HIGH-DIMENSIONAL PERCEPTION WITH THE DOUBLE MACHINE LEARNING LENS MODEL

Raymond V. Li AND **Jeremy C. Biesanz**

DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF BRITISH COLUMBIA, CANADA

October 14, 2025

Funding: Preparation of this manuscript was supported by Social Sciences and Humanities Research Council (SSHRC) of Canada Grant 435-2020-0203 to Jeremy C. Biesanz.

Competing Interests: The authors declare no competing interests.

Ethical Approval: This study was approved by the Behavioural Research Ethics Board of the University of British Columbia (No. H20-01189).

Informed Consent: Informed consent was obtained from all participants.

Corresponding Author:

Raymond V. Li

Department of Psychology, University of British Columbia

E-mail: raymond.li@psych.ubc.ca

THIS ARTICLE IS BASED ON THE FIRST AUTHOR'S MASTER'S THESIS, SUBMITTED TO THE UNIVERSITY OF BRITISH COLUMBIA.

HIGH-DIMENSIONAL PERCEPTION WITH THE DOUBLE MACHINE LEARNING LENS MODEL

Abstract

Traditional perceptual models are ill-equipped for the high-dimensional data, such as text embeddings, central to modern psychology and AI. We introduce the Double Machine Learning Lens Model (DML-LM), a framework that utilizes machine learning to handle such data. We applied this model to analyze how a modern AI and human perceivers judge social class from 9,513 aspirational essays written by 11-year-olds in 1969. A systematic comparison of 45 analytical approaches revealed that regularized linear models using dimensionality-reduced language embeddings significantly outperformed traditional dictionary-based methods and more complex non-linear models. Our top model accurately predicted human ($R^2_{CV} = .61$) and AI ($R^2_{CV} = .56$) social class perceptions, capturing over 85% of the total accuracy. These results suggest that "unmodeled knowledge" in perception may be an artifact of insufficient measurement tools rather than an unmeasurable intuitive process. We find that both AI and humans use many of the same textual cues (e.g., grammar, occupations, cultural activities), only a subset of which are valid. Both appear to amplify subtle, real-world patterns into powerful, yet potentially discriminatory heuristics, where a small difference in actual social class creates a large difference in perception.

Key words:

double machine learning, lens model, large language models, artificial intelligence, perception accuracy

Introduction

From snap judgments in job interviews to the algorithmic screening of online profiles, both humans and artificial intelligence are increasingly tasked with forming consequential impressions from complex, high-dimensional data. Humans can evaluate others in as little as 100 milliseconds, with these rapid assessments predicting outcomes from hiring decisions to electoral results (Ambady & Rosenthal, 1992; Todorov et al., 2005). Similarly, modern AI systems now analyze vast streams of text, image, and behavioral data to make their own judgments about personality, competence, and risk. Despite this parallel function, we lack a unified theoretical and methodological framework to deconstruct and compare the judgmental processes of both human and artificial perceivers on equal terms.

The challenge of formally modeling judgment is not new. For decades, researchers have relied on Brunswik's (1955) lens model, and its statistical formulation in Tucker's (1964) Lens Model Equation (LME), to understand how perceivers use available cues to achieve accuracy. This framework, which underpins foundational theories like Funder's (1995) Realistic Accuracy Model, has been invaluable for demonstrating how traits are perceived via relevant and detectable cues across diverse domains (Wallace & Biesanz, 2021; Rule & Ambady, 2008). However, the LME was conceived for a low-dimensional world, requiring researchers to hand-select a small number of potential cues. It is ill-equipped to handle the "curse of dimensionality" (Hastie et al., 2009) inherent in modern data, such as the thousands of features generated by large language model (LLM) embeddings, which have become a cornerstone for quantifying language (Grimmer et al., 2022).

To bridge this critical gap, we introduce the Double Machine Learning Lens Model (DML-LM). By integrating the causal inference principles of Double Machine Learning (Chernozhukov et al., 2018) with the classic LME, our approach provides a robust and flexible

method for analyzing how accuracy is achieved in high-dimensional settings. This innovation allows researchers to finally leverage the full richness of modern data representations, from complex text embeddings to video features, within a structured and interpretable accuracy model.

Central to our framework is a unifying and interpretable statistic: the Percentage of Mediated Accuracy (PoMA). PoMA provides a single, intuitive metric that quantifies the proportion of total judgmental accuracy—whether from a human or an AI—that can be statistically explained by a given set of cues. This allows for a principled, apples-to-apples comparison between different cue sets (e.g., traditional dictionaries vs. LLM embeddings) and different analytical models. By using PoMA, researchers can systematically evaluate how well their models capture the information used in a perceptual judgment, providing a clear path forward for dissecting the mechanisms of human intuition and algorithmic decision-making alike.

Theoretical and Methodological Framework

The Lens Model

By using multiple regressions to model both perceiver judgments and the validity measure (e.g., perceived intelligence and someone's actual intelligence quotient) as a linear combination of behavioural cues, Tucker (1964) quantitatively formalized Brunswik's (1955) lens model as the lens model equation. Specifically, by regressing our validity variable X and judgment variable Y on a vector of behavioural cues Z (where k is the number of cues), we obtain the following regression equations:

$$\hat{X} = \beta_{0,X} + \sum_{i=1}^{k} \beta_{i,X}\, Z_i \qquad \hat{Y} = \beta_{0,Y} + \sum_{i=1}^{k} \beta_{i,Y}\, Z_i \tag{1}$$

The residuals are defined as:

$$e_X = X - \hat{X} \qquad e_Y = Y - \hat{Y} \tag{2}$$

The first parameter of interest from Tucker's lens model is called achievement accuracy (r_a), is defined as the observed correlation between judgment scores and validity scores. From the above equations, the four components of the achievement correlation can be readily derived. The variance in the validity variable X and judgment variable Y can be decomposed into two components: the variance explained by the model, $\text{Var}(\hat{X})$, and the residual variance, $\text{Var}(e_X)$, which are orthogonal under ordinary least squares (OLS) regression. This yields us our traditional R^2 values, interpreted as the proportion of explained variance. R^2 can also be expressed in terms of $\text{Var}(e)$ which is mathematically equivalent to the former definition under OLS, but in general, represents how much better at the model is at predicting the outcome when compared to the average (a negative R^2 means the average would have been a better predictor than the model).

$$R^2_X = \frac{\text{Var}(\hat{X})}{\text{Var}(X)} = 1 - \frac{\text{Var}(e_X)}{\text{Var}(X)} \tag{3}$$

R^2_Y measures the degree to which the set of cues Z are used by judges, whereas R^2_X quantifies the degree to which the set of Z cues are linearly related to the validity measure. The square root of these measures (i.e., the multiple correlation R) are called the response consistency and environmental predictability coefficients respectively (Karelaia & Hogarth, 2008; Osterholz et al., 2021). Researchers also use these coefficients as measures of how valid the cues are and how much they are used by perceivers (Borkenau et al., 2016). Since these measures reflect how well the models can predict validity measure and perceiver ratings, they reflect the degree to which the

entire set of cues is relevant and available (with respect to the validity measure), or detectable and utilized (with respect to the perceiver ratings; Funder, 1995).

Following those, two other statistics are computed. The matching coefficient G (also called cue sensitivity) represents the degree of alignment between a perceiver's usage of behavioural cues and the target itself, and is computed as the correlation between predicted values of both regression models, $\text{Cor}(\hat{X}, \hat{Y})$. For example, if perceivers generally use the tone of a warm voice as an indicator for agreeableness, and a warm voice tone is in fact associated with a target's agreeableness, then G will be relatively high. The unmodelled knowledge coefficient C is computed as the correlation between the residuals of both regression models, $\text{Cor}(e_X, e_Y)$. If perceivers had high accuracy achievement r_a of strangers' agreeableness using warm voice tone as a valid indicator, but warm voice tone wasn't assessed and included in the regression model as a cue, then the unmodelled knowledge coefficient C might be high. Importantly, C can also indicate that cues are being used in a non-linear or non-additive manner that is not captured by the linear model.

By applying scaling terms to G and C, we can decompose total achievement r_a into two components: modeled knowledge (GR_XR_Y) and unmodeled knowledge ($C\sqrt{1-R_X^2}\sqrt{1-R_Y^2}$). Using these scaled coefficients, we can compute the proportion of achievement that is modeled and not modeled by dividing by r_a. For instance, perceivers might have an achievement of 0.5 for assessing agreeableness in strangers, where 60% (0.3) is modelled through warm voice tone, and 40% (0.2) is left unexplained by unknown factors (e.g., smiling, nodding). This is how Tucker (1964) formally quantified Brunswik's (1955) lens model into the lens model equation:

$$\underbrace{r_a}_{\text{Total Achievement}} = \underbrace{GR_XR_Y}_{\text{Modeled Achievement}} + \underbrace{C\sqrt{1-R_X^2}\sqrt{1-R_Y^2}}_{\text{Unmodeled Achievement}} \tag{4}$$

Connecting the Lens Model and Classical Mediation

Although lens models are a staple for researchers conducting accuracy studies, equivalent models like the classical mediation model are used in this context as well. For instance, Stavrova & Haarmann (2020) fit several accuracy models using various text-based emotions via the Linguistic Inquiry Word Count (LIWC; Tausczik & Pennebaker, 2010) to mediate the relationship between self- and other-rated life satisfaction to identify valid and utilized textual cues. Mediation models offer a structure that is conceptually similar to the lens model, especially when considering multiple mediators. In a multiple mediator model, the relationship between an independent variable (X) and a dependent variable (Y) is transmitted through a set of k mediators ($M_1, M_2, ..., M_k$). Excluding intercepts, the model is given by the following equations. 1) The effect of X on each mediator M_i (Path a_i): $\hat{M}_i = a_i X \quad (\text{for } i = 1, ..., k)$. 2) The effect of all mediators on Y and the direct effect of X on Y (Paths b_i and τ'): $\hat{Y} = \sum_{i=1}^{k} b_i M_i + \tau' X$. 3) The total effect of X on Y (Path τ): $\hat{Y} = \tau X$.

In a way that is analogous to the lens model's decomposition of achievement, the mediation framework decomposes the total effect into a direct effect and the sum of all specific indirect effects. This decomposition is given by the expression:

$$\underbrace{\tau}_{\text{Total Effect}} = \underbrace{\tau'}_{\text{Direct Effect}} + \underbrace{\sum_{i=1}^{k} a_i b_i}_{\text{Indirect Effect}} \tag{5}$$

Where τ is the Total Effect, τ' is the Direct Effect, and $\sum_{i=1}^{k} a_i b_i$ is the Total Indirect Effect. Thus, by fitting a mediation model using a set of cues as mediators, the classic mediation framework can also decompose accuracy (the total effect, τ) into accuracy that is mediated through the various indirect paths ($a_i b_i$), or accuracy that is not captured by the mediators in linear models (the direct path, τ').

A Helpful Heuristic and Unifying Statistic: Percentage of Mediated Accuracy (PoMA)

To find common language between these two frameworks with disparate terminology, and to lay the groundwork for incorporating double machine learning later in this paper, we capitalize on this percentage of mediated accuracy concept, which we will term as PoMA. For the multiple mediation model, it is well known that $\text{PoMA}_{\text{mediation}}$ can be calculated through the product-of-coefficients approach ($\frac{\sum a_i b_i}{\tau}$) or through the difference-in-coefficients approach ($\frac{\tau - \tau'}{\tau}$). The former and the latter are equivalent under OLS but not under non-linear conditions like probit, logistic, or most machine learning models (MacKinnon, 2008). Since the lens model uses a standardized residual correlation (C) and regression models that yield partial coefficients with respect to other cues Z (not equivalent to mediation's a_i or b_i), some scaling has to be done to equate a $\text{PoMA}_{\text{lens}}$ with a $\text{PoMA}_{\text{mediation}}$. The difference-in-coefficients method yields the most promising equivalence.[1]

$$\text{PoMA} = 1 - \frac{\tau'}{\tau} = 1 - \frac{C\frac{\text{SD}(e_Y)}{\text{SD}(e_X)}}{r_a \frac{\text{SD}(Y)}{\text{SD}(X)}} = 1 - \frac{C\sqrt{\frac{1-R_Y^2}{1-R_X^2}}}{r_a} \tag{6}$$

[1]The product-of-coefficients approach doesn't have a simple formula to convert between frameworks, since the lens model uses zero-order regression coefficients (a marginal effect), rather than partial effects yielded by the path b_i. Simply put, the product of marginal coefficients is not equivalent to the product of partial coefficients. In the single mediator model, this inequality is most clearly expressed by the following expression: $\frac{ab_{\text{partial}}}{\tau} \neq \frac{ab_{\text{marginal}}}{\tau} = \frac{GR_X R_Y \frac{\text{SD}(Y)}{\text{SD}(X)}}{r_a \frac{\text{SD}(Y)}{\text{SD}(X)}}$ where b_{marginal} comes from regressing Y on the cue without controlling for X), while traditional mediation partials out the effect of X from the mediator, yielding b_{partial}.

