Knowledge without Wisdom: Measuring Misalignment between LLMs and Intended Impact

Michael Hardy
Stanford University
hardym[$\alpha\tau$]stanford[$\circ$]edu

Yunsung Kim
Stanford University
yunsung[$\alpha\tau$]stanford[$\circ$]edu

Abstract

LLMs increasingly excel on AI benchmarks, but doing so does not guarantee validity for downstream tasks. This study contrasts LLM alignment on benchmarks, downstream tasks, and, importantly the intended impact of those tasks. We evaluate the performance of leading LLMs (i.e., generative pre-trained base models) on difficult-to-verify tasks of the teaching and learning of schoolchildren. Across all LLMs, inter-model behaviors on disparate tasks correlate higher than they do with expert human behaviors on target tasks. These biases shared across LLMs are poorly aligned with downstream measures of teaching quality and often *negatively aligned with the intended impact* of student learning outcomes. Further, we find multi-model ensembles, both unanimous model voting and expert-weighting by benchmark performance, further exacerbate misalignment with learning. We measure that selection of LLM and/or prompting strategy only reliably accounts for 15% of all measured misalignment error and that variation in misalignment error is shared across LLMs, suggesting that common pretraining accounts for much of the misalignment in these tasks. We demonstrate methods for robustly measuring alignment of complex tasks and provide unique insights into practical applications of LLMs in high-noise contexts.

1 Introduction

> *Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?* (Eliot, 1934)

Large language models (LLMs) now exhibit striking competence on benchmarks that operationalize *knowledge*: answering questions, reproducing domain vocabulary, and generating fluent explanations. We have also seen rapidly growing optimism about using LLMs for tasks that require more than static Q&A, such as scientific discovery and hypothesis generation (Xin et al., 2025; Yamada et al., 2025; Gottweis et al., 2025; Swanson et al., 2025; Cong et al., 2025). Yet recent work in this area highlights failure modes that are not well-captured by standard evaluation: overconfidence in interpreting evidence, brittle multi-step inference, and performance plateaus that appear tied to shared pretraining distributions rather than to idiosyncratic architectures (Song et al., 2025; Alampara et al., 2025; Mirza et al., 2025; Kim et al., 2025; Kalai et al., 2025). Such limitations increase when tasks move downstream, away from "correct answers" and toward some *intended impact* in the world.

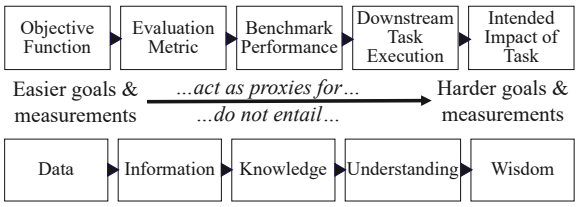


Figure 1: Cascading levels of inference found in LLM development and evaluation and a parallel adaption of Ackoff's progression (Ackoff, 1989; Rowley, 2007)

The research of LLMs-in-scientific-discovery studies illustrates a gap between benchmarks and downstream tasks that is not reserved for prestigious tasks that are difficult for scientific experts. This paper uses similarly rigorous methods to study what might be considered a simpler domain than frontier scientific discovery—elementary classrooms—but one that makes the same scientific point sharply: impressive language competence does not guarantee that a model's judgments align with the implied construct of interest. Classroom instruction is an archetypal high-stakes, high-noise setting where quality may be inferred from unstructured discourse text and where the ultimate objective is delayed: student learning. Yet, in ways similar to many application areas, annotated classroom

discourse is effectively absent from the Internet text[1] that dominates LLM pretraining, raising generalization questions about what is attributable to pretraining or LLM idiosyncrasies: *what shared behaviors do LLMs exhibit when asked to perform an exceptionally uncommon task with respect to their training data?*

1.1 From proxy evaluation to intended impact

The hierarchies of inference in generative AI evaluations can be represented as a series of cascading *proxies*: easier measures acting in place of the true desired outcome (Fig. 1). In children's education settings, common proxy criteria, such as (non-student) user preference (Jurenka et al., 2024; Kornell et al., 2024), are potentially informative but incomplete, because proxies can be optimized without improving schooling's intended impact: student growth. This concern mirrors the literature in other disciplines such as coding and scientific-discovery, where strong performance on question-answering benchmarks (e.g., GPQA, Rein et al. 2023; MMLU-Pro, Wang et al. 2024) can coexist with weaknesses on tasks that are required to deploy real code or conduct real science (Becker et al., 2025; Song et al., 2025; Alampara et al., 2025; Mirza et al., 2025; Kim et al., 2025; Kalai et al., 2025). We posit that there are many other domains with congruent gaps. In education, we observe that models may reproduce the *language* of effective pedagogy without tracking the features of instruction that causally support learning.

As it is for such application areas, standard AI benchmarks, which are typically focused on question-answering or tasks with discrete solutions, are ill-suited for the nuanced, generative, and high-stakes nature of educational applications (Gehrmann et al., 2022; Hu and Levy, 2023; Zhou et al., 2023a, 2024; Kim et al., 2024; Wu and Aji, 2023; Reuel et al., 2024; Hardy, 2026). In this setting, one could build AI systems that *act* pedagogically sound while failing to identify teaching practices that actually improve achievement, risking deploying technologies that are not only ineffective but potentially harmful to student learning (Bastani et al., 2024; Shein, 2024; Rismanchian et al., 2026; Lee et al., 2024).

To avoid these pitfalls, we use two external criteria that, to our knowledge, have only been linked by one other LLM study (Hardy, 2025a): (i) ***Downstream Task*, which is expert human observation annotations** using real-world instructional instruments, and (ii) ***Intended Impact* through value-added measures (VAMs) of long-term student achievement gains** for those same classrooms. The latter is considered the "gold standard" for measuring impact on student learning. Methodologically, we treat both as alignment targets: alignment with the *downstream task* (expert ratings of teaching practice) and alignment to the *intended impact* (predicting which classrooms produce greater learning gains). A primary contribution at the intersection of LLMs and classrooms, it evaluates LLMs using outcome-based criteria rather than human preference alone (see sections 4.2 and 4.1).

High-level Trends Across analyses in the present study, a consistent picture emerges that parallels recent results in other disciplines: models can converge on confident, mutually reinforcing judgments that are poorly tethered to the underlying target. In classrooms, "knowledge" of pedagogical concepts does not reliably translate into the "wisdom" to discern what is relevant to human student learning.

1.2 Contributions to High-noise Contexts

The study provides several important contributions to the current study of applied LLMs. First, it directly quantifies a novel gap between LLM execution of downstream tasks and the *intended impact* (see section 5.2). Then, using ensembling, we test whether LLM idiosyncratic competence (by weighting models by "pedagogy expertise" on benchmarks) or shared pretraining (by using consensus/unanimity) mitigates misalignment. Finally, we decompose the variance in misalignment attributable to the two levers practitioners most often control, model choice and prompt choice, with a method that can be generalized to other alignment studies.

A key contribution is the general methodological framework we use for measuring LLM alignment in high-noise contexts where strong experimental controls may not be feasible:

1. **Correlate behaviors across space of generalization**: Demonstrate correlation of behaviors across the types of models, prompting and post-training techniques, environments, and/or harnesses across which we hope to generalize desired downstream tasks. § 3.1
2. **Measure best-proxy alignment**: Measure (mis)alignment on downstream tasks.

[1] see Limitations in 6.

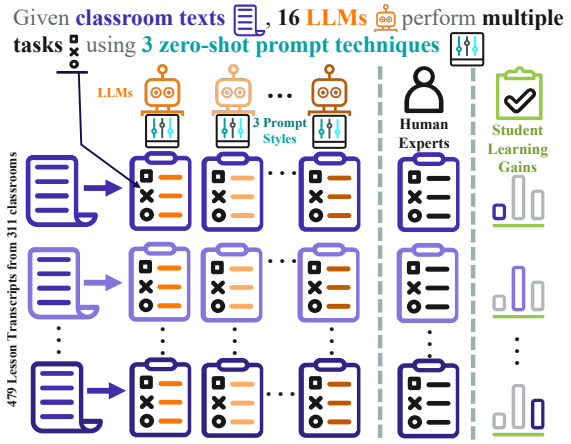


Figure 2: **Task Data and Experimental Design**. Each LLM is provided classroom transcripts. Using several prompting techniques for each model, LLMs place an ordinal rating on the quality on an aspect of teaching and learning. This is done across seven distinct tasks. We evaluate for alignment of LLM values, and not for accuracy, when comparing the relative ranking provided by each LLM on each task with human experts and with student learning gains for the class.

Kendall's τ is a simple, strong option that can be estimated with any datatype. § 3.2

3. **Determine impact alignment**: Data representing intended impact may have time-delays after the original downstream tasks or may have grain-size mismatches. For high-noise contexts, establish real-world baselines using best available predictors to be able to discern meaningful and practical reference points for comparison. § 3.2
4. **Decompose the error:** decompose misalignment error variance across the space of generalization (Meehl, 1990; Brennan, 2001b) to quantify contributions to the error. § 3.3

2 Experimental Design

Using transcripts from primary school mathematics classrooms, we prompt a suite of 16 leading LLMs to assign ordinal ratings based on a rubric (Fig. 2) where each transcript is evaluated across multiple observation dimensions. We then measure the directional alignment between (a) LLM scores with expert human ratings on the same dimensions and (b) LLM scores with VAMs. Because primary school discourse is effectively absent from pretraining, we can measure how well models generalize to a mismatched distribution.

For each classroom lesson in our test set, we provide multiple LLMs with a transcript segment. The models are prompted to perform seven distinct tasks, each focused on a different dimension of teaching and learning (details in Section 4). We first use the bias-corrected squared distance correlation dCor_n^2 to measure any deviation from independence between the ratings. Our core alignment analysis compares pairwise directionality: we assess whether expert human ratings or student learning data also lesson pairs in the same order as LLMs. This approach, which evaluates the alignment of second-moment rank-based information, provides a direct and robust measure of alignment that is insensitive to variations in rating distributions and allows for authentic baseline comparisons. We then use multiple ensemble methods to amplify/attenuate shared misalignment signals. To understand the sources of misalignment, we structurally decompose observed misalignment errors by the variables in our space of generalization: LLM, prompting strategy, item/task, and class transcript.

3 Methods

Even among other high-noise contexts, evaluation of educational applications is particularly challenging (Kraft, 2020; Jurenka et al., 2024) and, at best, is based on noisy instruments (McCaffrey et al., 2009; Kane and Staiger, 2012; Kane et al., 2013; Hardy, 2024) measuring latent constructs (Messick, 1995; Hill et al., 2012a) and questionable data quality (Ho and Kane, 2013; Xu et al., 2024). Our methodology is designed to measure the alignment between the relative ordering of teaching quality as judged by LLMs, expert humans, and student learning outcomes. We eschew direct comparisons of absolute rating scores, which are susceptible to noise and idiosyncratic scale use by both humans and models. Instead, we treat LLM ordinal outputs as Thurstonian indicators of their pedagogical values by focusing on pairwise concordance, a robust measure of directional agreement, inspired by research on AI safety and utility (Mazeika et al., 2025; Huang et al., 2025).

3.1 Measuring dependence with dCor_n^2

To measure any amount of dependence between observations, including nonlinear and nonmonotonic, we use the Bias Corrected Squared Distance Correlation (Székely et al., 2007; Szekely and Rizzo, 2014) to determine the strength of the relationships between tasks and raters. We report the average squared correlations disaggregated by relationship

type in Figure 3. Inter-LLM specific patterns of shared behavior are more visually striking in Figs. 10 and 11. Additional details are in Appendix E.1.

3.2 Measuring alignment with Kendall's τ

To formalize this concept of alignment, we employ Kendall's τ, a nonparametric and robust measure of concordance (Kendall, 1945; Bishara and Hittner, 2017). We reconstruct its formulation here to motivate it as an intuitive and precise measure of alignment between two sets of scores, such as those from an LLM, x, and an outcome metric, y (e.g., expert scores or student learning gains). Consider a set of n lessons. For any pair of distinct lessons, indexed by i and j, we can evaluate whether the ratings from source x and source y agree on their relative order:

LLM X rates lesson i as better than lesson j: $x_i > x_j$. Does this align with human experts Y, $y_i > y_j$? Does this align with student learning Z associated with each lesson, $z_i > z_j$?

Adding brackets as indicator functions we get: $x_{ij} = [x_j > x_i] - [x_j < x_i]$ and $y_{ij} = [y_j > y_i] - [y_j < y_i]$ where the alignment between x and y for two lessons is simply the product $x_{ij}y_{ij}$. All pairwise comparisons between lesson ratings from LLM X and some outcome Y are collected into the antisymmetric matrices $X = (x_{ij})$ and $Y = (y_{ij})$, respectively. Measuring alignment between X and Y is the aggregation of all pairwise directional alignments (is lesson i better than j) across all lessons:[2] $\tau_{XY} = \langle X, Y \rangle_{\mathrm{F}} / \|X\|_{\mathrm{F}} \|Y\|_{\mathrm{F}}$ where $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{F}}$ are the Frobenius inner product and Frobenius norm, respectively. This formulation mathematically mirrors the needs of the alignment question, and does so by reducing the sensitivity to noise in the ratings. This scale-independent approach allows us to establish meaningful real-world baselines, such as using a teacher's years of experience or a prior VAM as the ratings (Fig. 4).

3.3 Decomposing error variance

Aggregate misalignment metrics (correlations, mean error) do not reveal *why* LLMs fail: the same overall error can be produced by (i) fixable implementation choices (model selection, prompting), (ii) shared, systemic biases that persist across models, or (iii) transcript- and construct-specific difficulties. To localize failure modes, we structurally decompose the observed misalignment error into fully crossed variance components associated with each facet of our evaluation design.

Let $c \in \mathcal{C}$ index classroom transcript segments (observations), $i \in \mathcal{I}$ rubric items, $m \in \mathcal{M}$ foundation models, and $p \in \mathcal{P}$ prompt families. For each cell (c, i, m, p) we observe a standardized LLM score $\widetilde{X}_{cimp}$ and an aligned (pre-standardized) value-added outcome $\widetilde{Y}_c$. The (unsigned) misalignment error is the squared difference

$$\hat{e}_{cimp} = (\widetilde{X}_{cimp} - \widetilde{Y}_c)^2. \quad (1)$$

We then fit a fully crossed random-effects model (Generalizability Theory; Brennan, 2001b) that partitions $\hat{e}_{cimp}$ into main effects and interactions among $\{\textsc{Obs} = c, \textsc{Item} = i, \textsc{LLM} = m, \textsc{Prompt} = p\}$:

$$\hat{e}_{cimp} = \mu + \sum_{\emptyset \neq \alpha \subseteq \{c,i,m,p\},\, |\alpha| \leq 3} \nu_\alpha + \eta_{cimp}, \quad (2)$$

where each $\nu_\alpha \sim \mathcal{N}(0, \sigma^2_\alpha)$ is a mean-zero random effect for facet set α, and $\eta_{cimp} \sim \mathcal{N}(0, \sigma^2_\eta)$ is the cell-specific remainder comprised of the four-way interaction confounded with residual noise.

For each component k we report its *variance share* $\pi_k = \sigma^2_k / \sigma^2_{\mathrm{tot}}$ (Eq. 6), which quantifies how much of the misalignment landscape is attributable to (a) developer-controlled levers (LLM, PROMPT, and their interactions) versus (b) evidence- and construct-conditioned effects (OBS, ITEM, and interactions). Appendix D provides the full model specification, estimation details, code, sign-preserving variants, and a decision-study analysis for how many models/prompts are needed to stably recover the shared misalignment signal.

4 Data

All data used in this study originate from publicly available sources, ensuring the reproducibility of our findings. The core dataset is from the National Center for Teacher Effectiveness (NCTE) Main Study (Kane et al., 2015), a landmark project that collected extensive data on teaching and learning over three years. The NCTE dataset comprises observations of roughly 350 4th and 5th-grade mathematics teachers across four U.S. school districts. It is one of two educational datasets that contains measures of teaching practice, authentic classroom artifacts, and student learning outcomes at scale (VAMs).

[2] Confidence intervals are computed using the correction from (Fieller et al., 1957)

4.1 Classrooms and downstream task observation instruments

Our primary input for the LLMs are anonymized transcripts (Demszky and Hill, 2022) of video-recorded classroom lessons using the test set defined by (Wang and Demszky, 2023). Human raters (Kane et al., 2015) rated lessons by watching the videos. These same lessons were previously evaluated by teams of expert human raters using two validated, multi-dimensional observation instruments currently used in the field: **Mathematical Quality of Instruction (MQI):** A content-specific framework for evaluating the richness and precision of mathematics instruction (Hill et al., 2008) and **Classroom Assessment Scoring System (CLASS):** A framework for assessing general dimensions of classroom quality, including behavior management and class climate (Pianta et al., 2008). The 63 MQI raters were recruited for their mathematics instruction expertise and underwent rigorous certification and continual calibration to ensure scoring reliability, a practice also used with the 19 CLASS raters (Blazar et al., 2017; Kane et al., 2015). The expert scores from these instruments serve as our first target for LLM alignment.[3]

4.2 VAMs: Value-added to student learning

To connect model outputs to the intended impact of teaching, we use value-added measures (VAMs) of student learning. VAMs are widely considered the gold standard for statistically estimating a teacher's causal effect on student achievement gains (Bacher-Hicks et al., 2017, 2019; Kane and Staiger, 2012). A VAM quantifies how much a teacher's students grew academically over a school year compared to their expected growth, controlling for prior achievement, context, peer effects, and other student-level covariates. The NCTE dataset provides multiple high-quality VAM scores for each teacher. Following established practice (Kane and Staiger, 2012), we use stacked VAMs (Apdx. E.2, Kane et al., 2015) for each teacher-year corresponding to the observed lesson estimate of a teacher's contribution to student learning. Crucially, our evaluations assume that improved teaching practices are positively associated with improved gains to student learning, in aggregate and on average, even if all sources of noise cannot be removed. To our knowledge, this is the first study to use VAMs as a benchmark for evaluating generative LLMs.

[3]Human rater data:https://www.icpsr.umich.edu/web/ICPSR/studies/36095/datadocumentation; Transcripts are available for 1,600 of the lessons (Demszky and Hill, 2022) https://github.com/ddemszky/classroom-transcript-analysis; replication test set and prompts https://github.com/rosewang2008/zero-shot-teacher-feedback. Full replication LLM output for this study can be found https://drive.google.com/file/d/1fP7xyKasJ4Ui6di-SlY3TLdNQCeg59Tr/view?usp=sharing

5 Results and Discussion

5.1 The convergent bias of foundation models

A noteworthy result is the striking *behavioral homogeneity* of LLMs when evaluating classroom transcripts. As summarized in Figure 3, different models' ratings are substantially more correlated with one another than with expert human ratings, both within the same task and across different instructional tasks.

Two patterns are especially notable. First, **LLM-LLM agreement is consistently higher than LLM-human agreement**. Second, **within- and between-model inter-task correlations are high**, indicating that a model's outputs for distinct instructional constructs (e.g., language support vs. remediation) tend to move together more than expert ratings do. In other words, when confronted with authentic classroom discourse, models appear to rely on a shared latent heuristic of "good teaching" that is not strongly anchored to the constructs that human observers are trained to distinguish.

This convergence is plausibly explained by what these systems share: an autoregressive pretraining objective and large-scale Internet text. Authentic elementary classroom discourse is largely absent from such corpora, forcing generalization under distribution shift. The resulting shared bias echoes emerging evidence that, as models scale, their representations and judgments can become increasingly correlated across developers and architectures (Kim et al., 2025; Ren et al., 2024; Huang et al., 2025). The following section investigates whether these model convergences are aligned.

