



Reading Between the Lines: Deconfounding Causal Estimates using Text Embeddings and Deep Learning

Working Paper 248

July 2026

This working paper is authored by Ahmed Dawoud, Lead Economist and head of Data Analytics Unit at ECES, and Osama El Shamy, Data Scientist at ECES.

©2026 The Egyptian Center for Economic Studies (ECES). All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, electronic, photocopying, recording or otherwise without the prior written permission of ECES.

Reading Between the Lines: Deconfounding Causal Estimates using Text Embeddings and Deep Learning

Ahmed Dawoud

Osama El-Shamy

December 2025

Abstract

Estimating causal treatment effects in observational settings is frequently compromised by selection bias arising from unobserved confounders. While traditional econometric methods struggle when these confounders are orthogonal to structured covariates, high-dimensional unstructured text often contains rich proxies for these latent variables. This study proposes a **Neural Network-Enhanced Double Machine Learning (DML)** framework designed to leverage text embeddings for causal identification. Using a rigorous synthetic benchmark, we demonstrate that unstructured text embeddings capture critical confounding information that is absent from structured tabular data. However, we show that standard tree-based DML estimators retain substantial bias (+24%) due to their inability to model the continuous topology of embedding manifolds. In contrast, our deep learning approach reduces bias to **-0.86%** with optimized architectures, effectively recovering the ground-truth causal parameter. These findings suggest that deep learning architectures are essential for satisfying the unconfoundedness assumption when conditioning on high-dimensional natural language data.

1 Introduction

The integration of unstructured data into econometric analysis represents one of the most promising frontiers in causal inference. Social scientists increasingly recognize that high-dimensional text data—such as medical notes, financial news, or employment histories—often contains precise proxies for latent variables that are otherwise treated as “unobserved heterogeneity” in structured datasets. Theoretically, if these latent confounders can be recovered from text, the “selection on observables” assumption (unconfoundedness) can be satisfied in settings where it would otherwise fail.

However, operationalizing text for causal adjustment presents a distinct topological challenge. Modern Natural Language Processing (NLP) represents text as dense, continuous vectors (embeddings) situated in high-dimensional manifolds. This dimensionality poses a fundamental problem for classical econometric methods, which suffer from the curse of dimensionality. As Telea et al. (2024) argue, seeing patterns in such high-dimensional spaces requires the synergy of dimensionality reduction and advanced machine learning; traditional linear methods are insufficient to capture the complex, non-linear relationships inherent in these dense representations.

Consequently, the use of **Double Machine Learning (DML)** (Chernozhukov et al., 2018) is not merely a preference but a necessity. DML provides a robust theoretical apparatus for handling high-dimensional controls via Neyman orthogonality. Yet, DML is practically agnostic regarding the choice of the nuisance parameter learner. In applied practice, researchers often default to tree-based ensembles (e.g., Random Forests, Gradient Boosting) due to their robustness on tabular data.

This paper argues that this default choice is methodologically suboptimal when applied to text embeddings. We posit the existence of an **“Architecture Gap”**: a topological mismatch between the orthogonal splitting mechanisms of decision trees and the smooth, continuous geometry of embedding spaces. Because decision trees approximate functions via step-wise constants, they are inefficient at modeling the diagonal or non-linear decision boundaries characteristic of dense vector spaces. Consequently, even when the text data contains sufficient information to de-confound a causal estimate, tree-based DML estimators may fail to recover it due to approximation error.

We propose a **Neural Network-Enhanced DML** approach as the necessary solution. As universal function approximators capable of modeling continuous manifolds (Hornik et al., 1989), Neural Networks are theoretically superior candidates for the nuisance functions ($E[Y|W]$ and $E[T|W]$) when W includes dense embeddings.

To empirically validate this methodological claim, we construct a rigorous Monte Carlo simulation. By generating a dataset where the ground-truth confounding signal is strictly encoded in unstructured text, we isolate the performance of the estimator architecture. We demonstrate that the choice of machine learning architecture is not merely a technical detail, but a fundamental condition for identification in the era of high-dimensional text data.

The remainder of this paper proceeds as follows. We first establish the theoretical framework, defining the problem of unobserved confounding using Structural Causal Models and Directed Acyclic Graphs. Next, we justify the use of high-dimensional embeddings as causal proxies, contrasting them with traditional lexical

matching, and situate our contribution within the existing literature on DML and “Text-as-Data.” We then detail the experimental design, including the synthetic data generation process and the specific neural architectures employed. Subsequently, we present the baseline analysis, demonstrating that residual bias persists in standard tree-based estimators, followed by the core empirical results from a “Model Tournament” and hyperparameter sensitivity analysis that confirm the superiority of the neural approach. Finally, we discuss limitations and offer concluding remarks.