Challenges from High-Dimensional Data

While this percentage-based interpretation provides valuable insights in accuracy studies, applying these models to modern, high-dimensional data presents significant challenges. When the number of potential cues exceeds the number of observations, a common scenario with text embeddings containing hundreds or thousands of features, traditional regression approaches fail. Even when the sample size is not surpassed, the assumption that high-dimensional relationships between cues, perceptions, and judgments are strictly linear and without interaction becomes untenable (for more on known limitations of mediation and equivalent models, see MacKinnon, 2008). These modeling limitations have prevented researchers from leveraging rich, high-dimensional representations of text and other complex stimuli in accuracy research.

Furthermore, using machine learning algorithms introduces a statistical complexity for the classic lens model decomposition. Unlike OLS regression, the predictions and residuals from machine learning models are not guaranteed to be independent, especially under cross-validation (Hastie et al., 2009). This violates a core assumption that allows for the simple decomposition of accuracy into modeled and unmodeled components seen in the classic Lens Model Equation. To maintain mathematical integrity, the equation must be generalized to account for these additional covariance terms, a technical expansion that is detailed in Appendix A.

For our purposes, however, the primary goal remains conceptual: to partition total accuracy into the portion explained by our cues and the portion that remains unexplained. The PoMA statistic continues to serve this purpose well. Therefore, our proposed framework adopts an estimation strategy to achieve a robust estimate of this fundamental split, focusing on the unmodeled relationship after the influence of all textual cues has been partialled out.

Double Machine Learning

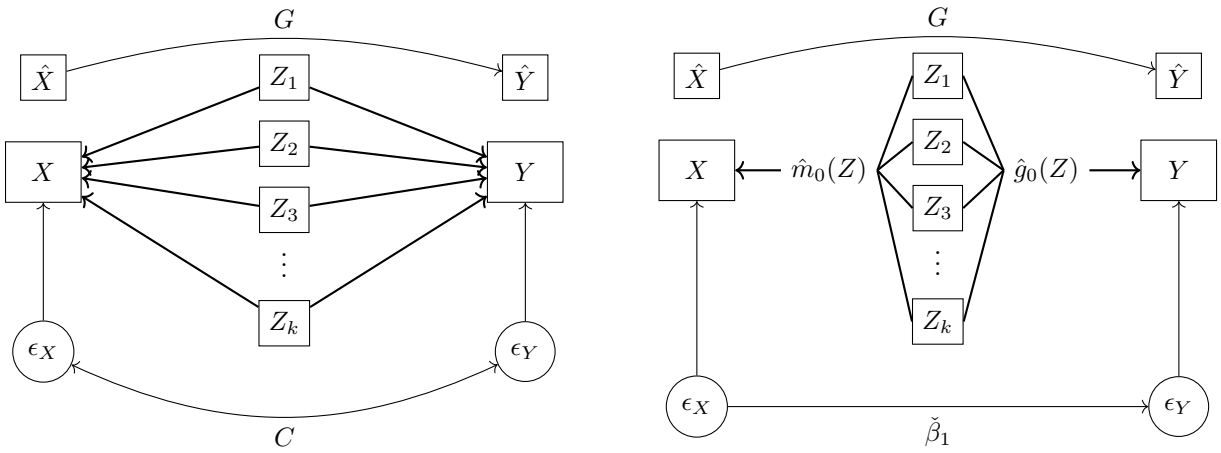
From the domain of econometrics, double machine learning has emerged as a method for partialling out high-dimensional confounding variables (a large numbers of potential confounders, possibly exceeding the sample size) to enable unbiased estimation of regression coefficients (Chernozhukov et al., 2018). Chernozhukov demonstrated that using non-parametric machine learning algorithms m_0 and g_0 (e.g., random forests, neural networks, etc.) to directly partial out high-dimensional confounding variables Z on an independent variable X and dependent variable Y can result in a biased estimate of the slope parameter β_1 (denoted as θ in their paper).

Chernozhukov et al. (2018) optimizes for an alternative score function while using a technique called Neyman orthogonalization to prevent regularization biases from biasing the estimation of β_1, where $e_X = X - \hat{m}_0(Z)$ ($\hat{m}_0$ estimated using an auxiliary sample I^c). By implementing Neyman orthogonalization alongside a cross-fitting procedure (sample-splitting and then swapping the roles of main and auxiliary samples to produce multiple estimates for averaging into a final statistic $\check{\beta}_1$), they found that $\sqrt{n}(\check{\beta}_1 - \beta_1) \xrightarrow{p} 0$. In sum, removal of regularization biases through orthogonalization and cross-fitting made β_1 estimation robust to errors in estimating m_0 and g_0, producing an unbiased and consistent estimator $\check{\beta}_1$ for regressing Y on X while addressing high-dimensional confounders (denoted as $\check{\theta}$ in their paper). Theorem 3.1 and 3.2 in Chernozhukov's paper outline the point estimate and asymptotic variance of $\check{\beta}_1$, translated using our notation below:

$$\check{\beta}_1 = \left(\frac{\frac{1}{n} \sum_{i \in I} e_{X_i} (Y_i - \hat{g}_0(Z_i))}{\frac{1}{n} \sum_{i \in I} e_{X_i}^2} \right) \tag{7}$$

$$\hat{\sigma}^2_{\check{\beta}_1} = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(e_{X_i} e_{Y_i} - \check{\beta}_1 e^2_{X_i} \right)^2}{\left(\frac{1}{n} \sum_{i=1}^{n} e^2_{X_i} \right)^2} \tag{8}$$

Figure 1. Comparison of the Traditional Lens Model and the Double Machine Learning Framework



Note. The left panel illustrates the classic Brunswik lens model, where a validity measure (X) and a judgment (Y) are modeled as linear combinations of k cues (Z). The right panel depicts the Double Machine Learning (DML) framework, which uses machine learning models (m_0 and g_0) to partial out high-dimensional confounders (Z) to estimate the direct effect ($\check{\beta}_1$) of X on Y.

Connecting DML to the Lens and Mediation Model

Although lens and mediation models are not necessarily concerned about confounders, DML does share many similarities with the lens model in terms of structure. Both frameworks use a set of variables Z to predict an X and Y variable. Both yield a direct effect akin to a regression coefficient (Chernozhukov's $\check{\beta}_1$ and the lens model's $C\sqrt{\frac{1-R^2_Y}{1-R^2_X}}$, comparable to the mediation framework's τ').

We can apply the difference-in-coefficients approach to get an estimate of PoMA for the double machine learning approach, and although these estimates will not yield the same values as

the product-of-coefficients approach under non-linear models, this coefficient can be easily interpreted as the reduction in total effect after using cross-fitted machine learning methods to control for a set of high-dimensional confounders. Thus, PoMA_{DML} would be interpreted as the proportion of accuracy reduced by controlling cues with machine learning. In the causal mediation framework, this is a function of what is referred to as the controlled direct effect (not to be confused with the natural direct effect; Imai et al., 2010). By establishing a baseline total effect of β_1 (a simple regression coefficient from regressing Y on X), we can express this proportion with the following equation:

$$\text{PoMA}_{\text{DML}} = 1 - \frac{\check{\beta}_1}{\beta_1} \tag{9}$$

The Present Study

The study of perceptual accuracy - how well one person can infer the traits or states of another - has long been a cornerstone of psychological science. However, traditional methods are often ill-equipped to handle the high-dimensional, unstructured data, such as text or images, that are increasingly central to modern research. To address this challenge, the present study introduces the DML-LM, a method for deconstructing judgment processes in flexible and high-dimensional settings, enabling the investigation of both human and artificial perceivers of complex data types.

Our primary aims are designed to first establish the performance of various text representations, learning algorithms, and dimensionality reduction techniques, and then to apply a balanced approach (in terms of parsimony and predictiveness) to a rich, real-world dataset. This involves three key steps:

1. Which text representation is most effective? First, we must first identify the highest-quality inputs. To do this, we leverage our largest available dataset ($n = 9,513$) to compare the efficacy of different text representation methods, from traditional LIWC dictionaries to modern language model embeddings (all-MiniLM-L6-v2 and NV-Embed-v2). This comparison, focused on the AI Accuracy model, allows us to determine which method best captures the cues relevant to perceptual accuracy.

2. What is the optimal analytical pipeline? We then conduct a comparison of 45 analytical models, systematically contrasting various machine learning algorithms (OLS, Ridge, Lasso, XGBoost, and Random Forest) and dimensionality levels. This allows us to identify the most effective and robust combination of techniques for predicting outcomes and maximizing the explained accuracy (PoMA) within the DML-LM.

3. How do we apply the DML-LM to deconstruct judgment? With an optimal pipeline established, we apply the full framework twice to analyze 9,513 aspirational essays written by 11-year-olds in 1969. We conduct the core analyses of AI Accuracy (on the full dataset), human accuracy (on a subset with complete human ratings, $n = 547$), and their consensus to deconstruct how humans and AI perceives social class from text. This serves as a powerful case study for uncovering the specific linguistic mechanisms that drive algorithmic bias and the "kernel-of-truth" phenomenon.

By achieving these aims, we demonstrate the utility of the DML-LM as a comprehensive and flexible toolkit for dissecting the perceptual models of both human and artificial agents. This research paves the way for a new generation of studies into person perception, stereotypes, and algorithmic fairness that are more accurate and transparent.

Methods

Data Source

This study utilizes data from the National Child Development Study (NCDS), a highly influential and ongoing longitudinal cohort study that has followed the lives of all individuals born in a single week of March 1958 across England, Scotland, and Wales. The richness and breadth of the NCDS make it an unparalleled resource for life-course research. For the present study, we analyzed 9,513 aspirational essays collected in 1969 when the cohort members were 11 years old. Participants were given the prompt: "Imagine you are now 25 years old. Write about the life you are leading, your interests, your home life, and your work." The resulting essays averaged 197.7 words in length ($SD = 105.6$). These texts provide a unique window into the hopes and societal understandings of children from that era. Crucially, extensive demographic information, including the social class validity used in this study, was collected concurrently through home interviews with parents and standardized assessments with the children, providing a rich context for the essay data. All study procedures were approved by the University of British Columbia's Behavioral Research Ethics Board (BREB #H20-01189).

Measures

Social Class Validity

The ground-truth measure of each child's social class was based on the Registrar General's Social Classes (RGSC) classification of their father's occupation, as recorded in 1969. The RGSC was the United Kingdom's official social classification system for much of the 20th century and provides a reliable indicator of socioeconomic standing for the period. This five-point ordinal scale ranks occupations by their perceived standing in the community: Class I (professional

occupations, e.g., doctor, lawyer), Class II (managerial/technical, e.g., teacher, manager), Class III (skilled, both manual and non-manual, e.g., carpenter, clerk), Class IV (partially skilled, e.g., farm worker, bus conductor), and Class V (unskilled, e.g., laborer, cleaner). For our analyses, these classes were reverse-coded (1 to 5) so that higher values indicate higher social class ($M = 2.92, SD = 0.89$).

Social Class Judgment

To generate a set of perceptual judgments, we used a leading quantized, open-source large language model (AWQ Qwen 2.5 32B; Qwen et al., 2024) to rate each of the 9,513 essays. The use of a quantized model allows for efficient processing on local hardware of the large dataset without a prohibitive loss in performance (Lin et al., 2023). The AI was prompted with an adapted version of the well-validated MacArthur Scale of Subjective Social Status (see Appendix B). This prompt frames the task in intuitive terms, asking the AI to place the essay writer's family on a 10-point ladder representing their standing in society, from the "worst off" at the bottom to the "best off" at the top. The resulting AI ratings demonstrated substantial consistency with averaged ratings from 10 human raters (randomly selected from an overall pool of 600 human raters) who performed the same task ($r(545) = .65$), suggesting the AI's perceptual model shares significant variance with human social perception. The achievement accuracy, or the zero-order correlation between the AI's judgments and the actual social class validity, was modest but significant ($r(9511) = .24, p < .001$), indicating the AI was capturing some signal of social class from the text, and to a similar magnitude as human perceivers as well ($r(545) = .26, p < .001$). See Appendix C for details on our 600 human perceivers and Appendix D for the distributions of social class judgments and the validity measure.

