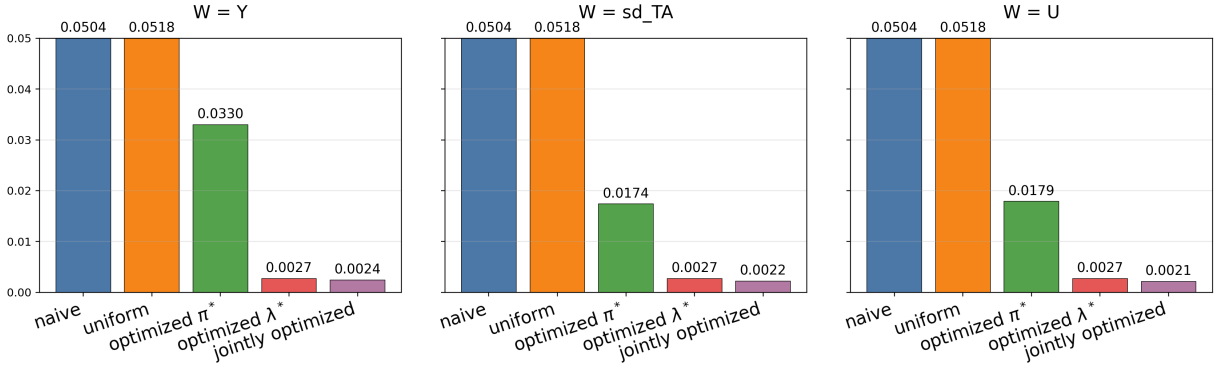
To make the distinct roles of λ and π transparent, we specialize to the case of a single model–prompt pair ($M = 1$). In this setting, λ is no longer an ensemble weight across prompts; instead,

Figure 1: MSE of the slope estimator under the smooth bias design



This figure reports the Monte Carlo mean squared error (MSE) of the slope estimator $\hat{\beta}$ under the smooth bias design: $b_1(U) = \delta_0$, $b_2(U) = \delta U$, $b_3(U) = \delta \tanh(U)$ with $\delta = 1$. The three panels correspond to different choices of labeling information W_i: outcome level ($W_i = Y_i$), model disagreement ($W_i = \text{std}(\tilde{T}_{i,m})$), and oracle index ($W_i = U_i$). Bars show the MSE of five estimators: naive GenAI regression (no bias correction), debiased with uniform π and λ, optimized π with uniform λ, uniform π with optimized λ, and jointly optimized (λ^*, π^*). Values exceeding the y-axis bound $[0, 0.05]$ are labeled above the panel. The data generating process uses $U_i \sim \mathcal{N}(0,1)$, $T_i = U_i$, $Y_i = 0.5 + 1.5T_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0,1)$, $N = 3000$, $r_N = 0.10$, $M = 3$ model–prompt pairs, and 200 Monte Carlo replications.

