Measurement in the Age of AI: Discovery, Definition, and Observation

Melissa Dell   Ashesh Rambachan

NBER Methods Lecture 2026

Slides posted on the NBER Methods Lecture website.

AI and Measurement

Our starting point

Artificial intelligence (AI) is transforming measurement.

- AI systems can convert **high-dimensional, unstructured data** (*e.g.*, text, images, audio, and video) into **low-dimensional variables** for statistical analysis.
- This can often be done **cheaply and at massive scale**.
- Understanding the implications of AI predictions for empirical analysis therefore requires examining how AI **reshapes the measurement process**.

Outline

Measurement Is Representation

Core idea

Measurement maps complex, high-dimensional reality into lower-dimensional representations.

- It preserves some features of reality while discarding or transforming others.
- Measurement is therefore **not passive observation**; it is a **disciplined simplification** of reality.

Why Simplification Is Necessary

Reality is vastly high-dimensional:

"There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy."

— *Shakespeare, Hamlet*

Performing any given task will require abstraction:

"To think is to forget differences, generalize, make abstractions."

— *Borges, "Funes the Memorious"*

Measurement formalizes this simplification: it specifies which features of reality are preserved in the variables we analyze.

"The Map Is Not the Territory" — Alfred Korzybski

Borges imagines cartographers who pursue perfect representation, until they produce:

> *a map "whose size was that of the Empire"*
>
> — *Borges, "On Exactitude in Science"*

But the **perfect map is useless**: because it no longer reduces complexity, it ceases to function as a map.

Measurement as Disciplined Simplification

- Selectivity does **not** make measurement merely subjective.
- The relevant question is whether a given simplification **preserves the features needed to support useful inference** about the underlying high-dimensional world, for the research question at hand.

In short

Measurement is a **disciplined, fallible, and purpose-oriented** attempt to learn from reality through simplified representations.

Measurement as a Projection

Measurement specifies and implements a function $m(\cdot)$ mapping high-dimensional reality X into a lower-dimensional representation that can be interpreted and analyzed:

The Measurement Function

$$X \xrightarrow{m(\cdot)} m(X)$$

Because X is high-dimensional, there may be many ways to measure a given concept:

$$X \longrightarrow m_j(X), \qquad j = 1, \ldots, J$$

Different m_j may **preserve different features** and **support different empirical conclusions**.

Measurement Stages

Colloquially, we often think of measurement as the act of observing a quantity with an instrument or procedure, but end-to-end measurement involves three distinct stages:

1. **Discovery**: Given the complexity of the world, what features exist and are worth measuring?

Each stage of measurement involves different types of uncertainty, and they have traditionally been examined through different disciplinary and methodological lenses.

Measurement Stages

Colloquially, we often think of measurement as the act of observing a quantity with an instrument or procedure, but end-to-end measurement involves three distinct stages:

1. **Discovery**: Given the complexity of the world, what features exist and are worth measuring?
2. **Construct definition**: How should the relevant concept be operationally defined?

Each stage of measurement involves different types of uncertainty, and they have traditionally been examined through different disciplinary and methodological lenses.

Measurement Stages

Colloquially, we often think of measurement as the act of observing a quantity with an instrument or procedure, but end-to-end measurement involves three distinct stages:

1. **Discovery**: Given the complexity of the world, what features exist and are worth measuring?
2. **Construct definition**: How should the relevant concept be operationally defined?
3. **Observation**: How can that definition be implemented at scale?

Each stage of measurement involves different types of uncertainty, and they have traditionally been examined through different disciplinary and methodological lenses.

Stage 1: Discovery

Uncertainty

Which patterns, dimensions, or relationships exist and are worth measuring?

Methodological perspectives:

- Theory and contextual knowledge
- Exploratory analysis
- Unsupervised learning and data-driven discovery

Stage 2: Construct Definition

Uncertainty

How should a phenomenon of interest be translated into an operational construct for a given research setting?