5.2 Perils of proxy alignment

Figure 4 juxtaposes two forms of alignment for each model and task: pairwise concordance correlations with expert ratings ($\tau_{S_f X}$, x-axis) and with student learning gains ($\tau_{S_f Y}$, y-axis). The central empirical finding is a **systematic disconnect** between these axes. Models that appear more aligned to expert judgments are *not* correspondingly aligned to learning, and in many cases, are

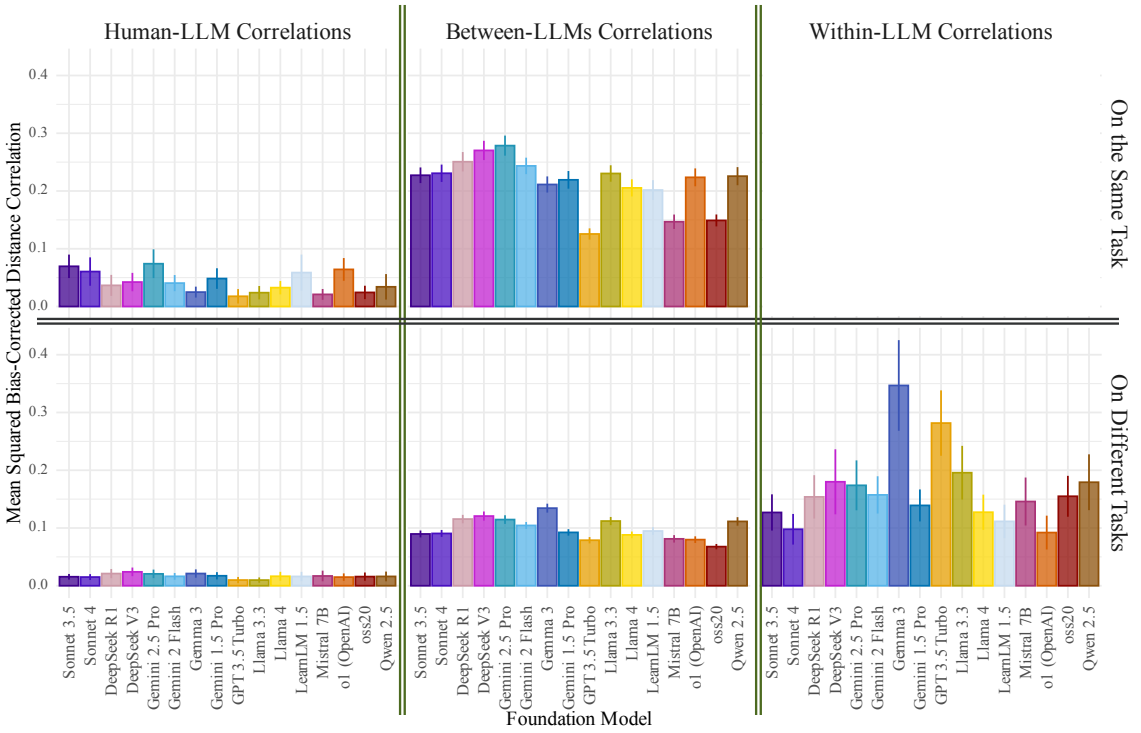


Figure 3: **Mean Inter-task Bias Corrected Squared Distance Correlations** dCor_n^2 : between LLMs and human raters across different evaluation tasks. **Top row: Same-task Correlation** Mean inter-rater distance correlations across transcripts for the same task and (**bottom row: different task correlation**) for different tasks using the same transcript. (**left: correlations with humans**) Mean inter-rater distance correlations with expert human raters, (**center: correlations with other LLMs**) with other LLMs, and (**right: intramodel intertask correlations**) each LLM with itself. The top right is omitted for redundancy. Lines are standard errors for each estimated mean. Means and SEs were computed under Fisher's z transformation and back transformed to preserve variance. The complete correlation matrix for each task and model are found in Figs. 10 and 11

more negatively associated with learning outcomes. The magnitudes of the baselines are consistent with the literature (McCaffrey et al., 2009; Kane and Staiger, 2008, 2012) (see Appendix C.3).

This pattern constitutes a particularly consequential failure mode: **proxy alignment without impact alignment**. A system can appear to "do the job"—produce plausible, rubric-concordant scores—while selecting classrooms that are *worse* on the objective schooling ultimately values. This mirrors the caution from scientific-discovery evaluations: benchmark success can obscure fragility in the behaviors that matter when evidence is messy and objectives are implicit rather than explicitly labeled (Song et al., 2025; Zhou et al., 2023a, 2024).

We also observe the converse: some smaller or older models occasionally show slightly better $\tau_{S_f Y}$ while exhibiting weaker $\tau_{S_f X}$. Qualitative inspection of failures suggests these cases often reflect *task noncompliance*: rather than applying the intended rubric, models may latch onto superficial transcript features that, coincidentally, correlate with higher achievement gains in this sample. The implication is that neither "sounding pedagogical" nor matching expert observation scores is sufficient evidence that the model has learned a transferable representation of effective instruction.

We analyze whether additional test-time "reasoning" variants yield improvements analogous to those experienced in many complex tasks. Comparing paired models sharing the same base (e.g., DeepSeek-R1 vs. DeepSeek-V3.1; and similarly GPT-3.5-class vs. reasoning-oriented variants), we find **no measurable improvement** on either axis of alignment in Figure 4. Findings comparing the chain-of-thought prompt were similar. For classroom evaluation, additional reasoning context window alone does not appear to repair the core mismatch between model judgments and constructs that predict learning gains.

5.3 Ensembling exacerbates misalignment

A natural response to noisy model behavior is ensembling. We evaluate two conceptually oppo-

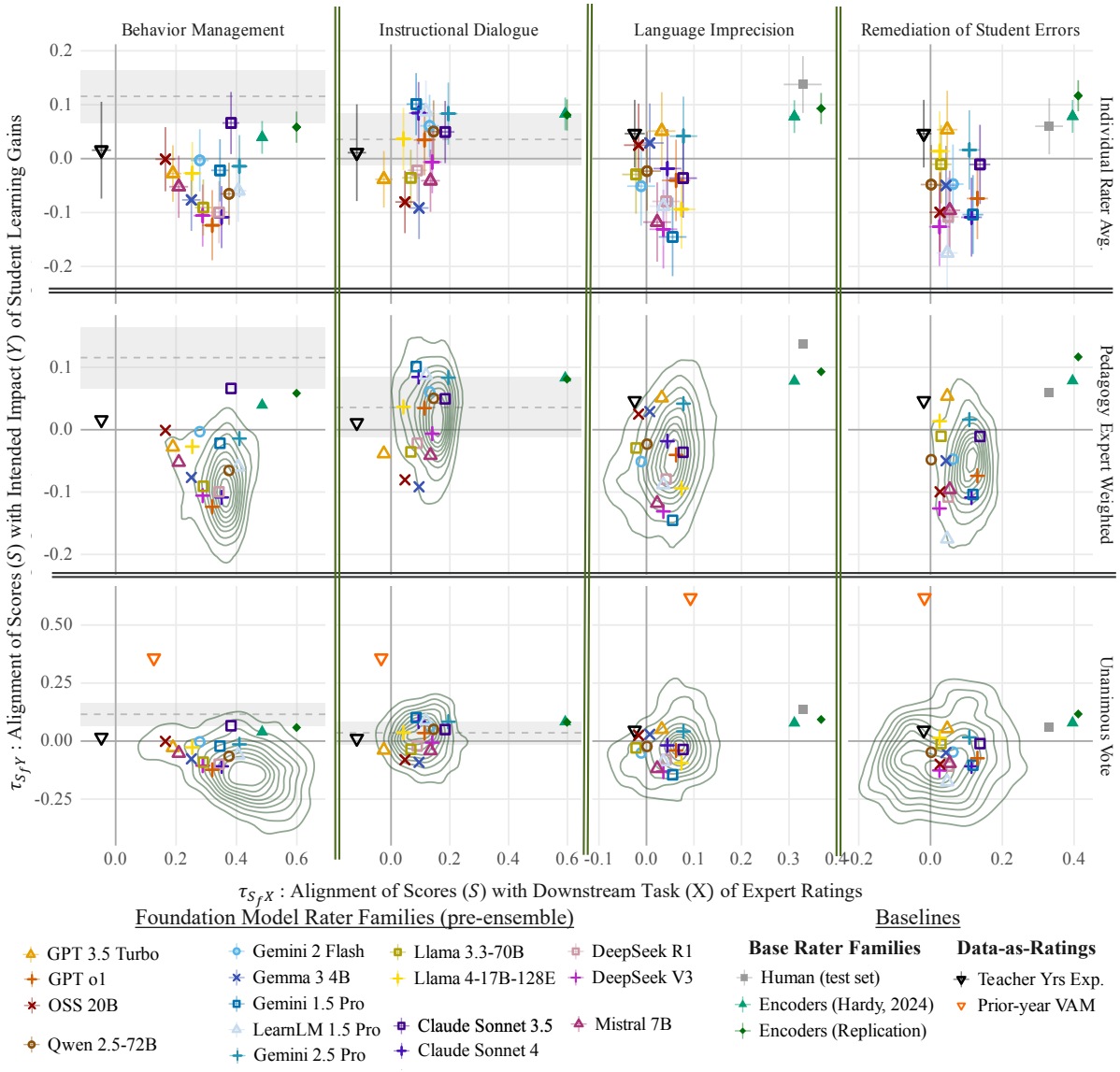


Figure 4: **(Mis)alignment with Downstream Task (Teaching) and Intended Impact (Learning)**: The **x-axes** measure the alignment of scores (S_f) from each LLM, ensemble, and baseline f with expert human ratings on downstream tasks (X) on the quality of teaching skills in a given lesson: $\tau_{S_f X}$. Similarly, the **y-axes** measure alignment with the value-added to learning via student achievement gains (Y): $\tau_{S_f Y}$. Each color-shape combination represents a different rater family or baseline. Each column represents a specific instructional rating task, from (**left** to **right**): `CLBM`, `CLINSTD`, `LANGIMP`, and `REMED`. Each row represents a distinct implementation scenario: **Top row (individual rating models)** The 95% CI are shown for individual models. **Middle row (Pedagogy Expertise Weighted Ensembles)** and **Bottom row (Unanimous Vote Ensembles)**: The alignment correlations of all possible model-prompt ensembles are represented by the density contours. The innermost contours represent the alignment regions' greatest ensemble performance density. For the bottom row ensembles, we amplify the shared signal by only measuring observations where all three models are in agreement (See Appendix E.3). **Baselines**: the gray line and shaded regions on `CLBM` and `CLINSTD` represent the estimate and confidence intervals from the human expert rating alignment with student learning gains on the study's joint test set. The green Encoder models (data from (Hardy, 2024) and our own replications) are LMs but not LLMs, and they merely serve as a deep learning baseline for possible alignment based on transcripts. The **baseline of Teacher Experience** represent pairwise comparisons that always rank more experienced teachers higher than less experienced ones. For better plotting details, the "oracle" **VAM baseline** which puts a higher value to teachers with higher prior year VAMs, is only displayed on the bottom row. Random baselines (both uniform and stratified) predictably clustered around the origin and, with no real-world example of this as a valid baseline, we exclude them from the plots.

Table 1: **Misalignment Shares**: Proportion of Variation in Squared Error by Source. Row partitions separate main, second-order, and higher-order effect posterior distributions.

Facet of Error Variation	median	MAD	MAP	95% HDI
ITEM	0.08	0.08	0.03	[0.01,0.62]
LLM	0.07	0.04	0.06	[0.02,0.18]
OBS	0.02	0.01	0.02	[0,0.03]
PROMPT	0.02	0.03	0.00	[0,0.58]
ITEM:OBS	0.04	0.01	0.05	[0.01,0.06]
LLM:ITEM	0.03	0.01	0.03	[0.01,0.06]
LLM:OBS	0.00	0.00	0.00	[0,0]
LLM:PROMPT	0.02	0.01	0.01	[0,0.04]
PROMPT:ITEM	0.01	0.01	0.00	[0,0.05]
PROMPT:OBS	0.00	0.00	0.00	[0,0]
LLM:ITEM:OBS	0.19	0.03	0.21	[0.05,0.23]
LLM:PROMPT:ITEM	0.03	0.01	0.03	[0.01,0.05]
LLM:PROMPT:OBS	0.14	0.02	0.15	[0.04,0.17]
PROMPT:ITEM:OBS	0.02	0.00	0.02	[0.01,0.03]
LLM:PROMPT:ITEM:OBS + ϵ	0.24	0.04	0.26	[0.06,0.29]

site approaches: (i) a *pedagogy-expertise-weighted* ensemble, where model votes are weighted by pedagogical benchmark performance, highlighting LLM uniqueness, and (ii) a *unanimous-vote* ensemble, where we score only those instances where all models agree, emphasizing shared signal (§E.3).

Neither strategy improves alignment with student learning. Instead, both frequently **worsen** $\tau_{S_f Y}$, especially on core instructional dimensions such as remediation of student errors and behavior management (Figure 4). This result has two implications. First, it suggests that benchmark-measured pedagogical "knowledge" does not translate into reliable recognition of effective pedagogy in authentic discourse (i.e., the benchmark construct is not externally valid for this downstream setting). Second, it indicates that when models agree, they may be amplifying a shared but flawed heuristic; consensus is not evidence of correctness with correlated errors (Chen et al., 2024; Zhu et al., 2025).

5.4 Persistent artifacts of autoregression

What the decomposition tells us (and why we need it). Aggregate metrics (e.g., mean error, correlation) cannot distinguish *fixable* misalignment from *structural* misalignment. Our fully crossed variance decomposition over OBS (c), ITEM (i), LLM (m), and PROMPT (p) localizes where misalignment lives: in developer-accessible choices (LLM, PROMPT), in the evidence and construct being scored (OBS, ITEM), or in their interactions. This turns a vague diagnosis ("the model disagrees with VAM") into an actionable one ("error is dominated by transcript-conditioned interactions, so model shopping and prompt iteration will not reliably fix it").

Model choice and prompt choice are weak levers. Table 1 shows that the main effects of LLM and PROMPT explain only a small portion of the squared-error landscape (median shares 0.07 and 0.02), and even their interaction remains small (median 0.02). If misalignment were caused by a few deficient models or a single poor prompting recipe, these components would be large and stable. Instead, the decomposition implies a more sobering practical conclusion: swapping models and prompts may change outcomes locally, but it is unlikely to yield a *reliable* improvement in validity against intended impact.

Prompting is brittle, not corrective. Although the typical prompt effect is small, its posterior is long-tailed (wide HDI despite a low median), indicating *prompt brittleness*: prompts can occasionally inject large error without delivering consistent gains across transcripts. The risk is asymmetric: prompt changes can produce dramatic "wins" on a handful of cases while silently degrading performance elsewhere, an often missed failure mode.

Misalignment concentrates in context-conditioned interactions. The dominant variance shares occur in higher-order interactions that condition on the classroom text, especially LLM$\times$ITEM$\times$OBS (0.19) and LLM$\times$PROMPT$\times$OBS (0.14), together with a substantial cell-specific remainder (0.24). This is the signature of a transcript-conditional failure mode: models behave unpredictably on particular kinds of instructional evidence. Scientifically, that pattern is consistent with LLMs relying on latent proxies (fluency, affect, participation cues, stylistic norms) that are only loosely coupled to student learning gains and whose influence varies with the segment.

The direction of error is more shared than the magnitude. Squared error $\hat{e}$ hides whether models over- or under-estimate impact. When we preserve direction using the signed quadratic error $\hat{e}^{\pm}$ (Appendix D.2), the ITEM$\times$OBS share increases substantially (Table 5). This indicates that models often drift in the *same direction* on the same item–segment pair, even when they disagree stochastically about how large the error will be. In other words, LLMs may look noisy at the surface, yet

still share a coherent (and undesirable) inductive bias about what "good teaching" looks like in text.

Implication: validity fails in regimes where users most need reliability. Taken together, these findings explain why surface-level interventions underperform. The misalignment is largely not a property of a single model or prompt; it is an evidence-conditioned artifact that persists across models and is revealed only when we ask about intended impact. For high-stakes applications, this motivates a different optimization target: reducing the *shared* misalignment component (the part that survives averaging over models/prompts), rather than chasing occasional prompt-dependent gains on fixed transcript sets.

6 Conclusion

This paper studies a common but under-instrumented problem in contemporary NLP: LLMs can exhibit strong internal agreement and high apparent competence while failing to align with the outcomes a domain actually values. Here, we evaluate LLM-based scoring of classroom instruction against two external criteria: expert human observations (the downstream task) and value-added measures (VAM) of student learning gains (the intended impact). Across leading models and prompting strategies, we find that (i) models converge behaviorally, (ii) their convergence is misaligned with intended impact, as alignment with the downstream task is not a strong enough proxy, and (iii) common ensembling and prompt engineering do not reliably repair the gap .

A key methodological contribution is a structural decomposition that makes these failures measurable in high-noise regimes. Rather than treating misalignment as a scalar, we model the observed error as arising from multiple, simultaneously operating facets—ITEM, OBS, LLM, and PROMPT—and their interactions in a fully crossed design. This variance decomposition changes what an evaluation can claim. It supports a practitioner-relevant statement of the form: *how much of the observed error is plausibly removable by changing implementation choices (model/prompt), and how much persists as transcript-conditional, shared behavior that will remain even if one "shops" for a better model?* In our data, the dominant error structure concentrates in higher-order interactions involving the classroom evidence, with comparatively small and brittle contributions from prompt and model main effects. The implication is concrete: when misalignment is concentrated in evidence-conditioned interactions rather than in stable model or prompt effects, iterative prompt engineering and model substitution are unlikely to yield reliable validity with respect to intended outcomes.

More broadly, the decomposition offers an evaluation template for other AI and NLP problems where (a) labels are noisy, (b) the unit of prediction mismatches the unit of impact, or (c) the true target is only observable through delayed, aggregated, or confounded measurements. Examples include human-centered summarization (where user decisions matter more than ROUGE), clinical decision support (where outcomes may occur weeks or months later), content moderation (where harms are downstream and partially unobserved), and policy or educational tools (where success is causal and context-dependent). In such domains, chasing marginal benchmark gains can be rational while still missing the central scientific question: *does the system improve the world it is deployed in?*

We therefore argue for a shift in evaluation practice. First, NLP evaluation should more often include an explicit intended-impact target—however noisy—and should report uncertainty and sensitivity rather than only point estimates. Second, when noise is unavoidable, the correct response is not to abandon measurement, but to use designs and estimands that are robust to it: multifacet decompositions, decision studies, hierarchical rater models, residualization and mediation analyses, and pre-registered stress tests. Third, the community should treat "high noise" not as a disqualifier for rigorous work, but as a signal that the domain is real: many high-stakes applications are difficult precisely because outcomes are multi-causal, delayed, and imperfectly measured. We need to listen more closely to measure the noise.

Finally, our results suggest a cautionary principle for applied LLM evaluation: *knowledge without wisdom is detectable*. When models exhibit strong consensus yet their shared error correlates poorly or negatively with intended impact, scaling and superficial alignment interventions are unlikely to suffice. Progress will require methods and training signals that explicitly bind model judgments to causal, downstream consequences, and evaluation frameworks that can reveal when apparent competence fails to translate into real-world benefit. In short, LLM evaluation should demonstrate the wisdom that our very smart LLMs may lack.