Text Representation Methods

To quantify the textual cues within the essays that could potentially mediate the AI's accuracy, we compared three distinct and representative methods, spanning from traditional to state-of-the-art.

Linguistic Inquiry and Word Count. Representing a classic, theory-driven approach, we used the LIWC-22 software to analyze the essays. This method works by counting the percentage of words in a text that fall into 118 predefined psychological and linguistic categories (e.g., analytic, affect, power; Boyd et al., 2022). The strength of LIWC lies in its interpretability, but its reliance on fixed dictionaries means it cannot capture semantic meaning, context, or novel language use. A small percentage (0.07%) of data points were invalid or missing upon exiting LIWC-22, primarily involving the Tone variable. We used scikit-learn's Iterative Imputer to resolve these data points to refrain from using listwise deletion.

All-MiniLM-L6-v2 Embeddings. Using knowledge distillation (training a language model by having it learn from the self-attention module of a larger language model), Wang et al. (2020) created the popular sentence-transformer model, MiniLMv2 - capable of retaining comparable performance to the teacher model, BERT-base. We used a six-layer version of this lightweight embedding model to generate a 384-dimensional vector embedding for each essay. Consisting of only 22.7M parameters, this model is capable of running on most modern machines locally, while still capturing the semantic meaning of the entire essay in a dense numerical vector. This model is optimized for speed, making it a practical choice for many applications, and provides a strong baseline for the performance of compact language models.

NV-Embed-v2 Embeddings. To assess the capabilities of a larger, state-of-the-art model, we also generated 4,096-dimensional vector embeddings for each essay using NV-Embed-v2, a 7.85B

parameter embedding model (Lee et al., 2025). The larger parameter count, along with a novel latent attention mechanism for better pooled embedding output provide the theoretical capacity to capture more subtle and complex nuances of language, potentially leading to higher predictive accuracy at the cost of increased computational resources.

By including these three methods, our analysis is designed to span the full spectrum of text representation, from classic, interpretable dictionaries to efficient, modern embeddings and large-scale, high-performance models.

Data Analytic Plan

Our analytical approach sought to systematically uncover a model that balances parsimony and predictive power. This involved three major phases: first, establishing an optimal methodological pipeline; second, applying the full DML-LM framework to the data; and third, performing supplementary and interpretive analyses to validate and understand the results.

Phase 1: Establishing the Optimal Analytical Pipeline

It was necessary to identify the most effective combination of text representations and learning algorithms given many permutations. To do this, we conducted a 45 model comparison crossing a set of 3 text features, 5 learning algorithms, and 3 dimensionalities. This comparison leveraged our largest available dataset ($n = 9,513$) consisting of AI Accuracy data, with performance robustly estimated using a five-fold cross-validation procedure for all models to ensure generalizability. Our three distinct text representation methods (LIWC dictionaries, all-MiniLM-L6-v2 embeddings, and NV-Embed-v2 embeddings) were evaluated under five learning algorithms (OLS, Ridge Regression, Lasso Regression, XGBoost, and Random Forest) chosen to span a spectrum of complexity and assumptions. These were applied to the data at three levels of dimensionality (Full dimensionality, 200 principal components, and the top six

most predictive principal components) to examine how model performance is retained as models move towards parsimony.

Learning Algorithms

The predictive performance of several statistical learning algorithms was evaluated. These models were chosen to span a range of functional forms and complexities, from a simple linear baseline to powerful, non-linear ensemble methods.

Ordinary Least Squares. OLS was included as a simple, interpretable baseline. It identifies the optimal linear model by minimizing the sum of the squared differences between observed and predicted outcomes (i.e., the sum of squared residuals). While its coefficients are easily interpretable, the reliability of OLS depends on strict statistical assumptions, including linearity in parameters, independence of errors, and homoscedasticity (Greene, 2003). In high-dimensional settings, where the number of features is large relative to the number of observations, OLS models are highly susceptible to overfitting and exhibit high variance, providing a crucial benchmark against which the performance gains from more complex models can be measured (Hastie et al., 2009).

Ridge and Lasso Regression. To directly address the challenges of high-dimensionality within a linear framework, two regularized regression models were included. Both methods augment the OLS loss function with a penalty term to constrain the size of the coefficients, thereby reducing model variance. To ensure an optimal level of regularization was applied, we used cross-validating versions of each model. Ridge regression, which uses an L_2 penalty ($\lambda \sum_{j=1}^{p} \beta_j^2$) to handle multicollinearity (Hoerl & Kennard, 1970), was implemented with RidgeCV, which uses efficient leave-one-out cross-validation to select the best regularization strength. In contrast, Lasso regression, which uses an L_1 penalty ($\lambda \sum_{j=1}^{p} |\beta_j|$) to perform automatic feature selection

(Tibshirani, 1996), was tuned via an internal 5-fold cross-validation through LassoCV.

XGBoost. As a powerful, non-linear model, Extreme Gradient Boosting (XGBoost) was included. XGBoost is a highly efficient implementation of the gradient boosting framework that sequentially builds decision trees to correct prior errors (Chen & Guestrin, 2016). To provide a fair, out-of-the-box baseline, the model was run with the default hyperparameters from the xgboost library without specific tuning, allowing for a direct comparison against other standard models without extensive, model-specific optimization. For computational efficiency, the GPU tree construction method was used.

Random Forest. A second non-linear ensemble method, Random Forest, was also included. A Random Forest operates by constructing a multitude of decision trees at training time and outputting their mean prediction (Breiman, 2001). We used an implementation with 100 trees and allowed trees to grow to their maximum depth, since 100 trees is a robust number for stabilizing the ensemble's predictions, and using deep, fully-grown trees is standard for Random Forest, as the model's variance and overfitting are controlled by aggregating across the decorrelated trees created through bagging and random feature subspaces (Hastie et al., 2009).

Feature Sets of Decreasing Dimensionality

To assess model performance under varying levels of complexity, three distinct feature sets were created. The first set utilized the full dimensionality of the features to establish a performance baseline. The second set was a reduced feature space, where Principal Component Analysis (PCA) was employed to reduce the original features to 200 components for the embedding methods, and 30 components for LIWC. To ensure features with larger scales did not disproportionately influence the analysis, particularly regarding the coefficient shrinkage in lasso and ridge regression, all data was standardized before and after the PCA transformation. This

number of components was chosen as a way to reduce the subspace to a reasonable dimensionality and level of explained variance (94.7% for MiniLM, 73.7% for NV-Embed, and 61.7% for LIWC), thereby balancing signal retention with noise reduction (Jolliffe, 2016). Finally, a highly parsimonious set with only 6 components was created to test model performance under conditions of extreme simplicity, a choice justified by its analogy to traditional psychological studies that rely on a small handful of core predictors (Cooksey, 1996).

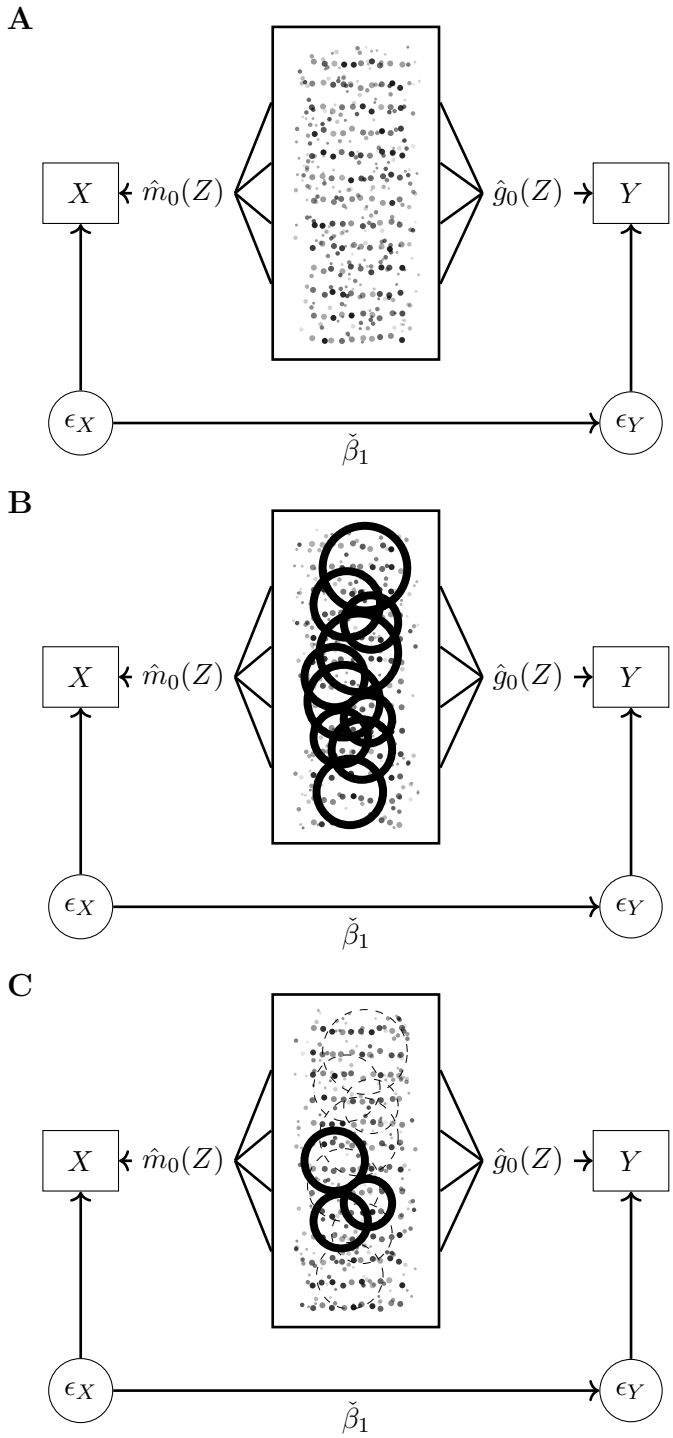
Performance Evaluation

Performance was systematically evaluated using metrics central to the DML-LM, including judgmental consistency, environmental predictability, and PoMA. For robustness and comparability with the classic literature, these were supplemented with traditional lens model statistics, such as the matching coefficient and the residual correlation. To ensure our estimates are generalizable and not susceptible to overfitting, all models were evaluated using a five-fold cross-validation procedure. We report the mean cross-validated R^2 across all folds, supplemented by the simple value-prediction correlation (e.g., $\text{Corr}(X, \hat{X})$), which offers a more stable and intuitive measure of predictive accuracy, particularly in settings where R^2 values can be negative (Hastie et al., 2009).

Phase 2: Applying the DML-LM to Decompose Human-AI Accuracy

Upon establishing a model pipeline that balanced parsimony, computational efficiency, and predictive power, we applied the DML-LM across three analyses. First, we used the DML-LM to fit an AI-accuracy model utilizing the full dataset ($n = 9,513$), regressing AI ratings on the validity measure. Second, we fit a human-accuracy model utilizing the subset of data with complete human ratings ($n = 547$), regressing human ratings on the validity measure. Finally, we

Figure 2. Conceptual Illustration of Dimensionality Reduction Techniques for the DML-LM



Note. A conceptual illustration of how different approaches handle the high-dimensional cues (Z). **Panel A** represents the full, unstructured feature set. **Panel B** illustrates dimensionality reduction via Principal Component Analysis (PCA), where the original features are combined into a smaller set of broader components. **Panel C** depicts a feature selection approach like Lasso, which identifies and retains a sparse subset of the most important original features.

applied the DML-LM to fit a human-AI consensus model by regressing AI ratings on human ratings, an approach analogous to studies that reconcile data from different sources like parents and teachers (Fleece & Teglasi, 2024). This allowed us to understand the model's shared cue utilization and the extent to which unmeasured cues were shared between human and AI ratings. The same performance metrics from first analytic phase are reported across all three applications of the DML-LM.

Phase 3: Interpretive Analysis of Core Cues

To move from prediction to explanation, we conducted a two-part interpretive analysis involving post-lasso coefficients and topic modeling to understand the magnitude of cue utilization and validity, as well as the substance of specific principal components.