Methodological perspectives:

- Theory and domain expertise
- Piloting and qualitative validation
- Externally available proxies
- Formal literatures in psychometrics and survey design

Stage 3: Observation

Uncertainty

How accurately can the construct be observed at scale and how do errors propogate to the target parameter?

Methodological perspectives:

- Measurement error
- Semiparametric inference with missing data

Measurement Has Often Been Treated as Fixed

- Measurement costs frequently prevent researchers from comparing different measurement functions $m(X)$.
- Researchers have therefore relied on externally constructed measures as **fixed proxies** for theoretically meaningful concepts.

Consequence

The focus is **justifying the proxy** rather than **comparing alternative measurement functions**.

Outline

AI as a Measurement Tool

Deep neural networks are flexible implementations of measurement functions:

$$X \xrightarrow{m_{\mathrm{AI}}(\cdot)} m_{\mathrm{AI}}(X)$$

- They learn structured representations from high-dimensional inputs such as text, images, audio, and video.
- These representations can then be interpreted and used in statistical analysis.
- This is changing the measurement landscape by making it more feasible to convert unstructured data into variables for analysis.

From Rules to Learned Representations

- Before neural networks, researchers often extracted low-dimensional features from unstructured data using **hand-engineered rules**.
- Neural networks instead **learn representations from empirical examples**.
- Deep neural networks use millions to billions of learned parameters to transform high-dimensional inputs into lower-dimensional representations, such as continuous vectors or predicted labels.

How Neural Networks Learn What to Preserve

Neural networks are trained to perform specific tasks:

- predicting the next word in a sequence;
- classifying an image;
- identifying objects;
- matching text to labels.

Their parameters are adjusted through **gradient descent**.

Key idea

With sufficient training data and an appropriate objective, neural networks learn **which features of reality are useful to preserve** for the task.

Why Neural Networks Are Powerful Measurement Tools

- They can **generalize** beyond the task on which they were trained.
- Across many tasks, neural networks **outperform human-engineered rules**.
- Neural networks are **highly scalable**. AI can therefore substantially **lower the cost** of converting unstructured data into low-dimensional variables for statistical analysis.

Fixed Costs May Still Be Substantial

In some settings, credible AI measurement requires **substantial fixed investments**.

To achieve sufficiently accurate predictions, researchers may need to:

- develop high-quality training data;
- fine-tune customized models.

Constructing the measurement function m may remain costly, even if applying m to very large datasets is relatively cheap.

When Candidate Measures Become Cheap

In other settings, off-the-shelf or lightly adapted models perform sufficiently well.

This makes it inexpensive to construct **many candidate measures** of the same broad concept.

$$X \begin{cases} \longrightarrow m_1(X) \\ \longrightarrow m_2(X) \\ \quad \vdots \\ \longrightarrow m_J(X) \end{cases}$$

Key implication:

The bottleneck shifts from developing any scalable measure to **choosing among many plausible measures**, which may yield different estimates of the target parameter.

A Parallel with the Empirical Revolution

- These challenges closely parallel those created by the rise of **personal computing** in the 1990s.
- Cheap computation enabled researchers to estimate not just one empirical specification, but hundreds, thousands, or even millions (Sala-I-Martin, 1997).

This forced the field to ask

- What should discipline the choice among possible specifications?
- What makes a specification credible?

AI Creates an Analogous Problem

In response to cheap computation, literatures developed around:

- causal inference,
- robustness,
- multiple testing,
- related topics.

AI-enabled measurement raises an analogous problem

By expanding the set of feasible measurement functions, AI underscores the importance of assessing **what makes measurement credible** and **how to choose among multiple** m.

AI Expands the Measurement Choice Set

When measurement is costly, researchers often:

- inherit a proxy;
- adopt an existing operational definition;
- implement the only scalable procedure available.

AI can expand measurement possibilities at each stage of the pipeline:

1. **Discovery:** identify patterns or dimensions not specified in advance.
2. **Construct definition:** generate and compare alternative operationalizations.
3. **Observation:** implement a chosen definition at scale.

Discovery

- Discovery is often guided by **theory**, **contextual knowledge**, or incremental advances in an **existing literature**.
- This works well when the problem is sufficiently understood for **pre-specified measurement** to be well motivated.