Limitations

Our central aim is methodological: to measure alignment between LLM-generated evaluative scores and both expert ratings (a downstream task target) and student learning gains (an intended-impact target) under realistic noise. This design necessarily inherits the constraints of education measurement. Value-added measures (VAM) are time-delayed measures of causal impact; transcript segments are partial, lossy views of instruction; and expert rubrics, while informative, exhibit nontrivial rater disagreement. These limitations do not invalidate the study—they define the regime we seek to operate in—but they bound the claims we can make. Accordingly, we interpret variance decomposition results as statements about *where observed misalignment concentrates* under this measurement system, not as a definitive census of all sources of pedagogical effectiveness. We also emphasize that our estimates are conditional on the sampled items, segments, models, and prompt families; extending the universe of items, observation windows, or prompting mechanisms could change component magnitudes, even if the core inference about controllable versus systemic error structure persists.

In the presence of low reliabilities observed by human annotators, we echo that there is still substantial and meaningful information, even with the measurement error found throughout education contexts (Ho and Kane, 2013; McCaffrey et al., 2015; Hardy, 2025a). The test set (Wang and Demszky, 2023) may have unobserved confounding factors in its construction. To account for this, we also investigated other methods of estimating these effects given the complexity of the relationships. In the extended Appendices F and G, we demonstrate a multi-stage residualization of confounding and mediating effects. We are pleased to report that our conclusions of this paper are robust to and even are strengthened by these additional tests. For additional information, context, and tests for working with high noise data, see also Appendix C.4 and (Brennan, 2001a; McCaffrey et al., 2009; Brennan, 2013; Hardy, 2024; Casabianca, 2025).

The transcript data used in this work contain only fourth- and fifth-grade mathematics classrooms from the United States. Furthermore, the associated ratings pertain solely to a subset of rating items on a specific rubric, which may introduce limitations when addressing other tasks of classroom instructional support for children. While there is no evidence to suggest that findings would be different in other primary classrooms, the data make generalization to all classrooms not demonstrable in the current study.

Meaningful representation of student classroom learning is absent on the internet largely for laws protecting children's privacy. The text for our data was only anonymized and made public in 2022 (Demszky and Hill), and it is not "crawlable"; nor is it easy to link these data with the (not crawlable) annotations and value-added measures of student learning (VAMs) (Kane et al., 2015). The linking of these two sources makes it 1 of 2 extant datasets having classroom interactions with both expert ratings on real teaching instruments and VAMs. The second dataset ("MET Project") is not publicly available. Part of our interest in doing this research is that conducting these studies is extremely expensive, and economic drivers exclude this line of work from studies that would measure impact (see the exclusion of education, social work, therapy, and other fields from studies like, for instance, Patwardhan et al. 2025).

One purpose of this study is to measure the extent to which LLMs have the capacity to conduct tasks in downstream applications ***responding to classroom teaching and learning***, where these challenges not only are particularly prominent but also bear significant real-world consequences. Authentic educational content, particularly for school-age children, is effectively non-existent on the Internet and therefore absent in the pretraining data of large LLMs.[4] See preliminary Appendix Ethical Considerations section A.3 for implications.

Ethical Considerations

Due to page requirements, see extended Ethical Considerations in the preliminary Appendix A.

Acknowledgments

We thank Ben Domingue and Chris Piech for their time and insights. We are also grateful for the feedback from members of Stanford Trustworthy AI Research (STAIR) Lab, Stanford Lytics Lab, Piech Lab, Stanford's Social NLP working group, and Stanford's "Researching, Presenting and Publishing Work in AI & Education" Group from Spring 2025. Finally, we thank the reviewers for their constructive feedback.

[4] Part of this underrepresentation is a result of the protection of data and educational records of minors

References

Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21(1):10.

Daron Acemoglu and Pascual Restrepo. 2022. Tasks, Automation, and the Rise in U.S. Wage Inequality. *Econometrica*, 90(5):1973–2016.

Russell Ackoff. 1989. From data to wisdom. *Journal of applied systems analysis*, 16(1).

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology.

Elena Aguilar. 2013. Developing a Work Plan: How Do I Determine What to Do? In *The art of coaching: effective strategies for school transformation*, pages 119–144. Jossey-Bass, A Wiley Brand, San Francisco.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5(10):952–961.

Andrew Bacher-Hicks, Mark Chin, Thomas Kane, and Douglas Staiger. 2017. An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys. Technical Report w23478, National Bureau of Economic Research, Cambridge, MA.

Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73:101919.

Peter Riley Bahr. 2007. Double Jeopardy: Testing the Effects of Multiple Basic Skill Deficiencies on Successful Remediation. *Research in Higher Education*, 48(6):695–725.

Nishant Balepur, Jie Huang, and Kevin Chang. 2023. Expository Text Generation: Imitate, Retrieve, Paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.

Paul Bambrick-Santoyo. 2016. *Get better faster: a 90-day plan for coaching new teachers*. Jossey-Bass, A Wiley Brand, San Francisco, CA.

Paul Bambrick-Santoyo. 2018. *Leverage leadership 2.0: a practical guide to building exceptional schools*. Jossey-Bass, San Francisco, CA.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2024. Generative AI Can Harm Learning.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1).

Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. 2025. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. *arXiv preprint*. ArXiv:2507.09089 [cs].

Ruha Benjamin. 2019. *Race after technology: abolitionist tools for the New Jim Code*. Polity, Cambridge, UK Medford, MA.

Alexander Bick, Adam Blandin, and David J. Deming. 2024. The Rapid Adoption of Generative AI.

Anthony J. Bishara and James B. Hittner. 2017. Confidence intervals for correlations when data are not normal. *Behavior Research Methods*, 49(1):294–309.

David Blazar, David Braslow, Charalambos Y. Charalambous, and Heather C. Hill. 2017. Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments. *Educational Assessment*, 22(2):71–94. _eprint: https://doi.org/10.1080/10627197.2017.1309274.

David Blazar and Cynthia Pollard. 2022. Challenges and Tradeoffs of "Good" Teaching: The Pursuit of Multiple Educational Outcomes. Technical report, Annenberg Institute at Brown University. Publication Title: EdWorkingPapers.com.

Kathryn Bonney, Cory Breaux, Cathy Buffington, Emin Dinlersoz, Lucia S. Foster, Nathan Goldschlag, John C. Haltiwanger, Zachary Kroff, and Keith Savage. 2024. Tracking Firm Use of AI in Real Time: A Snapshot from the Business Trends and Outlook Survey.

Ulrich Boser, Matthew M. Chingos, and Chelsea Straus. 2015. The Hidden Value of Curriculum Reform. Technical report, Center for American Progress, Washington, D.C.

Megan Brenan. 2021. K-12 Parents Remain Largely Satisfied With Child's Education. Section: Education.

Robert L. Brennan. 2001a. Advanced Topics in Univariate Generalizability Theory. In Robert L. Brennan, editor, *Generalizability Theory*, Statistics for Social Sciences and Public Policy, pages 141–177. Springer, New York, NY.

Robert L. Brennan. 2001b. *Generalizability Theory*. Springer, New York, NY.

Robert L. Brennan. 2001c. Multifacet Universes of Generalization and D Study Designs. In Robert L. Brennan, editor, *Generalizability Theory*, Statistics for Social Sciences and Public Policy, pages 95–139. Springer, New York, NY.

Robert L Brennan. 2003. Coefficients and Indices in Generalizability Theory. Technical Report 1, Center for Advanced Studies in Measurement and Assessment.

Robert L. Brennan. 2013. *Generalizability Theory*. Springer Science & Business Media. Google-Books-ID: nbHbBwAAQBAJ.

John Seely Brown and Kurt VanLehn. 1980. Repair Theory: A Generative Theory of Bugs in Procedural Skills. *Cognitive Science*, 4(4):379–426.

Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M. Douglas, Jim A. C. Everett, Gerd Gigerenzer, Christine Greenhow, Daniel A. Hashimoto, Julianne Holt-Lunstad, Jolanda Jetten, Simon Johnson, Chiara Longoni, Pete Lunn, and 12 others. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *arXiv preprint*. ArXiv:2401.05377.

Jodi M Casabianca. 2025. Psychometrics is all you need.

Jodi M. Casabianca, Brian W. Junker, and Richard J. Patz. 2016. Hierarchical Rater Models. In *Handbook of Item Response Theory*. Chapman and Hall/CRC. Num Pages: 18.

Charalambos Y. Charalambous and Seán Delaney. 2019. 13 Mathematics Teaching Practices and Practice-Based Pedagogies. Brill. Section: International Handbook of Mathematics Teacher Education: Volume 1.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024. Are More LLM Calls All You Need? Towards Scaling Laws of Compound Inference Systems. *arXiv preprint*. ArXiv:2403.02419 [cs].

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *arXiv preprint*. ArXiv:1706.03741.

Elizabeth Chu, Andrea Clay, and Grace McCarty. 2021. Fundamental 4: Pandemic Learning Reveals the Value of High-Quality Instructional Materials to Educator-Family-Student Partnerships. Technical report, Center for Public Research and Leadership, New York, NY.

Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. 2024. Building Machines that Learn and Think with People. *arXiv preprint*. ArXiv:2408.03943 [cs].

Le Cong, David Smerkous, Xiaotong Wang, Di Yin, Zaixi Zhang, Ruofan Jin, Yinkai Wang, Michal Gerasimiuk, Ravi K. Dinesh, Alex Smerkous, Lihan Shi, Joy Zheng, Ian Lam, Xuekun Wu, Shilong Liu, Peishan Li, Yi Zhu, Ning Zhao, Meenal Parakh, and 14 others. 2025. LabOS: The AI-XR Co-Scientist That Sees and Works With Humans. ISSN: 2692-8205 Pages: 2025.10.16.679418 Section: New Results.

Roderic N. Crooks. 2024. *Access Is Capture: How Edtech Reproduces Racial Inequality*. Univ of California Press. Google-Books-ID: q1ANEQAAQBAJ.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, and 21 others. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv preprint*. ArXiv:2011.03395 [cs, stat].

Linda Darling-Hammond. 2014. What Can PISA Tell Us about U.S. Education Policy? *New England Journal of Public Policy*, 26(1).

Linda Darling-Hammond, Lisa Flook, Channa Cook-Harvey, Brigid Barron, and David Osher. 2020. Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2):97–140. _eprint: https://doi.org/10.1080/10888691.2018.1537791.

Dorottya Demszky and Heather Hill. 2022. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. *arXiv preprint*. ArXiv:2211.11772 [cs].

Dorottya Demszky and Heather Hill. 2023. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Paul Denny, Sumit Gulwani, Neil T. Heffernan, Tanja Käser, Steven Moore, Anna N. Rafferty, and Adish Singla. 2024. Generative AI for Education (GAIED): Advances, Opportunities, and Challenges. *arXiv preprint*. ArXiv:2402.01580.

EdReports. 2021. Review Tools.

EdReports. 2023. State of the Instructional Materials Market: Use of Aligned Materials in 2022. Technical report, EdReports.org.

T. S. (Thomas Stearns) Eliot, 1888-1965. 1934. *The rock, a pageant play written for performance at Sadler's Wells theatre 28 May - 9 June 1934*. London, Faber, England, United Kingdom.

Virginia Eubanks. 2019. *Automating inequality: how high-tech tools profile, police, and punish the poor*, first picador edition edition. Picador St. Martin's Press, New York.

E. C. Fieller, H. O. Hartley, and E. S. Pearson. 1957. TESTS FOR RANK CORRELATION COEFFICIENTS. I. *Biometrika*, 44(3-4):470–481.

Barbara Foorman, Nicholas Beyler, Kelley Borradaile, Michael Coyne, Carolyn A. Denton, Joseph Dimino, Joshua Furgeson, Lynda Hayes, Juliette Henke, Laura Justice, Betsy Keating, Warnick Lewis, Samina Sattar, Andrei Streke, Richard Wagner, and Sarah Wissel. 2016. Foundational Skills to Support Reading for Understanding in Kindergarten through 3rd Grade. Educator's Practice Guide. NCEE 2016-4008. Technical report, What Works Clearinghouse. ERIC Number: ED566956.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *arXiv preprint*. ArXiv:2202.06935 [cs].

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. Towards an AI co-scientist. *arXiv preprint*. ArXiv:2502.18864 [cs].

Richard Greiner. 1909. Ueber das Fehlersystem der Kollektiv-maßlehre. In *Zeitschrift für Mathematik und Physik*, volume 57, pages 121–158, 225–260, 337–373. B. G. Teubner, Leipzig.

Jason Grissom, Susanna Loeb, and Benjamin Master. 2013. Effective Instructional Time Use for School Leaders: Longitudinal Evidence from Observations of Principals. *Educational Researcher*, 42(8)(42(8)):433.

Zaretta Hammond. 2015. *Culturally responsive teaching and the brain: promoting authentic engagement and rigor among culturally and linguistically diverse students*. Corwin, a SAGE company, Thousand Oaks, California. OCLC: ocn889185083.

John D. Hansen and Justin Reich. 2015. Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, 350(6265):1245–1248.

Michael Hardy. 2024. "All that Glitters": Approaches to Evaluations with Unreliable Model and Human Annotations. *arXiv preprint*. ArXiv:2411.15634 [cs].

Michael Hardy. 2025a. Measuring Teaching with LLMs. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 367–384, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Michael Hardy. 2025b. "All that Glitters": Techniques for Evaluations with Unreliable Model and Human Annotations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2250–2278, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael Hardy. 2026. Autoscoring Anticlimax: A Meta-analytic Understanding of AI's Short-answer Shortcomings and Wording Weaknesses. *arXiv preprint*. ArXiv:2603.04820 [cs].

Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. 2024. Fantastic Copyrighted Beasts and How (Not) to Generate Them. *arXiv preprint*. ArXiv:2406.14526.

Carolyn J. Heinrich, Jennifer Darling-Aduana, Annalee Good, and Huiping (Emily) Cheng. 2019. A Look Inside Online Educational Settings in High School: Promise and Pitfalls for Improving Educational Opportunities and Outcomes. *American Educational Research Journal*, 56(6):2147–2188.

Heather C. Hill, Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, 26(4):430–511.

Heather C. Hill, Charalambos Y. Charalambous, David Blazar, Daniel McGinn, Matthew A. Kraft, Mary Beisiegel, Andrea Humez, Erica Litke, and Kathleen Lynch. 2012a. Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. *Educational Assessment*, 17(2-3):88–106. _eprint: https://doi.org/10.1080/10627197.2012.715019.

Heather C. Hill, Charalambos Y. Charalambous, and Matthew A. Kraft. 2012b. When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, 41(2):56–64.

Andrew D. Ho and Thomas J. Kane. 2013. The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540957.

Wayne Holmes. 2022. *Artificial Intelligence and Education: A Critical View Through the Lens of Human Rights, Democracy and the Rule of Law*, 1st ed edition. Council of Europe, Namur.

Juliana Menasce Horowitz. 2022. Parents Differ Sharply by Party Over What Their K-12 Children Should Learn in School.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. *arXiv preprint*. ArXiv:2309.16349.

Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations. LingBuzz Published In:.

Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, and Xiuruo Zhang. 2025. Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions.

Anders Humlum and Emilie Vestergaard. 2024. The Adoption of Chatgpt.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. *arXiv preprint*. ArXiv:2205.05256 [cs].

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, and 51 others. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. *arXiv preprint*. ArXiv:2509.04664 [cs].

Thomas Kane, Heather Hill, and Douglas Staiger. 2015. National Center for Teacher Effectiveness Main Study: Version 4.

Thomas Kane and Douglas Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Technical Report w14607, National Bureau of Economic Research, Cambridge, MA.

Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540959.

Thomas J. Kane, Antoniya M. Owens, William H. Marinell, Daniel R. C. Thal, and Douglas O. Staiger. 2016. Teaching Higher: Educators' Perspectives on Common Core Implementation. Technical report.

Thomas J. Kane and Douglas O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. Technical report, Bill & Melinda Gates Foundation. Publication Title: Bill & Melinda Gates Foundation ERIC Number: ED540960.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Julia H. Kaufman, V. Darleen Opfer, Michelle Bongard, and Joseph D. Pane. 2018. Changes in What Teachers Know and Do in the Common Core Era: American Teacher Panel Findings from 2015 to 2017. Technical report, RAND Corporation.

Camilla Kempe, Anna-Lena Eriksson-Gustavsson, and Stefan Samuelsson. 2011. Are There any Matthew Effects in Literacy and Cognitive Development? *Scandinavian Journal of Educational Research*, 55(2):181–196. _eprint: https://doi.org/10.1080/00313831.2011.554699.

Maurice George Kendall. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93.

Maurice George Kendall. 1945. THE TREATMENT OF TIES IN RANKING PROBLEMS. *Biometrika*, 33(3):239–251.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated Errors in Large Language Models. *arXiv preprint*. ArXiv:2506.07962 [cs] version: 1.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. *arXiv preprint*. ArXiv:2405.00623.

René F. Kizilcec. 2024. To Advance AI Use in Education, Focus on Understanding Educators. *International Journal of Artificial Intelligence in Education*, 34(1):12–19.

Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160:108352.

Ben Kornell, Alex Sarlin, Sarah Morin, and Laurence Holt. 2024. The Edtech Insiders Generative AI Map.

Eliza Kosoy, Soojin Jeong, Anoop Sinha, Alison Gopnik, and Tanya Kraljic. 2024. Children's Mental Models of Generative Visual and Text Based AI Models.

Matthew A. Kraft. 2020. Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4):241–253.

Holly Kurtz, Sterling Lloyd, Alex Harwin, Victor Chen, and Yukiko Furuya. 2020. Early Reading Instruction. Technical report, Editorial Projects in Education, Bethesda, MD.

Team LearnLM, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Mike Schaekermann, and 27 others. 2024. LearnLM: Improving Gemini for Learning. *arXiv preprint*. ArXiv:2412.16429 [cs].

Daniel Lee, Matthew Arnold, Amit Srivastava, Katrina Plastow, Peter Strelan, Florian Ploeckl, Dimitra Lekkas, and Edward Palmer. 2024. The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, 6:100221.

Maxime Lelièvre, Amy Waldock, Meng Liu, Natalia Valdés Aspillaga, Alasdair Mackintosh, María José Ogando Portela, Jared Lee, Paul Atherton, Robin A. A. Ince, and Oliver G. B. Garrod. 2025. Benchmarking the Pedagogical Knowledge of Large Language Models. *arXiv preprint*. ArXiv:2506.18710 [cs].

Doug Lemov. 2021. *Teach like a champion 3.0: 63 techniques that put students on the path to college*, third edition edition. Jossey-Bass, a Wiley imprint, Hoboken, NJ.

Doug Lemov and Norman Atkins. 2015. *Teach like a champion 2.0: 62 techniques that put students on the path to college*, second edition edition. Jossey-Bass, San Francisco, CA.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic Evaluation of Language Models. *arXiv preprint*. ArXiv:2211.09110 [cs].

Peter Liljedahl, Tracy Johnston Zager, and Laura Wheeler. 2021. *Building thinking classrooms in mathematics: 14 teaching practices for enhancing learning: Grades K-12*. Corwin Mathematics. Corwin, Thousand Oaks, California London New Delhi Singapore.

Jing Liu and Julie Cohen. 2021. Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, 43(4):587–614.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Opportunities and Challenges in Neural Dialog Tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Madaio, Su Lin Blodgett, Elijah Mayfield, and Ezekiel Dixon-Román. 2021. Beyond "Fairness:" Structural (In)justice Lenses on AI for Education. *arXiv preprint*. ArXiv:2105.08847 [cs].