Post-Lasso Coefficients. First, to obtain non-shrunken estimates of the specific weight of each principal component in our final model, we employed a post-Lasso OLS approach. While the main Lasso procedure is ideal for feature selection and building robust predictive models, its coefficients are intentionally shrunk toward zero and are thus biased, a property that improves out-of-sample prediction but makes the coefficients themselves unsuitable for direct inferential interpretation (Tibshirani, 1996). To generate estimates suitable for inference, the post-selection OLS method refits a standard OLS model using only the subset of features that were selected by the Lasso algorithm. This technique yields asymptotically unbiased coefficients and permits valid statistical inference for the selected predictors, overcoming the challenges posed by Lasso's original, shrunken estimates (Belloni & Chernozhukov, 2013; Hastie et al., 2009). The validity of this naïve post-selection approach is further supported by recent work, where Zhao et al. (2021) established that with sufficiently large samples, the variable set selected by the lasso can be considered deterministic with high probability. This finding provides an asymptotic justification for applying

standard inferential tools directly to the post-selection OLS model to achieve valid inference.

BERTopic Model for Topic-Variable Associations. To translate the opaque principal components into interpretable constructs, we first selected the 15 most impactful components for analysis. This selection was based on the magnitude of their post-Lasso OLS feature weights in a model predicting a composite of the validity measure, AI ratings, and human ratings. Next, using a methodology inspired by BERTopic (Grootendorst, 2022), we generated topic-based descriptions for the positive and negative poles of these 15 components. We began by using UMAP to reduce the essay embeddings to three dimensions, which we then clustered into hundreds of high-resolution topics with HDBSCAN . For each topic cluster, we extracted representative keywords using a class-based TF-IDF algorithm and calculated the cluster's mean score on our main variables (the 15 PCs, validity, and AI/human judgments). A topic was defined as "high" or "low" on a variable if its mean score fell in the top or bottom quartile of all topic means, respectively. Finally, to visualize these relationships (see Figure 5), we plotted the topic clusters in the 3D UMAP space and drew lines connecting the highest- and lowest-scoring cluster for each PC. To clarify the thematic content of each PC's poles, we also generated word clouds based on the vocabulary of its six most extreme topic clusters (the top three and bottom three).

Results

The Macro View: Which Models Performed Best?

Language Model Embeddings Surpass Dictionary Methods

A clear performance hierarchy emerged among the text representation methods. State-of-the-art language model embeddings significantly outperformed the traditional LIWC dictionary-based approach across all key metrics. The optimal model, combining NV-Embed features with Lasso or Ridge regression and 200 principal components, achieved the highest

cross-validated environmental predictability ($R^2_X = .11$) and judgmental consistency ($R^2_{AI} = .56$). The smaller MiniLM embeddings also proved highly effective, performing about four-fifths of the NV-Embed embedding performance ($R^2_X = .09$, $R^2_{AI} = .44$). Both embedding methods substantially exceeded the performance of the best LIWC-based model ($R^2_X = .06$, $R^2_{AI} = .34$).

Regularized Linear Models Outperform Non-Linear and Unregularized Approaches

Across all feature sets, regularized linear models (Lasso and Ridge) demonstrated the most robust and balanced performance. In contrast, the more flexible, non-linear XGBoost model consistently failed to generalize to the real-world criterion, producing negative environmental predictability scores (e.g., R^2_X = -0.03). The value-prediction correlations provide crucial context for this result: despite the negative R^2, the correlation between XGBoost's predictions and actual social class was still positive (e.g., $\text{Corr}(X, \hat{X}) = .22$ for NV-Embed). This indicates that the model's predictions were directionally correct but had such high variance that they performed worse than a simple baseline model predicting the mean. At the other extreme, unregularized OLS regression on high-dimensional features suffered from severe overfitting, producing a catastrophic R^2_X of -0.94 with the full NV-Embed feature set. These results illustrate the utility of combining regularization with high-quality embeddings for producing optimal perception models.

Table 1. Performance Metrics for Five Machine Learning Algorithms Across Text Representations

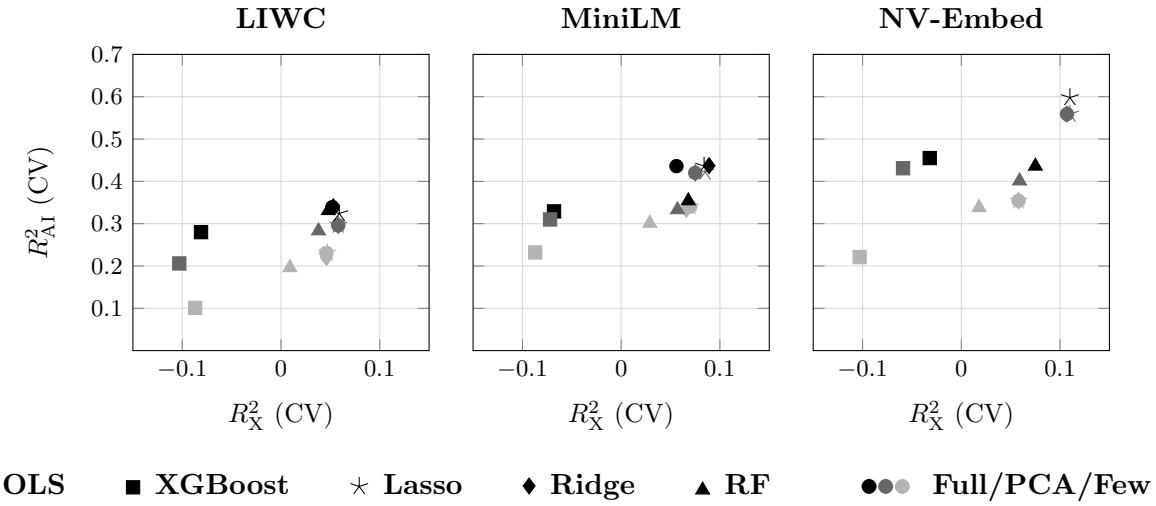
Learner	Text	Dim	$\check{\beta}$	SE	p	R^2_{AI}	R^2_{X}	G	C	$r_{X,\hat{X}}$	$r_{AI,\hat{AI}}$	PoMA
OLS	None	N/A	0.368	0.015	<0.001	N/A	N/A	N/A	N/A	N/A	N/A	0.00%
NV-Embed Models												
OLS	NV-Embed	Full	0.033	0.010	<0.001	0.254	−0.936	0.243	0.046	0.147	0.647	91.00%
OLS	NV-Embed	PCs	0.053	0.011	<0.001	0.559	0.107	0.708	0.049	0.332	0.748	85.64%
OLS	NV-Embed	Few	0.166	0.013	<0.001	0.354	0.058	0.945	0.131	0.242	0.595	54.77%
XGBoost	NV-Embed	Full	0.058	0.012	<0.001	0.455	−0.032	0.417	0.052	0.219	0.677	84.21%
XGBoost	NV-Embed	PCs	0.070	0.012	<0.001	0.431	−0.059	0.400	0.063	0.187	0.659	80.85%
XGBoost	NV-Embed	Few	0.187	0.013	<0.001	0.221	−0.103	0.428	0.145	0.111	0.497	49.09%
Lasso	NV-Embed	Full	0.038	0.011	<0.001	0.598	0.110	0.725	0.037	0.333	0.773	89.60%
Lasso	NV-Embed	PCs	0.054	0.011	<0.001	0.559	0.110	0.744	0.050	0.332	0.748	85.28%

(continued on next page)

Learner	Text	Dim	$\check{\beta}$	SE	p	R^2_{AI}	R^2_{X}	G	C	$r_{X,\hat{X}}$	$r_{AI,\hat{AI}}$	PoMA
Lasso	NV-Embed	Few	0.166	0.013	<0.001	0.354	0.058	0.945	0.131	0.242	0.595	54.76%
Ridge	NV-Embed	Full	0.038	0.010	<0.001	0.353	−0.678	0.270	0.040	0.164	0.677	89.70%
Ridge	NV-Embed	PCs	0.053	0.011	<0.001	0.559	0.107	0.708	0.049	0.332	0.748	85.69%
Ridge	NV-Embed	Few	0.166	0.013	<0.001	0.354	0.058	0.945	0.131	0.242	0.595	54.76%
Random Forest	NV-Embed	Full	0.122	0.012	<0.001	0.436	0.075	0.712	0.101	0.274	0.660	66.93%
Random Forest	NV-Embed	PCs	0.147	0.013	<0.001	0.401	0.059	0.652	0.120	0.243	0.633	59.92%
Random Forest	NV-Embed	Few	0.161	0.013	<0.001	0.338	0.018	0.631	0.128	0.134	0.581	56.23%
MiniLM Models												
OLS	MiniLM	Full	0.099	0.012	<0.001	0.436	0.056	0.666	0.080	0.269	0.661	73.12%
OLS	MiniLM	PCs	0.114	0.013	<0.001	0.420	0.075	0.732	0.091	0.284	0.649	69.08%
OLS	MiniLM	Few	0.148	0.013	<0.001	0.340	0.070	0.959	0.114	0.266	0.583	59.75%
XGBoost	MiniLM	Full	0.087	0.013	<0.001	0.329	−0.068	0.404	0.071	0.180	0.585	76.34%
XGBoost	MiniLM	PCs	0.095	0.013	<0.001	0.310	−0.072	0.403	0.077	0.171	0.568	74.06%
XGBoost	MiniLM	Few	0.144	0.013	<0.001	0.232	−0.087	0.456	0.111	0.146	0.508	60.95%
Lasso	MiniLM	Full	0.105	0.013	<0.001	0.436	0.084	0.772	0.087	0.291	0.660	71.49%
Lasso	MiniLM	PCs	0.115	0.013	<0.001	0.422	0.085	0.810	0.094	0.293	0.649	68.80%
Lasso	MiniLM	Few	0.148	0.013	<0.001	0.340	0.070	0.959	0.114	0.266	0.583	59.80%
Ridge	MiniLM	Full	0.108	0.013	<0.001	0.437	0.089	0.789	0.089	0.299	0.661	70.72%
Ridge	MiniLM	PCs	0.110	0.013	<0.001	0.420	0.075	0.732	0.091	0.284	0.649	69.99%
Ridge	MiniLM	Few	0.165	0.013	<0.001	0.334	0.066	0.923	0.127	0.258	0.578	55.24%
Random Forest	MiniLM	Full	0.145	0.013	<0.001	0.354	0.068	0.685	0.113	0.261	0.595	60.65%
Random Forest	MiniLM	PCs	0.152	0.013	<0.001	0.333	0.057	0.690	0.117	0.239	0.577	58.73%
Random Forest	MiniLM	Few	0.154	0.013	<0.001	0.301	0.029	0.649	0.118	0.170	0.549	58.20%
LIWC Models												
OLS	LIWC	Full	0.173	0.013	<0.001	0.339	0.052	0.770	0.136	0.236	0.583	52.96%
OLS	LIWC	PCs	0.200	0.014	<0.001	0.296	0.058	0.826	0.151	0.242	0.544	45.51%
OLS	LIWC	Few	0.231	0.014	<0.001	0.231	0.046	0.907	0.167	0.217	0.481	37.29%
XGBoost	LIWC	Full	0.133	0.013	<0.001	0.280	−0.081	0.412	0.106	0.158	0.547	63.78%
XGBoost	LIWC	PCs	0.151	0.013	<0.001	0.206	−0.103	0.385	0.116	0.139	0.486	59.00%
XGBoost	LIWC	Few	0.189	0.014	<0.001	0.101	−0.087	0.431	0.136	0.136	0.383	48.48%
Lasso	LIWC	Full	0.189	0.013	<0.001	0.324	0.059	0.817	0.145	0.243	0.569	48.58%
Lasso	LIWC	PCs	0.201	0.014	<0.001	0.297	0.058	0.829	0.151	0.241	0.545	45.45%
Lasso	LIWC	Few	0.231	0.014	<0.001	0.231	0.047	0.905	0.167	0.217	0.481	37.25%
Ridge	LIWC	Full	0.174	0.013	<0.001	0.340	0.053	0.776	0.136	0.237	0.584	52.58%
Ridge	LIWC	PCs	0.200	0.014	<0.001	0.296	0.058	0.826	0.151	0.242	0.544	45.62%
Ridge	LIWC	Few	0.236	0.014	<0.001	0.220	0.046	0.899	0.170	0.217	0.469	35.87%
Random Forest	LIWC	Full	0.167	0.013	<0.001	0.331	0.048	0.711	0.130	0.219	0.575	54.64%
Random Forest	LIWC	PCs	0.182	0.014	<0.001	0.283	0.038	0.644	0.138	0.195	0.532	50.48%
Random Forest	LIWC	Few	0.217	0.014	<0.001	0.196	0.009	0.611	0.157	0.095	0.443	41.01%

Note. This table compares five machine learning algorithms (OLS, XGBoost, Lasso, Ridge, Random Forest) based on text representation (NV-Embed, MiniLM, LIWC), learner type, and dimensionality reduction approach. Performance is evaluated using cross-validated (CV) R-squared values for AI judgments (R^2_{AI}) and the validity measure (R^2_X), lens model statistics (G, C), value-prediction correlations (r), and the Percentage of Mediated Accuracy (PoMA). Dim = dimensionality/features; Full = all features; PCs = principal components (200 for embeddings, 30 for LIWC); Few = top 6 features selected by feature importance.