Figure 2: MSE of the slope estimator under the opposite bias design

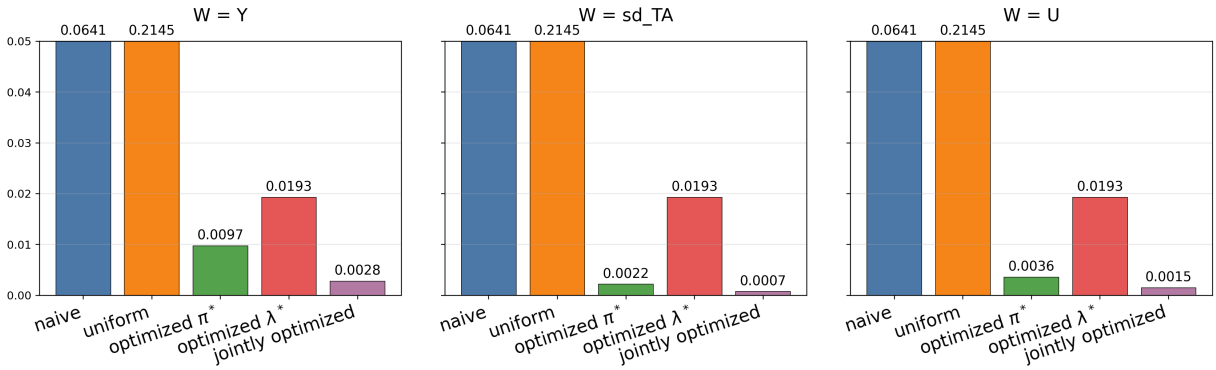


This figure reports the Monte Carlo MSE of $\hat{\beta}$ under the opposite bias design: $b_1(U) = \delta_0$, $b_2(U) = 1.5\delta U$, $b_3(U) = -\delta U$ with $\delta = 1$. The opposing bias directions of models 2 and 3 create an opportunity for adaptive weighting to partially cancel the biases, which is reflected in the large efficiency gains from optimizing λ. All other simulation parameters are identical to Figure 1: $N = 3000$, $r_N = 0.10$, $M = 3$, and 200 replications. The three panels correspond to labeling information $W_i \in \{Y_i, \text{std}(\tilde{T}_{i,m}), U_i\}$.

it governs the overall contribution of the single generated proxy.

Specifically, as shown in the asymptotic variance in Theorem 2, the weight λ acts as a *global* shrinkage parameter: it determines how strongly the single GenAI proxy is used overall. It scales the GenAI moment so that it approximates the human moment as well as possible on average. When

Figure 3: MSE of the slope estimator under the tail-spike bias design



This figure reports the Monte Carlo MSE of $\hat{\beta}$ under the tail-spike bias design, where $M(U) = \mathbf{1}\{|U| > 2.5\}$ and $b_1(U) = \delta_0 + \delta M(U)|U|$, $b_2(U) = \delta M(U)|U|$, $b_3(U) = \delta M(U)(U^2 - 2)$ with $\delta = 1$. The GenAI models are accurate for typical inputs ($|U| \leq 2.5$) but hallucinate badly in the tails of the distribution. Targeted labeling with the oracle index $W_i = U_i$ can directly identify the tail region and concentrate labels there, yielding large efficiency gains. All other parameters are as in Figure 1.

the LLM is highly accurate, λ^* is close to 1; when it is unreliable, λ^* shrinks toward 0. In contrast, the policy $\pi(\cdot)$ is a *local* allocation rule: it does not change the GenAI approximation itself, but decides where to spend the human-labeling budget. To minimize the asymptotic variance, it assigns larger labeling probability to those regions of W_i where the conditional residual is large.

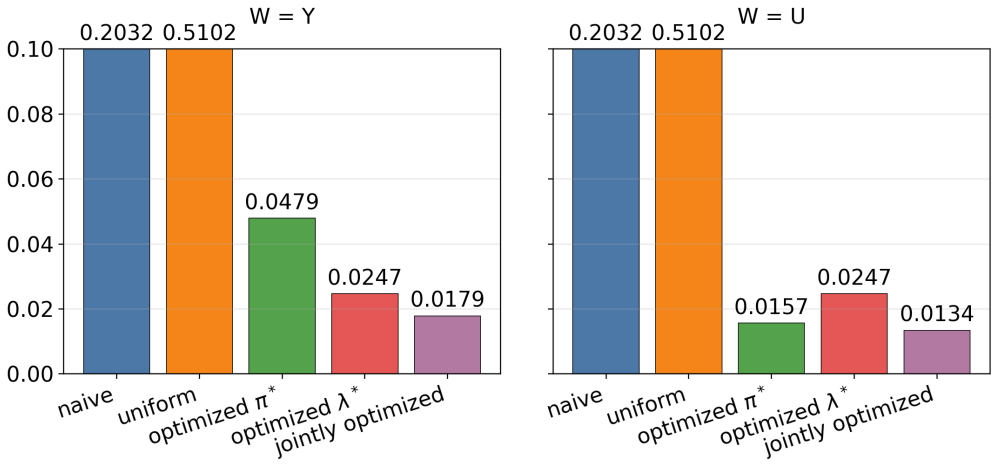
We now conduct a simulation that illustrates these distinct roles in finite samples, using $M = 1$ under the tail-spike bias design. Figure 4 shows the MSE results. Optimizing λ alone (with uniform π) provides a global adjustment: it scales the GenAI contribution to minimize the average approximation error. Optimizing π alone (with $\lambda = 1$) provides a local adjustment: it concentrates labels where the GenAI proxy is least reliable without changing the proxy itself. Joint optimization combines both mechanisms and achieves the lowest MSE.