Eran Malach. 2024. Auto-Regressive Next-Token Predictors are Universal Learners. *arXiv preprint*. ArXiv:2309.06979 [cs].

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. 2025. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. *arXiv preprint*. ArXiv:2502.08640 [cs].

Daniel F. McCaffrey, Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy*, 4(4):572–606.

Daniel F. McCaffrey, Kun Yuan, Terrance D. Savitsky, J. R. Lockwood, and Maria O. Edelen. 2015. Uncovering Multivariate Structure in Classroom Observations in the Presence of Rater Errors. *Educational Measurement: Issues and Practice*, 34(2):34–46.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve.

Paul E. Meehl. 1990. Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, 1(2):108–141.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749. Place: US.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Mehrdad Asgari, Juliane Eberhardt, Amir Mohammad Elahi, Hani M. Elbeheiry, María Victoria Gil, Christina Glaubitz, Maximilian Greiner, Caroline T. Holick, Tim Hoffmann, and 16 others. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi,

Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2024. OLMoE: Open Mixture-of-Experts Language Models. *arXiv preprint*. ArXiv:2409.02060.

Dianna R. Mullet. 2018. A General Critical Discourse Analysis Framework for Educational Research. *Journal of Advanced Academics*, 29(2):116–142.

Allen Nie, Yash Chandak, Miroslav Suzara, Malika Ali, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2024. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters Exam Performances. *arXiv preprint*. ArXiv:2407.09975 [cs, stat].

Amber M. Northern and Michael J. Petrilli. 2019. Dear teachers, most of the popular lessons you found online aren't worth using. *The Thomas B. Fordham Institute*.

James O'Donnell. 2024. Here's how ed-tech companies are pitching AI to teachers.

V. Darleen Opfer, Julia H. Kaufman, and Lindsey E. Thompson. 2017. Implementation of K–12 State Standards for Mathematics and English Language Arts and Literacy: Findings from the American Teacher Panel. Technical report, RAND Corporation.

Christopher Michael Ormerod and Alexander Kwako. 2024. Automated Text Scoring in the Age of Generative AI for the GPU-poor. *arXiv preprint*. ArXiv:2407.01873 [cs] version: 1.

Zachary A. Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint*. ArXiv:2302.06871 [cs].

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. 2025. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. *arXiv preprint*. ArXiv:2510.04374 [cs].

Thomas M. Philip, Megan Bang, and Kara Jackson. 2018. Articulating the "How," the "For What," the "For Whom," and the "With Whom" in Concert: A Call to Broaden the Benchmarks of our Scholarship. *Cognition and Instruction*, 36(2):83–88. _eprint: https://doi.org/10.1080/07370008.2018.1413530.

Robert C. Pianta, Jay Belsky, Nathan Vandergrift, Renate Houts, and Fred J. Morrison. 2008. Classroom Effects on Children's Achievement Trajectories in Elementary School. *American Educational Research Journal*, 45(2):365–397.

Robert C. Pianta and Bridget K. Hamre. 2009. Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, 38(2):109–119.

Morgan Polikoff. 2019. The Supplemental Curriculum Bazaar: Is What's Online Any Good? Technical report, Washington, D.C.

Morgan Polikoff, Elaine Lin Wang, Shira Korn Haderlein, Julia H. Kaufman, Ashley Woo, Daniel Silver, and V. Darleen Opfer. 2020. Exploring Coherence in English Language Arts Instructional Systems in the Common Core Era. Technical report, RAND Corporation.

Justin Reich. 2020. *Failure to Disrupt: Why Technology Alone Can't Transform Education*. Harvard University Press.

Justin Reich and Mizuko Ito. 2017. From Good Intentions to Real Outcomes: Equity by Design in Learning Technologies. Technical report, Digital Media and Learning Research Hub, Irvine, CA.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint*. ArXiv:2311.12022 [cs].

Cheng Ren, Zachary Pardos, and Zhi Li. 2024. Human-AI Collaboration Increases Skill Tagging Speed but Degrades Accuracy. *arXiv preprint*. ArXiv:2403.02259 [cs].

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. *arXiv preprint*. ArXiv:2411.12990 [cs] version: 1.

Sina Rismanchian, Peter Liu, Gabe A Orona, Duncan Pritchard, and Shayan Doroudi. 2026. Artificial Integrity: Concerning Patterns of AI Usage Among Undergraduate Students.

Kate Rix. 2023. New Report Flunks Teacher Prep Programs on the Science of Reading.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond Flesch-Kincaid: Prompt-based Metrics Improve Difficulty Classification of Educational Texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Jennifer Rowley. 2007. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 33(2):163–180.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Lydia Saad. 2022. Americans' Satisfaction With K-12 Education on Low Side. Section: Education.

Johnny Saldaña. 2016. *The coding manual for qualitative researchers*, 3. edition edition. SAGE, Los Angeles, Calif. London New Delhi Singapore Washington DC.

Jon Saphier, Mary Ann Haley-Speca, and Robert Gower. 2008. *The skillful teacher: building your teaching skills*, 6th ed edition. Research for Better Teaching, Acton, Mass.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? *arXiv preprint*. ArXiv:2304.15004 [cs].

Daniel L. Schwartz, Jessica M. Tsang, and Kristen P. Blair. 2016. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*, first edition edition. Norton books in education. W.W. Norton & Company, New York.

Pranab Kumar Sen. 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324):1379–1389.

Esther Shein. 2024. The Impact of AI on Computer Science Education. *Communications of the ACM*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context Pretraining: Language Modeling Beyond Document Boundaries. *arXiv preprint*. ArXiv:2310.10638.

Jim Short and Stephanie Hirsh. 2020. The Elements: Transforming Teaching through Curriculum-Based Professional Learning | Professional Learning for Educators | Carnegie Corporation of New York. Technical report, Carnegie Corporation of New York, New York, NY.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Andrew F. Siegel. 1982. Robust regression using repeated medians. *Biometrika*, 69(1):242–244.

Emily J. Solari, Nicole Patton Terry, Nadine Gaab, Tiffany P. Hogan, Nancy J. Nelson, Jill M. Pentimonti, Yaacov Petscher, and Sarah Sayko. 2020. Translational Science: A Road Map for the Science of Reading. *Reading Research Quarterly*, 55(S1):S347–S360.

Zhangde Song, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M. Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, Yi Qu, Yinkai Wang, Haorui Wang, Jeff Guo, Jingru Gan, Parshin Shojaee, Di Luo, Andres M. Bran, Gen Li, Qiyuan Zhao, and 37 others. 2025. Evaluating Large Language Models in Scientific Discovery. *arXiv preprint*. ArXiv:2512.15567 [cs].

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. *arXiv preprint*. ArXiv:2402.05070.

Keith E. Stanovich. 1986. Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly*, 21(4):360–407.

David Steiner. 2017. Curriculum Research: What We Know and Where We Need to Go. Technical report, StandardsWork.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2025. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723.

Gabor J. Szekely and Maria L. Rizzo. 2014. Partial Distance Correlation with Methods for Dissimilarities. *arXiv preprint*. ArXiv:1310.2926 [stat].

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. *arXiv preprint*. ArXiv:2306.06941.

TNTP. The Opportunity Myth.

TNTP. 2024. The Opportunity Makers. Technical report, The New Teacher Project, New York, NY.

Jutta Treviranus. 2022. Learning to learn differently. In *The Ethics of Artificial Intelligence in Education*. Routledge. Num Pages: 22.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When Are Combinations of Humans and AI Useful? *arXiv preprint*. ArXiv:2405.06087 [cs].

Luc Van der Gun and Olivia Guest. 2024. Artificial Intelligence: Panacea or Non-Intentional Dehumanisation? *Journal of Human-Technology Relations*, 2.

Kurt VanLehn. 1990. *Mind Bugs : The Origins of Procedural Misconceptions*. Learning, Development, and Conceptual Change. The MIT Press, Cambridge, Mass.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2). ArXiv:1903.08008 [stat].

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv preprint*. ArXiv:2306.07899.

Rose E. Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. *arXiv preprint*. ArXiv:2306.03090 [cs].

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *arXiv preprint*. ArXiv:2406.01574 [cs].

Michael Rhys Morgan Ward, Sara Delamont, and Sara Delamont, editors. 2020. *Handbook of qualitative research in education*, second edition edition. Edward Elgar Publishing, Cheltenham. OCLC: on1197746807.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2024. Language Models Learn to Mislead Humans via RLHF.

Cedric Deslandes Whitney and Justin Norman. 2024. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1733–1744, New York, NY, USA. Association for Computing Machinery.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, 4.

Minghao Wu and Alham Fikri Aji. 2023. Style Over Substance: Evaluation Biases for Large Language Models. *arXiv preprint*. ArXiv:2307.03025.

Hongliang Xin, John R. Kitchin, and Heather J. Kulik. 2025. Towards agentic science for advancing scientific discovery. *Nature Machine Intelligence*, 7(9):1373–1375.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The Promises and Pitfalls of Using Language Models to Measure Instruction Quality in Education. *arXiv preprint*. ArXiv:2404.02444 [cs].

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint*. ArXiv:2504.08066 [cs].

Jianhao Yan, Yun Luo, and Yue Zhang. 2024. RefuteBench: Evaluating Refuting Instruction-Following for Large Language Models.

Chunpeng Zhai, Santoso Wibowo, and Lily D. Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. *arXiv preprint*. ArXiv:2401.06730 [cs].

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023a. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. *arXiv preprint*. ArXiv:2302.13439 [cs].

Xiaofei Zhou, Christopher Kok, Rebecca M. Quintana, Anita Delahay, and Xu Wang. 2023b. How Learning Experience Designers Make Design Decisions: The Role of Data, the Reliance on Subject Matter Expertise, and the Opportunities for Data-Driven Support. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, pages 132–143, New York, NY, USA. Association for Computing Machinery.

Alan Zhu, Parth Asawa, Jared Quincy Davis, Lingjiao Chen, Boris Hanin, Ion Stoica, Joseph E. Gonzalez, and Matei Zaharia. 2025. BARE: Leveraging Base Language Models for Few-Shot Synthetic Data Generation. *arXiv preprint*. ArXiv:2502.01697 [cs] version: 3.

A Ethical considerations

This paper is about measurement: how to detect, quantify, and localize misalignment when LLM scores are used as stand-ins for human judgment and, more importantly, for downstream outcomes.

But our case study is not measurement in the abstract. Classroom evaluation is a high-stakes domain in which errors are not evenly distributed: mismeasurement can change students' learning opportunities, shape teachers' careers, and amplify existing inequities. The purpose of this section is therefore twofold. First, it explains why the empirical patterns we uncover—strong inter-model consensus, weak connection to intended impact, and limited efficacy of prompt/model tweaks—are ethically consequential in K–12 settings. Second, it provides a roadmap for what "responsible next steps" look like when the technical work reveals that validity cannot be assumed.

Our findings shift the ethical question from "Is an LLM accurate enough?" to "Accurate *for what*, and by what evidence?" In education technology, it is tempting to treat alignment with expert rubrics as sufficient validation. Our results warn against that shortcut: systems can look reasonable to adults (and even agree with each other) while remaining misaligned with student learning gains. This mismatch creates a distinctive risk profile: deployment can produce confident, scalable judgments with unclear or negative educational benefit. The ethical challenge is thus not merely model bias in the usual demographic sense, but *validity risk*: using an apparently coherent evaluator whose shared inductive biases reward proxies for quality rather than the outcomes schools are responsible for delivering.

For readers asking what to do with these education-specific implications, the remainder of this section is organized around "next-step" operational problems and workable frameworks. We first discuss how to understand educational needs and the limits of what transcript-based scoring can justify, especially for children. We then address equity and accountability: why equal access to an AI tool is not evidence of equal benefit, and what monitoring is required before scaling. Finally, we outline practical safeguards for researchers and developers—including intended-impact evaluation, pre-deployment decision studies, and harm-aware piloting—that translate the measurement methods in this paper into responsible practice in public education contexts.

A.1 Understanding educational needs of children

Foundation models' generality and adaptability as next-token predictors (Malach, 2024) have energized the education technology (edtech) sector. We, like many developers, are eager for a world where all children get access to high quality education. With increased hype, why is there yet very little evidence of these models meaningfully improving student learning in K12 contexts? Recent work has even shown that generative AI deployed as conversational agents can harm student learning (Bastani et al., 2024; Nie et al., 2024; Shein, 2024). However, developers of education technology (edtech) have not been deterred (Kornell et al., 2024; O'Donnell, 2024), further increasing the high adoption rates of generative AI in educators (Bonney et al., 2024; Humlum and Vestergaard, 2024; Bick et al., 2024). So we feel the need to elaborate on the pertinent ethical considerations for edtech in public education spaces.

Answering whether a LLM is good enough or better than the alternative is difficult (D'Amour et al., 2020; Hutchinson et al., 2022). Evaluating model pedagogical outputs in K12 education is even harder than for most other disciplines (Denny et al., 2024; Macina et al., 2023; Wollny et al., 2021), where understanding the impact is paramount. One challenge in evaluating for K12 impact is the paucity of datasets that allow for quantitative evaluation of model performance with LLMs underperforming (Jurenka et al., 2024; Tack et al., 2023; Ormerod and Kwako, 2024). Unfortunately, poor model performance on existing evaluable K12 datasets is often ignored and can be hidden by changing the nature of the evaluation (Röttger et al., 2024; Schaeffer et al., 2023). For example, instead of asking the model to accurately autoscore student work—a critical task for K12 impact—one could ask it to write plausible feedback which might then be evaluated indirectly using surveys of relative preference by a few raters.

Failure of any educational resource is generally not a binary characteristic, but rather measures of extent along *continua of quality*. For a given group of children, an excellent human tutor would facilitate each child learning at their fullest capacity, however defined, whereas a weaker tutor may only be able to help some students learn some of the content some of the time, much like some non-humans (Collins et al., 2024; Pardos and Bhandari, 2023; Vaccaro et al., 2024). If the human tutor were replaced by a textbook, there would still a very small subset of children who could fully benefit from that intervention, but most students would not be as well served. These continua of quality exist for all educational resources, e.g., lesson plans, reme-

diation activities, curricula, formative assessments, systems for classroom climate, IEPs, feedback to students, etc.

The Paradox of Free Advice Similarly, all AI-generated content will be imperfect, and the degree of quality available can be difficult to recognize. Freely available generative AI introduces a challenge that we will call the **Paradox of Free Advice**:

Those needing more guidance are also those that are less discerning of the quality of guidance or support offered.

Such individuals may turn to generative AI tools, which can be speciously compelling and confident, even when such models are not deserving of our trust (Kim et al., 2024; Wu and Aji, 2023; Yan et al., 2024; Zhou et al., 2024). The implications are that, for example, while an automated service may save a teacher time, it may result in a child losing learning by spending time or resources in less efficacious interventions (Acemoglu and Restrepo, 2022; Holmes, 2022). In fact, some practices in Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2023) exacerbate this even further: RLHF-based model tuning simultaneously leads to less accurate outputs and yet increased human confidence that the less accurate output is trustworthy (Wen et al., 2024). If even the most expert humans have shown worryingly bad tendencies in collaborative decision making with AI (Agarwal et al., 2023), how will children and public school teachers fare? The *Paradox of Free Advice* can affect researchers and scientists as well. Failing to have access to or invest in high quality evaluations utilizing true expertise (Hosking et al., 2024) for the expected human judgments can result in evaluation and reporting of research findings that deepen gaps in quality when convenience samples of human experts rely on these same systems to respond to researcher requests (Veselovsky et al., 2023).

Biased performance is even harder for individual users to see on single use cases, since GPT models can often appear reasonable and trustworthy even when they are less accurate (Klingbeil et al., 2024; Wen et al., 2024; Zhou et al., 2023a). Children, like adults, become more trusting of these tools with exposure (Kosoy et al., 2024), an observation that demands deeper exploration into biases, as the models have already been associated with behavioral changes in postsecondary learning contexts (Abbas et al., 2024; Nie et al., 2024; Zhai et al., 2024).

Inequity along Continua of Quality Educational tools targeted for use in compulsory public schooling contexts have the intrinsic responsibility to provide equitable learning for all children (and to clearly communicate otherwise if that cannot be demonstrated). For this paper, **equality** is defined as equal access to resources, opportunities, tools, or systems, and, in contrast, **equity** is defined as equal access to the *benefits* of those resources, opportunities, tools, and systems. In other words, merely offering all students access to a particular educational resource is not sufficient evidence of equitable outcomes for students, and, in many cases, may exacerbate existing gaps (Benjamin, 2019; Crooks, 2024; Eubanks, 2019; Kasneci et al., 2023; Madaio et al., 2021). In general, over the last couple decades, edtech has not provided evidence that it can equitably improve student outcomes in K12 public education and, in many cases, has even produced results which have widened preexisting inequities in learning (Hansen and Reich, 2015; Reich, 2020; Reich and Ito, 2017). This increased inequity often arises as an example the "Matthew Effect," a term for the disparity-widening effects of accumulated advantage that make it easier to benefit further from new advantage. This phenomenon appears in many education contexts (Bahr, 2007; Kempe et al., 2011; Reich, 2020; Stanovich, 1986). As a pre-LLM edtech example, asynchronous educational content learning experiences, such as those found in massive open online courses (MOOCs), used to support high schoolers recovering credits towards graduation have widened gaps for those most in need of support (Heinrich et al., 2019): those who needed more help who, in fact, got less help from the intervention. As with the example of tutors of varying quality, poorly executed K12 edtech solutions can exacerbate such issues, and thus an even greater need to be able to identify tasks where evaluating output quality is more challenging or uncertain.

Principle of Precision in education Equitable and individualized K12 learning requires precision. Recognizing that most uses of generated K12 content, often branded as saving teachers time, are, in fact, *making instructional decisions*—choices about what and how to teach—not just generating ideas for top performers. Imprecision in instructional decisions degrades its quality, resulting in inefficient teaching and unforced losses to a child's

learning time. A critical component for helping learners "catch up" is to make instructional decisions that maximize time spent on things students need with precision, rather than on content that they already know. AI use can help with writing when precision is not critical, but has led to worsened decision-making outcomes (Vaccaro et al., 2024) and can further marginalize learners whose needs are not well represented by the mean of the distribution (Treviranus, 2022).

Precise instructional decisions are needed for efficiently and effectively resolving student misconceptions, as a practical example. Using an imperfect but illustrative analogy from computer programming, we could consider confusion and misconceptions as "bugs" in a student's reasoning (Brown and VanLehn, 1980; VanLehn, 1990). Both fixing and not creating new bugs require expert precision, which may not be a strength of the current GPT models when supporting struggling coders.

A.2 Demands placed on educators

Public K12 educators have deep insight into their specific students and their idiosyncratic teaching tendencies. But critically, the average K12 teacher is not an education expert at the level needed and assumed by researchers and developers for many edtech questions, resources, and products. This is in no way saying that researchers and developers should not consult educators. On the contrary, K12 educators know things about the day-to-day dynamics within their classroom that researchers, developers, and designers can learn from (Kizilcec, 2024). This hyperlocalized expertise about the children in their custody may explain why most parents feel like their schools are headed in the right direction while most think that overall schooling is not headed in the right direction (Brenan, 2021; Horowitz, 2022; Saad, 2022). However, hyperlocalized educator expertise does not mean that:

1. the educator is effective and that their teaching practices result in significant learning gains for students;
2. the educator's best practices and insights generalize to other (a) students, (b) classrooms, (c) teachers, (d) content areas, (e) grades/ages, (f) schools, or (g) contexts;
3. the task or idea the educator is being asked to do, evaluate, or provide feedback on is a task about which they are expert and about which they can accurately identify the conditions needed for its generalization; nor
4. the educator is aware of the extent to which they have the expertise described here.