Figure 3. Model Performance in Predicting AI Judgments and the Social Class Validity



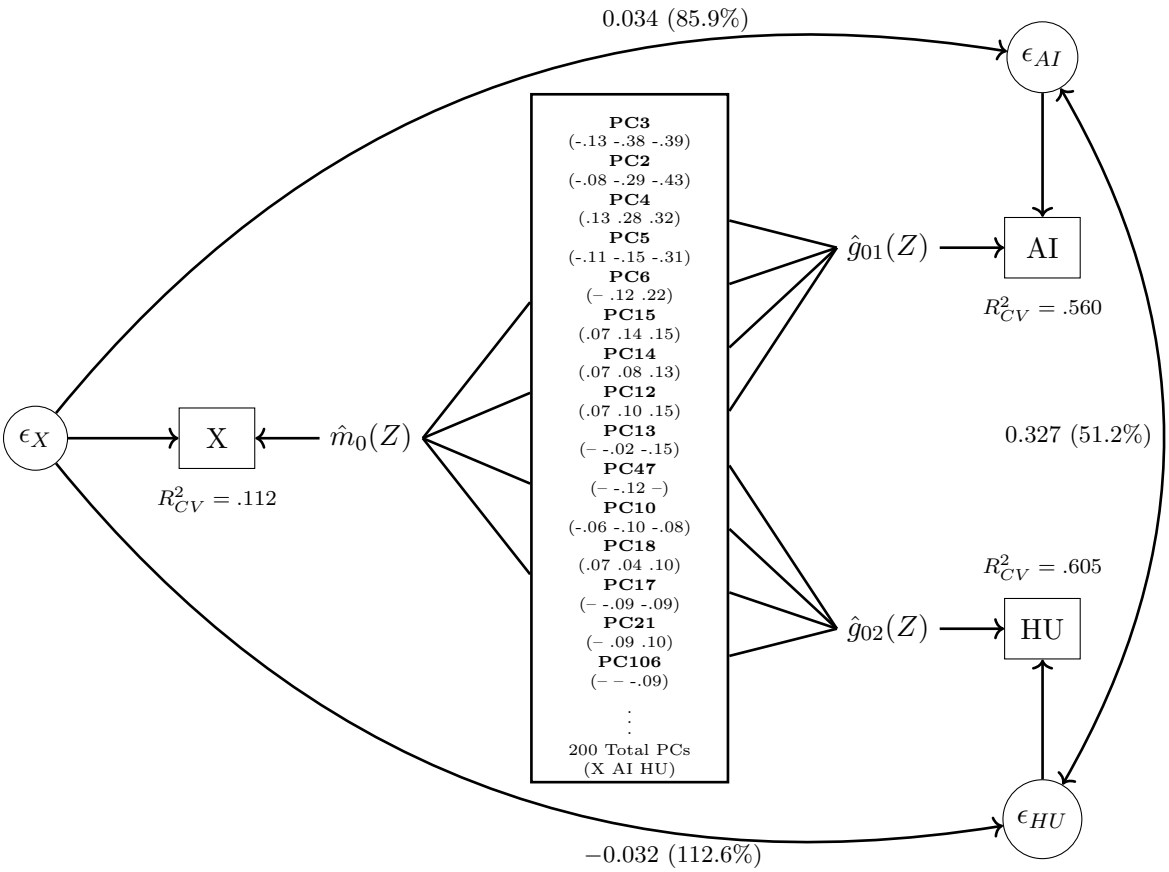
Note. The scatterplots compare the cross-validated (CV) judgmental consistency (R^2_{AI}) against the environmental predictability (R^2_X) for five learning algorithms (OLS, XGBoost, Lasso, Ridge, Random Forest [RF]) across three text representation methods (LIWC, MiniLM, NV-Embed). Marker colour corresponds to the dimensionality of the features used, where black is the full dimensionality, dark grey is the complete set of 200 PCs (or 30, in the case of LIWC), and light grey is the set of top 6 PCs. Two data points representing models with full NV-Embed features were omitted to improve axis scaling: OLS ($R^2_X = -0.94$) and Ridge ($R^2_X = -0.68$).

Meso-Level Analysis: The Trade-off Between Parsimony and Explanatory Power

The Impact of Dimensionality on Predictive Performance

The analysis revealed a trade-off between model parsimony and explanatory power. For instance, reducing the 4,096 NV-Embed dimensions to 200 PCs often brought minimal performance reductions (and in OLS, it drastically improved the overfitting issue). However, by selecting only the top 6 most predictive components, performance was reduced by a noticeable amount. For instance, R^2_{AI} fell by 39% for NV-Embed, 21% for MiniLM, and 20% for LIWC when comparing models using lasso regression with the full set versus top six PCs. Thus, the subsequent analyses opted to analyze the full set of 200 components, although it was insightful to discover that six components alone can capture the majority of the variance explained.

Figure 4. The DML-LM Decomposing AI and Human Judgment



Note. The model displays the results of the DML-LM analysis using 200 Principal Components (PCs) as cues (Z). The diagram shows the variance in the validity (X) explained by the cues ($R^2_{CV} = .112$), the variance in AI and Human (HU) judgments explained by cues ($R^2_{CV} = .560$ and .605, respectively), and the direct effects after accounting for the cues. The percentages represent the Proportion of Mediated Accuracy (PoMA). The values in parentheses below each Principal Component (e.g., PC3) correspond to the post-LASSO coefficients for the validity measure (X), AI judgments, and Human (HU) judgments, respectively.

By applying the NV-Embed 200 component model using post-lasso learners onto the human-perception subset of the data, we see that human perceptions ($R^2_{HU} = .61$) are explained comparably, even marginally superior to the AI perceptions ($R^2_{AI} = .56$), and then the validity measure remain weakly explained ($R^2_X = .11$). Percentage of mediated accuracy was very large for both human and AI perceptions, at 112.6% and 85.9% respectively. Importantly, the direct effect between humans perception of social class and the social class validity became non-significant after controlling for the cues ($\check{\beta} = -0.03, SE = .02, z = -1.45, 95\%\text{CI} = [-0.076, 0.011], p = .15$).

For AI perception, the direct effect was small but significant ($\check{\beta} = 0.03, SE = .01, z = 4.63, 95\%\text{CI} = [0.02, 0.05], p < .001$) Furthermore, the covariation between humans and AI perception was reduced by 51.2% by controlling for these 200 principal components, indicating that the current cue set well explains the cues that used by both humans and AI, but do not fully explain their shared use. On a similar note, the policy similarities were fairly high between humans and AI ($G = .88$), and were a touch lower for the AI and humans in terms of the validity measure (G = .75 and .71 respectively). On a similar note to the PoMAs, only small amounts of residual correlation remains ($C = .05$ and $C = .01$ for AI and humans respectively).

Table 2. Optimal DML-LM using Post-Lasso Regression to Decompose Human and AI Social Class Judgment

Panel A: Total Effects (OLS)					
Path	β	*SE*	t	95% CI	n
X → AI	.239	.010	24.02***	[.220, .259]	9,513
X → HU	.255	.041	6.16***	[.174, .336]	547
HU → AI	.669	.032	21.02***	[.607, .732]	547
Panel B: Direct Effects (DML)					
Path	$\check{\beta}$	*SE*	z	95% CI	PoMA
X → AI	.034	.007	4.63***	[.019, .048]	85.9%
X → HU	-.032	.022	-1.45	[-.076, .011]	112.6%
HU → AI	.327	.061	5.35***	[.207, .446]	51.2%
Panel C: Model Performance					
Outcome	Feat.	R^2_{CV}	R^2_{IS}	C	G
X	160/200	.112	.145	—	—
AI	193/200	.560	.580	.049	.749
HU	131/200	.605	.798	.010	.709
HU–AI Relationship				.156[a]	.877[b]

Note. X = validity measure; AI = AI judgment; HU = human judgment. Panel A: OLS total effects. Panel B: DML with LASSO-selected PCs from 186 language features. PoMA = proportion of mediated accuracy. Panel C: post-LASSO estimates with PCs from 4,096-dim NV-Embed. Feat. = number of features selected/total; CV = cross-validation; IS = in-sample. ***p < .001.
[a] HU-AI residual correlation.
[b] HU-AI policy similarity.

Micro-Level Analysis: Deconstructing Judgment in the DML-LM

To understand the cue utilization and validity of the judgment processes of the AI and human raters, we investigated the post-lasso cue utilization and validity coefficients. In particular, the top 9 coefficients for each of the validity, human, and AI measures with the largest magnitude were rank ordered, selected, and then visualized, resulting in a set of 15 unique principal components.

Qualitative Interpretation of Valid and Utilized Components. Components 2, 3, 4, and 5 were the variables with largest post-lasso coefficients for all three of social class validity, AI-perceived social class, and human-perceived social class. Based on the Extreme-3-Topic-Word-Clouds, PC2 and 3 appear to measure misspellings (e.g., wood, wen, wont, woud, illegible fiche), and also football, associated with lower social class validity and perception, versus teaching (e.g., teacher, teaching, class), is associated with higher perceived social class.

PC4 appears to differentiate between traditionally male and female interests. At the high end of PC4, we see football again, which is paradoxically associated with both low and high perceived social class (e.g., manchester, cricket, footballer topic was high in perceived social class, but scotland, football, play, was associated with lower social class perceptions). For traditionally female interests, we see nursing, hospital ward, female, daughter, babies, doctor, and so on. These topics were associated with higher perceived social class. Although PC4 globally was related to social class validity, these topics in particular were primarily related to perceptions of social class.

At the low end of PC5, we see topics that are associated with high levels of social class involving planes or being an air hostess, whereas staying home and being "maried" (poor spelling), with boys or girls and children was associated with lower perceived social class. Thus, this component appears to contrast stay-at-home responsibilities versus being an airline professional, and paradoxically, even though there were illegible fiches involved, university was also

mentioned, and thus was still associated with higher levels of social class validity and perception.