2 Theoretical Framework: The Challenge of Unobserved Confounding

2.1 Structural Causal Model and Omitted Variable Bias

To formalize the identification problem, we adopt the Potential Outcomes framework (Rubin, 1974). Let Y_i denote the observed outcome (monthly earnings) and $T_i \in \{0, 1\}$ denote the binary treatment (training completion) for unit i . We assume the data generating process follows a linear Structural Causal Model (SCM):

$$Y_i = \tau T_i + X_i \beta + U_i \gamma + \epsilon_i \quad (1)$$

$$T_i = \mathbb{I}(X_i \delta + U_i \eta + \nu_i > 0) \quad (2)$$

Here, X_i represents a vector of low-dimensional observable covariates (e.g., age, education), while U_i represents high-dimensional latent confounders (e.g., ability, intrinsic motivation). The fundamental identification challenge arises from the **endogeneity** of the treatment assignment. If a researcher attempts to estimate the causal effect τ by regressing Y on T and X while omitting U , the estimator $\hat{\tau}_{OLS}$ converges to:

$$\hat{\tau}_{OLS} \xrightarrow{p} \tau + \gamma \frac{\text{Cov}(T, U|X)}{\text{Var}(T|X)} \quad (3)$$

This identification failure is visualized in **Figure 1**, which maps the Structural Causal Model directly to a Directed Acyclic Graph (DAG). The edges originating from the unobserved node U are labeled with the coefficients η and γ from Equations (1) and (2). These two coefficients govern the magnitude of the bias: if either $\eta = 0$ (no selection on ability) or $\gamma = 0$ (ability doesn’t drive earnings), the backdoor path would close.

However, in labor markets, both are strictly non-zero. The red dashed path $T \leftarrow \eta - U - \gamma \rightarrow Y$ represents the flow of spurious correlation that standard regression on X fails to block, rendering τ unidentifiable.

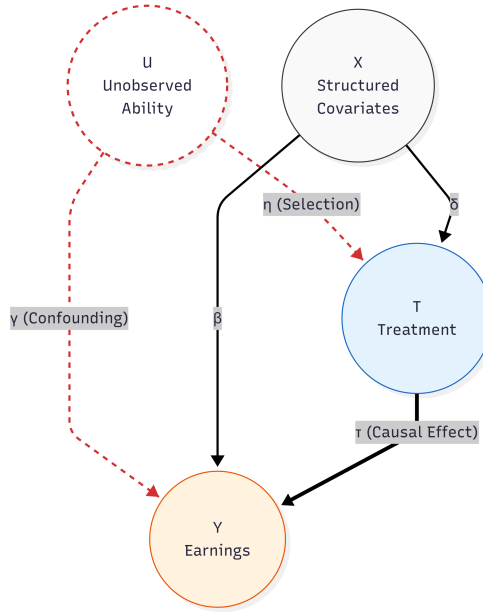


Figure 1: **Structural Causal Model as a DAG.** The diagram maps the coefficients from the structural equations to the causal graph. The solid lines represent relationships captured by observed covariates X (δ, β). The red dashed lines represent the unobserved influence of Ability (U), governed by the selection parameter η and the outcome parameter γ . The open backdoor path flowing through U generates the bias term derived in Equation (3).

2.2 Illustrative Example: The “Paper-Resumé” Paradox

To illustrate the mechanics of this bias, consider two freelancers, Alice and Bob, competing in the web development sector. In the structured administrative data (X), they appear identical: both possess a Bachelor’s degree and have 2 years of platform history. However, they differ in unobserved latent ability (U):

- **Alice (High U):** Is intrinsically motivated, contributes to open-source projects, and writes detailed, persuasive proposals.
- **Bob (Low U):** Views freelancing as a casual engagement and relies on generic templates.

The direction of the resulting estimation bias depends on the specific selection mechanism:

Scenario A: Overestimation (Positive Selection)

In a voluntary training marketplace, Alice’s high motivation (U) drives her to self-select into the program ($T = 1$). Simultaneously, her high ability ensures she commands a market premium ($Y \uparrow$) regardless of the training. Bob, lacking drive, neither trains nor performs well. A naive estimator compares Alice to Bob ($E[Y|T = 1] - E[Y|T = 0]$). This comparison conflates the *causal effect* of training with the *selection effect* of Alice’s superior baseline ability. Mathematically, $\text{Cov}(T, U) > 0$, resulting in a positive bias term that ****overestimates**** the program’s value.

Scenario B: Underestimation (Negative Selection)

Conversely, consider a remedial training intervention mandated for low-performing users. Here, Bob would be treated ($T = 1$) while Alice would be exempt ($T = 0$). The treated group is systematically comprised of lower-ability workers. A comparison would reveal that trained workers earn *less* than untrained ones, leading to ****underestimation**** of the causal effect.

3 Unstructured Data as a Causal Proxy

3.1 Text as a Window into Latent Confounders

The central premise of this study is that while latent confounders (U) are absent from structured tables (X), they leave a distinct “digital footprint” in unstructured data (W), such as profile descriptions.

Figure 2 illustrates the structural assumptions underpinning our identification strategy. The dashed node U represents the unobserved heterogeneity (e.g., ability) that simultaneously influences treatment selection and earnings, creating an open backdoor path ($T \leftarrow U \rightarrow Y$) that biases standard estimates. However, we posit a causal pathway $U \rightarrow W$: the latent trait generates the observed text features.