These criteria of expected expertise can easily be extended by replacing "educator" in the above with "researcher" or "technologist". Finding qualified subject matter experts is critical to high-quality solutions (Zhou et al., 2023b). As important as working with and understanding the needs of K12 educators is for finding solutions, edtech research and products are typically not built for a handful of specific teachers with hyperlocalized expertise, but for education contexts generally. Ho and Kane found that school administrators—with more general expertise—were more discerning of differences in teaching quality and produced ratings that were 12%–25% more reliable compared to other teachers with similar/different teaching certifications, respectively (Ho and Kane, 2013). In education literature this is at least in part because school administrators have more generalizing power from regular exposure to different classrooms. However, even school administrators benefit from localized expertise: Ho and Kane found that administrators from the same school as the instruction being evaluated were 18% more reliable than non-local administrators.

Expanding the sample of or crowdsourcing across educators may not lead to quality annotations, effective solutions, or generalization across contexts (Macina et al., 2023). For some K12 materials and practices, a majority of teachers (and even professors of education) may not be implementing practices that are known in research to be more effective, such as the disparity between known effective practices for early literacy instruction and common practices in schools and teacher prep programs (Foorman et al., 2016; Kurtz et al., 2020; Rix, 2023; Solari et al., 2020).

The Principle of Precision applies across continua of quality in education content creation, which represents most current applications of GPT models as educator assistants. While there is clear evidence about the importance of using high quality instructional materials (Boser et al., 2015; EdReports, 2023; Kaufman et al., 2018; Opfer et al., 2017; Steiner, 2017; TNTP), most of these materials are either not easily accessible or have a large enough presence online for the purposes of model training. Unfortunately, most of the K12 instructional materials easily accessible online by and popular among educators are not high quality (Northern and Petrilli, 2019; Polikoff, 2019). Although

automatically generating materials may appear to save teachers time, it may cost students learning time if the ignores the important attributes of instructional materials such as curricular coherence, meeting criteria for "high quality", and teachers' abilities to both recognize those criteria and use curricula effectively (Chu et al., 2021; EdReports, 2021; Kane et al., 2016; Polikoff et al., 2020; Short and Hirsh, 2020; TNTP, 2024). Poorer quality materials and the *Paradox of Free Advice* would logically lead to the exacerbating Matthew Effect (Acemoglu and Restrepo, 2022; Capraro et al., 2024). While techniques for content generation are improving (Balepur et al., 2023; Rooein et al., 2024), the criteria for high quality instructional materials, the coherence in the learning, and the capacity of the teacher to recognize and deliver such content are challenges not yet addressed.

A.3 Dearth of Data

Studies about the quality of teaching and learning are expensive (Grissom et al., 2013; Liu and Cohen, 2021; Jurenka et al., 2024), with only two major studies having measures of both teaching and learning: the MET study (Kane et al., 2013; Kane and Staiger, 2012) and the NCTE Main Study (Kane et al., 2015), the latter of which is the source of data for this study. Even then, no transcripts or artifacts of classroom discourse are meaningfully annotated with respect to reliable measure of a) the ***quality of teaching***, using authentic instruments designed for humans or b) the ***quality of learning***. In fact, the majority of instructional materials that would be available on the Internet for pre-training are of low quality (Polikoff, 2019; Northern and Petrilli, 2019; EdReports, 2023; TNTP, 2024).

Many critical education tasks are poorly represented in training data for GPT models. Authentic natural language found in K12 public education contexts is rare on the internet, especially data with meaningful labels or interpretations, because these data involve children and typically have more stringent protections (e.g. FERPA and COPPA in the United States). Thus, a limitation noted by this paper is the need to create more datasets, a limitation shared with many fields, including LLM-as-scientist studies (Song et al., 2025; Alampara et al., 2025; Mirza et al., 2025).

A.4 Potential solutions to challenges

Maximizing Education Expertise For improving expertise in the field of edtech products, we recommend that researchers and developers acquire the needed expertise with much more intentionality by selecting few experts who, when combined, jointly maximize expertise relevant to the target audiences and contexts of generalization. If improvements to student outcomes are desired, K12 experts should have a demonstrated track record of positively impacting student outcomes, and the generalizable insights they make should correspond with the contexts of their track record. If an expert does not allow for generalization into some content or context of interest, experts should be added. Non-K12 subject matter experts in academia or industry should likewise represent the breadth of the content and should include expertise in the assessment of intended student outcomes. We make the claim that the skills and capacity needed to lead initiatives that are high-quality, impactful to student outcomes, equitable, and scalable are exceedingly rare. This skill set is not found in an average K12 educator, so greater effort must be made by all.

School leaders working with teachers to improve the quality of instruction typically evaluate the teacher's proficiency in a range of competencies (typically measured during in-class observation and evaluation on a teaching rubric; see Aguilar 2013; Bambrick-Santoyo 2016, 2018), then determine which competencies are most important to improve first (i.e., which change will have the biggest impact on student learning), and then provide supportive feedback and coaching. Without strong expertise and measures of student learning, it is challenging for practitioners to prioritize instructional needs and aligned practices from among the many elements of good teaching (Saphier et al., 2008; Darling-Hammond, 2014; Hammond, 2015; Lemov and Atkins, 2015; Lemov, 2021; Liljedahl et al., 2021; Darling-Hammond et al., 2020; Schwartz et al., 2016) and for researchers to empirically quantify the impact of good teaching practices (Pianta and Hamre, 2009; Charalambous and Delaney, 2019; Blazar and Pollard, 2022; Jurenka et al., 2024).

Measuring what matters Additionally, we would like to add that perhaps the most important direction for evaluating future edtech work is improving the ability of models or systems to accurately recognize the state of student learning. The field cannot correct what it cannot detect. And, at present, the Paradox of Free Advice applies to developers and researchers in the presence of GPT

model speciousness and current practices for K12 expertise. Meaningful, rapid iteration on edtech tools and products cannot be without reliable and valid measures of student learning.

Dealing with data deficiency Distributional pluralistic alignment of models in new tasks is a difficult due to limited knowledge of how to calibrate models to be more representative (Sorensen et al., 2024). Research in pretraining data provenance is a promising area of research for addressing biases. For example, studies have shown that performant models can be constructed by curating smaller, more manageable datasets (Shi et al., 2024) and then training individual "expert" models from appropriate data which can be combined into a final model (Muennighoff et al., 2024). Such techniques have been used to address copyright violations (He et al., 2024).

However, training on authentic education data, such as transcripts from learning, may not be free of bias or may not support improved model performance toward student learning outcomes without some meaningful interpretation. Being able to identify whether a particular artifact is high-quality is key to improving performance. One possible approach to explore is to collaborate with qualitative researchers who have spent countless hours coding studies (Ward et al., 2020; Saldaña, 2016; Mullet, 2018; Philip et al., 2018) to combine smaller dataset studies. The meaningful curation and combination of such data could serve as a starting point for use of authentic K12 data during pretraining.

Synthetic substitutions Even when faced with the challenge of a lack of quality data, we recommend minimizing the use of synthetic K12 educational content. Use of such content would risk perpetuating gaps in equitable learning as models trained thusly increasingly and irreversibly forget the original distributional tails (Shumailov et al., 2024; Whitney and Norman, 2024; Van der Gun and Guest, 2024). A more expertly built smaller model, even if limited in scope (see for example Hardy), which is more robustly evaluated, could support the improved annotation of existing K12 artifacts or act as a decision-making member multi-agent ensemble.

B Datasets for Supporting Classrooms

Meaningful representation of student classroom learning is absent on the internet largely for laws protecting children's privacy. The text for the data was only made anonymized and public in 2023 (Demszky and Hill, 2023) and is not "crawlable" nor is it easy to link these data with the (also not crawlable) annotations and VAMs (Kane et al., 2015). It is the linking of these two sources that makes it 1 of 2 extant datasets having classroom interactions with both expert ratings on real teaching instruments and VAMs. The second dataset (the "MET Project") is not publicly available. Part of our interest in doing this research is because conducting these studies is extremely expensive. (At least 2 organizations are trying to collect more data that meet these criteria.) Regardless, these datasets are notoriously challenging; Xu et al. use ½ page to articulate challenges with this data. Ho and Kane give a measure of success for humans at reliability of 0.65–our initial target for using LLMs. LLMs initially demonstrated high consistency (increasing reliability), but then we discovered the misalignment. Ideally success would lead to ratings that improve human reliability (Hardy, 2025b) and alignment with robust checks to avoid unhelpful LLM contributions.

B.1 Observation Instruments

For each of the observation instruments, the abbreviation codes used in this study are listed with the expanded names in Table 2. The distributions of scores across all items for all rater families are in Figure 5. The CLASS rubric has 12 items on a scale from 1 to 7, rated at 15 minute intervals. The MQI rubric has 13 items on a scale from 1 to 3, rated at 7.5 minute intervals.

The mathematics-specific MQI and general CLASS frameworks are explicitly and implicitly capturing different dimensions of classroom instruction (Blazar et al., 2017) for human raters.

The MQI Framework (13 items) The MQI framework is intended to detect specific classroom practices, assuming that much of the classroom discourse will not be relevant for each item and are one-inflated. The 13 MQI items[5] within the dataset have at least two human raters per classroom observation. The MQI instrument has four instructional domains that capture information on the quality of teaching and learning: Richness of Mathematics, Working with Students, Student Participation,

[5]Instruments for classroom instruction are composed of multiple items that represent distinct instructional dimensions to be evaluated

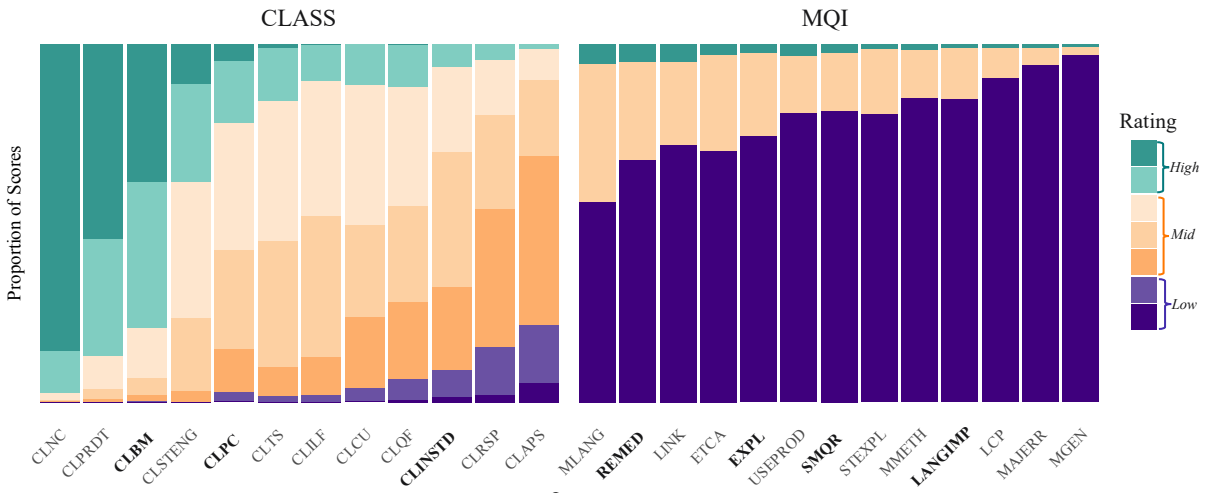


Figure 5: Proportions of human rater scores by item.

and Errors and Imprecisions. This paper will focus on one item from each of the four MQI domains: teacher explanations (EXPL), remediation of student errors (REMED), student questioning and reasoning (SMQR), and imprecision in mathematical language (LANGIMP), respectively. For ease of interpretation in this study, LANGIMP is reverse-coded, so higher scores are better. (For additional information about the MQI instrument, see (Hill et al., 2008; Hardy, 2024; Hill et al., 2012b; Kane and Staiger, 2012)).

The CLASS Framework (12 items) The CLASS items capture broader constructs and assume that most of the classroom discourse is relevant. These differences can be seen visually in Figure 5. The 12 CLASS items have one raters per classroom observation. Prior work has shown that the CLASS items generally have higher human agreement than MQI items (Kane and Staiger, 2012; Kane et al., 2015). Three items from the CLASS instrument will be analyzed, following prior work (Wang and Demszky, 2023): behavior management (CLBM), instructional dialogue (CLINSTD), and positive classroom climate (CLPC).

B.2 Replication Study Details

The present study contains two replication studies; we used the same settings, sampling, and data as specified in each original (Wang and Demszky, 2023; Hardy, 2025b). We did not alter the prompts, samples, nor settings when replicating the original studies. Footnote 3 has a link to the full replication test set and access to the full prompts used. Wang & Demsky (2023) also provide a link on the first page of their study (as well as many appendix pages of prompt templates). We used the same hyperparameter settings (temperature 0, completion) as found in the original code.

Additionally, for verifiability and replication, Footnote 3 also has the outputs of each prompt in the present study (where failure modes can be identified and claims verified), with lookup columns that match the data from Wang & Demszky.

B.2.1 Test Set and Tasks

Building on prior work, we adopt the zero-shot test set and prompts established by (Wang and Demszky, 2023). This test set consists of a stratified sample of lessons from the NCTE data. Our evaluation covers seven distinct rating tasks (i.e., instrument items): four from the MQI framework (teacher explanations, remediation of student errors, student questioning, and precision of language) and three from the CLASS framework (behavior management, instructional dialogue, and positive climate). For the measures of alignment, we maximize content differences by selecting one item from each of the four empirical factors identified in (Blazar et al., 2017; Kane et al., 2015). We select the items with the highest interrater reliability (Cohen's κ in Appendix of (Kane et al., 2015)) where available, otherwise, test set item with the largest factor loading.

B.2.2 Models and Prompts

We selected 4 of the LLMs that performed the best at the time of the experiment, chosen from the HELM leaderboard (Liang et al., 2023) and the

"Pedagogy Benchmark" (Lelièvre et al., 2025), and then prioritized diversity of models. We also included Google's *LearnLM* due to its stated focus on education (LearnLM et al., 2024). Following prior work (Liang et al., 2023; Wang and Demszky, 2023), we query each model using three distinct prompting strategies for each task: (1) a **base prompt** with the core instructions, (2) a **chain-of-thought** prompt encouraging step-by-step reasoning (Wei et al., 2023), and (3) a prompt that acts like a **retrieval-augmented generation (RAG)** by including additional, relevant task-specific rubric information. All for LLM calls are available online. [6] All model predictions are publicly available to support future research for reproducibility (see Footnote 3).

Below are three examples for three different tasks for each of the three prompt template families from Wang and Demszky (where all examples of other prompt templates can be found): Base prompt for the Remediation of Student Errors (REMED) task, a RAG-like, additional-details prompt template for the Language Imprecision (LANGIMP) task, and the reasoning/chain-of-thought template for the Classroom instructional dialogue (CLINSTD) task.

REMED (Base):

```
Consider the following classroom transcript.

Transcript: {transcript}

Based on the classroom transcript, rate the teacher's degree of remediation of student errors and difficulties on a scale of 1-3 (low-high). This means that the teacher gets at the root of student misunderstanding, rather than repairing just the procedure or fact. This is more than a simple correction of a student mistake.

Rating (only specify a number between 1-3):
```

LANGIMP (simple RAG-like/extra information)

```
Consider the following classroom transcript.

Transcript: {transcript}

Based on the classroom transcript, rate the teacher's imprecision in language or notation on a scale of 1-3 (low-high). The teacher's imprecision in language or notation refers to problematic uses of mathematical language or notation. For example, errors in notation (eg. mathematical symbols), in mathematical language (eg. technical mathematical terms like "equation") or general language (eg. explaining mathematical ideas or procedures in non-technical terms). Do not count errors that are noticed and corrected within the segment.

Explanation of ratings:
1: Brief instance of imprecision. Does not obscure the mathematics of the segment.
2: Imprecision occurs in part(s) of the segment or imprecision obscures the mathematics but for only part of the segment.
3: Imprecision occurs in most or all of the segment or imprecision obscures the mathematics of the segment.

Rating (only specify a number between 1-3):
```

CLINSTD (Reasoning)

```
Consider the following classroom transcript.

Transcript: {transcript}

Please do the following.
1. Think step-by-step how you would rate the instructional dialogue of the teacher on a scale of 1-7 (low-high). Instructional dialogue captures the purposeful use of content-focused discussion among teachers and students that is cumulative, with the teacher supporting students to chain ideas together in ways that lead to deeper understanding of content. Students take an active role in these dialogues and both the teacher and students use strategies that facilitate extended
```

[6] https://github.com

```
dialogue.
2. Provide your rating as a number between 1 and 7. Format your answer as: Reasoning: Rating (only specify a number between 1-7):

Reasoning:
```

Additionally, while not the focus of this study, we replicated the SOTA models of (Hardy, 2025b) to have confidence that the misalignment we observed were not the result of an impossible task using only transcripts. These encoders and those from (Hardy, 2025b) are shown as baselines in Fig. 4. This results in 103,148 total observations across models, tasks and prompts. Additionally, while not the focus of this study, we replicated the SOTA models of (Hardy, 2025b) to have confidence that the misalignment we observed were not the result of an impossible task using only transcripts. These encoders and those from (Hardy, 2025b) are shown as baselines in Fig. 4.

C Challenges in VAM signal

Human-VAM alignment for a given 15 minute rating on one dimension of teaching will necessarily be small. But across thousands of observations, we would expect better teaching to be aligned with better outcomes. To test this before the main study, we performed 6 robustness checks, excluded only for brevity, including highly robust estimators between expert ratings and VAM. Summary below:

C.1 Thiel-Sen Estimator

Thiel-Sen (TS) estimator (Sen, 1968) is a non-parametric method very closely related to Kendall's τ. It computes the median of all slopes between point pairs: $\hat{\beta} = \text{median}\left(\frac{y_j - y_i}{x_j - x_i}\right)$. We would expect a positive median slope, $\hat{\beta} < 0$ as a direct analogue of the expected relationship of Kendall's τ (Kendall, 1938).

C.2 Repeated Median Estimator

Siegel's Repeated Median Estimator (RME) is a second highly robust estimator of linear relationships with the highest breakdown point of 50% (Siegel, 1982). It estimates the slope of the regression line $y = A + Bx$ for a set of points. For our purposes, if humans raters fail to have a positive RME slope with respect to VAM on an item, the item either has too high a proportion of bad data or no meaningful relationship. $\hat{B} = \underset{i}{\text{median}}\ \underset{j \neq i}{\text{median}}\ (Y_j - Y_i)/(X_j - X_i)$ where the inner median is the median of the slopes connected to this observation, similar to the Thiel-Sen Estimator, and the outer median is the median slope across all items for the inner slopes. For signal to be robustly identified, we expert humans ratings should have a positive relationship with VAM, measured by τ, the 90% lower CI of τ, TS, and RME. We also test that τ is greater than the τ generated by random sampling and we perform a quartile test for difference between rating Q1 and Q4 and VAM (Kane and Staiger, 2008), as shown in Table 3.

Given the tests in Table 3, we take the position that, while the signal is faint, it is indeed positive and distinguishable from the noise. One way to contextualize how we can extract the signal over many observations is to consider that VAMs have about the same accuracy as predicting year-over-year ERA[7] of professional baseball pitchers (McCaffrey et al., 2009). This task is harder, as would be the predicting of pitcher ERA from only watching a few innings of one game: on its own, there isn't enough signal, but in aggregate we assume that there is.