Figure 5 displays the audit propensity curves. Under the uniform design, $\pi(W)$ is flat. When π is optimized with oracle information $W = U$, the propensity spikes in the tail region $|U| > 2.5$ where the bias is concentrated, precisely targeting the observations where human labels are most valuable.

8 Conclusion

This paper proposes Generative-AI Powered Inference (Gen-PI), a general framework for valid estimation and inference when covariates are produced by generative AI models. The framework

Figure 4: Slope MSE under tail-spike bias with a single model–prompt pair



This figure reports the Monte Carlo MSE of $\hat{\beta}$ under the tail-spike bias design with $M = 1$ model–prompt pair, $S = 1$ seed, and audit rate $r_N = 0.05$. The DGP uses $(\alpha_0, \beta_0) = (0.5, 2.5)$ with $b(U) = \mathbf{1}\{|U| > 2.5\}\, U$. Two panels correspond to $W_i = Y_i$ (feasible) and $W_i = U_i$ (oracle). Five estimators are compared: naive GenAI (no correction), uniform design ($\lambda = 1$, $\pi \equiv r_N$), π-only optimization ($\lambda = 1$, π^*), λ-only optimization ($\pi \equiv r_N$, λ^*), and joint optimization (λ^*, π^*). This design isolates the contributions of λ (global shrinkage) and π (local allocation), as discussed in Section 7. $N = 3000$, 200 Monte Carlo replications.

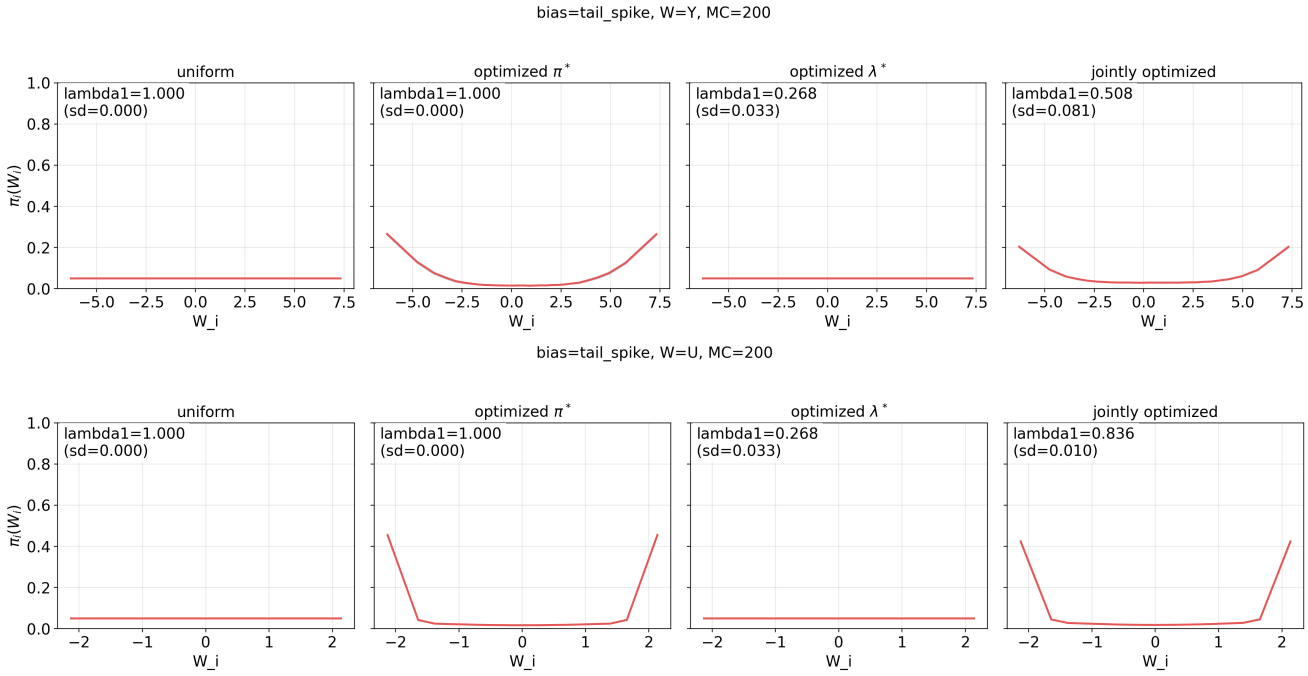
treats LLM outputs as imputations, constructs bias-correction terms using a calibration set of human-labeled observations, and combines multiple model–prompt pairs through adaptive weighting. We establish consistency and asymptotic normality under general conditions, derive the optimal calibration-set design and model–prompt weights, and provide practical guidance for implementing the framework in empirical research.