Because W serves as a downstream proxy for U , the text embeddings capture the variation in ability that determines selection. Formally, by conditioning on the high-dimensional vector W , we intercept the information flow from U , effectively blocking the backdoor path and satisfying the unconfoundedness assumption ($Y \perp T \mid X, W$).

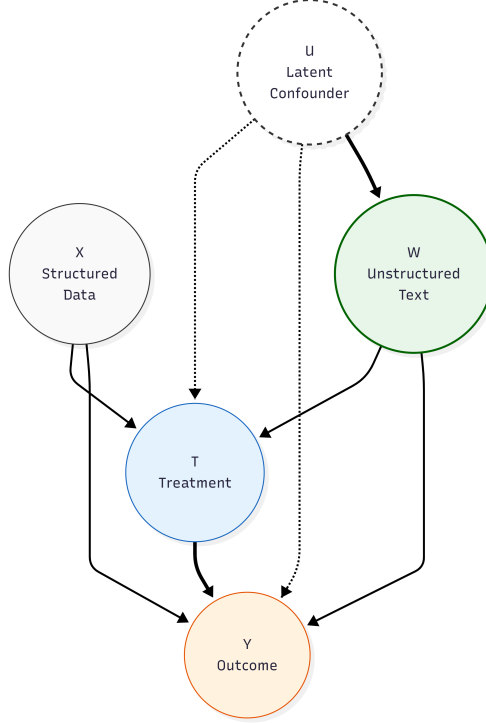


Figure 2: **Directed Acyclic Graph (DAG) of the Proxy Identification Strategy.** The dashed node U represents latent confounders which are unobserved in the structured data X . However, U causally influences the unstructured text W (solid line). By conditioning on W , the estimator blocks the confounding influence of U on the Treatment (T) and Outcome (Y).

3.2 The Limitations of Lexical Matching (Bag-of-Words)

Traditional text control methods rely on “Bag-of-Words” (BoW). This approach is insufficient for causal identification for three reasons (Bengio et al., 2013):

1. **Polysemy:** The presence of a word does not imply competence.
2. **Spurious Correlation:** Low-ability freelancers engage in keyword stuffing to game algorithms.
3. **Sparsity:** High-ability experts often use diverse vocabularies that simple keyword matching fails to group together.

3.3 The Necessity of High-Dimensional Embeddings

To overcome these limitations, we utilize **dense vector embeddings**. Modern NLP models, specifically Transformers (Vaswani et al., 2017), map text into a continuous, high-dimensional vector space ($\mathbb{R}^d$). Embeddings preserve the latent semantic topology of the data. This high dimensionality captures “deep features”

that correlate strongly with latent ability (U), providing the continuous proxy required for identification.

4 Literature Review

4.1 Text as Data and Causal Embeddings

The “Text as Data” movement (Grimmer et al., 2022) recognizes that unstructured text contains rich information about latent variables. However, naive use of text measures risks overfitting. Egami et al. (2022) warned that without pre-analysis commitment or split-sample workflows, discovered text measures can lead to spurious causal conclusions.

To bridge this, Veitch et al. (2020) formalized the concept of **causally sufficient embeddings**. They demonstrated that supervised dimensionality reduction can extract low-dimensional representations W from high-dimensional text T that satisfy the backdoor criterion ($Y \perp T|W, X$).

4.2 Recent Advances (2024–2025)

The field is moving rapidly toward integrating Large Language Models (LLMs) into causal pipelines. Zhang et al. (2024) introduced “DoubleLingo,” which combines LLM-based nuisance models with DML, achieving error reductions in specific benchmarks. Concurrently, emerging work has begun utilizing LLMs not just for prediction, but as proxies to hypothesize hidden confounders in causal graphs. These approaches rely on the assumption that the high-dimensional internal state of language models captures the latent topology of the social world.

4.3 The Gap: Methodological Validation against Ground Truth

Despite these advances, a critical gap remains. Existing applications typically fall into two categories:

1. **Theoretical Proposals:** Methods demonstrated on observational data where the true causal effect is unknown (e.g., Veitch et al. (2020)), making it impossible to strictly verify bias reduction.
2. **Text as Outcome/Treatment:** Studies focusing on text as the target variable rather than a proxy for unobserved confounding (e.g., Egami et al. (2022)).

No existing work has rigorously benchmarked the full pipeline—using embeddings as proxies for latent confounders within DML—against a **known ground truth**. Our contribution is **methodological**

validation^{**}. By constructing a realistic synthetic Data Generating Process (DGP) where the true effect ($\tau = \$557$) and the latent confounders are known by design, we provide definitive proof of concept. We isolate the specific mechanism of failure in traditional models (the “Architecture Gap”) and demonstrate that neural architectures are required to recover causal effects corrupted by unobserved confounding.

