Bayesian LLM Finetuning for Statistical Inference

1 Introduction

Large language models (LLMs) have increasingly been employed in social science research. A common workflow involves three stages: (1) fine-tuning pre-trained LLMs on domain-specific datasets to improve performance on specialized tasks, (2) using the fine-tuned models to extract latent variables or classify text data, and (3) incorporating these outputs as outcomes or covariates in downstream statistical inference.

However, this approach ignores uncertainty quantification in the LLM output estimations, leading to invalid statistical inference. As shown in Appendix A, ignoring fine-tuning uncertainty creates four distinct types of inferential problems: biased coefficient estimates, invalid standard errors and confidence intervals, biased point predictions, and invalid prediction intervals.

This is especially problematic because fine-tuning on small, domain-specific datasets creates systematic overconfidence and measurement errors that violate classical assumptions. Classical measurement error theory assumes that errors are independent of both the true values being measured and other variables in the analysis, leading to predictable attenuation bias that shrinks coefficient estimates toward zero. However, fine-tuning-induced measurement errors are systematically correlated with the constructs being measured—for instance, when a model trained on imbalanced data consistently overpredicts certain categories. Unlike classical measurement error which predictably attenuates coefficients toward zero, this bias can be either upward or downward depending on the covariance between fine-tuning-induced measurement errors and the underlying constructs of interest.

Empirical evidence shows that fine-tuned LLMs exhibit overconfidence when adapted to small datasets, with parameter-efficient methods like LoRA being particularly susceptible to this issue (Citation: To-be-updated). The limited size of typical annotation datasets (often hundreds or thousands of examples) relative to the model's billions of parameters creates conditions where models memorize training patterns and become artificially overconfident about unseen cases (Citation:

To-be-updated).

To illustrate the problem, consider a researcher using a fine-tuned LLM to measure political stance from social media posts, then using these measures as covariates in a regression predicting voting behavior. If the LLM was fine-tuned on a small dataset with predominantly liberal examples, it will systematically overpredict liberal stance, especially for conservative or ambiguous posts. This creates negatively correlated measurement error where errors are larger for conservative content, leading to biased coefficient estimates, invalid standard errors, and overconfident predictions about voting outcomes.

In this research, we propose a framework that properly accounts for uncertainty from LLM fine-tuning in downstream statistical inference. Our approach jointly estimates LLM fine-tuning parameters and statistical model parameters within a Bayesian framework, rather than the conventional three-stage approach that treats LLM outputs as deterministic point estimates. Specifically, we obtain posterior distributions over fine-tuning parameters using tractable approximations, then propagate this uncertainty to statistical models, ensuring valid statistical inference. (We don't have a solution yet)

A key advantage of our approach is that it enables joint estimation of fine-tuning parameters specifically for the downstream statistical task. Traditional workflows treat fine-tuning and statistical analysis as separate stages: models are first fine-tuned on domain-specific classification or regression tasks, then the resulting features are used in entirely different downstream analyses. This separation ignores the potential mismatch between the fine-tuning objective and the research question. Our unified framework allows the fine-tuning process to be jointly optimized with the statistical model, enabling the LLM to learn representations that are specifically tailored to the downstream statistical task rather than just the intermediate labeling task. This joint estimation not only improves predictive performance but also ensures that the uncertainty quantification reflects the specific requirements of the statistical analysis, leading to more reliable statistical inference for the research question of interest.

2 Problem Setting

2.1 Sources of Uncertainty

We focus specifically on uncertainty arising from the fine-tuning stage rather than the pre-training uncertainty inherent in the base LLM. Pre-training uncertainty is essentially unidentifiable without access to the massive proprietary training corpus and computational resources, while fine-tuning uncertainty is both tractable and often dominates when adapting to small domain datasets (Citation: To-be-updated).

Fine-tuning pre-trained LLMs often employs parameter-efficient methods like Low-Rank Adaptation (LoRA) (Hu et al., 2022). Standard LoRA fine-tuning proceeds as follows: the pre-trained weights W_0 are adapted using low-rank decomposition $W = W_0 + BA$ where $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$ with rank $r \ll \min(m, n)$. During fine-tuning, only the LoRA parameters $\theta_{\text{LoRA}} = [A, B]$ are optimized on a domain-specific dataset to obtain point estimates $\hat{A}, \hat{B}$.