C.3 VAMs in other Research

Prior rater-to-VAM studies that used a much stronger signal by aggregating human rater scores and employing the far more forgiving Pearson correlation found that the significant overall relationships for MQI items were 0.09 and for CLASS items were 0.18 (Kane and Staiger, 2012), with statistical significance despite the noise, compared to the Kendall τ_b 0.11/0.03 and 0.14/0.06 reported for the humans in Figure 4. Concerned with the question of text-only signal identification, we also did the replication of (Hardy, 2024) seen as the green diamonds in Figure 4.

C.4 High-noise Context Measurement

C.4.1 Failure Mode of LLM Behavioral Homogeneity

Behavioral homogeneity, in various forms, has been in literature and should not be surprising given what we know about pretraining (McCoy et al., 2023). Behavioral homogeneity is the first of the five studies, and should not be seen as the core finding: similar to contemporary studies on LLM

[7] Earned Run Average (ERA) is the average number of earned runs a pitcher allows per nine innings, calculated as (Earned Runs × 9) / Innings Pitched. It is the primary statistic for measuring pitcher effectiveness.

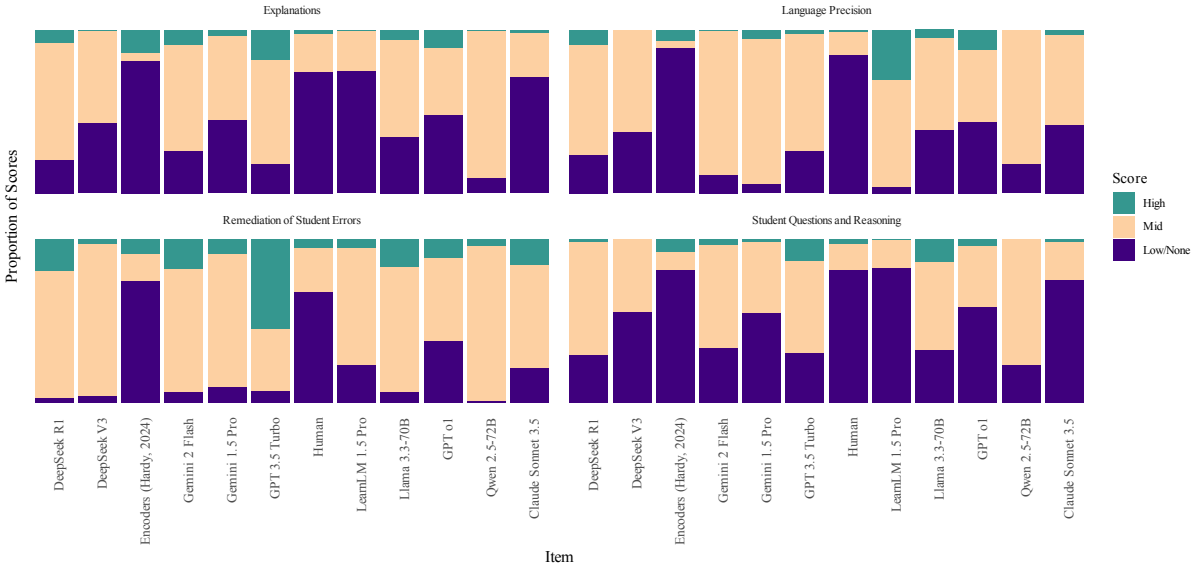


Figure 6: Proportions of rater scores by MQI item.

correlated errors (Kim et al., 2025), we first demonstrate their existence. A key idea of the alignment studies herein is highlighting the importance of scrutinizing our evaluations and working on problems that are not easily verifiable. There need to be methods for measuring, even what may be already known, in contexts where measurement has previously been too difficult.

We posit that the knowledge of the types of failure modes found within this study have been in the literature for a long time, but that it may not be a shared belief across NLP researchers as LLMs increase in size and in ultracrepidarianism. K12 education is not the only domain affected by rushing research because the tasks of interest are very difficult to verify. Efforts to understand the economic impact[8] of LLMs on professionals, such as GDPval (Patwardhan et al., 2025), have systematically excluded classrooms and contexts involving children. Education is not alone in this either; social work, therapy, and other sectors are also affected beyond the "top 9". The issues of being able to produce meaningful measurements where F1 scores above 0.95 affect much more than education.

Thus, an interesting contribution of this paper is methodological: robust and principled methods for measuring outcomes in high-noise contexts. In our opinion, it is less interesting that our first finding is a possibly predictable failure mode and more interesting that our methods allowed us to detect it, given the high noise contexts. This may provide hope for other domains where downstream tasks are difficult to measure and intended impacts even more so. Detecting these failure modes, and then measuring their augmentation as LLMs are used in ensemble, is novel in these contexts.

[8]We note that this study evaluates downstream task, but does not, in fact, evaluate the ultimate relationship of the economic impact of the quality of the downstream task outputs.

D Variance Decomposition Model

This appendix formalizes the structural decomposition used to attribute observed misalignment error to facets of a fully crossed evaluation design, introduced in Section 3.3 and discussed in Section 5.4. The aim is not merely to report error, but to identify which portions of error are (i) developer-controllable (LLM, PROMPT), (ii) systemic and shared across implementations (ITEM, OBS, and their interactions), and (iii) irreducibly idiosyncratic at the granularity of a single score (cell-specific residual variation). While our application concerns classroom instruction, the design and estimands apply broadly to LLM evaluation whenever outputs are compared to noisy, multifaceted targets.

We present the framework with sufficient generality that it can be applied to any fully crossed evaluation design, not only in education but wherever LLM outputs are compared against noisy, multifaceted ground truth. We are not aware of prior use of this method in the ACL literature, and we hope it proves useful to researchers seeking to disentangle controllable from irreducible sources of error in high-noise evaluation contexts. We hope

Table 2: CLASS and MQI item descriptions and corresponding abbreviations from the test set of Wang and Demszky. †denotes items that are reverse coded due to being negatively worded with respect to the construct of teacher ability. **Bolded** items are the focus items of the present alignment study. Bracketed item descriptions are the names of the factors with highest identified loadings by Blazar et al. per category in Appendix 2.b of the original study (Kane et al., 2015). While in the present impact alignment study we only evaluated four items, we performed a full replication of all seven items, the responses and outputs of which can be found in the online data. Variance decompositions for all seven items on the test set can be found in Appendix D.

Item	Item Name	[Factor] Item Description
MQI		
EXPL	*Teacher Explanations*	[Mathematical Instruction] Teacher explanations that give meaning to ideas, procedures, steps, or solution methods.
LANGIMP†	*Imprecision in Language or Notation*	[Mathematical Errors] Imprecision in language or notation, with regard to mathematical symbols and technical or general mathematical language.
REMED	*Remediation of Student Errors and Difficulties*	[Mathematical Instruction] Remediation of student errors and difficulties addressed in a substantive manner.
SMQR	*Student Mathematical Questioning and Reasoning*	[Mathematical Instruction] Student mathematical questioning and reasoning, such as posing mathematically motivated questions, offering mathematical claims or counterclaims.
CLASS		
CLPC	*Classroom Positive Climate*	[General Instruction] The relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions.
CLBM	*Behavior Management*	[Classroom Organization] The use of effective methods to encourage desirable behavior and prevent and redirect misbehavior.
CLINSTD	*Instructional Dialogue*	[General Instruction] The purposeful use of dialogue across the class to facilitate students' understanding of content including structured, cumulative questioning and discussion which guide and prompt students.

Table 3: Robustness Test Results Across Different Tasks

Task	$\tau > 0$	Low. 90% CI > 0	Thiel Sen > 0	RME > 0	$T >$ Rand	$T \geq T$. Exp	$Q_4 > Q_1$	Robust. Pass Rate
CLBM	0.07, Y	0.06***, Y	0.013, Y	0.008, Y	-0.04, Y	0.0004, Y $(-0.04, 0.04)$	0.02**, Y	100 (7/7)
CLINSTD	0.02, Y	0.003*, Y	0.003, Y	0.0002, Y	0.004, Y	0.0004, Y $(-0.04, 0.04)$	0.02**, Y	100 (7/7)
LANGIMP	0.12, Y	0.10***, Y	0.06, Y	0.05, Y	0.03, Y	0.04, Y (0.000, 0.08)	0.03**, Y	100 (7/7)
REMED	0.02, Y	0.001*, Y	0.008, Y	0.004, Y	-0.01, Y	0.04, Y (0.000, 0.08)	0.01, N	71 (5/7)

Note: Y indicates the test passed, N indicates failure. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the methods will support others in being able to understand how much of the error they are seeing in their models is irreducible in contexts where their signal is clouded by many facets of variation (Brennan, 2001b,a, 2003). Education happened to be the content matter of this study, but the study is about the ability to measure in high-noise contexts, with noisy labels and mismatched grain sizes. We hope this work helps to create a middle ground between "gold standard" labels and qualitative research: the "pewter standard" labels, perhaps.

D.1 Model specification

Let the index set of the fully crossed design be $\mathcal{D} = \mathcal{C} \times \mathcal{I} \times \mathcal{M} \times \mathcal{P}$, where $\mathcal{C}$ denotes observed classroom transcript segments ($|\mathcal{C}| = N_C$),

$\mathcal{I}$ rubric items ($|\mathcal{I}| = N_I$), $\mathcal{M}$ foundation models ($|\mathcal{M}| = N_M = 16$), and $\mathcal{P}$ prompting strategies ($|\mathcal{P}| = N_P = 3$). For each $(c, i, m, p) \in \mathcal{D}$, let $\widetilde{X}_{cimp}$ be the standardized[9] LLM rating and $\widetilde{Y}_c$ the (pre-standardized) stacked value-added measure for classroom c. The misalignment error is

$$\hat{e}_{cimp} \;=\; \big(\widetilde{X}_{cimp} - \widetilde{Y}_c\big)^2. \tag{3}$$

Random-effects decomposition. We fit a fully crossed random-effects model over all main effects and interactions:

$$\hat{e}_{cimp} = \mu \;+ \sum_{\emptyset \neq \alpha \subseteq \{c,i,m,p\},\, |\alpha| \leq 3} \nu_\alpha \;+\; \eta_{cimp}, \tag{4}$$

where each ν_α is a mean-zero random effect indexed by the corresponding facet(s), $\nu_\alpha \sim \mathcal{N}(0, \sigma^2_\alpha)$, mutually independent across α, and $\eta_{cimp} \sim \mathcal{N}(0, \sigma^2_\eta)$ is the cell-specific remainder. Because each (c, i, m, p) cell is observed once, the four-way interaction is not separately identifiable from the residual; hence we report the combined term as

$$\sigma^2_{cimp+\epsilon} \;:=\; \sigma^2_{\text{LLM:PROMPT:ITEM:OBS}} + \sigma^2_\epsilon.$$

Expanded Random-effects. More explicitly, we can reformulate Eq. 4 in expanded form the disaggregated sum of random effects corresponding to all main effects and interactions in the fully crossed $I \times M \times P \times C$ design:

$$\begin{aligned}
\hat{e}_{cimp} \;=\; \mu &+ \underbrace{\nu_c + \nu_i + \nu_m + \nu_p}_{\text{main effects}} \\
&+ \underbrace{\nu_{ci} + \nu_{cm} + \nu_{cp} + \nu_{im} + \nu_{ip} + \nu_{mp}}_{\text{two-way interactions}} \\
&+ \underbrace{\nu_{cim} + \nu_{cip} + \nu_{cmp} + \nu_{imp}}_{\text{three-way interactions}} \\
&+ \;\varepsilon_{cimp},
\end{aligned} \tag{5}$$

where μ is the grand mean, each random effect is normally distributed with mean zero, $\nu_\alpha \sim \mathcal{N}(0, \sigma^2_\alpha)$ for each index set $\alpha \subseteq \{c, i, m, p\}$, $|\alpha| \geq 1$, and $\varepsilon_{cimp} \sim \mathcal{N}(0, \sigma^2_\varepsilon)$ is the residual. All random effects are mutually independent.

[9]The findings are robust to standardization types. We standardize here by dividing by two standard deviations (instead of one) to better align the relative scales.

Variance share Let $\mathcal{S}$ denote the set of modeled variance components (all σ^2_α plus σ^2_η). The *variance share* of component α is

$$\pi_\alpha \;=\; \frac{\sigma^2_\alpha}{\sigma^2_{\text{total}}},\; \sigma^2_{\text{total}} = \sum_\alpha \sigma^2_\alpha + \sigma^2_\varepsilon = \sum_{k \in \mathcal{S}} \sigma^2_k. \tag{6}$$

Posterior estimates of π_α for all 15 components appear in Table 1 (summary) and Table 4 (full fit statistics including mean, SD, frequentist-style CI, and $\hat{R}$).

D.2 Sign-preserving estimations for shared behaviors in errors

The squared-error formulation in Eq. 3 and reported in Table 4 discards the direction of misalignment (whether an LLM over- or under-rates relative to VAM). To recover this information and better estimate the degree to which LLM errors move together, we fit two additional models using the *signed magnitude*:

$$\hat{e}^{\pm}_{cimp} \;=\; \big|\widetilde{X}_{cimp} - \widetilde{Y}_c\big| \;\cdot\; \big(\widetilde{X}_{cimp} - \widetilde{Y}_c\big), \tag{7}$$

which preserves the sign while retaining quadratic scaling. We estimate it twice: once using the items from the present study $\hat{e}^{\pm}$ and a second estimation, $\hat{e}^{\pm}_{(+)}$, includes all of the items from the original replication study (for a total of seven) to provide more observations of behavior. The results of these decompositions are in Tables 5 and 6, respectively.

From this latter estimation we find that at most $56.4\% \pm 10.2\%$, $\hat{R} = 0.99$ of the variance in signed error is attributable to factors involving the LLM or prompt, reinforcing the conclusion that roughly half of the misalignment variance arises from facets outside a developer's control. Note that these additional three items did not undergo robustness checks and were not used in the main body of the study because they had weaker loadings on constructs than other items in their same factor category (Blazar et al., 2017). See the parallel analyses performed on the human data in Table 8.

D.3 Bayesian estimation details

We estimate Eq. 4 using Bayesian multilevel modeling (`brms`). Each standard deviation parameter receives a weakly informative half-t prior:

$$\sigma_k \sim \text{Student-t}^+(3, 0, 2.5) \qquad \forall k \in \mathcal{S}.$$

Convergence is assessed via rank-normalized split-$\hat{R}$; all reported models satisfy $\hat{R} \approx 1$ (see tables).

A frequentist REML fit with the identical formula yields consistent component ordering and comparable point estimates, providing a second check that the variance partition is not an artifact of Bayesian regularization. $\hat{R}$ values were computed using rank-normalized split chains following Vehtari et al. (2021). The formula and all hyperparameters (iterations, chains, cores, thinning) are shown in Figure 7.

D.4 Generalizability and decision studies

Having estimated variance components, we use the Generalizability Theory decision-study (D-study) framework (Brennan, 2001c) to ask a forward-looking question: *how many LLMs and prompting strategies would one need to average over in order to obtain a stable estimate of the shared, persistent error signal?*

D.4.1 Generalizability Design

In a generalizability framework, the three prompt families (base, chain-of-thought, pseudo-RAG) are treated as sampled operationalizations from a broader universe of commonly used strategies. The key claim is not that these specific prompts failed, but that variance attributable to the prompt facet—including its interactions with LLM—is small relative to other sources. This rests on the full variance decomposition, not on the prompt main effect alone.

We define the *object of measurement* as the item–observation combination (i, c), and the *instrumentation facets* as LLM (m) and prompt (p). Averaging over n_m models and n_p prompts, the relative and absolute reliability coefficients are given by the following.

Definition D.1 (Relative reliability)**.** Let S denote the set of all variance components estimated in Eq. 5, and let $S_\alpha = \{\alpha \in S\}$ be the subset containing the object of measurement, α. The expected relative reliability for n_m models and n_p prompts is

$$E\widehat{\rho}^2_\alpha \;=\; \frac{\sigma^2_\alpha}{\displaystyle\sum_{k \in S_\alpha} \frac{\sigma^2_k}{n_k} \;+\; \frac{\sigma^2_\varepsilon}{n_m\, n_p}}, \tag{8}$$

where n_k is the product of the sample sizes of the instrumentation facets appearing in index set k (e.g., for interaction $k = \{i, c, m\}$, $n_k = n_m n_i n_c$ where $\alpha = ic$ is the object of measurement and thus $n_i = 1$ and $n_c = 1$).

Practically, we can look at the join task-transcript ic object of measurement (as the level of unit for a single score outside of developer control), i.e., $S_\alpha = \{\alpha \in S : \{i, c\} \subseteq \alpha\}$.

Definition D.2 (Absolute reliability)**.** The absolute error variance includes all components. The generalizability coefficient (index of dependability) is

$$\widehat{\Phi}_\alpha \;=\; \frac{\sigma^2_\alpha}{\displaystyle\sum_{k \in S} \frac{\sigma^2_k}{n_k} \;+\; \frac{\sigma^2_\varepsilon}{n_m\, n_p}}. \tag{9}$$

Both coefficients are computed by sampling directly from the posterior of the variance components (rather than from plug-in point estimates), yielding fully Bayesian credible intervals.

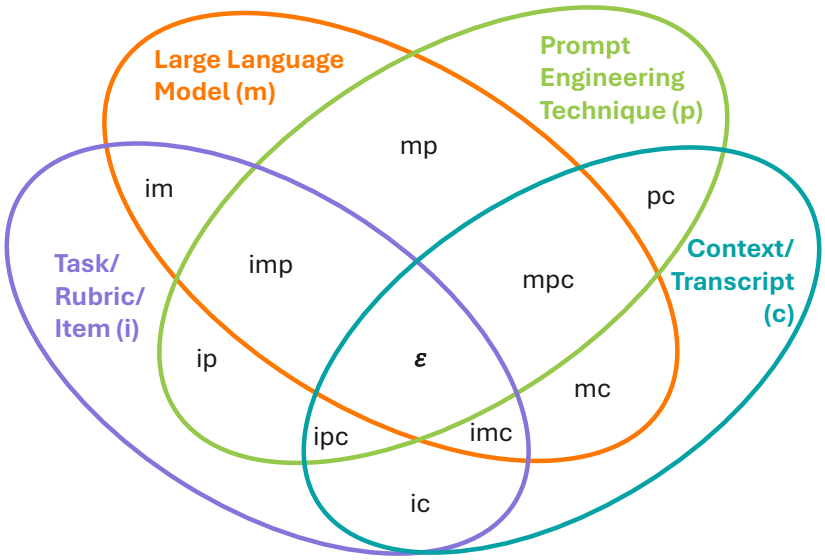
D.4.2 Signal Detectability

Figures 8 and 9, display $E\widehat{\rho}^2$ and $\widehat{\Phi}$ as functions of n_m and n_p, showing rapid saturation: even modest numbers of models and prompts suffice to recover the shared error signal. They represent different aspects of the present study's ability to capture meaningful underlying signal across LLMs for cases in $\hat{e}$. In both when $N_{\text{LLM}} = 16$ and $N_{\text{Prompt Strat.}} = 3$, the reliabilities of $E\widehat{\rho}^2$ and $\widehat{\Phi}$ are reported in Table 7. This suggests that the estimated item-transcript scores in this study achieve approximately the target level of consistency expected for these types of data (0.65, see Ho and Kane 2013).