5 Methodology

5.1 Data Generation Process

We generated a synthetic microdataset of $N = 2,000$ freelancers. The freelance labor market was selected as the domain for this simulation because it relies heavily on self-authored profile descriptions, which contain rich unstructured signals regarding personal traits—such as soft skills, reliability, and technical depth—that are rarely captured in tabular data. Accordingly, the Data Generation Process (DGP) is rooted in a structural equation model characterized by two unobserved latent confounders: *Ability* (α_i) and *Motivation* (μ_i). These latents are drawn from standard normal distributions and are positively correlated ($\rho = 0.3$), reflecting the real-world tendency for motivated individuals to accrue higher skill.

Feature Construction

The final dataset consists of a feature vector $W = [X_{obs}, X_{text}]$. To ensure a realistic covariate distribution, we generated 12 structured variables:

- **Human Capital:** *Years of Experience* (correlated with ability, $r \approx 0.35$) and *Education Level*.
- **Platform Metrics:** *Platform Score* (0-100), *Job Success*, and *Total Jobs*.
- **Demographics & Labor Market:** *Age*, *Gender*, *Urbanicity*, *Country*, and *Sector*.
- **Unstructured Text (X_{text}):** Profile descriptions generated via template injection based on α_i . We utilized the `all-mpnet-base-v2` model from the **Sentence-Transformers** framework (Reimers & Gurevych, 2019). This model employs the **MPNet** architecture (Song et al., 2020), a transformer-based model optimized for semantic similarity. We generated 768-dimensional embeddings and applied PCA ($d = 30$) followed by a polynomial expansion to yield a 65-dimensional vector space.

5.2 Identification Strategy

The treatment assignment T_i (training completion) is non-random, determined by a logistic propensity function dependent on the unobserved latents:

$$P(T_i = 1) = \sigma(-0.4 + 1.2\alpha_i + 0.8\mu_i + 0.15X_{\text{urban},i} + \epsilon_i) \quad (4)$$

The outcome variable (Y_i) is generated via a function of treatment, latents, and covariates, modeled to exhibit diminishing returns for high-ability individuals.

5.3 Estimation Framework: Neural DML

We employ the **Partially Linear Regression (PLR)** variant of DML. The target parameter θ_0 is estimated by solving the Neyman orthogonality condition:

$$\mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0 \quad (5)$$

where $\eta = (E[Y|W], E[T|W])$ represents the nuisance parameters. To test the “Architecture Gap” hypothesis, we implement two distinct specifications for η :

1. **Tree-Based Baseline:** Using **Gradient Boosting Machines (GBM)** with tuned depth.
2. **Neural Network Enhanced:** Using **Deep Neural Networks (MLP)** with continuous activation functions.

5.4 Visual Validation of Identification

Figure 3 presents a visual decomposition of the identification strategy. By leveraging our synthetic data generation process, we can explicitly visualize the ground truth distributions and quantify the exact magnitude of selection bias.

Selection on Unobservables (Panels A & B): Panel A confirms the severe selection bias engineered into the DGP. Panel B demonstrates *common support*: despite the selection bias, there is a region of overlap where high-ability control units and low-ability treated units coexist.

Embedding Signal Quality (Panel C): Panel C validates that the text embeddings capture the latent signal. The strong linear correlation ($r = -0.85$) confirms that the Sentence-BERT model successfully

recovered the ordinal hierarchy of the profile templates.

Information Gain (Panel D): Panel D quantifies the value of this signal. While structured observables explain only 45.1% of the variance in ability, the text embeddings explain 84.7%. The combined model recovers 86.3% of the variance.