The risk

If these are the only tools available, researchers risk "looking for the keys under the lamppost" — searching only for concepts they already know to specify.

Data-Driven Discovery

The researcher specifies a **representational procedure** rather than a **predefined construct**.

Examples include:

- embeddings;
- clustering algorithms;
- dictionary learning methods;
- autoencoders.

These procedures summarize high-dimensional data by identifying **structure not specified in advance**.

What Discovery Procedures Learn

Discovery procedures can identify:

- latent patterns;
- clusters;
- dimensions;
- categories;
- recurring components.

Some methods **group similar observations** together. Others learn **continuous dimensions** along which observations vary.

From Discovery to Measurement

The discovered structure can enter measurement in two ways:

1. **Input into construct definition:** use the discovered structure to formulate a more precise concept, measured on a separate sample to reduce post-selection inference concerns.
2. **Measurement output:** the discovered structure may itself constitute the measurement output. Interpretation proceeds through interpretability methods, high-dimensional statistical analysis, and external validation.

When Discovery Is Useful

These methods are useful when researchers:

- are unsure what to look for;
- suspect existing conceptual categories are incomplete;
- want to discipline discovery by pre-specifying a **data-driven procedure** rather than the object to be measured.

Example

Pre-registering a procedure to identify recurring themes in open-ended survey responses, rather than pre-specifying a fixed coding scheme ex ante.

Construct Definition

Constructin Definition

How should a given concept be **operationalized** for a particular research setting?

This is difficult because many concepts of interest are:

- multidimensional;
- not directly observable;
- closely related to other phenomena.

Example: a measure of **trust** — a high-dimensional, latent concept — should capture the dimension of trust relevant to the research question, not simply proxy institutional quality, income, or other correlated variables.

AI Does Not Define Constructs on Its Own

The Role of AI

AI does **not** determine construct definition on its own.

Rather, by reducing the costs of data construction, highly performing zero-shot or lightly-adopted AI systems expand:

- the set of **candidate operationalizations** researchers can examine;
- the **empirical evidence** available for choosing among them.

AI needs to be combined with theory and domain expertise that clarify what the construct should capture and predict, what it should be distinct from, and where it should be stable.

Observation

The Role of AI

In observation, AI serves as a **scalable measurement technology**.

Once a researcher has specified an operational definition, an AI system can apply it repeatedly to high-dimensional inputs at **low marginal cost**.

This makes it possible to:

- classify large text corpora;
- extract information from vast image collections;
- more generally, construct low-dimensional features from unstructured data at scale.

AI for Observation

AI is highly scalable, but different implementation choices — *e.g.*, models, prompts, training data, or input distributions — may produce **different measurements**.

Errors are unlikely to be classical. Systematic biases can arise from:

- network architecture;
- the distribution of training data;
- implementation details;
- nonlinear transformations at every level of the neural network;
- binary or multiclass classification.

These biases can propagate to estimates of the target parameter.

Implication

Validation is **central** to AI-enabled measurement.

Outline

The Goal of Measurement

Measurement does not aim to **reproduce reality**, so validation cannot show that a measure perfectly captures a latent concept or recovers "fundamental truth."

It aims to construct **simplified representations** that preserve the features needed to draw useful inferences about the high-dimensional world.

In principle, validation would assess whether a measure achieves this goal. In practice, this is difficult for many social science measures, because the underlying reality is:

- complex;
- partially unobservable;
- multi-causal;
- not easily manipulated.

The Role of Validation

Validation anchors measurement to criteria that can be

stated explicitly · interpreted · audited · debated · refined

- Ultimately, measurement is judged by whether it provides a **useful guide to reality**.
- Because validation cannot usually assess this ultimate goal directly, it builds an empirical case through **observable and interpretable criteria**.
- It states an **explicit measurement criterion** and assesses how well the measure behaves relative to it.