D.4.3 Interpretation

These results imply that prompt-induced shifts are largely additive and limited in magnitude rather than transformative. While unexplored prompts may exist, the observed interaction structure suggests that large corrective effects are improbable unless they represent qualitatively new mechanisms outside the sampled strategy universe. The inference is thus about *effect-size stability*, not prompt exhaustiveness. In long-context transcript settings, identifying the exact source of brittleness for any single prompt is a Sisyphean task; the generalizability framework instead supports a probabilistic claim: across a realistic and theory-informed prompt universe, prompt choice is unlikely to produce large alignment corrections relative to the error that is shared across all models.

We note that the shared item–observation error is confounded with measurement error in the "true" item–classroom score relative to VAM. Nonetheless, the preponderance of evidence in the present



I×M×P×C Bayesian Variance Design

```
brm(formula = sqdiffVAM ~
(1|ITEM) +
(1|OBS_CHAP)+
(1|PROMPT) + (1|LLM) +
(1|ITEM:OBS_CHAP)+
(1|LLM:PROMPT) +
(1|LLM:ITEM) +
(1|LLM:OBS_CHAP)+
(1|LLM:PROMPT:ITEM) +
(1|LLM:PROMPT:OBS_CHAP)+
(1|LLM:ITEM:OBS_CHAP) +
(1|PROMPT:ITEM)+
(1|PROMPT:ITEM:OBS_CHAP) +
(1|PROMPT:OBS_CHAP),
data = df, iter = 3500,
      chains = 3,
      cores = 6, thin=20)
```

Figure 7: **Bayesian Error Variance Decomposition**: (*left*) fully crossed facet diagram for sources of variance. (*right*) corresponding `brms` code listing.

study indicates that the LLM-correlated component of this error is undesirable. Future work could incorporate hierarchical rater models (Casabianca et al., 2016), as in Hardy (2024), to estimate the true score as a latent parameter, or apply noise-control methods such as those illustrated in Appendix F.

Prompt Brittleness We see distributional brittleness in prompting technique in the form of a skewed, long-tailed posterior distribution (see median and upper HDI bound, Tables 1, 4, and Apdx D). Practically, this means that certain prompting techniques can *occasionally* inject large amounts of error with no expected contribution. A dramatic improvement from a prompt change would not be expected to generalize to new situations.

D.5 Supplementary theory: identifiability and interpretation

Proposition D.3 (Four-way interaction non-identifiability)**.** *In a fully crossed $I \times C \times M \times P$ design with one observation per cell, the variance of the four-way interaction term ν_{icmp} is not separately identifiable from the residual variance.*

Proof sketch. With one observation per cell, any cell-specific deviation from the sum of lower-order effects can be equivalently represented as either a four-way interaction draw or residual noise. The likelihood depends only on their sum, implying non-identifiability. □

Proposition D.4 (Controllable-variance upper bound)**.** *Let $\mathcal{S}_{\text{dev}} := \{k \in \mathcal{S} : m \in k \text{ or } p \in k\}$ be the set of components involving developer-controlled facets (*LLM *or* PROMPT*). Then the maximum fraction of error variance that can be affected by changing model and/or prompt (holding*

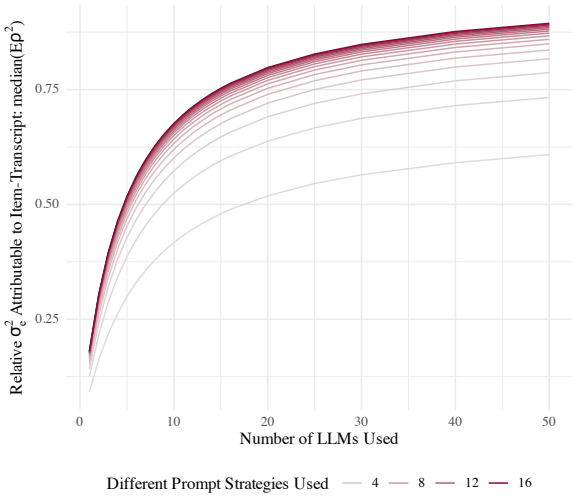


Figure 8: **Reliability of Relative Shared Error Signal for ITEM×OBS**: Bayesian Decision Study for item-transcript scores as the object of study across varying numbers of LLMs and Prompting Techniques required to remove LLM and prompt-specific idiosyncrasies from the error signal. Each value is calculated directly from sampling the posterior and extracting the median. For this study, where $N_{\text{LLM}} = 16$ and $N_{\text{Prompt Strat.}} = 3$, the reliability $E\hat{\rho}^2_{ic}$ is 0.66 ± 0.02 ($\hat{R} = 1.00$). This suggests that the estimated item-transcript scores in this study achieve approximately the target level of consistency expected for these types of data (0.65, see Ho and Kane 2013).

Table 4: **Bayesian Squared Error Variance Decomposition** ($\hat{e}$) Parameter Estimates and Fit Statistics

variable	median	MAD	MAP	HDI	mean	SD	CI	$\hat{R}$
ITEM	0.08	0.08	0.03	[0.01,0.62]	0.14	0.17	[0.01,0.54]	0.99
LLM	0.07	0.04	0.06	[0.02,0.18]	0.08	0.05	[0.02,0.16]	1.00
OBS	0.02	0.01	0.02	[0,0.03]	0.02	0.01	[0.01,0.03]	1.01
PROMPT	0.02	0.03	0.00	[0,0.58]	0.08	0.15	[0,0.46]	1.00
ITEM:OBS	0.04	0.01	0.05	[0.01,0.06]	0.04	0.01	[0.02,0.06]	1.00
LLM:ITEM	0.03	0.01	0.03	[0.01,0.06]	0.03	0.01	[0.01,0.05]	1.00
LLM:OBS	0.00	0.00	0.00	[0,0]	0.00	0.00	[0,0]	1.00
LLM:PROMPT	0.02	0.01	0.01	[0,0.04]	0.02	0.01	[0.01,0.03]	1.00
PROMPT:ITEM	0.01	0.01	0.00	[0,0.05]	0.01	0.01	[0,0.03]	1.01
PROMPT:OBS	0.00	0.00	0.00	[0,0]	0.00	0.00	[0,0]	1.00
LLM:ITEM:OBS	0.19	0.03	0.21	[0.05,0.23]	0.18	0.05	[0.07,0.23]	0.99
LLM:PROMPT:ITEM	0.03	0.01	0.03	[0.01,0.05]	0.03	0.01	[0.01,0.05]	1.01
LLM:PROMPT:OBS	0.14	0.02	0.15	[0.04,0.17]	0.13	0.03	[0.05,0.17]	1.00
PROMPT:ITEM:OBS	0.02	0.00	0.02	[0.01,0.03]	0.02	0.01	[0.01,0.03]	1.00
LLM:PROMPT:ITEM:OBS + ϵ	0.24	0.04	0.26	[0.06,0.29]	0.22	0.06	[0.09,0.28]	1.00

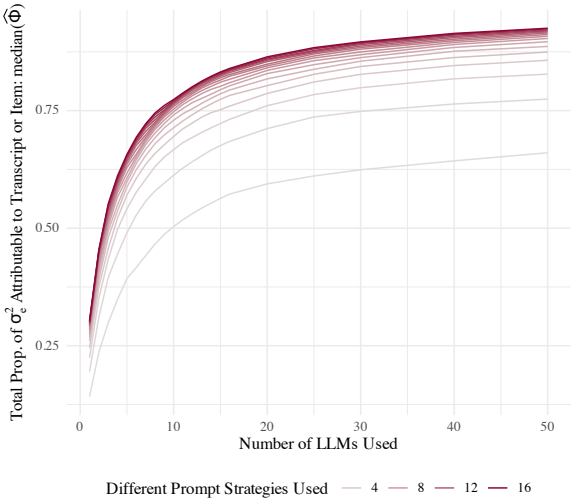


Figure 9: **Reliability of Absolute Shared Error Signal**: Bayesian Decision Study for any of item, transcript, or their interaction as the objects of measurement ($\sigma^2_\alpha = \sigma^2_i + \sigma^2_c + \sigma^2_{ci}$) across varying numbers of LLMs and Prompting Techniques required to remove LLM and prompt-specific idiosyncrasies from the error signal. Each value is calculated directly from sampling the posterior and extracting the median. For this study, where $N_{\text{LLM}} = 16$ and $N_{\text{Prompt Strat.}} = 3$, the reliability $\widehat{\Phi}$, Equation 9, of the mean rating across models and prompts shows is 0.73 ($\hat{R} = 1.00$).

the object facets fixed) is upper bounded by

$$\Pi_{\text{dev}} := \sum_{k \in \mathcal{S}_{\text{dev}}} \pi_k.$$

Interpretation. Π_{dev} is an *optimistic* ceiling: it counts as "controllable" even those interactions (e.g., LLM×OBS) whose direction may not generalize across new classroom segments. A small Π_{dev} therefore implies that prompt/model changes cannot be expected to deliver large, reliable alignment improvements.

Corollary D.5 (When prompt engineering is predictably effective)**.** *If π_p and $\sum_{k:p\in k} \pi_k$ are both large and have concentrated posteriors (narrow HDIs), then prompt changes can be expected to yield reliable error reductions across contexts; conversely, small medians with wide HDIs indicate brittle prompt behavior.*

D.6 Parallel estimation on human expert error

To aid readers less familiar with interpreting variance decompositions, we provide a parallel analysis that replaces the LLM×PROMPT instrumentation facets with individual human rater identifiers. This is a *different* estimation from Eq. 5—the facets, and therefore what counts as "controllable" versus "systemic," change accordingly. The results (Table 8) are not directly comparable in absolute terms, but the *proportional* distribution of variance is instructive.

D.6.1 Expected pattern for expert raters

If human raters are performing the rating task competently, and if items vary in their contribution to VAM (Kane et al., 2013; Hardy, 2025b), then most observed human error should concentrate on the item facet and item–observation interaction. This

Table 5: **Bayesian Signed Squared Error Variance Decomposition** ($\hat{e}^{\pm}$) Parameter Estimates and Fit Statistics

variable	median	MAD	MAP	HDI	mean	SD	CI	$\hat{R}$
ITEM	0.03	0.04	0.01	[0,0.34]	0.07	0.09	[0,0.22]	1.00
LLM	0.09	0.04	0.07	[0.03,0.24]	0.10	0.05	[0.03,0.19]	1.01
OBS	0.03	0.01	0.03	[0.01,0.05]	0.03	0.01	[0.01,0.04]	1.01
PROMPT	0.00	0.01	0.00	[0,0.5]	0.06	0.13	[0,0.47]	1.05
ITEM:OBS	0.12	0.02	0.12	[0.06,0.15]	0.11	0.02	[0.07,0.14]	1.05
LLM:ITEM	0.06	0.02	0.06	[0.03,0.1]	0.06	0.02	[0.03,0.1]	1.02
LLM:OBS	0.00	0.00	0.00	[0,0]	0.00	0.00	[0,0]	1.02
LLM:PROMPT	0.01	0.01	0.01	[0,0.03]	0.01	0.01	[0,0.03]	1.01
PROMPT:ITEM	0.00	0.00	0.00	[0,0.03]	0.00	0.01	[0,0.02]	1.02
PROMPT:OBS	0.00	0.00	0.00	[0,0]	0.00	0.00	[0,0]	1.00
LLM:ITEM:OBS	0.20	0.02	0.21	[0.1,0.24]	0.19	0.03	[0.11,0.23]	1.06
LLM:PROMPT:ITEM	0.06	0.01	0.06	[0.03,0.08]	0.06	0.01	[0.03,0.08]	1.06
LLM:PROMPT:OBS	0.13	0.02	0.14	[0.07,0.16]	0.13	0.02	[0.07,0.15]	1.05
PROMPT:ITEM:OBS	0.02	0.00	0.02	[0.01,0.03]	0.02	0.00	[0.01,0.03]	1.06
LLM:PROMPT:ITEM:OBS + ϵ	0.17	0.02	0.17	[0.08,0.19]	0.16	0.03	[0.09,0.19]	1.05

is precisely what we find.

D.6.2 Results

The rater main effect and the observation main effect each account for 0% of variance in squared error—individual expert biases are negligible in the absence of interactions. For unsigned error, $\hat{e}$ he item main effect absorbs 33.8%, and the item–observation interaction an additional 25.2%, for a combined 59.0% attributable to the rubric dimension and its interplay with the specific lesson segment. This is the region where "true score" mismatch between expert ratings and VAM is most expected (Kane et al., 2013). Rater-specific biases on particular items or lesson segments contribute a combined 17.3%.

This finding becomes much more pronounced when sign is included in the error signal $\hat{e}^{\pm}$, and $\hat{e}^{\pm}_{(+)}$. Here, the combined variance in error attributable to the item alone increases to 69.3%. If the items themselves are misaligned, we would expect humans to be highly correlated and sensitive to the addition of such items. We also see this dramatic variability in the LLMs with the addition of the three additional items.

D.6.3 Contrast with LLM error structure

For LLMs, the analogous item-plus-item×observation share is only 15% of total error variance. The error structure of expert humans is thus concentrated where theory predicts it should be—on the constructs being measured—whereas LLM error disperses across high-order interactions involving model and prompt, indicating instability rather than construct-related variation. Proportions absorbed by the residuals are similar between LLMs and humans, so the variation in shares is comparable.

The code for the estimations of the models can be found in code Listing D.6.3

```
lme4::lmer(errSCORE ~
          (1|ITEM)
        + (1|OBS_CHAP)
        + (1|RATERID)
        + (1|ITEM:OBS_CHAP)
        + (1|RATERID:ITEM)
        + (1|RATERID:OBS_CHAP),
        data = df,
control=lmerControl(optimizer="bobyqa"))
```

E Measures and Metrics

E.1 Measuring Dependence with Distance Correlation

In this appendix, we provide a self-contained technical overview of the bias-corrected squared distance correlation, the primary statistic used in our analysis to quantify the dependence between raters. We first motivate the need for a measure beyond traditional linear correlation and then formally define the statistic and its properties.

E.1.1 Motivation: Beyond Linear Correlation

A common method for measuring the association between two random variables, U and V, is the Pearson product-moment correlation coefficient, $\rho(U, V)$. While widely understood, Pearson's ρ

Table 6: **Bayesian Signed Squared Error Extended Variance Decomposition** ($\hat{e}^{\pm}_{(+)}$) Parameter Estimates and Fit Statistics for all replication study items. Note the large difference in Item variation when other items are included and the decrease in Prompting volatility.

variable	median	MAD	MAP	HDI	mean	SD	CI	$\hat{R}$
ITEM	0.29	0.13	0.27	[0.14,0.63]	0.33	0.14	[0.15,0.59]	0.99
LLM	0.04	0.03	0.04	[0.01,0.16]	0.05	0.04	[0.01,0.12]	0.99
OBS	0.07	0.02	0.07	[0.03,0.09]	0.07	0.02	[0.04,0.09]	1.00
PROMPT	0.00	0.00	0.00	[0,0.23]	0.02	0.06	[0,0.13]	1.01
ITEM:OBS	0.09	0.02	0.09	[0.04,0.12]	0.08	0.02	[0.05,0.11]	1.01
LLM:ITEM	0.04	0.01	0.04	[0.02,0.07]	0.04	0.02	[0.02,0.07]	1.00
LLM:OBS	0.01	0.00	0.01	[0,0.01]	0.01	0.00	[0.01,0.01]	1.00
LLM:PROMPT	0.01	0.01	0.01	[0,0.03]	0.01	0.01	[0,0.03]	1.01
PROMPT:ITEM	0.00	0.00	0.00	[0,0.01]	0.00	0.00	[0,0.01]	1.00
PROMPT:OBS	0.00	0.00	0.00	[0,0]	0.00	0.00	[0,0]	1.02
LLM:ITEM:OBS	0.12	0.02	0.13	[0.06,0.16]	0.12	0.03	[0.07,0.16]	0.99
LLM:PROMPT:ITEM	0.05	0.01	0.05	[0.02,0.07]	0.05	0.01	[0.03,0.07]	1.00
LLM:PROMPT:OBS	0.04	0.01	0.05	[0.02,0.06]	0.04	0.01	[0.03,0.05]	0.99
PROMPT:ITEM:OBS	0.03	0.01	0.03	[0.01,0.03]	0.02	0.01	[0.01,0.03]	0.99
LLM:PROMPT:ITEM:OBS + ϵ	0.15	0.03	0.16	[0.07,0.2]	0.15	0.03	[0.09,0.19]	0.99

Table 7: **Decision Study**: performing a similar set of estimations to the LLM variance decomposition, except using the human experts on same test sets. We estimate the coefficients for generalizability and dependability for each of the same decompositions, $\hat{e}$, $\hat{e}^{\pm}$, and $\hat{e}^{\pm}_{(+)}$ for two units of measurement

Metric	Unit of Measure	$\hat{e}$	$\hat{e}^{\pm}$	$\hat{e}^{\pm}_{(+)}$
$E\hat{\rho}^2$	$\mathcal{C} \times \mathcal{I}$	0.657	0.837	0.819
$\widehat{\Phi}$	$\mathcal{C} \times \mathcal{I}$	0.551	0.790	0.781
$E\hat{\rho}^2$	$\{\mathcal{C} \times \mathcal{I}, \mathcal{C}, \mathcal{I}\}$	0.819	0.839	0.946
$\widehat{\Phi}$	$\{\mathcal{C} \times \mathcal{I}, \mathcal{C}, \mathcal{I}\}$	0.732	0.809	0.931

Table 8: **Human Error Variance Decomposition**: performing a similar set of estimations to the LLM variance decomposition, except using the human experts on same test sets. We estimate the percentage of human expert-VAM error for each of the same decompositions, $\hat{e}$, $\hat{e}^{\pm}$, and $\hat{e}^{\pm}_{(+)}$ using Restricted Maximum Likelihood with `lmer` (Bates et al., 2015).

Facet	Pct($\hat{e}$)	Pct($\hat{e}^{\pm}$)	Pct($\hat{e}^{\pm}_{(+)}$)
ITEM	33.76	63.75	69.27
OBS	0.00	0.00	1.10
RATERID	0.00	0.00	0.00
ITEM:OBS	25.22	13.73	11.84
RATERID:OBS	12.85	8.01	3.69
RATERID:ITEM	4.42	3.27	3.35
RATERID:ITEM:OBS + ϵ	23.76	11.24	10.75

is designed to detect only *linear* relationships. It can be zero even when the variables share a strong, deterministic, but nonlinear relationship (e.g., $V = U^2$ for a symmetric U).

In the context of this study, we are comparing rating distributions from different sources (LLM-LLM, LLM-human). There is no *a priori* reason to assume that the relationship between two sets of ratings is linear. For example:

- One model might use a compressed range of scores compared to another, leading to a curvilinear relationship.
- Two raters might agree on clear-cut cases (very high or very low quality instruction) but diverge in their assessment of moderately effective instruction, producing a non-monotonic relationship.

Relying on a linear measure in such cases would systematically underestimate the true degree of behavioral correspondence between the raters.

To overcome this limitation, we employ the **distance correlation**, a non-parametric measure of dependence introduced by Székely et al. (2007). Distance correlation is designed to capture any type of statistical dependence between two random vectors of arbitrary and not necessarily equal dimension. Its central property, which makes it exceptionally powerful, is that the population distance correlation is zero *if and only if* the variables are statistically independent.

E.1.2 The Bias-Corrected Squared Distance Correlation

We now formally define the squared distance correlation and its bias-corrected sample estimator, which we denote dCor_n^2. We focus on the squared value, as it is more convenient for statistical inference and bias correction.