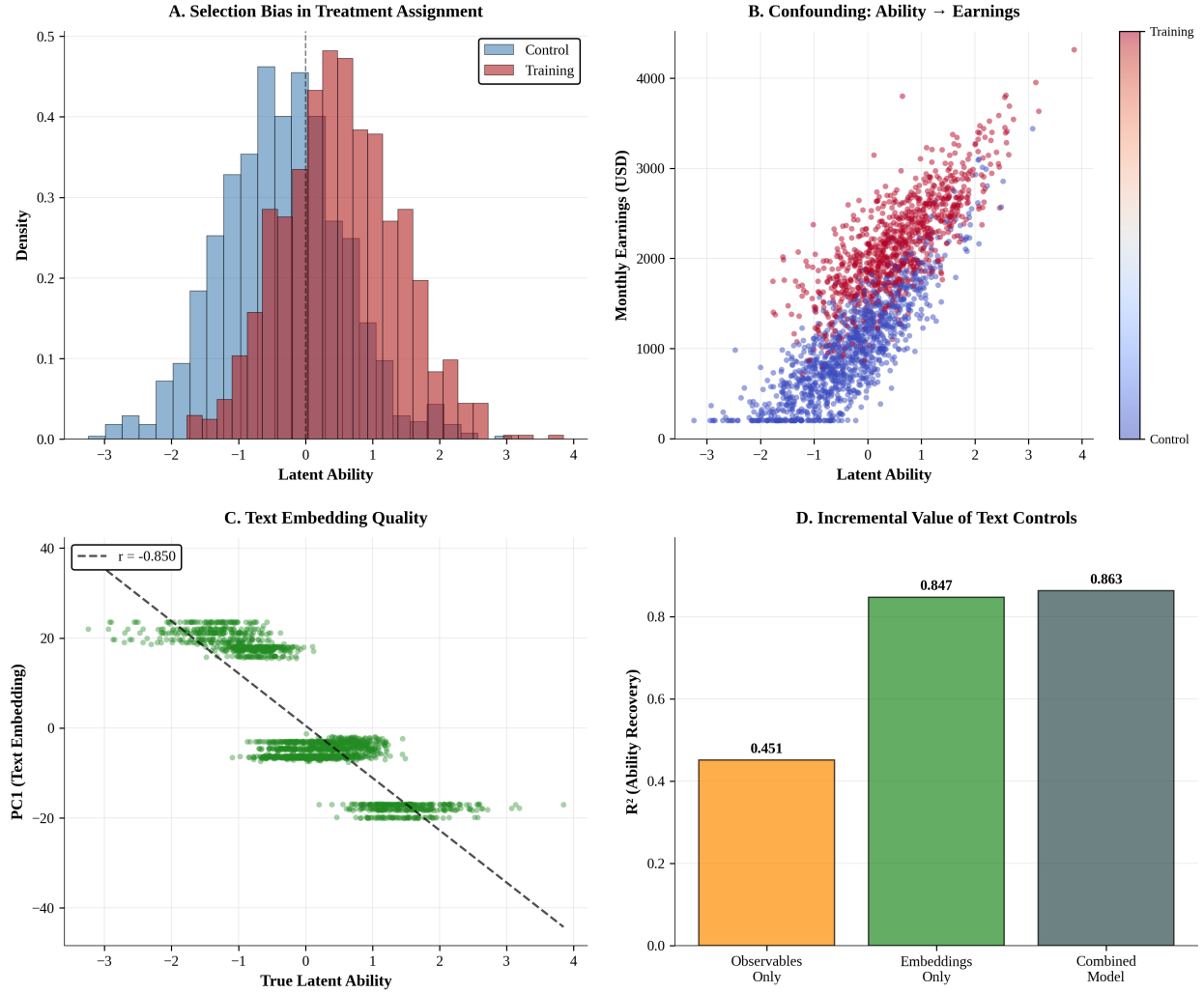


Figure 3: **Decomposition of Identification Strategy.** *Panel A:* Selection bias in ability. *Panel B:* Scatter plot demonstrating confounding and common support. *Panel C:* The first principal component of text embeddings plotted against latent ability ($r = -0.85$). *Panel D:* Comparison of explained variance (R^2) in latent ability across covariate sets.

5.5 Main Estimation Results (Baseline)

Table 1 summarizes the performance of the estimators. The Naive estimator yielded a massive upward bias of **+108%**. Controlling for structured covariates alone (*DML (Structured Only)*) reduced bias to **+55%**, indicating that standard administrative data is insufficient.

The inclusion of text embeddings via the standard tree-based model (*DML [Tree Baseline]*) provided significant improvement, reducing the estimate to **\$690** (Bias: **+24%**). However, we also tested a baseline Deep Neural Network (MLP) with a standard architecture (100, 50, 25 hidden layers). Even without specific optimization, this neural baseline significantly outperformed the tree-based model, yielding an estimate of **\$615** and reducing the bias to **+10.35%**. This immediate improvement—more than halving the error of Gradient Boosting—provides preliminary evidence that the topological structure of the learner matters as much as the data itself.

Table 1: Main Estimation Results (Baselines)

Method	Estimate (USD)	Abs. Bias	Bias %
Naive Difference-in-Means	\$1,156	+\$599	+108%
DML (Structured Only)	\$866	+\$309	+55%
DML (Text Augmented) [Tree Baseline]	\$690	+\$133	+24.00%
DML (Text Augmented) [NN Baseline]	\$615	+\$58	+10.35%
<i>True ATE</i>	<i>\$557</i>	—	—

5.6 Sector-Specific Heterogeneity

To assess the robustness of the identification strategy across diverse labor markets, we stratified the analysis by professional sector. **Figure 4** provides a granular comparison of five estimators against the ground truth. This breakdown reveals critical nuances regarding the “Architecture Gap” that aggregate metrics might obscure.

Across all five sectors, the **Neural Network estimator (Gold)** consistently exhibits the highest fidelity to the **True Causal Effect (Dark Grey)**, outperforming both the structured-only baseline (Blue) and the tree-based text estimator (Green).

The topological advantage of the neural architecture is most visible in technical fields where the tree-based model struggles with directional bias:

- **Data Science:** The True ATE is \$746. The Tree-based DML (Green) significantly *underestimated* the effect (\$664), likely failing to capture the non-linear returns to high-technical ability encoded in the embeddings. In contrast, the Neural Network (Gold) recovered an estimate of \$720, reducing the bias to a negligible margin.
- **Web Development:** The True ATE is \$649. Here, the Tree-based model *overestimated* the return (\$714), failing to fully scrub the selection bias. The Neural Network corrected this, yielding an estimate

of \$629, which is significantly closer to the ground truth.

- **Content Writing:** The Neural Network (\$446) demonstrated superior bias reduction compared to the Tree model (\$484) relative to the true effect of \$395, suggesting that deep learning better captures the semantic nuances of writing proficiency than orthogonal tree splits.