However, when these methods are applied to domain-specific datasets with limited examples, they exhibit pronounced overconfidence issues (Yang et al., 2023; Wang et al., 2024). The limited size of typical annotation datasets (often hundreds or thousands of examples) relative to the model's billions of parameters creates conditions where models memorize training patterns and become artificially overconfident about unseen cases.

This fine-tuning uncertainty is particularly consequential for the downstream statistical inference because it violates classical measurement error assumptions. Unlike independent measurement errors that lead to attenuation bias, fine-tuning errors can be systematically correlated with both the true constructs of interest and other variables in the analysis, leading to bias in unpredictable directions.

2.2 Setup and Notation

In our framework, observable components are: (1) raw text data for the downstream statistical task $\{x_i\}_{i=1}^n$, (2) structured covariates $Q \in \mathbb{R}^{n\times p}$ and outcome variables $Y \in \mathbb{R}^n$, (3) a domain-specific fine-tuning dataset $D_{\text{train}} = \{(\tilde{x}_j, \tilde{z}_j)\}_{j=1}^N$ used to adapt the pre-trained LLM, and (4) the LoRA parameters θ_{LoRA}.

The unobservable components include: (1) the true latent constructs z_i that we aim to

measure from text, (2) the massive pre-training dataset and computational resources needed to characterize uncertainty in the base model parameters W_0.

The estimable quantities in our framework are: (1) approximate posterior distributions over fine-tuning parameters $p(\theta_{\text{LoRA}}|D_{\text{train}})$, (2) joint posterior distributions over both LoRA fine-tuning parameters and downstream statistical model parameters, and (3) posterior predictive distributions that properly account for uncertainty propagation.

2.3 Bayesian Approach to LoRA Uncertainty

Standard LoRA fine-tuning treats the fine-tuned parameters as fixed, leading to deterministic feature extraction: $z_i = f_{\text{LLM}}(x_i; W_0 + \hat{B}\hat{A})$ where $\{x_i\}_{i=1}^n$ is raw text data and $z_i \in \mathbb{R}^d$. However, this approach ignores uncertainty in the fine-tuning parameters, particularly problematic when fine-tuning on small datasets.

Bayesian fine-tuning (Yang et al., 2023; Wang et al., 2024) addresses this by characterizing the full posterior distribution over parameters:

$$p(\theta_{\text{LoRA}}|D_{\text{train}}) \propto p(D_{\text{train}}|\theta_{\text{LoRA}}, W_0)p(\theta_{\text{LoRA}}) \tag{1}$$

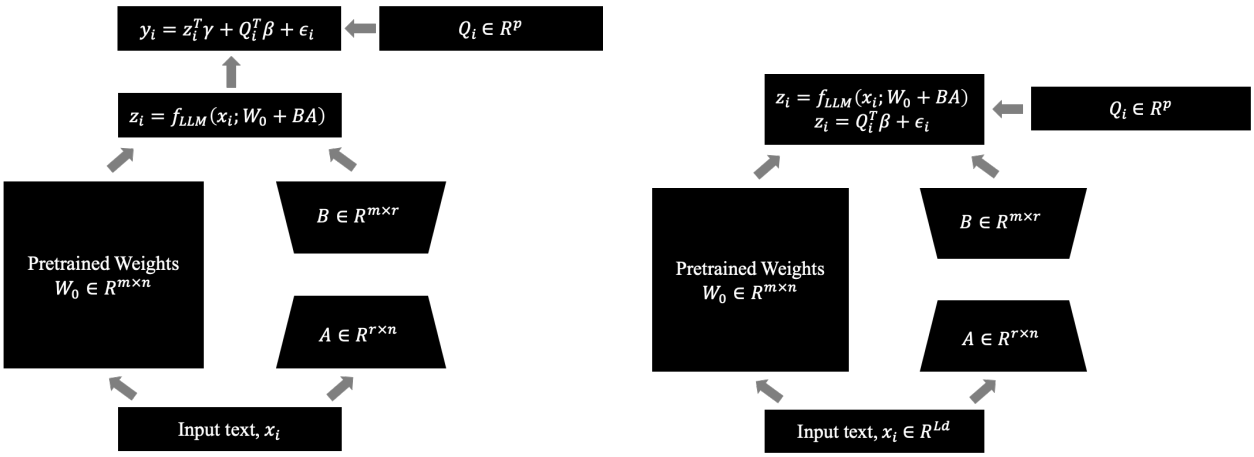
where $p(D_{\text{train}}|\theta_{\text{LoRA}}, W_0)$ is the likelihood of the fine-tuning data and $p(\theta_{\text{LoRA}})$ is the prior over parameters. For feature extraction, instead of deterministic outputs z_i, we obtain posterior predictive distributions that integrate over parameter uncertainty:

$$p(z_i|x_i, D_{\text{train}}) = \int p(z_i|x_i, W_0 + \theta_{\text{LoRA}})p(\theta_{\text{LoRA}}|D_{\text{train}})d\theta_{\text{LoRA}} \tag{2}$$

This posterior predictive distribution properly reflects uncertainty from the fine-tuning process, providing calibrated feature representations that account for the limited size of domain-specific training data.

2.4 Case A: Inference with LLM Outputs Used as Covariates

We first consider the case where LLM outputs are used as covariates in downstream statistical models. When covariates are measured with error, this leads to attenuation bias—coefficient estimates



(a) Case A: LLM outputs as covariates (b) Case B: LLM outputs as outcomes

Figure 1: Overview of the two cases of LLM uncertainty propagation in statistical inference

are biased toward zero—if errors are independent of the input data and can result in underestimating the true effects of LLM-derived variables on outcomes of interest. However, this classical measurement error framework does not apply to fine-tuned LLMs, where prediction errors systematically depend on the training dataset used for fine-tuning. Fine-tuning on small, domain-specific datasets often creates overconfidence issues that lead to systematic prediction biases rather than independent measurement errors. The resulting measurement error violates the independence assumption, meaning bias can occur in either direction depending on how the limited training data relates to the broader application domain.

The standard downstream regression with LLM outputs as covariates performs $Y_i = z_i^T\gamma + Q_i^T\beta + \varepsilon_i$, where $Q \in \mathbb{R}^{n\times p}$ represents structured covariates, and $Y \in \mathbb{R}^n$ is an outcome variable. This approach considers z_i, LLM outputs, as deterministic data, thus ignoring uncertainty in the LoRA parameters θ_{LoRA} from fine-tuning and leading to invalid statistical inference.

We propose a unified Bayesian framework that addresses this by jointly modeling both the feature extraction and statistical modeling stages. Rather than treating z_i as observed covariates, we recognize that z_i are uncertain predictions from a fine-tuned LLM with parameter uncertainty. The joint model specifies:

$$\text{Feature extraction:} \quad p(z_i|x_i, \theta_{\text{LoRA}}, W_0) \tag{3}$$

$$\text{Downstream regression model:} \quad Y_i|z_i, Q_i, \gamma, \beta \sim N(z_i^T\gamma + Q_i^T\beta, \sigma^2) \tag{4}$$

$$\text{LoRA parameter posterior:} \quad p(\theta_{\text{LoRA}}|D_{\text{train}}) \propto p(D_{\text{train}}|\theta_{\text{LoRA}}, W_0)p(\theta_{\text{LoRA}}) \tag{5}$$

The joint inferential target:

$$p(\gamma, \beta, \sigma^2, \theta_{\text{LoRA}}|Y, Q, D_{\text{train}}) \propto \tag{6}$$
$$\prod_i p(Y_i|z_i, Q_i, \gamma, \beta, \sigma^2)p(z_i|x_i, \theta_{\text{LoRA}}, W_0)p(\theta_{\text{LoRA}}|D_{\text{train}})p(\gamma, \beta, \sigma^2)$$

This joint posterior properly accounts for uncertainty propagation from fine-tuning through to statistical inference, enabling valid statistical inferences that reflect the true uncertainty in LLM-derived measurements.