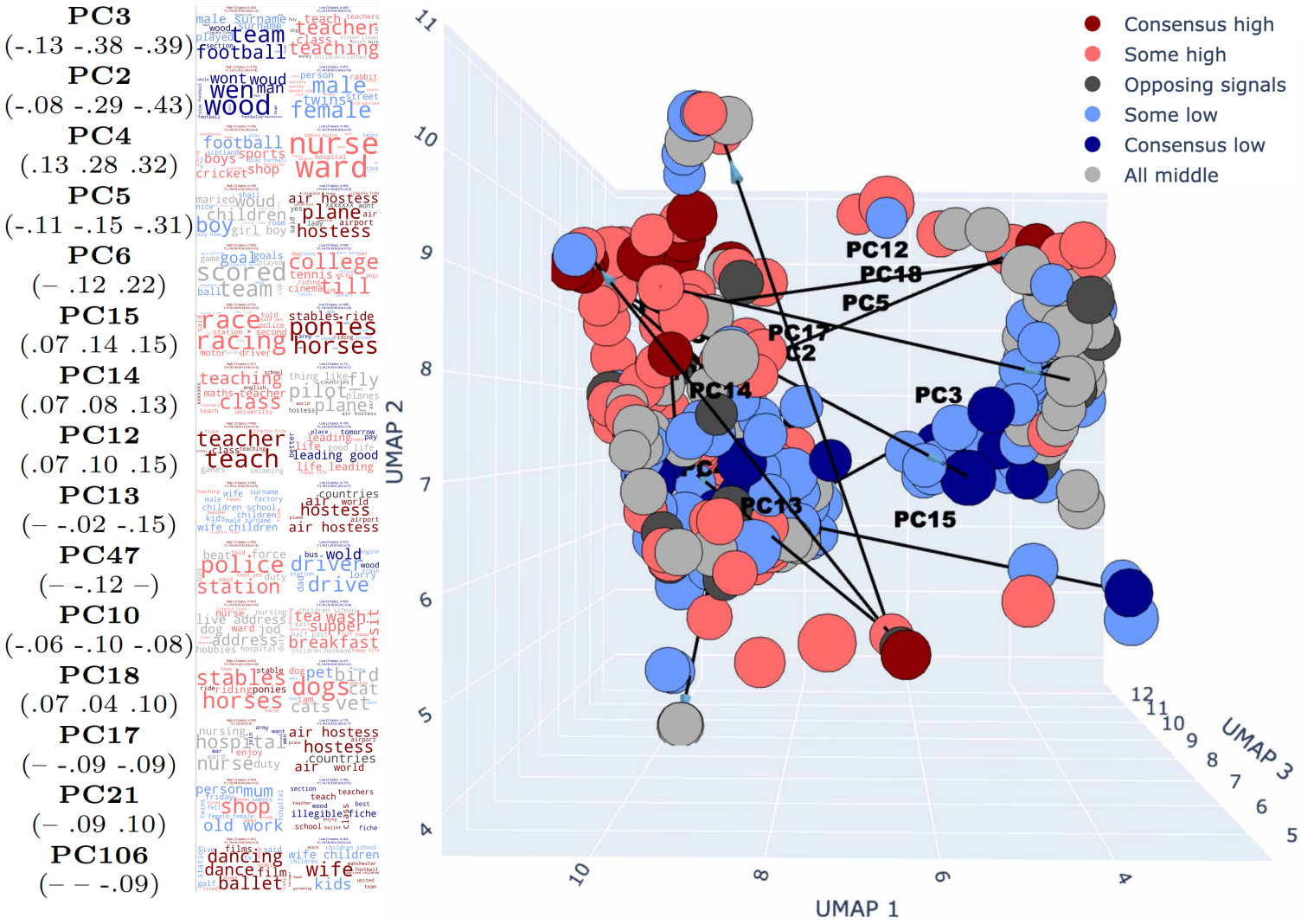
Other interesting themes that emerge include college, tennis, racing, ponies, stables, horses, police, meals, dogs, dancing, ballet, and films for instance, appear to be associated with higher levels of social class, whereas goals (perhaps related to football), and a symphony of basic vocabulary (male, female, boy, girl, army, wife, children, school pet, mum), and also specific professions (lorry driver, train stations) and, perhaps paradoxically, aspirational living (leading good, better pay, leading a good life) appear to be associated with lower levels of social class. In particular, the particularly valid signals for low social class seem to be related to poor grammar, whereas the particularly valid signals for social class appear to be higher culture activites (ballet, horse riding), and being a teacher, university, or being an airline professional. These results do not claim to be exhaustive, but rather a mere lens into some of the dominant topics that emerge at high and low levels of social class validity and perception. For a full list of coefficients for the top 15 unique components, see Table 3. Figure 5 provides a visualization of word clouds based on different components, along with a UMAP+HDBSCAN visualization of topic clusters with lines connecting the highest and lowest scores for each topic clusters based on each of the top 15 PCs to help visualize the general effect of each component.

Table 3. Principal Component Predictors of Social Class: Validity and Judgment Coefficients

PC	Model	β	SE	t	95% CI
PC3	Validity	−0.127	0.010	−13.15***	[−0.146, −0.108]
	AI	−0.376	0.006	−58.51***	[−0.389, −0.364]
	Human	−0.385	0.033	−11.74***	[−0.449, −0.321]
PC2	Validity	−0.080	0.010	−8.21***	[−0.099, −0.061]
	AI	−0.293	0.007	−43.32***	[−0.306, −0.280]
	Human	−0.428	0.034	−12.40***	[−0.495, −0.360]
PC4	Validity	0.125	0.010	12.91***	[0.106, 0.144]
	AI	0.280	0.007	42.13***	[0.267, 0.293]
	Human	0.319	0.035	9.20***	[0.251, 0.387]
PC5	Validity	−0.107	0.010	−11.03***	[−0.126, −0.088]
	AI	−0.153	0.007	−22.62***	[−0.166, −0.140]
	Human	−0.305	0.034	−8.85***	[−0.372, −0.237]
PC6	Validity	–	–	–	–
	AI	0.119	0.007	17.60***	[0.106, 0.132]
	Human	0.224	0.034	6.57***	[0.157, 0.291]
PC15	Validity	0.070	0.010	7.13***	[0.051, 0.089]
	AI	0.143	0.007	21.17***	[0.130, 0.156]
	Human	0.148	0.034	4.31***	[0.081, 0.216]
PC14	Validity	0.072	0.010	7.42***	[0.053, 0.091]
	AI	0.079	0.007	11.67***	[0.066, 0.092]
	Human	0.130	0.035	3.75***	[0.062, 0.198]
PC12	Validity	0.066	0.010	6.81***	[0.047, 0.085]
	AI	0.101	0.007	14.94***	[0.088, 0.114]
	Human	0.154	0.035	4.44***	[0.086, 0.222]
PC13	Validity	–	–	–	–
	AI	−0.017	0.007	−2.50*	[−0.030, −0.004]
	Human	−0.151	0.035	−4.36***	[−0.219, −0.083]
PC47	Validity	–	–	–	–
	AI	−0.119	0.007	−17.66***	[−0.132, −0.106]
	Human	–	–	–	–
PC10	Validity	−0.056	0.010	−5.74***	[−0.075, −0.037]
	AI	−0.100	0.007	−14.82***	[−0.113, −0.087]
	Human	−0.076	0.035	−2.19*	[−0.144, −0.008]
PC18	Validity	0.066	0.010	6.85***	[0.047, 0.085]
	AI	0.044	0.007	6.50***	[0.031, 0.057]
	Human	0.104	0.035	3.00**	[0.036, 0.172]
PC17	Validity	–	–	–	–
	AI	−0.086	0.007	−12.76***	[−0.099, −0.073]
	Human	−0.088	0.035	−2.54*	[−0.156, −0.020]
PC21	Validity	–	–	–	–
	AI	0.085	0.007	12.58***	[0.072, 0.098]
	Human	0.103	0.035	2.97**	[0.035, 0.171]
PC106	Validity	–	–	–	–
	AI	–	–	–	–
	Human	−0.085	0.035	−2.46*	[−0.153, −0.017]

Note. Post-LASSO coefficients from principal component analysis. Validity = Social Class; AI = AI judgment; Human = Human judgment. Dashes indicate predictors not selected by LASSO. ***$p < .001$, **$p < .01$, *$p < .05$.

Figure 5. Interpreting Social Class Perception: UMAP Visualization of Essay Topic Clusters



Note. 3D UMAP projection of essay embeddings, clustered into topics via HDBSCAN. Cluster colors indicate the level and consensus (i.e., agreement between social class validity, AI, and human ratings) of social class judgments, as detailed in the legend. Left: For each of the 15 predictive Principal Components (PCs), the corresponding word cloud is generated from the combined content of the three lowest-scoring (right) and three highest-scoring (left) topic clusters. Lines then connect each PC to its single highest- and lowest-scoring cluster on the plot.

Discussion

The DML-LM Reveals Similar Kernel-of-Truth Mechanisms Across Human and AI Perceivers

A central finding of this study is that AI and human judgments of social class are largely anchored to the same kernels-of-truth, and that they and have a surprising level of consensus in the way they use cues for perception. The broad levels of consensus aligns with the magnitude reported in recent human-AI comparison literature (e.g., Rathje et al., 2024), but this study helped to illustrate that many cues that inform our perceptions are similar, although not identical. Our results also provide strong evidence for the kernel-of-truth hypothesis (Campbell, 1967), showing how weak signals of an objective validity measure like social class can be magnified by both humans and AI into the large perceptual differences, aligning with social class stereotype accuracy results found in other research (Bjornsdottir, 2025; Eagly & Hall, 2025).

Re-evaluating "Unmodeled Knowledge" in the Age of Big Data

The high proportion of judgmental accuracy explained by our framework challenges the long-held role of 'unmodeled knowledge' in person perception. Historically, large residuals in accuracy models have been interpreted as evidence that perception is a largely ambiguous and intuitive process that defies quantification, a view advanced across phenomenological, qualitative, and critical traditions (e.g., Dreyfus, 1972; Merleau-Ponty, 2012; Polkinghorne, 1995). Our results, however, suggest that this explanatory gap may be an artifact of prior measurement limitations rather than an indication of an inherently unmeasurable process. This aligns with Funder's (**1995**) Realistic Accuracy Model, which posits that accuracy depends on the detection of valid, context-sensitive cues. While previous research has highlighted the immense difficulty of exhaustively coding such cues (e.g., Borkenau et al., 2016) and often only found larger effect sizes in experimental settings (see Karelaia & Hogarth, 2008), our findings demonstrate that

high-dimensional methods can successfully capture this complexity, reframing "unmodeled knowledge" as previously under-measured knowledge.

Furthermore, our findings speak to Meehl's (1990) influential critique that researchers should prioritize the magnitude of effects over mere statistical significance. The present study identified numerous significant predictors, yet our recommended approach emphasizes examining the most dominant and interpretable components (e.g., PC2, 3, 4, and 5) that explain the most variance and carry the strongest, most theoretically coherent weights. By concentrating on these substantial effects, we refrain from over-interpreting minor cues whose statistical significance may be an artifact of high statistical power, thereby addressing Meehl's concern.

This approach represents a paradigm that differs from building theory upon isolated and often weak effects. Instead, we first optimize for prediction, allowing complex signals to emerge from the data in an unsupervised manner. We acknowledge that these predictively strong components are composite measures that can be difficult to interpret, as evidenced by some incoherence in the thematic visualizations. Therefore, our interpretation of the dominant themes is offered as a partial but principled account of the model's primary drivers. This allows us to begin with the premise that the phenomenon is complex, and from there, identify the powerful signals that provide a robust, data-driven foundation for theory. This is a step toward a more meaningful understanding of the mechanisms of perception.

The Bias-Variance Tradeoff in a Text Perception Setting

Our findings provide a practical illustration of the bias-variance tradeoff with somewhat surprising results. Highly flexible, low-bias models like XGBoost demonstrated poor generalization due to high variance; they fit noise in the training data, as evidenced by negative cross-validated R^2 scores indicating predictions worse than the mean. Similarly, unregularized

OLS on high-dimensional features failed due to extreme overfitting. This was contrary to our initial ideas that flexible machine learning models would provide superior performance in such a high-dimensional space, such as that of the 4,096 embeddings. In contrast, regularized linear models (Lasso and Ridge) succeeded by introducing a small amount of bias to substantially reduce variance. This constraint on coefficient estimates prevented overfitting and yielded far superior predictive performance on new data. For psychological datasets like ours, characterized by moderate size, inherent noise, and class imbalance, the stability and generalizability offered by regularization proved more effective than the high flexibility of more complex models. This suggests that high-dimensional language features combined with principal component analysis may perform particularly well with linear models, despite an initial sentiment that a flexible model may be the best option.

Language Model Embeddings Model Text Better Than LIWC For Perception

Consistent with a growing body of research (Koutsoumpis et al., 2022; Schwartz et al., 2013), our study confirms that modern language model embeddings significantly outperform traditional dictionary-based methods like LIWC. This performance gap stems from the inherent limitations of LIWC's static, low-dimensional approach. Unlike embeddings, dictionary methods are not context-aware and cannot capture meaning derived from word co-occurrence, a principle central to distributional semantics (Firth, 1957). Furthermore, LIWC's closed-vocabulary design struggles to model misspellings, grammatical errors, or novel phrases (Schwartz et al., 2013), while embeddings leverage the high-dimensional representations that have long been recognized as essential for robust text analysis (Landauer & Dumais, 1997). The practical consequence of these limitations is a significant loss of predictive power, a conclusion supported by findings that only a small fraction of LIWC variables validly predict ground-truth personality traits (Koutsoumpis et

al., 2022). While LIWC's simplicity ensures its value as a baseline, our results underscore that achieving a more powerful and valid understanding of language now necessitates the adoption of these more sophisticated techniques.

Significant performance variability exists even among different language embedding models. For instance, the smaller miniLM model demonstrated commendable performance, achieving approximately four-fifths of the predictive accuracy of the state-of-the-art NV-Embed model. The superior performance of NV-Embed is likely attributable to more advanced features, such as its novel latent attention mechanism and larger parameter count, which theoretically enable it to capture more nuanced linguistic complexities. Beyond quantitative metrics, our own qualitative observations also suggested that the thematic topics generated from the NV-Embed embeddings were more coherent and specific than those from the MiniLM model. This highlights a critical area for future research: systematically investigating how different embedding architectures, sizes, and training objectives influence their capacity to identify psychologically relevant constructs.

Broader Implications and Future Directions

Advancing Perception Theory with Multimodal Data

The present study demonstrates how textual cues reveal a "kernel-of-truth" in social class stereotypes. However, a comprehensive understanding of social perception requires moving beyond a single modality. The DML-LM introduced here is modality-agnostic, making them ideal for building a more holistic model of perception by integrating multiple channels of information such as text, audio, and video.