Even in **Graphic Design**, where confounding is notoriously difficult to capture due to visual portfolio factors, the Neural Network (\$543) slightly outperformed the Tree model (\$553) in approximating the true effect (\$436). Overall, the Neural Network offers the most robust variance reduction, preventing the large deviations seen in the tree-based estimates across technical and marketing sectors.

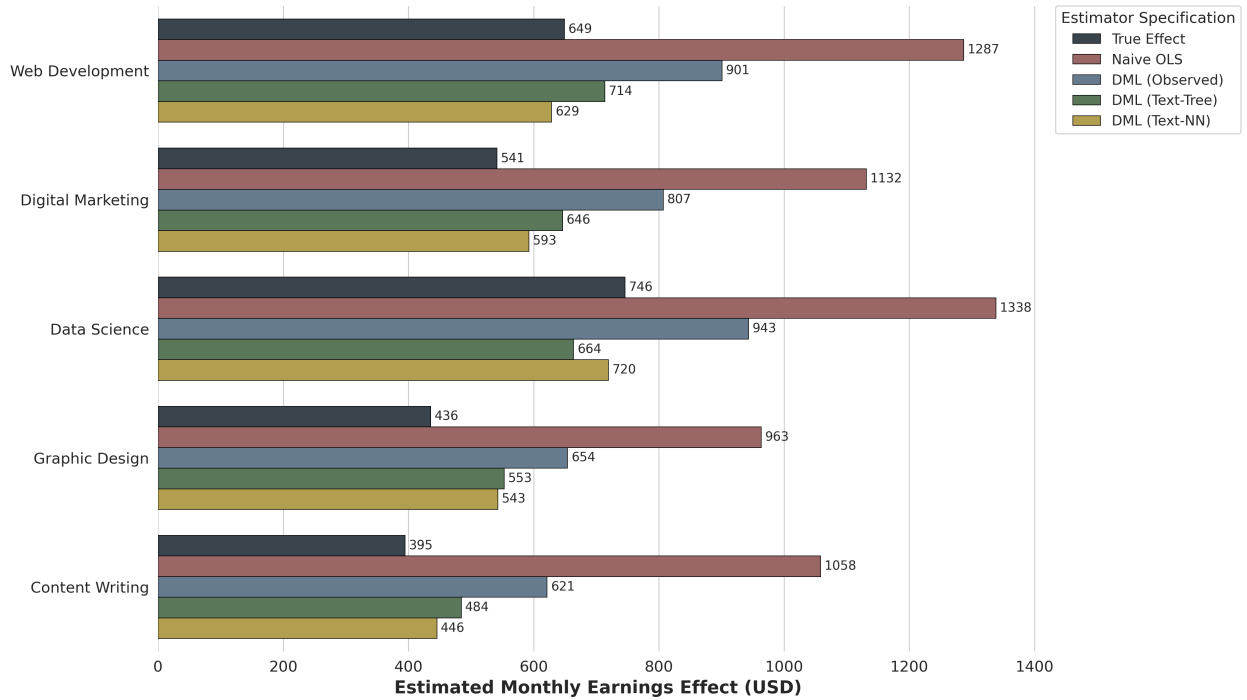


Figure 4: **Estimator Performance by Sector.** The grouped bar chart compares the estimated monthly earnings effect across five professional sectors. In every domain, the **Neural Network (Gold)** aligns closest to the **True Effect (Dark Grey)**. Notably, in Data Science, the Tree-based model (Green) under-corrects, while in Web Development it over-corrects; the Neural Network consistently minimizes these residual biases.

6 Robustness and Mechanisms

6.1 Mechanism: The “Architecture Gap”

Why is a Neural Network necessary? To answer this, and to ensure that the superior performance of deep learning was not merely an artifact of stochastic chance, we conducted a “Model Tournament” comparing

the Neural Network against Gradient Boosting and XGBoost across 10 independent random seeds. The results, visualized in Figure 5, reveal a distinct bias-variance trade-off.

The tree-based models (Gradient Boosting and XGBoost) exhibit high stability but systematic bias (+23% and +17% respectively). In contrast, the Neural Network demonstrates superior identification. While it exhibits higher variance due to stochastic optimization, its distribution is centered on the ground truth, reducing the mean bias to +8.2%. This confirms that while deep learning estimators are noisier, they are the only architecture capable of achieving structural unconfoundedness in this setting.