2.5 Case B: Inference with LLM outputs Used as Outcomes

We next consider the alternative case where LLM-generated measures serve as dependent variables in statistical models. Under classical assumptions, where measurement errors are independent of the regressors and mean-zero, errors in the dependent variable do not bias regression coefficients but do inflate the residual variance, leading to larger standard errors and reduced statistical power. However, this classical framework is often violated when LLMs are used as predictive instruments for outcomes. Fine-tuned LLMs, especially those trained on limited or domain-specific data, may exhibit systematic errors that correlate with the input variables or vary in structure across contexts. These violations can introduce endogeneity-like concerns, resulting in biased coefficient estimates and invalid inference. In particular, biased or structured prediction errors from LLMs may confound the relationship between covariates and outcomes, undermining the reliability of statistical conclusions drawn from such models.

The standard approach estimates $z_i = Q_i^T\beta + \varepsilon_i$, where z_i represents the LLM-generated measure of interest and Q_i are explanatory variables. However, researchers observe $\hat{z}_i$ rather than the true construct z_i, where $\hat{z}_i$ contains measurement error from the fine-tuning process. This

measurement error in the dependent variable contaminates all coefficient estimates β, and leads to inconsistent standard errors that invalidate hypothesis testing and confidence intervals.

Our unified Bayesian framework addresses this by jointly modeling the statistical relationship and the LLM measurement process. Rather than treating the observed LLM outputs $\hat{z}_i$ as the true dependent variable, our approach recognizes that we observe noisy measurements of an underlying latent construct z_i. The joint model specifies:

$$\text{Statistical relationship:} \quad z_i|Q_i, \beta, \sigma^2 \sim N(Q_i^T\beta, \sigma^2) \tag{7}$$

$$\text{LLM measurement process:} \quad \hat{z}_i|x_i, z_i, \theta_{\text{LoRA}} \sim p(\hat{z}_i|x_i, z_i, W_0 + \theta_{\text{LoRA}}) \tag{8}$$

$$\text{LoRA parameter posterior:} \quad p(\theta_{\text{LoRA}}|D_{\text{train}}) \propto p(D_{\text{train}}|\theta_{\text{LoRA}}, W_0)p(\theta_{\text{LoRA}}) \tag{9}$$

The joint inferential target:

$$p(\beta, \sigma^2, Z, \theta_{\text{LoRA}}|\hat{Z}, Q, x_i, D_{\text{train}}) \propto \tag{10}$$
$$\prod_i p(z_i|Q_i, \beta, \sigma^2)p(\hat{z}_i|x_i, z_i, \theta_{\text{LoRA}})p(\theta_{\text{LoRA}}|D_{\text{train}})p(\beta, \sigma^2)$$

where z_i are treated as latent variables to be estimated jointly with the regression parameters β and LoRA parameters θ_{LoRA}. This approach enables valid inference about the factors driving the latent construct while properly accounting for both the uncertainty in measuring z_i from text and the parameter uncertainty from fine-tuning.

References

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2025). Inference for regression with variables generated by ai or machine learning.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3. https://arxiv.org/abs/2106.09685

Yang, A. X., Robeyns, M., Wang, X., and Aitchison, L. (2023). Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*. https://arxiv.org/abs/2308.13111

Wang, Y., Shi, H., Han, L., Metaxas, D., and Wang, H. (2024). Blob: Bayesian low-rank adaptation by backpropagation for large language models. *Advances in Neural Information Processing Systems*, 37:67758–67794. https://proceedings.neurips.cc/paper_files/paper/2024/hash/2eebfcf61fdf4fd8ae1aa1c1dbc50066-Abstract-Conference.html

Zhang, J., Xue, W., Yu, Y., and Tan, Y. (2023). Debiasing ML-or AI-Generated Regressors in Partial Linear Models. *Available at SSRN 4636026*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4636026

A Consequences of Ignoring LLM Uncertainty Quantification

A.1 Sources of LLM Uncertainty

The LLM output uncertainty comes from two sources: (i) baseline uncertainty, reflecting errors inherent in the pre-trained model $f_{\text{LLM}}(x_i; W_0)$, and (ii) fine-tuning uncertainty, reflecting the variability induced by adapting to a small domain-specific dataset.