This direction builds on a rich tradition in lens model research, which has long shown that accuracy varies across different information channels (Osterholz et al., 2021). For social class, this means modeling the potent influence of visual and auditory cues. Visual information from the

face, for example, is strongly associated with class perceptions via stereotypes linking traits like competence and warmth to higher social standing (Bjornsdottir & Rule, 2017). Similarly, the voice acts as an "auditory face," with speech characteristics like accent and pitch serving as powerful, if stereotypical, markers of class (Kachel et al., 2018). Beyond social class, this approach can clarify how other traits, like extraversion, are expressed through a combination of controlled cues (e.g., stylish dress) and automatic ones (e.g., expressive body movements, cheerful voice) across different modalities (Hirschmüller et al., 2013). Applying the DML-LM to such high-dimensional, multimodal data would allow researchers to model how cues in one modality (e.g., a confident tone of voice) interact with or override cues in another (e.g., hesitant language). This would not only enable a direct comparison of how humans and AI weigh different information streams but also fuel the development of richer, more ecologically valid theories of perception that embrace the reality that trait expression is fundamentally context-dependent (Fleece & Teglasi, 2024).

Modeling Informant Discrepancy

In a similar way how different modalities matter, researchers often grapple with why different informants, such as parents and teachers, provide divergent ratings of the same person. For instance, fewer than half of children were classified similarly by both parents and teachers on key traits like behavioral inhibition (Fleece & Teglasi, 2024). Such discrepancies arise from informant-specific effects, where the perceived relationships between traits can be an artifact of a single rater's perspective (Biesanz & West, 2004). The DML-LM is uniquely positioned to address this challenge by moving beyond simple comparison. By treating each informant as a "perceiver," the DML-LM can use high-dimensional behavioral data (e.g., narrative accounts) to model their distinct judgment policies. This allows for a precise quantification of cue utilization variability

across informants, where one informant weighs a specific cue heavily while another discounts it (Hirschmüller et al., 2013).

A New Frontier for Auditing High-Stakes AI

Modeling informant discrepancy between humans and AI has never been more important, especially with AI beginning to encroach on high-stakes frontiers. For example, AI ought to be audited for the ways in which it uses cues in patient data when compared to expert physicians to reveal areas of discrepancy that may spur downstream inequalities (Ratwani et al., 2024). In human resources, AI-powered hiring tools have been deployed and leading to biased screening outcomes, and thus deserve auditing and understanding of the differential cue use and validity (Tilmes, 2022). In finance, this framework is especially salient. For example, Fuster et al. (2022) studied the U.S. mortgage market and found that shifting from traditional logistic regression models to Random Forest models for predicting a person's probability of defaulting resulted in disproportionately disadvantaged Black and Hispanic borrowers, increasing their interest rate by nearly double for Black and Hispanic borrowers compared to White non-Hispanic borrowers. Such outcomes may arise from the model's increased flexibility and enhanced ability to triangulate protected characteristics from permissible data. Through the DML-LM, researchers can go beyond the identification of the presence of bias, and investigate *how* machine learning models amplify biases, increasing transparency in the high-stakes frontiers like healthcare, employment, and finance.

Experimental Validation through Mechanistic Interpretability

While the DML-LM provides a powerful diagnostic lens, the next frontier is to move from correlational observation to causal validation. This requires delving into mechanistic

interpretability to reverse-engineer a model's internal computations. A primary obstacle is polysemanticity, where neurons represent multiple unrelated concepts due to superposition, a compression strategy where models encode more features than they have neurons (Elhage et al., 2022). Indeed, we see polysemanticity in our current results, where different principal components appear to be tapping into the same phenomenon (e.g., football for PC3 and PC4).

To overcome this, our framework can be integrated with techniques like sparse autoencoders, which decompose a model's internal activations into more monosemantic (single-concept) features that are functionally equivalent to the "cues" in our model (Cunningham et al., 2023). Once these latent features are identified, their causal role can be experimentally tested. For instance, using activation patching, a researcher could directly manipulate the model's internal state to activate a specific feature (e.g., "formal professional language") and measure its direct impact on the final judgment. If the manipulation reliably alters the outcome, it provides strong causal evidence that the feature is a mechanistic driver of the model's perception. This approach enables the automated discovery of entire causal circuits within the model, moving far beyond simple cue-judgment correlations.

Limitations

While the present study introduces a robust framework for analyzing high-dimensional perception, several limitations should be acknowledged. First, the analyses were conducted on a historical dataset of essays from the 1960s. Although this provided a unique opportunity to examine perceptions of social class, the language, societal norms, and class structures of that era may not be directly generalizable to contemporary contexts. The validity variable itself, a five-point measure of the father's occupational social class, was coarse and imbalanced, which likely constrained the predictive performance of all models and may have particularly

disadvantaged complex, high-variance models like XGBoost that are sensitive to noisy data.

Second, a central challenge in this work is the interpretability of high-dimensional features. While we demonstrated a viable approach for decoding principal components by combining topic modeling and extreme group analysis, this process remains inherently qualitative and requires careful consideration. A critical direction for future research is to develop and validate more systematic and rigorous methodologies for interpreting the otherwise opaque components derived from dimensionality reduction techniques. Establishing best practices for this process is essential for moving from statistical explanation to true theoretical understanding.

Third, while our high-dimensional text features capture a vast amount of information, there are still other domains of perception left unmeasured. One primary variable is time. For instance, Huang et al. (2023) found reaction time and judgment confidence during perception to be a relevant cue that improved classification performance. Future studies ought to examine how the DML-LM could capture time effects, reputation effects (e.g., McAbee & Connelly, 2016), and accuracy outside of a zero-acquaintance setting.

Conclusion

This study introduced the DML-LM as a novel framework for deconstructing and comparing the architecture of perception in the age of artificial intelligence. An application of this framework to the judgment of social class from text challenges a simple narrative of AI as an alien intelligence, revealing instead a remarkable degree of overlap with human perception. The findings indicated that both AI and human judgments are anchored to the same potent, thematic "kernels of truth" embedded in language.

The critical distinction, however, lies not in what is perceived but in how. The AI's judgment policy was found to be substantially less sparse, systematically integrating a much broader array

of textual cues than the more heuristic and parsimonious model employed by humans. While both perceivers amplify subtle, real-world patterns into powerful and potentially discriminatory heuristics, the AI's use of a wider set of cues results in a particularly rigid and impactful form of this amplification. The implication of this finding is profound: as AI systems are increasingly integrated into high-stakes domains, it is insufficient to know that they are often accurate. The frameworks presented here provide a necessary diagnostic lens, revealing that even when an algorithm's perception closely mirrors our own, subtle but significant differences in its process can lead to large and consequential differences in its real-world impact. This research represents a critical step toward a more nuanced understanding of AI and the development of more accurate, fair, and transparent decision-making systems.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT 4o and o1 (https://chat.openai.com/), Claude Opus 4 (https://claude.ai), and Gemini 2.5 Pro (https://gemini.google.com) in order to improve the clarity of writing, to assist with producing figures, and to help produce statistical code. After using these tools, the authors reviewed and edited the content and take full responsibility for the content of the publication.

Code Availability

The code used to analyzed in this study are available in the Open Science Framework repository at https://osf.io/wrn6e/. The analyses were conducted using the HAAM package, which is available on both GitHub (https://github.com/raymondli-me/haam) and OSF.

References

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology*, *19*(6), 586–592. https://doi.org/10.1037/0278-6133.19.6.586

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–283. https://doi.org/10.1037/0033-2909.111.2.256

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2). https://doi.org/10.3150/11-BEJ410

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the big five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, *72*(4), 845–876. https://doi.org/10.1111/j.0022-3506.2004.00282.x

Bjornsdottir, R. T. (2025). The Social Perception of Social Class: An Integrative Review. *European Journal of Social Psychology*, *55*(4), 547–553. https://doi.org/10.1002/ejsp.3168

Bjornsdottir, R. T., & Rule, N. O. (2017). The visibility of social class from facial cues. *Journal of Personality and Social Psychology*, *113*(4), 530–546. https://doi.org/10.1037/pspa0000091

Borkenau, P., Mosch, A., Tandler, N., & Wolf, A. (2016). Accuracy of Judgments of Personality Based on Textual Information on Major Life Domains. *Journal of Personality*, *84*(2), 214–224. https://doi.org/10.1111/jopy.12153

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22* (tech. rep.). University of Texas at Austin. Austin, TX. https://www.liwc.app

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. https://doi.org/10.1037/h0047470

Campbell, D. T. (1967). Stereotypes and the perception of group differences. *American Psychologist*, *22*(10), 817–829. https://doi.org/10.1037/h0025079

Centre for Longitudinal Studies. (2025). 1958 National Child Development Study (NCDS). https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

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. https://doi.org/10.1111/ectj.12097

Cooksey, R. W. (1996). The methodology of social judgement theory. *Thinking & Reasoning*, *2*(2-3), 141–173. https://doi.org/10.1080/135467896394483

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023, October). Sparse Autoencoders Find Highly Interpretable Features in Language Models. https://doi.org/10.48550/arXiv.2309.08600

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. Harper & Row.

Eagly, A. H., & Hall, J. A. (2025). The kernel of truth in gender stereotypes: Consider the avocado, not the apple. *Journal of Experimental Social Psychology*, *118*, 104713. https://doi.org/10.1016/j.jesp.2024.104713

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis* (pp. 1–32). Blackwell.

Fleece, H., & Teglasi, H. (2024). Behavioral Inhibition and Social Competence Through the Eyes of Parent and Teacher Informants. *Behavioral Sciences*, *14*(11), 1080. https://doi.org/10.3390/bs14111080

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, *77*(1), 5–47. https://doi.org/10.1111/jofi.13090

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Prentice Hall.

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.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based tf-idf procedure*. https://doi.org/10.48550/arXiv.2203.05794

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *104*(2), 335–353. https://doi.org/10.1037/a0030383

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.2307/1267351

Huang, J., Prijatelj, D., Dulay, J., & Scheirer, W. (2023). Measuring Human Perception to Improve Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 11382–11389. https://doi.org/10.1109/TPAMI.2023.3270772

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334. https://doi.org/10.1037/a0020761

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kachel, S., Simpson, A. P., & Steffens, M. C. (2018). "Do I Sound Straight?": Acoustic Correlates of Actual and Perceived Sexual Orientation and Masculinity/Femininity in Men's Speech. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1560–1578. https://doi.org/10.1044/2018_JSLHR-S-17-0125

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Koutsoumpis, A., Oostrom, J. K., Holtrop, D., Van Breda, W., Ghassemi, S., & De Vries, R. E. (2022). The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC). *Psychological Bulletin*, *148*(11-12), 843–868. https://doi.org/10.1037/bul0000381

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., & Ping, W. (2025, February). NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. https://doi.org/10.48550/arXiv.2405.17428

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2023). Awq: Activation-aware weight quantization for llm compression and acceleration. https://doi.org/10.48550/arXiv.2306.00978

MacKinnon, D. (2008). *Introduction to statistical mediation analysis* (1st ed.). Routledge. https://doi.org/10.4324/9780203809556

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, *123*(5), 569–591. https://doi.org/10.1037/rev0000035

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/PR0.66.1.195-244

Merleau-Ponty, M. (2012). *Phenomenology of perception* (D. A. Landes, Trans.). Routledge.

Osterholz, S., Breil, S. M., Nestler, S., & Back, M. D. (2021). Lens and dual lens models. In T. D. Letzring & J. S. Spain (Eds.), *The oxford handbook of accurate personality judgment* (pp. 45–60). Oxford University Press.

Polkinghorne, D. E. (1995). Narrative configuration in qualitative analysis. *International Journal of Qualitative Studies in Education*, *8*(1), 5–23. https://doi.org/10.1080/0951839950080103

Power, C., & Elliott, J. (2006). Cohort Profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, *35*(1), 34–41. https://doi.org/10.1093/ije/dyi201

Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., ... Bao, K., et al. (2024). *Qwen2.5 technical report*.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Rao, R., & Van Bavel, J. J. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Ratwani, R. M., Sutton, K., & Galarraga, J. E. (2024). Addressing AI Algorithmic Bias in Health Care. *Journal of the American Medical Association*, *332*(13), 1051. https://doi.org/10.1001/jama.2024.13486

Rule, N. O., & Ambady, N. (2008). The Face of Success: Inferences From Chief Executive Officers' Appearance Predict Company Profits. *Psychological Science*, *19*(2), 109–111. https://doi.org/10.1111/j.1467-9280.2008.02054.x

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach (T. Preis, Ed.). *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Stavrova, O., & Haarmann, L. (2020). How to tell a happy person: Accuracy of subjective well-being perception from texts. *Motivation and Emotion*, *44*(4), 597–607. https://doi.org/10.1007/s11031-019-09815-4

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, *58*(1), 267–288.