Validation in the Age of AI

Anchoring to explicit criteria is **complementary** to modern AI, addressing its limitations:

- Internal mappings from inputs to outputs are too complex to understand by inspecting parameters, and learned criteria are not expressed as explicit, auditable rules.
- Biases shift in complex ways with the inputs, and humans predict these shifts poorly (Vafa, Rambachan and Mullainathan, 2024).

Validation makes AI interpretable

Validation makes the black box interpretable **not by elucidating its inner workings**, but by **evaluating its outputs** against criteria specified by the researcher.

Validation Disciplines Measurement Choice

Validation also addresses the abundance of measures AI can cheaply produce. When many candidate measures can be constructed from the same reality, **measurement choices become objects of empirical comparison**.

Validation disciplines that comparison

It provides explicit, evaluable evidence for:

- choosing among candidate measures;
- correcting or refining measures;
- deciding when additional investment in measurement is needed.

What Validation Privileges

Validation evidence may come from

human-coded labels · scientific instruments · administrative data · theoretical predictions · algorithmic criteria

- Validation does **not** privilege humans over models as measurement instruments.
- It privileges **explicit, interpretable criteria** over implicit decision rules learned by a black-box model.
- Even when a neural network is prompted or trained to apply a criterion, its internal decision rules are not directly interpretable.

Imperfect Validation

Ambiguous edge cases and residual disagreement between annotations do **not** imply validation lacks value. Instead, they identify where refinement may be needed:

- the construct definition may be ambiguous;
- the measurement procedure may be underspecified;
- the validation criterion may need clarification.

Neural networks versus criteria

- A black-box neural network encodes **richer information** than a criterion.
- The question is: does this suggest a better interpretable construct, or reflect irrelevant features, confounding variation, or difficult to interpret patterns?

Outline

Validation Across the Measurement Pipeline

Because the three stages of measurement involve different sources of uncertainty, validation takes **different forms** at each stage.

1. **Discovery:** has the exploratory procedure uncovered meaningful structure?
2. **Construct definition:** does the operationalization capture the intended dimension of the concept?
3. **Observation:** does the measurement procedure accurately implement the chosen criterion?

Validating Discovery

In data-driven **discovery**, the researcher has not yet specified a fixed criterion against which the output can be validated.

The question

Has an exploratory procedure uncovered structure meaningful enough to:

- interpret;
- refine into a construct; or
- use as a measurement object?

Internal Validation and Stability

Internal validation

Do the data exhibit **separable structure** under a particular algorithm or representation?

In clustering, this can mean (Rousseeuw, 1987; Tibshirani, Walther and Hastie, 2001):

- within-cluster cohesion;
- between-cluster separation.

Stability

Meaningful structure should recur under reasonable perturbations of sampling, pre-processing, or implementation (Ben-Hur, Elisseeff and Guyon, 2002).

Multimethod Validation

A related concern: discovered structure may reflect the **method** used to elicit it rather than a meaningful dimension of reality.

Multitrait-multimethod framework (Campbell and Fiske, 1959)

A measure should:

- converge with measures of related concepts;
- remain distinct from nearby phenomena;
- be recoverable using alternative methods.

Multimethod Validation in AI Discovery

In AI-enabled discovery, multimethod validation asks whether the discovered structure can be **recovered through different measurement routes**. Researchers might vary:

- the representation model;
- the discovery algorithm;
- the method of eliciting or compiling the underlying raw data;
- the corpus or data source.

Interpretability in Discovery

- A discovered representation becomes useful for measurement only if researchers can **interpret** it, a concern in older traditions (e.g. principal components analysis) as well as modern representation learning.
- LLMs can often cheaply generate labels, summaries, and explanations for discovered structure.
- But these interpretations should **themselves be validated**, since the model's internal reasoning is not directly inspectable.

External Validation in Discovery

One approach to validating an interpretation is to collect a small **manual audit sample**.

A proposed interpretation gains credibility if the discovered representation exhibits **external patterns** — what it correlates with, predicts, or remains distinct from — that should follow if the interpretation is correct.

In short, the role of external validation is to **discipline the interpretation** of a discovered structure.