Let the latent construct of interest (e.g., sentiment, topic proportion, policy position) be denoted by z_i. When we apply a pre-trained LLM with parameters W_0 to text x_i, we obtain an approximation $z_i^{(0)} = f_{\text{LLM}}(x_i; W_0) = z_i + u_i^{(0)}$, where $u_i^{(0)}$ represents baseline error from the pre-trained model. After fine-tuning with LoRA on a small domain dataset D_{train}, we obtain $\tilde{z}_i = f_{\text{LLM}}(x_i; W_0 + \theta_{\text{LoRA}}) = z_i^{(0)} + u_i^{(ft)} = z_i + u_i^{(0)} + u_i^{(ft)}$, where $u_i^{(ft)}$ captures fine-tuning uncertainty.

In principle, both $u_i^{(0)}$ and $u_i^{(ft)}$ contribute to measurement error. However, in this work, we focus on $u_i^{(ft)}$ and consider $u_i^{(0)}$ as absorbed into the latent construct z_i. The motivation is twofold: (i) baseline error from pretraining is essentially unidentifiable without access to the massive training corpus, and (ii) fine-tuning error is tractable and often dominates uncertainty when adapting to small domain datasets. By specifying priors on the LoRA parameters, θ_{LoRA}, and conditioning on the domain-specific dataset D_{train}, its posterior distribution can be approximated, making this error both estimable and consequential for downstream inference.

With this simplification, we write the observed LLM-derived surrogate as $\tilde{z}_i = z_i + u_i$, where $u_i \equiv u_i^{(ft)}$ denotes the fine-tuning error. Importantly, u_i may be correlated with the latent construct z_i, the structured covariates Q_i, and even the regression error ε_i.

This violation of classical measurement error assumptions leads to four distinct types of inferential problems in both cases: biased coefficient estimates, invalid standard errors and confidence intervals, biased point predictions, and invalid prediction intervals. We analyze these issues separately for each case below.

A.2 Inference Issues of Ignoring Fine-Tuning Uncertainty

A.2.1 Case A: LLM Outputs as Covariates

We now show the statistical consequences of ignoring this uncertainty in downstream regression analysis. Consider the true model:

$$Y_i = z_i^\top \gamma + Q_i^\top \beta + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid z_i, Q_i] = 0, \tag{11}$$

where z_i represents the latent construct of interest extracted from text x_i, Q_i denotes covariates, and Y_i is the outcome variable. In practice, researchers observe the noisy LLM output $\tilde{z}_i = z_i + u_i$ and estimate the model using $\tilde{z}_i$ in place of the unobserved true values z_i, ignoring the uncertainty in the fine-tuning process.

Notation and residualization. To analyze the bias introduced by this substitution, we employ the residual maker $M_Q = I - Q(Q^\top Q)^{-1}Q^\top$ to partial out the covariates (assume Q has full column rank). For clarity, we first present the single-covariate case ($d = 1$); the multivariate generalization follows analogously. Let $\dot{z} = M_Q \tilde{z}$, $z^* = M_Q z$, $\dot{u} = M_Q u$, and $\dot{\varepsilon} = M_Q \varepsilon$ denote the variables after partialling out Q. The OLS estimator for γ becomes:

$$\hat{\gamma} = (\dot{z}^\top \dot{z})^{-1} \dot{z}^\top \dot{Y}, \tag{12}$$

where $\dot{Y} = \gamma\, z^* + \dot{\varepsilon}$ is the residualized model.