Tilmes, N. (2022). Disability, fairness, and algorithmic bias in AI recruitment. *Ethics and Information Technology*, *24*(2), 21. https://doi.org/10.1007/s10676-022-09633-2

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of Competence from Faces Predict Election Outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, *71*(6), 528–530. https://doi.org/10.1037/h0047061

Wallace, J. D., & Biesanz, J. C. (2021). Examining the consistency of the good target across contexts and domains of personality. *Journal of Personality*, *89*(2), 188–202. https://doi.org/10.1111/jopy.12574

Wang, W., Bao, H., Huang, S., Dong, L., & Wei, F. (2020). *Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers.* https://doi.org/10.48550/arXiv.2012.15828

Zhao, S., Witten, D., & Shojaie, A. (2021). In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference. *Statistical Science*, *36*(4). https://doi.org/10.1214/20-STS815

Appendices

Appendix A: Derivation of the Double Machine Learning Lens Model

Step 1: Observed Variables

We represent the validity (X) and judgment (Y) with cross-fit predictions and residuals:

$$X = \hat{X} + e_X, \quad Y = \hat{Y} + e_Y,$$

where $\hat{X} \approx \hat{m}_0(Z)$ and $\hat{Y} \approx \hat{g}_0(Z)$, each estimated via cross-fitting over covariates Z. Under double machine learning, the Neyman-orthogonal structure ensures that final correlation estimates remain robust to potential estimation errors from the regularization of $\hat{m}_0$ and $\hat{g}_0$.

Step 2: Covariance Decomposition

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(\hat{X} + e_X,\ \hat{Y} + e_Y) \\ &= \mathrm{Cov}(\hat{X}, \hat{Y}) + \mathrm{Cov}(\hat{X}, e_Y) + \mathrm{Cov}(e_X, \hat{Y}) + \mathrm{Cov}(e_X, e_Y).\end{aligned}$$

Step 3: Variance Decomposition

$$\mathrm{Var}(X) = \mathrm{Var}(\hat{X} + e_X) = \mathrm{Var}(\hat{X}) + \mathrm{Var}(e_X) + 2\,\mathrm{Cov}(\hat{X}, e_X),$$

$$\mathrm{Var}(\hat{X}) = \mathrm{Var}(X) - \mathrm{Var}(e_X) - 2\,\mathrm{Cov}(\hat{X}, e_X).$$

Define the cross-fit coefficient:

$$R^2_X = 1 - \frac{\mathrm{Var}(e_X)}{\mathrm{Var}(X)} \;=\; \frac{\mathrm{Var}(\hat{X})}{\mathrm{Var}(X)} + \frac{2\,\mathrm{Cov}(\hat{X}, e_X)}{\mathrm{Var}(X)}.$$

If $\hat{X}$ and e_X are not orthogonal, we set:

$$E_X = \frac{2\,\mathrm{Cov}(X, e_X)}{\mathrm{Var}(X)}.$$

Then $\mathrm{Var}(\hat{X}) = (R^2_X - E_X)\,\mathrm{Var}(X)$ and $\mathrm{Var}(e_X) = (1 - R^2_X)\,\mathrm{Var}(X)$.

Step 4: Modeled and Residual Covariances

$$\begin{aligned}
\mathrm{Cov}(\hat{X}, \hat{Y}) &= \mathrm{Cor}(\hat{X}, \hat{Y})\,\sqrt{\mathrm{Var}(\hat{X})}\,\sqrt{\mathrm{Var}(\hat{Y})} \\
&= G\,\sqrt{R^2_X - E_X}\,\sqrt{R^2_Y - E_Y}\,\sigma_X\,\sigma_Y, \\
\mathrm{Cov}(e_X, e_Y) &= \mathrm{Cor}(e_X, e_Y)\,\sqrt{\mathrm{Var}(e_X)}\,\sqrt{\mathrm{Var}(e_Y)} \\
&= C\,\sqrt{1 - R^2_X}\,\sqrt{1 - R^2_Y}\,\sigma_X\,\sigma_Y.
\end{aligned}$$

We define cross-covariance parameters:

$$C_U = \frac{\mathrm{Cov}(\hat{X}, e_Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \quad \text{and} \quad C_V = \frac{\mathrm{Cov}(e_X, \hat{Y})}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

They may be non-zero for flexible ML estimators.

Step 5: Overall Achievement

The correlation $r_a = \mathrm{Cor}(X, Y)$ can be written:

$$r_a = G\sqrt{R_X^2 - E_X}\sqrt{R_Y^2 - E_Y} \; + \; \big(\mathrm{Cov}(\hat{X}, e_Y) + \mathrm{Cov}(e_X, \hat{Y})\big)/(\sigma_X \sigma_Y) \\ + \; C\sqrt{1 - R_X^2}\sqrt{1 - R_Y^2},$$

which often is expanded as:

$$r_a \; = G\sqrt{R_X^2 - E_X}\sqrt{R_Y^2 - E_Y} + \; C_U \; + \; C_V \; + \; C\sqrt{1 - R_X^2}\sqrt{1 - R_Y^2}.$$

If the cross-covariances vanish and $E_X = E_Y = 0$ (pure OLS), it reduces to

$r_a = G\sqrt{R_X^2}\sqrt{R_Y^2} + C\sqrt{1 - R_X^2}\sqrt{1 - R_Y^2}$.

Step 6: Debiased Unmodeled Correlation

Under cross-fitting with standardized residuals, C corresponds to the double-machine-learning estimator:

$$C = \check{\beta}_1 = \frac{\sum_{i \in I} e_{X_i}\,[Y_i - \hat{g}_0(Z_i)]}{\sum_{i \in I} e_{X_i}^2},$$

providing an unbiased estimate of the residual relationship between X and Y after controlling for high-dimensional Z.

Appendix B: Large Language Model Configuration and Prompting Protocol

Model Configuration

For the automated generation of social class ratings, we employed the `Qwen/Qwen2.5-32B-Instruct-AWQ` large language model, a 32-billion parameter instruction-tuned model. The model was accessed through the `vLLM` inference library with the following technical specifications:

- **Quantization:** Activation-aware Weight Quantization (AWQ) with 16-bit floating-point precision (`float16`) for computational efficiency
- **Context Length:** Maximum of 2,048 tokens
- **Hardware Distribution:** Two GPUs with `tensor_parallel_size=2`
- **Memory Utilization:** 80% GPU memory limit
- **Sampling Temperature:** 0.1 for semi-deterministic outputs
- **Maximum Response Tokens:** 200 per generation
- **Batch Processing:** Essays processed in batches of 100 for computational throughput optimization
- **Optimization Settings:** Triton Flash Attention and `torch.compile` disabled for reproducibility

Prompting Protocol

Each essay was evaluated using a carefully structured multi-part prompt based on an adaptation of the MacArthur Scale of Subjective Social Status (Adler et al., 2000). The prompt instructed the model to place the child's family on a 10-rung societal ladder and return the rating as a JSON object.

System Message:

```
Return JSON with 'social_class_judgment' (1 to 10).
```

User Message Template:

```
Imagine that there is a ladder that pictures how society is set up.
At the top of the ladder are the people who are the best off | they
```

```
have the most money, the highest amount of schooling, and the jobs
that bring the most respect. At the bottom are people who are the
worst off | they have the least money, little or no education, no
job, or jobs that no one wants or respects. Where do you think the
family of the person who wrote this essay would be on this ladder?
Rate from 1 (bottom) to 10 (top).


[ESSAY TEXT]
```

Example Model Response:

```
json
{"social_class_judgment": 6}
```

Appendix C: Human Perceiver Data Collection

Participant Recruitment and Demographics

In 2021, we recruited 600 participants from the United Kingdom through the Prolific online research platform to evaluate the perceived social class of essays written by 11-year-old participants in the 1958 National Child Development Study (NCDS) (Power & Elliott, 2006; Centre for Longitudinal Studies, 2025). The study protocol received approval from the University of British Columbia's Behavioral Research Ethics Board (BREB #H20-01189).

Sample Characteristics:

- **Gender Distribution:** 304 female participants
- **Age:** Mean = 37 years ($SD = 13.06$), Range = 18–76 years
- **Nationality:** All current or former UK residents

- **Ethnic Composition:**
  - European ancestry: $n = 493$
  - East Asian: $n = 15$
  - Southeast Asian: $n = 9$
  - South Asian: $n = 31$
  - Middle Eastern: $n = 4$
  - Other backgrounds: $n = 48$
- **Compensation:** £6.91 per participant

Data Collection Procedure

The data collection protocol consisted of the following steps:

1. Participants completed a demographic questionnaire
2. Participants viewed an informational video about the NCDS to provide contextual understanding
3. Each participant evaluated 10 randomly assigned essays
4. Participants rated the perceived social class of each essay writer's family using an adapted MacArthur Scale

MacArthur Scale Instructions: Participants were presented with a ladder image and the following prompt:

> "Imagine that this ladder pictures how society is set up. At the top of the ladder are the people who are the best off—they have the most money, the highest amount of

> schooling, and the jobs that bring the most respect. At the bottom are people who are the worst off—they have the least money, little or no education, no job, or jobs that no one wants or respects. Where do you think the family of the person who wrote this essay would be on this ladder?"

Data Processing and Sample Size Determination

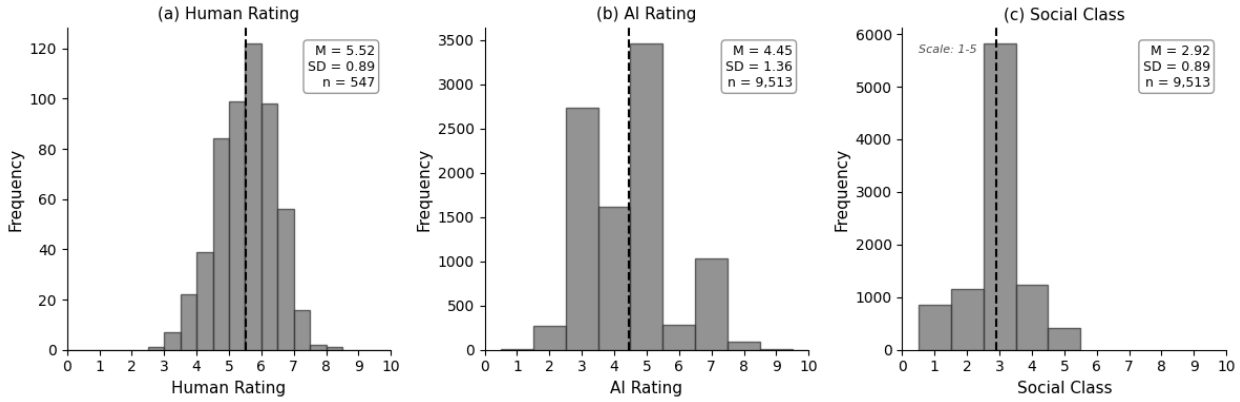
The original dataset comprised 10,511 essays written in 1969 by 11-year-old NCDS participants. From this corpus, 600 essays were randomly selected for human evaluation. Of these, 547 essays had corresponding social class validity data available for analysis.

Averaging Across Human Raters: Each essay received ratings from an average of 10 different perceivers. The final perceived social class score for each essay was calculated as the mean of these individual ratings, resulting in 547 essays with averaged human social class judgments.

Statistical Power: With a sample size of 547 essays, our study achieved 80% power to detect correlations of $r \geq .12$ between human ratings and social class validity measures, ensuring adequate sensitivity for detecting even small effect sizes in the accuracy of social class perception.

Appendix D: Distribution of Social Class Measures

Figure 6. Distributions of Human Judgments, AI Judgments, and Social Class Validity.



Note. The figure displays histograms for the three key variables in the study. Panel (a) shows the distribution of social class ratings from human perceivers ($n = 547$) on a 10-point scale. Panel (b) shows the distribution of social class ratings from the AI model for all essays ($n = 9,513$) on the same 10-point scale. Panel (c) shows the distribution of the ground-truth social class validity measure, based on the reverse-coded 5-point Registrar General's Social Class scale ($n = 9,513$). Dashed vertical lines indicate the mean for each distribution.