Discovery and Post-Selection Concerns

Because discovery is exploratory, it is vulnerable to:

- overfitting;
- cherry-picking;
- ex post narrative construction.

Sample splitting

Separate discovery from subsequent measurement: use one sample to identify candidate structures, then apply a refined measurement criterion, informed by discovery, on a **separate sample**.

When Discovered Structure Is the Measure

Alternatively, the discovered representation may itself be the **measurement output**.

In that case, credibility depends on evidence that the structure is:

- well-defined, stable, and robust;
- interpretable;
- externally meaningful.

Construct Definition

In **construct definition**, validity concerns how a theoretically meaningful concept should be concretely defined and operationalized in a specific setting.

Seminal work (Cronbach and Meehl, 1955) highlights the difficulty: many important social science concepts are **latent and high-dimensional**.

- trust;
- respect for authority;
- social capital;
- intelligence.

Key Insight

When direct observation is impossible, validation must be **indirect**.

The Nomological Network

Establishing construct validity for latent or high-dimensional concepts (Cronbach and Meehl, 1955)

A construct is validated not by a single observable variable, but by a **broader pattern of relationships** suggested by theory or the existing literature.

They propose validating constructs through a **nomological network**: theory or contextual knowledge imply what the construct should correlate with, predict, be distinct from, and how stable it should be.

Construct validity requires evidence of both

- **Convergence:** the measure relates to concepts it should relate to;
- **Discrimination:** the measure remains distinct from nearby but different concepts.

Formalizing Construct Validity

In econometric language

This logic is analogous to **partial identification under theoretically motivated moment restrictions**.

- No single restriction establishes that an operationalization captures the intended construct.
- But additional restrictions can **narrow the set** of plausible measures.

Construct Validity Is Important for Economics

Economics has historically emphasized measures for which direct validation is often plausible: *e.g.*, prices; schooling; employment; revenue.

As a result, economics has devoted **less attention** to construct validity than psychometrics or survey design. In many settings it is largely settled.

But this is changing. Economists increasingly study latent, high-dimensional concepts where construct validity is far more **contested**.

AI and Construct Validation

AI has the potential to **improve construct validity** by making empirical assessment of difficult-to-validate concepts more tractable.

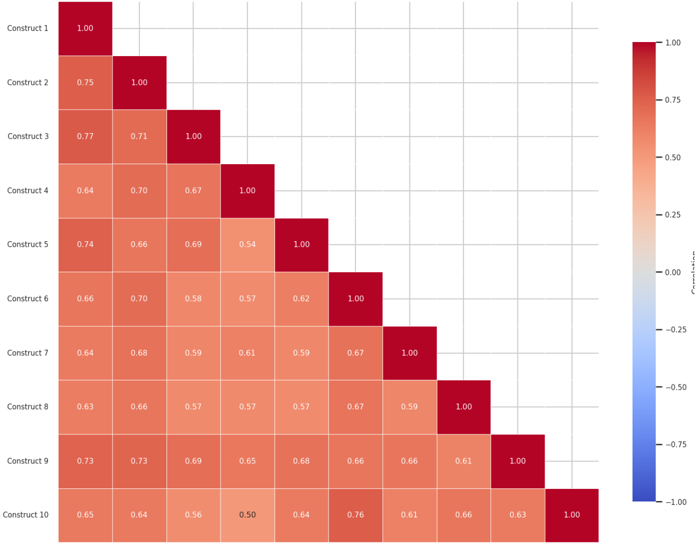
When zero-shot or lightly adapted systems perform sufficiently well, they lower the cost of:

- generating candidate operationalizations;
- measuring observable implications of a nomological network.

Key argument

AI makes it **essential** to develop empirically grounded methods for assessing construct validity.

AI and Constructs: Economic Policy Uncertainty



AI Moves Specification Search Upstream

- A researcher can repeat the downstream analysis using each resulting measure.
- These choices may produce **meaningfully different estimates** of the target parameter.
- AI moves the problem of searching across regression specifications (Sala-I-Martin, 1997) to **searching across representations** of the underlying reality.