Substituting $\dot{z} = z^* + \dot{u}$ yields

$$\hat{\gamma} = \frac{(z^* + \dot{u})^\top (\gamma z^* + \dot{\varepsilon})}{(z^* + \dot{u})^\top (z^* + \dot{u})}. \tag{13}$$

The numerator expands to $\gamma (z^*)^\top z^* + (z^*)^\top \dot{\varepsilon} + \gamma\, \dot{u}^\top z^* + \dot{u}^\top \dot{\varepsilon}$, while the denominator becomes $(z^*)^\top z^* + 2(z^*)^\top \dot{u} + \dot{u}^\top \dot{u}$. Taking probability limits and using $\mathrm{Cov}(z^*, \dot{\varepsilon}) = 0$, we obtain

$$\operatorname{plim} \hat{\gamma} = \gamma \cdot \frac{\mathrm{Var}(z^*) + \mathrm{Cov}(z^*, \dot{u})}{\mathrm{Var}(z^*) + \mathrm{Var}(\dot{u}) + 2\ \mathrm{Cov}(z^*, \dot{u})} + \frac{\mathrm{Cov}(\dot{u}, \dot{\varepsilon})}{\mathrm{Var}(\dot{z})}. \tag{A1}$$

This result reveals several important deviations from classical measurement error, with

consequences for both estimation and inference.

Biased coefficient estimates. Under classical measurement error, $u \perp (z, Q, \varepsilon)$ and $\mathbb{E}[u] = 0$, yielding the attenuation bias

$$\text{plim } \hat{\gamma} = \gamma \cdot \frac{\text{Var}(z^*)}{\text{Var}(z^*) + \text{Var}(\dot{u})} \in (0, 1) \cdot \gamma. \tag{A2}$$

However, fine-tuned LLMs typically violate these assumptions. Structured measurement error arises when $\text{Cov}(z^*, \dot{u}) \neq 0$ (e.g., overconfident patterns learned from small training data), so the bias in (A1) can be upward or downward. Endogeneity-type correlation emerges when LLM errors correlate with the regression error, $\text{Cov}(\dot{u}, \dot{\varepsilon}) \neq 0$, introducing the second bias term in (A1).

Invalid standard errors and confidence intervals. Standard OLS treats $\dot{z}$ as fixed and centers intervals at the pseudo-true limit in (A1) rather than the true γ, causing systematic mis-coverage. Moreover, variance is generally misspecified (often underestimated) because first-stage uncertainty from fine-tuning is ignored.

Biased point predictions. Plug-in predictions $\hat{Y}_i = \hat{\gamma}\, \tilde{z}_i + Q_i^\top \hat{\beta}$ suffer from (i) coefficient bias inherited from (A1) and (ii) input bias due to using $\tilde{z}_i$ in place of z_i. The net direction depends on the sign and magnitude of the biases in (A1).

Invalid prediction intervals. Starting from $Y_i = \gamma z_i + Q_i^\top \beta + \varepsilon_i$ and substituting $z_i = \tilde{z}_i - u_i$, we obtain $Y_i = \gamma \tilde{z}_i + Q_i^\top \beta - \gamma u_i + \varepsilon_i$. Thus,

$$\text{Var}(Y_i \mid Q_i, \tilde{z}_i) = \sigma^2 + \gamma^2 \, \text{Var}\big(u_i \mid x_i, D_{\text{train}}\big) - 2\gamma \, \text{Cov}\big(u_i, \varepsilon_i \mid Q_i, \tilde{z}_i\big). \tag{A3}$$

We model u_i as driven by (x_i, D_{train}), so we use $\text{Var}(u_i \mid x_i, D_{\text{train}})$ to reflect fine-tuning uncertainty rather than conditioning on Q_i. Omitting these components yields intervals that are systematically too narrow and overconfident.

A.2.2 Case B: LLM Outputs as Outcomes

Consider next the case where LLM-derived measures serve as dependent variables. The latent structural relationship is:

$$z_i = Q_i^\top \beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid Q_i] = 0, \tag{14}$$

where researchers observe $\hat{z}_i = z_i + u_i$ and estimate the model by regressing $\hat{z}_i$ on Q_i. The resulting OLS estimator is

$$\hat{\beta} = (Q^\top Q)^{-1} Q^\top \hat{z} = \beta + (Q^\top Q)^{-1} Q^\top \varepsilon + (Q^\top Q)^{-1} Q^\top u. \tag{15}$$