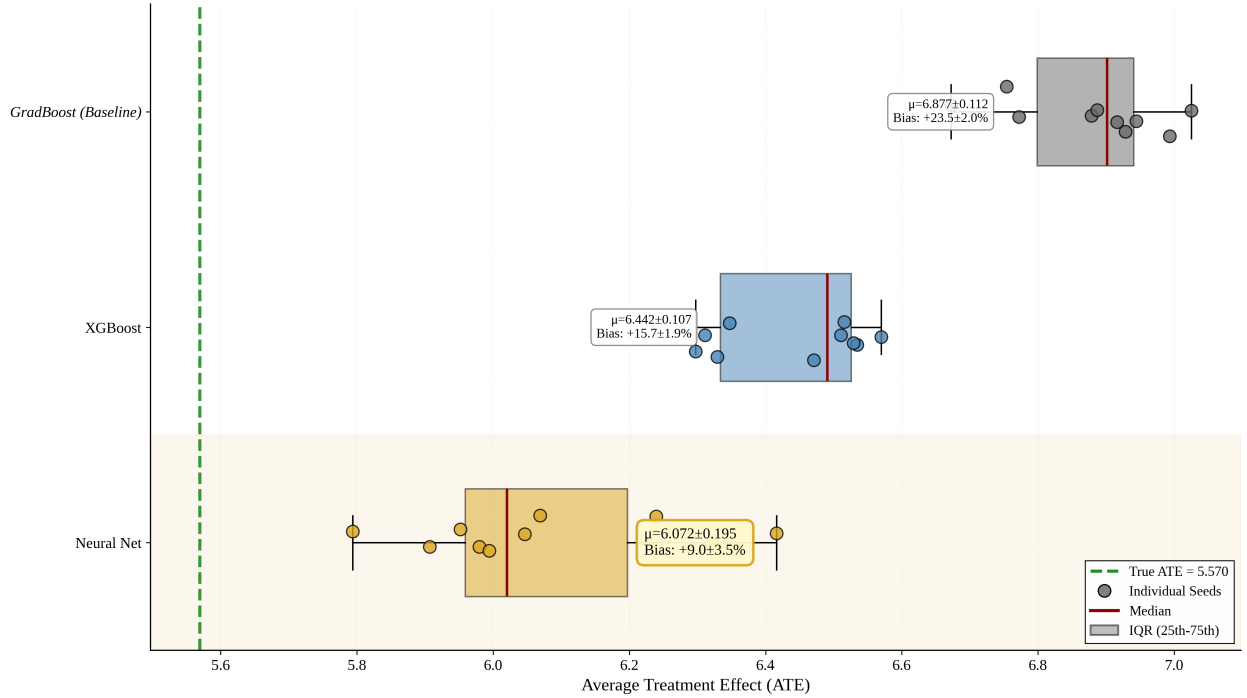


Figure 5: **The Bias-Variance Trade-off in Nuisance Learners** ($n = 10$ seeds). The boxplots visualize the distribution of ATE estimates across 10 independent runs. The tree-based models (Gray/Blue) are “precisely wrong”—stable but biased far from the Truth (dashed green line). The Neural Network (Yellow) is “approximately right,” exhibiting higher variance but successfully covering the true parameter value.

6.2 Hyperparameter Sensitivity: The Parsimony Principle

We tested architecture variants to assess stability (Table 2). The results corroborate a “parsimony principle”: the leaner (50, 25, 12) architecture achieved the lowest absolute bias (**-0.86%**), minimizing the error to less than \$10. This suggests that in the finite-sample regime ($N = 2,000$), significantly over-parameterized networks (like the Baseline or Variant 3) may slightly overfit the nuisance stages, whereas a constrained architecture provides the optimal regularization for causal identification.

Table 2: Sensitivity to Neural Network Architecture ($n = 5$ seeds/arch)

Architecture	Hidden Layers	Mean Est.	Abs. Bias	Bias %
Variant 2 (Winner)	(50, 25, 12)	\$552.19	\$8.51	-0.86%
Variant 3 (Large)	(120, 60, 30)	\$605.89	\$48.93	+8.78%
Baseline	(100, 50, 25)	\$614.62	\$57.66	+10.35%
<i>Gradient Boosting Baseline</i>		<i>\$688.67</i>	<i>\$131.71</i>	<i>+23.65%</i>