The risk

Without principled, widely accepted methods for construct validity, cheap measurement risks encouraging **p-hacking with "AI slop"**, while promoting nihilism about what can be learned from a high-dimensional world.

The Limitations of Construct Validity

- A common critique: construct validation depends on **sufficiently developed theory**, which we sometimes lack.
- If theory is weak, a measure's failure to align with a nomological network may reflect shortcomings of the **theory** more than of the measure.
- Even so, validation remains possible — imperfectly — when researchers can articulate provisional external criteria (*e.g.*, concepts a measure should predict, or remain distinct from).

When Discovery Comes First

If researchers lack credible external criteria, it may be better to treat **discovery as an end in itself**.

- Data-driven discovery can uncover stable, interpretable patterns without strong priors about a not-yet-understood construct.
- These patterns reveal what relationships exist in the data and can inform later developments of theory or construct definition.

Construct vs. Observational Validity

- **Construct validity:** does an operationalization capture the concept needed for the research question?
- **Observational validity:** does a procedure correctly implement that operationalization?

A useful, though imperfect, analogy

- **Observational** validity $\leftrightarrow$ **internal** validity: credibility conditional on a specified target (a construct definition, or an estimand in a given setting).
- **Construct** validity $\leftrightarrow$ **external** validity: whether that target answers the broader research question.

Validating Observation

The premise

Researchers have access to a costly but credible **criterion-based** measurement technology for a subset of observations.

- This ranges from human coding using a well-defined rubric to ground-station weather measurements from a scientific instrument.
- The criterion may be a flawed operationalization of the underlying concept, but it is **interpretable**.
- Observation treats this criterion as **fixed** (in the sample used for observation).

Observation at Scale

Credible criterion-based measurement is costly, so it can often be applied only to a **modest sample**. But researchers can use a cheaper black-box technology to leverage a **much larger** sample.

Examples:

- an LLM applied to news articles to measure economic policy uncertainty;
- a computer vision model applied to satellite imagery to predict weather;
- a machine learning model applied to survey responses to predict costly-to-observe behavior.

The scalable technology need not be AI, though AI-enabled prediction is the focus here.

Observational Validation Samples

Validation Sample

A validation sample contains observations for which **both** the scalable measure and the criterion-based measure are observed.

- Under appropriate assumptions, this sample characterizes the measurement error.
- An appropriately designed validation sample allows estimating the target parameter **consistently and efficiently** in the much larger sample, even under non-classical measurement error.

Observational Validation Is Not Sensitivity Analysis

- **Sensitivity analysis** asks whether conclusions change across different observational procedures.
- **Validation** links measures to an explicit criterion, providing a **benchmark** for interpreting robustness checks.

Why Agreement Across AI Measures Is Not Enough

With AI, researchers can cheaply generate many alternative measures, but these may share **correlated errors** due to:

- common model architectures and training procedures;
- overlapping training data;
- similar prompting choices; etc.

Takeaway

Agreement across AI-generated measures may reflect **shared bias** rather than measurement quality.

Can One AI Measure Instrument Another?

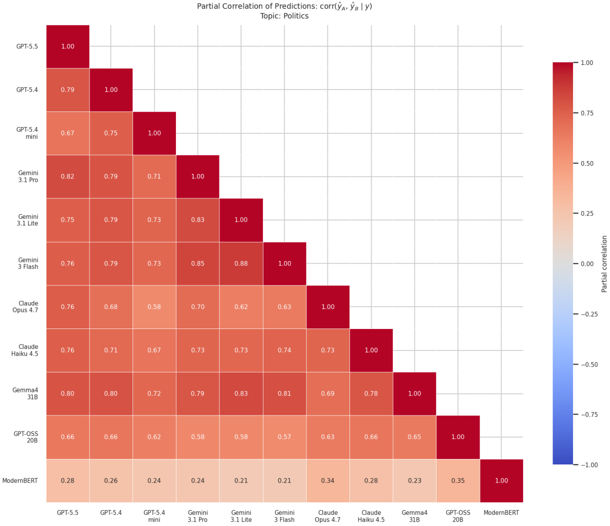
Correlated errors also weaken the case for using one AI-generated measure as an **instrument** for another: the instrument's errors must be independent of errors in the original measure.