Taking probability limits and applying $\mathbb{E}[Q_i \varepsilon_i] = 0$, we obtain

$$\text{plim } \hat{\beta} = \beta + \left(\mathbb{E}[Q_i Q_i^\top]\right)^{-1} \mathbb{E}[Q_i u_i]. \tag{B1}$$

Biased coefficient estimates. Expression (B1) shows that coefficient bias depends entirely on whether fine-tuning errors correlate with the covariates. When $\mathbb{E}[Q_i u_i] \neq 0$—for instance, when fine-tuning errors systematically depend on observable features correlated with Q_i—the estimator becomes biased. For a single regressor, this simplifies to:

$$\text{plim } \hat{\beta} = \beta + \frac{\text{Cov}(Q, u)}{\text{Var}(Q)}. \tag{B2}$$

Invalid standard errors and confidence intervals. Even when coefficient estimates remain unbiased ($\mathbb{E}[Q_i u_i] = 0$), the regression residuals become $\varepsilon_i + u_i$ because $\hat{z}_i = z_i + u_i$. Thus, the variance of residuals inflates to

$$\text{Var}(\varepsilon_i + u_i) = \text{Var}(\varepsilon_i) + \text{Var}(u_i) + 2\ \text{Cov}(\varepsilon_i, u_i),$$

making intervals wider if treated correctly. However, standard OLS formulas treat $\hat{z}_i$ as fixed rather than estimated with uncertainty, omitting the additional variance component $\text{Var}(\hat{\beta} \mid \hat{z})$ that accounts for first-stage estimation. The net effect on coverage depends on whether $\text{Var}(u_i) + 2\,\text{Cov}(\varepsilon_i, u_i)$ outweighs the ignored first-stage variance, but in general intervals are invalid because

they are centered at a biased limit or computed with a misspecified variance.

Biased predictions. When $\mathrm{Cov}(Q_i, u_i) \neq 0$, coefficient estimates are biased according to (B1), so predictions of the latent construct are systematically shifted. Formally,

$$\mathbb{E}[\hat{z}_j \mid Q_j] - \mathbb{E}[z_j \mid Q_j] = Q_j^\top \left(\mathbb{E}[Q_i Q_i^\top]\right)^{-1} \mathbb{E}[Q_i u_i],$$

which represents the asymptotic prediction bias under the linear-projection interpretation. More generally, prediction bias equals $\mathbb{E}[u_j \mid Q_j]$.

Invalid prediction intervals. Standard procedures use the inflated residual variance

$$\mathrm{Var}(\varepsilon_i + u_i) = \mathrm{Var}(\varepsilon_i) + \mathrm{Var}(u_i) + 2\ \mathrm{Cov}(\varepsilon_i, u_i)$$

instead of the true latent variance $\mathrm{Var}(\varepsilon_i)$. This produces prediction intervals for the latent construct z_i that overstate or misstate uncertainty, depending on the sign of $\mathrm{Cov}(\varepsilon_i, u_i)$. When $\mathrm{Cov}(\varepsilon_i, u_i) = 0$, intervals systematically overestimate uncertainty due to the extra $\mathrm{Var}(u_i)$ term.

A.3 Empirical Examples of Fine-Tuning Bias

For example, suppose z_i is the true stance of a political text, and the fine-tuned LLM was trained on a small, imbalanced dataset with mostly liberal examples. A researcher then uses these LLM-predicted stance measures as covariates in a regression predicting voting behavior (Y_i), such as whether individuals vote for conservative candidates. The fine-tuned model tends to systematically over-predict liberal stance, especially for borderline cases. In this case, the fine-tuning error u_i is larger (and systematically biased) when z_i is conservative, $\mathrm{Cov}(u_i, z_i) \neq 0$. Similarly, if a structured covariate Q_i captures demographic attributes (e.g., age or region) that affect writing style, fine-tuning errors may be larger for certain groups, leading to $\mathrm{Cov}(u_i, Q_i) \neq 0$. Finally, if the same unobserved factors affect both the downstream regression and misclassification by the LLM, then $\mathrm{Cov}(u_i, \varepsilon_i) \neq 0$. These violate the independence assumptions of classical measurement error, causing bias that can amplify rather than attenuate coefficient estimates.