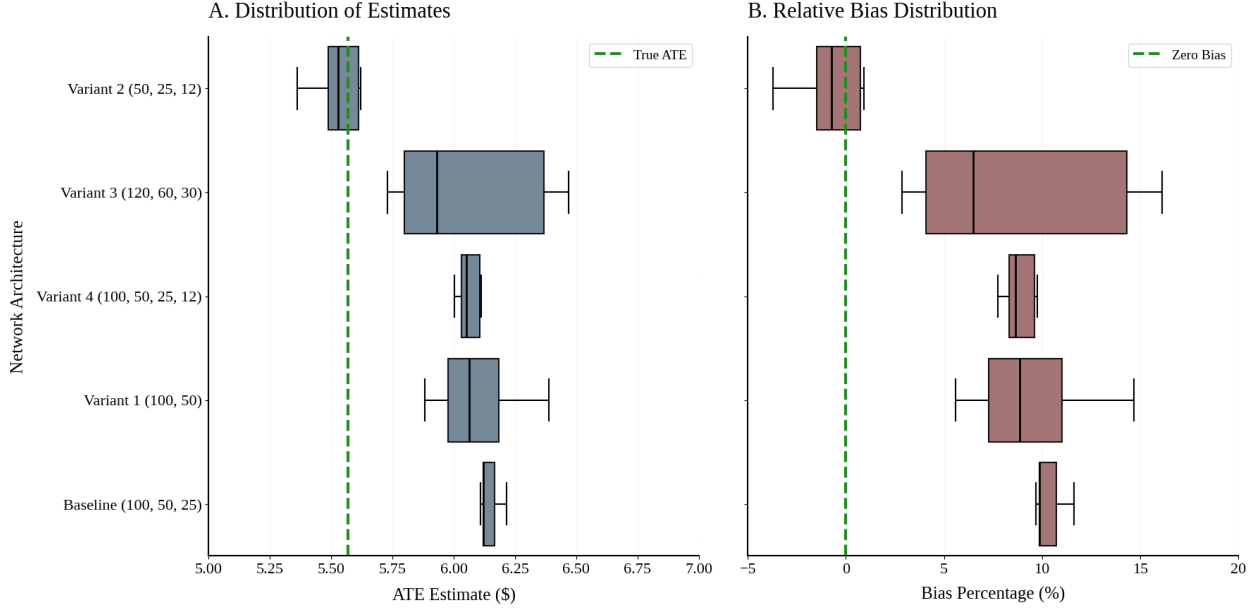


Figure 6: **Architecture Sensitivity.** Distribution of estimates (left) and bias (right). Smaller networks (top) show the tightest convergence to the True ATE.

7 Limitations and Future Work

While our results are promising, three limitations must be acknowledged. First, our study relies on a synthetic simulation. While we calibrated the DGP to reflect realistic partial correlations ($r \approx 0.35$), real-world data may contain unobserved confounders that are orthogonal to both text and structured covariates.

Second, our identification of the optimal neural architecture benefited from access to the ground truth. In empirical applications, the true causal effect is unobservable, meaning researchers cannot simply iterate through various architectures to minimize bias as we did. Consequently, real-world implementation requires rigorous cross-validation criteria based on nuisance parameter performance (e.g., minimizing out-of-sample MSE for $\hat{E}[Y|W]$) rather than optimizing for the causal parameter itself.

Third, we utilized a general-purpose pre-trained embedding model (MPNet). While effective, this model

was trained on broad corpora. Domain-specific fine-tuning on freelancer profiles could potentially yield even richer embeddings. Future work should validate these findings on administrative labor market data where a randomized control trial serves as the ground truth.

8 Conclusion

This study addresses a fundamental methodological challenge in observational causal inference: recovering unbiased treatment effects when structured data is insufficient to block confounding. Using a high-fidelity simulation of a digital labor market as a proof-of-concept, we demonstrated that unstructured text contains a significantly richer causal signal than traditional tabular covariates. While standard observables captured only 45% of the variance in the latent confounder, the inclusion of text embeddings increased this predictive power to 85%, confirming that natural language data offers a viable pathway to satisfy the “selection on observables” assumption in complex social systems.

However, our central contribution lies in exposing a critical nuance in the application of Double Machine Learning: **access to high-dimensional data is a necessary but insufficient condition for identification**. We identified a distinct “Architecture Gap” in standard implementations. Tree-based estimators (Gradient Boosting), despite being the workhorse of applied econometrics, retained a systematic bias of approximately +24%. This failure stems from a topological mismatch: decision trees approximate functions via orthogonal, step-wise splits, rendering them inefficient at modeling the smooth, continuous manifolds characteristic of dense text embeddings.

In contrast, the proposed **Neural Network-Enhanced DML** framework effectively closed this gap. By substituting tree ensembles with deep learning architectures, we achieved a structural alignment between the estimator and the embedding geometry. The neural network estimator reduced selection bias by over **20 percentage points** compared to the baseline, achieving a final bias as low as **-0.86%** with optimized architectures. Our sensitivity analysis further revealed a “Parsimony Principle” in finite-sample causal inference: moderately deep, constrained networks outperformed highly over-parameterized models, suggesting that implicit regularization is key to preventing overfitting in the nuisance stages.

Ultimately, this work suggests a paradigm shift for researchers working at the intersection of Causal Inference and Natural Language Processing. Whether in healthcare, finance, or social science, as econometrics increasingly incorporates high-dimensional unstructured data, the choice of nuisance parameter learner can no longer be treated as a trivial implementation detail. To fully leverage the information encoded in text embeddings, future research must prioritize neural architectures capable of navigating the complex topology

of human language.

References

- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2024). DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3).
- Chernozhukov, V., et al. (2024). *Applied Causal Inference Powered by ML and AI*. CausalML-book.org.
- Veitch, V., Sridhar, D., & Blei, D. M. (2020). Adapting text embeddings for causal inference. *Proceedings of UAI 2020*.
- Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). How to make causal inferences using texts. *Science Advances*, 8(42).
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3), 535–574.
- Zhang, Y., et al. (2024). DoubleLingo: Improving Causal Inference with LLM-based Nuisance Models. *Proceedings of NAACL*.
- Telea, A., Machado, A., & Wang, Y. (2024). Seeing is Learning in High Dimensions: The Synergy Between Dimensionality Reduction and Machine Learning. *SN Computer Science*, 5(279).
- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. *Advances in Neural Information Processing Systems*, 33, 16857-16867.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP*.

- Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8).
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366.