But if both measures rely on similar architectures, training data, prompts, and RLHF, their agreement may reflect **shared errors** rather than independent identifying variation.

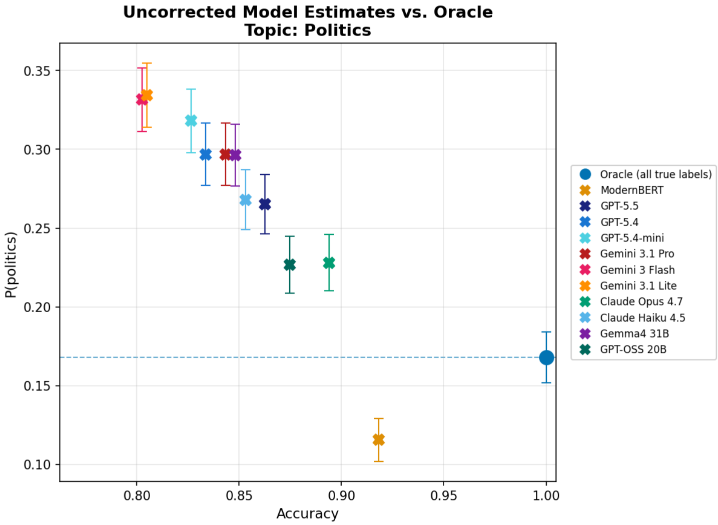
An emerging ML literature on **response homogenization** (Kirk et al. 2023; Jiang et al. 2025) argues that pre-training makes models broadly competent in similar ways, and RLHF pushes them toward similar preferred answers, with output variation driven more by prompting than by model choice.

Bauman et al. (2026) replicate 37 annotation tasks from 21 social-science studies: with prompt paraphrases, **virtually anything** can be made statistically significant. They find incorrect conclusions in $\approx$31% of hypotheses for state-of-the-art LLMs.

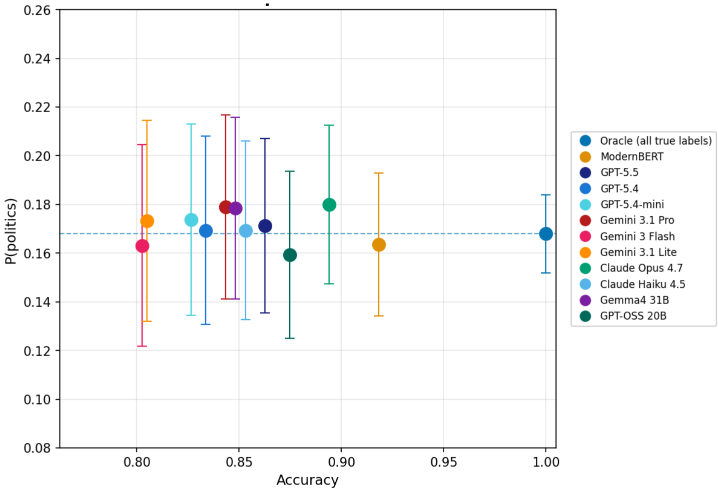
Prediction Error Is Highly Correlated



Estimates Can Be Systematically Biased



Debiased Estimates



Bibliography

Ben-Hur, Asa, Andre Elisseeff, and Isabelle Guyon. 2002. "A stability based method for discovering structure in clustered data." In *Biocomputing 2002*. 6–17. World Scientific.

Campbell, Donald T, and Donald W Fiske. 1959. "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological bulletin*, 56(2): 81.

Cronbach, Lee J, and Paul E Meehl. 1955. "Construct validity in psychological tests." *Psychological bulletin*, 52(4): 281.

Rousseeuw, Peter J. 1987. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, 20: 53–65.

Sala-I-Martin, Xavier X. 1997. "I Just Ran Two Million Regressions." *The American Economic Review*, 178–183.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the royal statistical society: series b (statistical methodology)*, 63(2): 411–423.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function."