Gen-PI delivers a precise legitimacy criterion for the use of synthetic data in structural statistics: valid inference is recoverable if and only if a gold-standard calibration set is available to rectify hallucination bias. This criterion provides clear guidance for the rapidly growing empirical literature that uses LLM-generated features. Researchers should allocate resources to collect calibration labels and use a formal bias-correction framework to ensure the validity of their inferences.

Our simulation results demonstrate that even a small calibration set (5–10% of the sample) suffices to eliminate hallucination bias and produce valid confidence intervals. The adaptive weighting scheme automatically adjusts the reliance on GenAI features based on their quality: when the generator is accurate, the estimator achieves efficiency close to the oracle; when the generator is unreliable, the estimator gracefully degrades to the calibration-only estimator.

Several directions for future research emerge from our framework. First, the extension to settings

Figure 5: MC-averaged audit propensity curves $w \mapsto \pi(w)$ under tail-spike bias



This figure displays the Monte Carlo averaged audit propensity curves as a function of the labeling information W_i. Top row: $W_i = Y_i$ (feasible proxy). Bottom row: $W_i = U_i$ (oracle index). In each row, the four panels from left to right correspond to: (1) uniform design ($\pi \equiv r_N$), (2) π-only optimization with $\lambda = 1$, (3) λ-only optimization with uniform π, and (4) joint optimization of (λ^*, π^*). Under the oracle index, the optimized π spikes in the tail region $|U| > 2.5$ where the hallucination bias is concentrated, precisely targeting the observations where human labels are most valuable. With the feasible proxy $W_i = Y_i$, the propensity also increases for extreme values of Y, which are correlated with the tail region, though the targeting is less precise. Same simulation parameters as Figure 4.

where the unstructured data themselves are generated (rather than just the features extracted from them) raises new challenges for the bias structure. Second, the integration of Gen-PI with active learning methods could enable adaptive calibration-set construction that iteratively targets the most informative observations. Third, the application of our framework to specific domains—including clinical trials, policy evaluation, and financial regulation—will require domain-specific guidance on calibration-set design and LLM pipeline configuration.

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.

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*.

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.

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.

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.

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.

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*.

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.

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.

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*.

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

A Appendix: Proofs

A.1 Proof of Theorem 1

Proof. For notational simplicity, define

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

By definition,

$$G_{\lambda,\pi}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \psi_i(\theta), \qquad \psi_i(\theta) \coloneqq b_i(\theta) + \frac{D_i}{\pi_i}\big(g_i(\theta) - b_i(\theta)\big),$$

where $\pi_i = \pi(W_i)$.

By Assumption 1.1,

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

Therefore,

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

Similarly, for each m,

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

which implies that $\mathbb{E}[\psi_i(\theta)] = G(\theta)$ and $\mathbb{E}[G_{\lambda,\pi}(\theta)] = G(\theta)$ for every $\theta \in \Theta$.

Step 1 – We first prove the pointwise convergence of $G_{\lambda,\pi}(\theta)$ 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, and therefore

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

Using Assumption 2.1 again, $\mathbb{E}\|\psi_i(\theta)\|^2 = O(r_N^{-1})$, and hence,

$$\mathrm{Var}(G_{\lambda,\pi}(\theta)) = \frac{1}{N}\mathrm{Var}(\psi_i(\theta)) = O\left(\frac{1}{Nr_N}\right) = O\left(\frac{1}{n}\right) = o(1).$$

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

Step 2 – Next, we prove the uniform convergence of $G_{\lambda,\pi}(\theta)$ on compact subsets.

Let $K \subset \Theta$ be any compact subset contained in the neighborhood of θ_0 on which Assumption 2.3 holds. Then 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

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

Therefore

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

where

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

We now show $\mathbb{E}[L_i] < \infty$. 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 $\mathbb{E}[L_i] < \infty$.

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 $j(\theta)$ such that $\|\theta - \theta_{j(\theta)}\| \le \delta$. Then

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

Taking the supremum over $\theta \in K$,

$$\sup_{\theta \in K} \|G_{\lambda,\pi}(\theta) - G(\theta)\| \leq \max_{1 \leq j \leq J} \|G_{\lambda,\pi}(\theta_j) - G(\theta_j)\| + \delta \cdot \frac{1}{N} \sum_{i=1}^{N} L_i + \delta \cdot \mathbb{E}[L_i].$$

The first term is $o_p(1)$ by Step 1 and the fact that $J < \infty$. The second term is $O_p(1)$ by the law of large numbers. Because $\delta > 0$ is arbitrary, we obtain

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

In particular, if Θ is compact and Assumption 2.3 holds globally on Θ, then

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

Step 3 – We prove the consistency of the estimator $\hat{\theta}_{\lambda,\pi}$ under compactness.

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 $\hat{\theta}_{\lambda,\pi} \xrightarrow{p} \theta_0$.

Step 4 – Finally, we prove the consistency results for the monotonicity case.

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

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

is contained in the neighborhood from Assumption 2.3. By Step 2,

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

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

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

so that the moment estimator can be written as

$$G_{\lambda,\pi}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \psi_i(\theta), \qquad \psi_i(\theta) := b_i(\theta) + \frac{D_i}{\pi_i}\big(g_i(\theta) - b_i(\theta)\big),$$

As established in the proof of Theorem 1, the moment function satisfies $\mathbb{E}[\psi_i(\theta)] = G(\theta)$, and in particular, $\mathbb{E}[\psi_i(\theta_0)] = 0$.

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

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

By Assumption 2.2, G is differentiable at θ_0, so

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

By Assumption 2.3, for θ 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, $\pi_i = r_N \rho(W_i)$ with $\rho(W_i) \ge \rho_{\min} > 0$, so

$$\mathbb{E}[r_N L_i^2] \lesssim \mathbb{E}[L_{\theta,H}(Z_i)^2] + \sum_{m=1}^{M} \mathbb{E}[L_{\theta,m}(Z_i)^2] < \infty.$$

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

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

Therefore, by Lemma 19.24 of Van der Vaart (2000), 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,

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

Combining the previous steps gives

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

In the following, we show that $\sqrt{n}\,G_{\lambda,\pi}(\theta_0) = O_p(1)$. Since

$$\sqrt{n}\,G_{\lambda,\pi}(\theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sqrt{r_N}\,\psi_i(\theta_0).$$

By independence across i and $\mathbb{E}[\psi_i(\theta_0)] = 0$,

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

Since

$$\mathbb{E}\left[\left(\frac{D_i}{\pi_i}\right)^2\middle| W_i\right] = \frac{1}{\pi_i},$$

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

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

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

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

Finally, we apply the central limit theorem to $\sqrt{n}\,G_{\lambda,\pi}(\theta_0)$ and get

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

where $\Omega_{\lambda,\pi} = \lim_{N\to\infty} \mathbb{E}[r_N \psi_i(\theta_0)\psi_i(\theta_0)^\top]$.

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

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

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

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

Because $r_N \to r$, $\rho(W_i) \geq \rho_{\min} > 0$, and $\mathbb{E}\|R_\lambda(W_i)\| \leq \mathbb{E}\|g_i - b_i\|^2 < \infty$, we have

$$\Omega_{\lambda,\pi} = r\mathrm{Var}(g_i) + E\left[\left(\frac{1}{\rho(W_i)} - r\right) R_\lambda(W_i)\right].$$

Combining the asymptotic linear expansion with the CLT above gives

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

This proves the theorem. □

A.3 Proof of Theorem 3

Proof. For notational simplicity, define

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

and assume $s_\lambda(w) \geq 0$. 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 $\rho(W_i) \in [\rho_{\min}, r_N^{-1}]$ almost surely.

Let

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

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,

$$h(0) = \max\{\rho_{\min}, r_N^{-1}\} \leq 1.$$

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 3,

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

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 $\rho_N^* \in \mathcal{R}_N$.

It remains to prove optimality. Let $\mu_\lambda = a_\lambda^{-2}$ when $a_\lambda > 0$. For any $\rho \in \mathcal{R}_N$, pointwise optimality of $x_{\mu_\lambda}(w) = \rho_N^*(w \mid \lambda)$ 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.4 Proof of Proposition 1

Proof. Denote $s_\lambda(w) = \sqrt{\mathrm{tr}(J_{\theta_0}^{-1} R_\lambda(W) J_{\theta_0}^{-\top})}$. The minimization target is

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

Define the inner optimum value $\phi(\lambda) = \mathbb{E}[s_\lambda(W)^2 / \pi^*(W \mid \lambda)]$. For any fixed λ, introduce a scalar multiplier μ for the equality constraint and a nonnegative measurable multiplier $\eta(W) \geq 0$ for the inequality constraint $\pi(W) \leq 1$. The KKT stationary condition is (pointwise in W):

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

If $\pi^*(W \mid \lambda) < 1$, complementary slackness implies $\eta(W) = 0$, giving $\pi^*(W \mid \lambda) = s_\lambda(W)/\sqrt{\mu} = c_\lambda s_\lambda(W)$. If $\pi^*(W \mid \lambda) = 1$, then $\eta(W) = s_\lambda(W)^2 - \mu \geq 0$.

Consider the map $\lambda \mapsto \mathcal{L}(\pi^*(\cdot \mid \lambda); \mu^*(\lambda), \eta^*(\cdot \mid \lambda), \lambda)$. Differentiating with respect to λ, the terms involving $\partial/\partial\pi$, $\partial/\partial\mu$, and $\partial/\partial\eta$ vanish by stationarity, the budget constraint, and complementary slackness, respectively. Therefore

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

Since $\widetilde{V}_{\mathrm{prof}}(\lambda) = r_N \, \phi(\lambda) - r \, \mathbb{E}[s_\lambda(W)^2]$,

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

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

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

Setting $\nabla_\lambda \widetilde{V}_{\mathrm{prof}}(\lambda^*) = 0$ and using the law of iterated expectations yields the stated first-order condition (11). □