Let $(X, Y) = \{(x_k, y_k) : k = 1, \ldots, n\}$ be a statistical sample of n paired observations from a joint distribution (U, V). In our case, x_k and y_k represent the scalar ratings given by two different raters to the k-th classroom transcript.

Sample Distance Matrices The computation begins by constructing Euclidean distance matrices for each variable. Let the $n \times n$ distance matrix for the X sample be $\mathbf{a}$, with entries:

$$a_{kl} = \|x_k - x_l\| = |x_k - x_l|, \tag{10}$$

and similarly for the Y sample, $\mathbf{b}$, with entries $b_{kl} = |y_k - y_l|$. The use of absolute differences is a specific case of the Euclidean norm for scalar ratings.

Double Centering and Sample Distance Covariance To make the measure invariant to rigid transformations (translation and orthogonal rotation), the distance matrices are *double-centered*. For each element a_{kl} in the distance matrix $\mathbf{a}$, we compute its centered counterpart A_{kl}:

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}, \tag{11}$$

where $\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}$ is the k-th row mean, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$ is the l-th column mean, and $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$ is the grand mean of the distance matrix. The same procedure is applied to $\mathbf{b}$ to obtain the centered matrix $\mathbf{B}$.

The *sample squared distance covariance*, $\mathrm{dCov}_n^2(X, Y)$, is the arithmetic average of the products of the corresponding centered distances:

$$\mathrm{dCov}_n^2(X, Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}. \tag{12}$$

The sample squared distance variances are the distance covariances of each variable with itself: $\mathrm{dCov}_n^2(X, X)$ and $\mathrm{dCov}_n^2(Y, Y)$.

E.1.3 The Bias-Corrected Estimator

The natural (but biased) sample estimator for the squared distance correlation is the ratio of the sample squared distance covariance to the product of the sample distance standard deviations. However, this estimator is positively biased for finite samples; it will be non-zero on average even for independent variables. This bias complicates inference, especially with moderate sample sizes.

To address this, Szekely and Rizzo (2014) introduced a bias-corrected estimator, which we denote $\mathrm{dCor}_n^2(X, Y)$. It is based on a U-statistic estimator of the squared distance covariance, which is unbiased. The computation involves a slightly different centering:

$$\tilde{A}_{kl} = a_{kl} - \frac{1}{n-2}\sum_{j=1}^{n} a_{kj} - \frac{1}{n-2}\sum_{i=1}^{n} a_{il} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^{n} a_{ij}. \tag{13}$$

The unbiased estimator of squared distance covariance is then:

$$\widetilde{\mathrm{dCov}}_n^2(X, Y) = \frac{1}{n(n-3)}\sum_{k\neq l} \tilde{A}_{kl}\tilde{B}_{kl}. \tag{14}$$

The bias-corrected squared distance correlation, dCor_n^2, is constructed as the ratio of these unbiased estimators:

$$\mathrm{dCor}_n^2(X, Y) = \frac{\widetilde{\mathrm{dCov}}_n^2(X, Y)}{\sqrt{\widetilde{\mathrm{dCov}}_n^2(X, X)\,\widetilde{\mathrm{dCov}}_n^2(Y, Y)}}. \tag{15}$$

Due to the bias correction, dCor_n^2 can take small negative values when the true dependence is zero or near-zero. In practice, negative estimates are typically treated as evidence of independence and can be reported as zero. The statistic is bounded above by 1.

E.1.4 Key Properties

The power of distance correlation is summarized in the following from Székely et al. (2007): Let U and V be random variables with finite first moments. The population distance correlation $\mathrm{dCor}(U, V) = 0$ if and only if U and V are statistically independent. This property guarantees that distance correlation will detect any departure from independence, linear or not. In contrast, Pearson correlation's analogous property holds only for bivariate normal distributions. By using the bias-corrected estimator dCor_n^2, we obtain a reliable and sensitive measure of the dependence between rater judgments, robust to the specific functional form of

their relationship and suitable for the sample sizes in our study. This makes it the ideal tool for uncovering the subtle but strong patterns of behavioral convergence among LLMs documented in Section 5.1.

E.2 Disattenuated Stacked Correlations for Underlying VAM

By stacking both VAM measures (STA and ALT in Eq. 16) (Kane and Staiger, 2008, 2012) at the teacher-year level, we can estimate the correlation of the underlying value-added to student learning through these two standardized measures. The correlation between the two VAMs is reduced by the fact that there is measurement error in both estimates for a given teacher-year.

We can use the noise-ceiling, the geometric mean of the product of implied reliabilities with the underlying value add, to disattenuate the correlation from some of the measurement error. This is because the correlation between either measure and the underlying value-add is the square root of the correlation between the two noisy measures (Kane and Staiger, 2012).

$$\tau_{S_j,Y} = \frac{\tau_{S_j\tilde{Y}}}{\sqrt{\tau_{\tilde{Y}_{STA},\tilde{Y}_{ALT}}}} \tag{16}$$

We use this relationship when measuring the correlations between classroom ratings and the underlying teacher value-add on student learning. This scalar transform for the y-axis has no effect on positions (only changing the y-axis tick marks) in Figure 4. The findings of the study are robust to (and would be strengthened by) additionally utilizing Greiner's equality, $\rho = \sin\left(\frac{\pi}{2}E[\tau]\right)$ (Greiner, 1909) when performing this transformation. We report the results without this transformation both for simplicity and to better preserve the alignment nature of τ.

E.3 Expert Ensembling

Conventional wisdom suggests that ensembling multiple models improves robustness and accuracy by leveraging diverse model strengths or averaging out independent errors. Our findings directly challenge this assumption for educational evaluation tasks. We tested two conceptually opposed ensemble strategies: (1) **pedagogy-expertise weighting**, which weights model votes by performance on an AI pedagogy benchmark (Lelièvre et al., 2025),[10] and (2) **unanimous voting**, which selects only cases where all models agree, distilling their shared consensus.

For a set of models $\mathcal{F} = \{f_1, \ldots, f_K\}$, the pedagogy-weighted ensemble score for transcript t on task j is:

$$S_{\text{weighted},tj} = \frac{\sum_{k=1}^{K} w_k \cdot S_{f_k,tj}}{\sum_{k=1}^{K} w_k} \tag{17}$$

where w_k represents model f_k's performance on a pedagogy knowledge benchmark (we test MMLU Education subset, MMLU Mathematics, and a specialized mathematics pedagogy benchmark). The unanimous ensemble restricts analysis to the subset $\mathcal{T}_{\text{unan}} = \{t : S_{f_1,tj} = S_{f_2,tj} = \cdots = S_{f_K,tj}\}$ of transcripts where all models assign identical ratings.

Figure 4 (middle and bottom rows) reveals that **both ensemble strategies not only fail to improve alignment with student learning but dramatically worsen it** for several critical instructional dimensions. For REMED (remediation of student errors), pedagogy-weighted ensembles (middle row) shift the alignment distribution downward: median $\tau_{S_{\text{weighted}}Y}$ decreases from approximately -0.15 (individual models) to -0.28 (weighted ensemble), a statistically significant degradation ($p < 0.01$, bootstrap test). Similarly, for CLBM (behavior management), unanimous voting ensembles (bottom row) shift dramatically lower $\tau_{S_{\text{unan}}Y} < -0.2$, compared to the more dispersed individual model distribution.

F Alternate Methods for Downstream Task Alignment: BLUPs

F.1 Teacher Skill Model Best Linear Unbiased Predictions (BLUPs)

A teacher skill model described allows us to capture the nested structure of the data and account for various sources of random variation, including teacher effects, rater bias, lesson-specific factors, and skill-specific effects. By modeling these sources of variance as random effects, we obtain more accurate estimates of the underlying teacher skill. The strength of this approach is that it helps isolate the parts of the variation attributable to specific teacher skill targeted by the rubric item, which is particularly important where there may be correlations across skills.

The detailed model specification is given by:

[10] Findings are robust to using other expertise benchmark scores such as MMLU-Pro and math-specific benchmarks

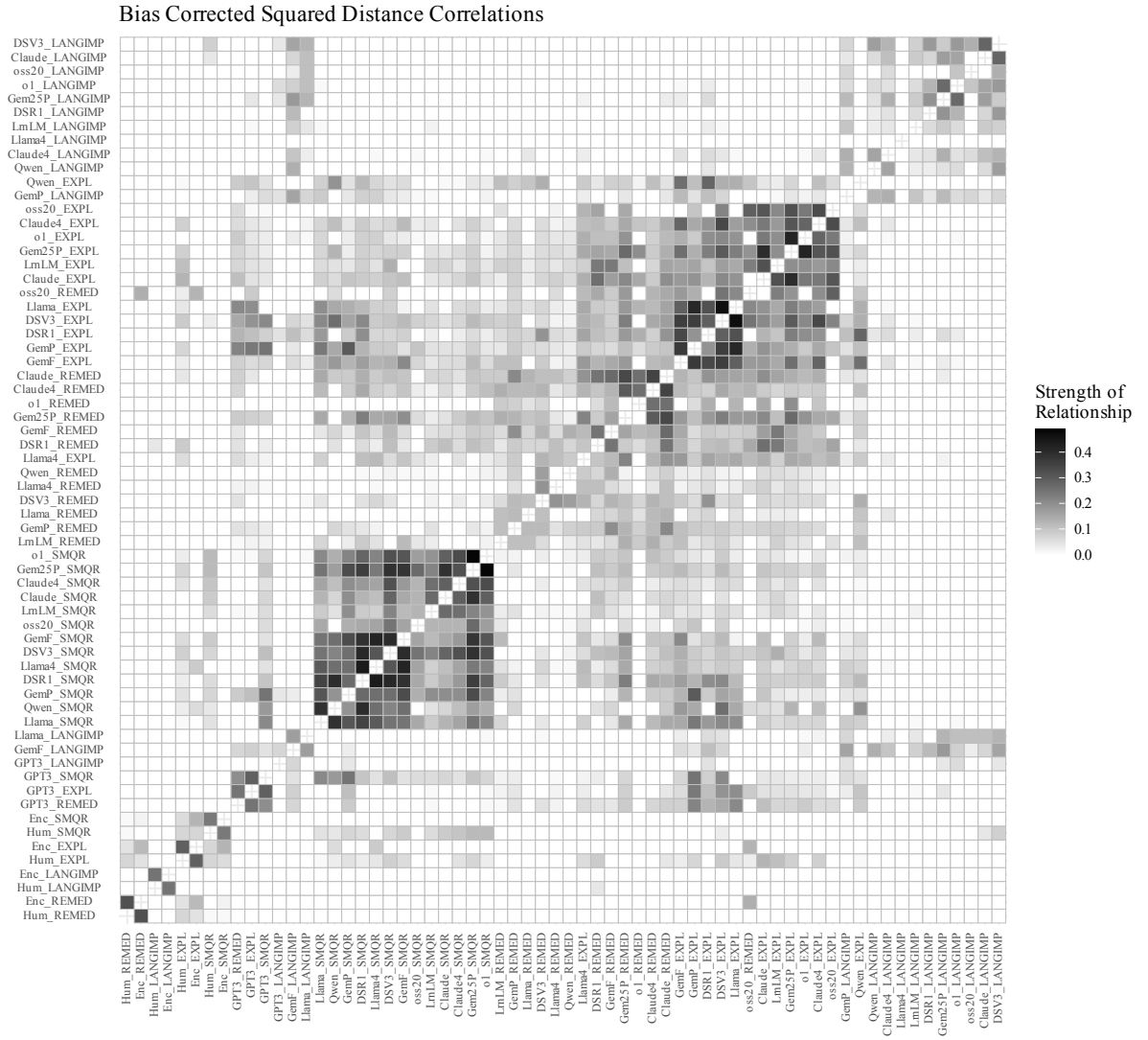


Figure 10: **Between and Within LLM Distance Correlations: MQI**, distance correlations nonparametric measure of dependence between and within rater families across MQI items. Correlation are conducted at the item-transcript level using pairwise-complete observations. Nonsignificant relationships (at $\alpha < 0.05$) are shown as blank after adjusting for family-wise error rate using the Bonferroni correction. Hierarchical clustering is done using complete linkages.

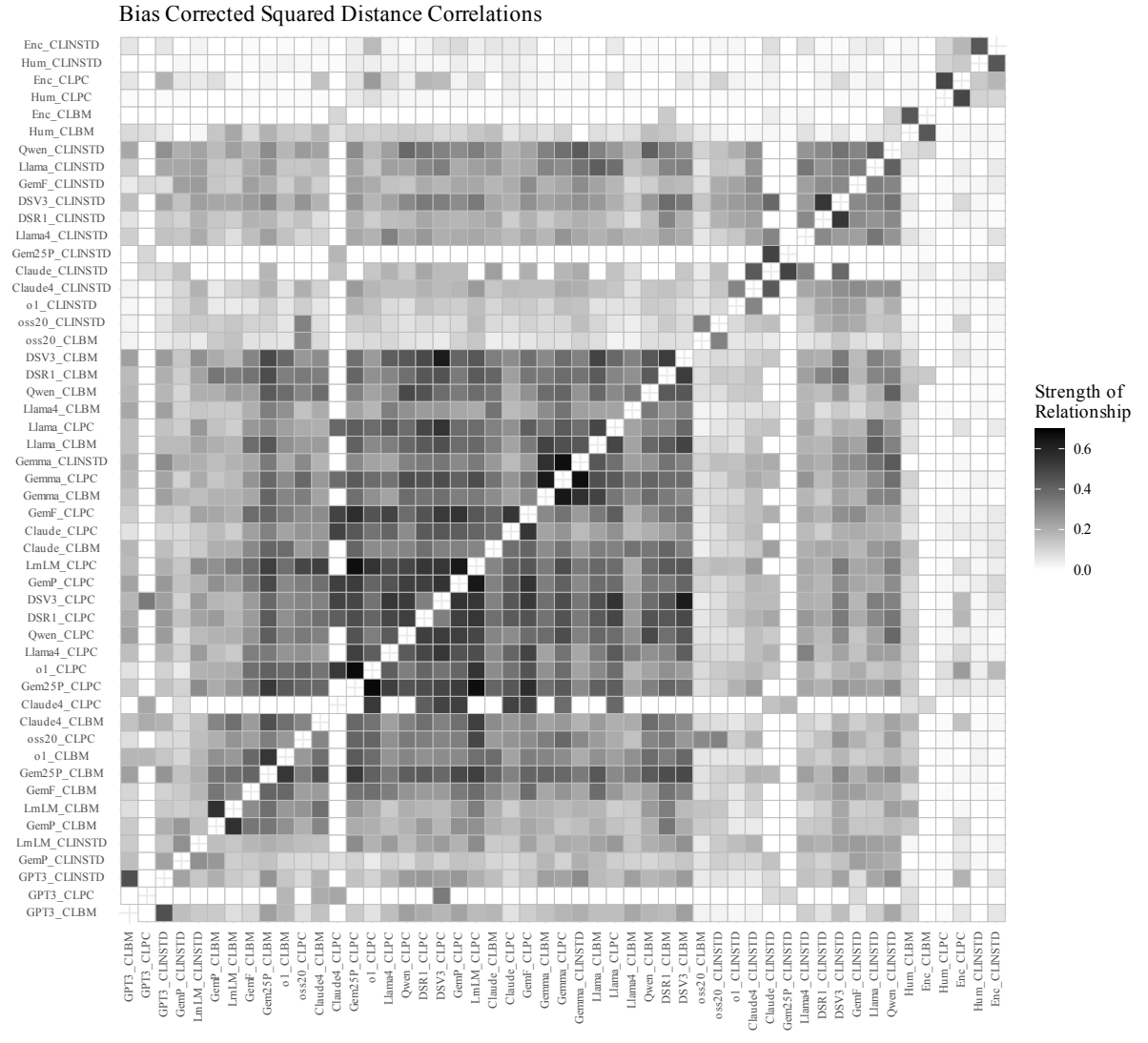


Figure 11: **Between and Within LLM Distance Correlations: CLASS**, distance correlations nonparametric measure of dependence between and within rater families across CLASS items. Correlation are conducted at the item-transcript level using pairwise-complete observations. Nonsignificant relationships (at $\alpha < 0.05$) are shown as blank after adjusting for family-wise error rate using the Bonferroni correction. Hierarchical clustering is done using complete linkages.

$$
\begin{aligned}
\mathrm{S}_{rjsli} = \mu + \nu_i + \nu_r + \nu_j + \nu_{l:i} + \nu_{s:l:i} + \\
\nu_{ji} + \nu_{jl:i} + \nu_{js:l:i} + \nu_{ri} + \nu_{rl:i} + \nu_{rs:l:i} \\
+ \nu_{rji} + \nu_{rjl:i} + \nu_{rj} + \epsilon_{rjs:l:i}
\end{aligned} \tag{18}
$$

where S_{rjsli} represents the score for Item j given by Rater r, during class segment s within lesson l taught by teacher i. Here, μ is the overall mean, and the random effects capture the hierarchical variance. Our focal component for evaluating model annotations at the classroom transcript level is $\nu_{js:l:i}$, which encapsulates the variation attributable to the observable teacher skill for each segment after accounting for other sources of variation. This allows us to estimate BLUPs for the individual transcript sections seen by the models, thereby isolating the variation associated with the specific task of interest, $\hat{\nu}_{js:l:i}$, which is used in the correlation analyses. A breakdown of the notation for the model is below:

- X_{rjsli} represents the rating for the i-th teacher, by the r-th rater, on the j-th skill, during the s-th segment, for the l-th lesson.
- μ is the overall grand mean.
- ν_i is the random effect for the i-th teacher. Interpretation: some teachers receive higher ratings than others. [11]
- ν_r is the random effect for the r-th rater: some raters are more lenient than others.
- ν_j is the random effect for the j-th skill item: some items are easier than others.
- ν_{rj} is the random effect for the interaction between rater and skill. Some raters score certain items higher.
- $\nu_{l:i}$ is the random effect for the l-th lesson within a teacher: some lessons receive higher ratings than others. Confounded with teacher overall score dependence on lessons.
- $\nu_{s:l:i}$ is the random effect for the s-th segment of the l-th lesson for the i-th teacher. A segment is the unit of one transcript: some lessons receive higher ratings than others. Confounded with lesson overall score dependence on segments.
- ν_{ji} is the random effect for the interaction between the skill and teacher: some teachers score higher on some skills. This component will be used to estimate BLUPs of the expected skill level for a teacher.
- $\nu_{jl:i}$ is the random effect for the interaction between skill and lesson: some lessons score higher on some skills.
- $\nu_{js:l:i}$ is the random effect for the interaction between skill and segment: some segments of lessons receive higher scores than others. This component will be used to estimate BLUPs at the transcript level.
- ν_{ri} is the random effect for the interaction between rater and teacher: some raters score certain teachers higher.
- $\nu_{rl:i}$ is the random effect for the interaction between rater and lesson. Some raters score certain lessons higher.
- ν_{rji} is the random effect for the interaction between rater, skill, and teacher: some raters score certain teachers higher on some items.
- ν_{rjli} is the random effect for the interaction between rater, skill, and lesson: some raters score certain lessons higher on some items.
- ϵ_{rjsli} is the residual error term.

[11] For holistic evaluations, this might be considered the "true score" of a teacher's overall performance.

F.1.1 Code Listing

Below is the code used for the estimation of the model defined by Equation 18 using `lme4` syntax. The model was estimated using restricted maximum likelihood (REML) with the BOBYQA optimizer. The variable names are the same found in the original study (Kane et al., 2015). Scores were minmax scaled to $[0, 1]$ prior to estimation.

```
+ (1 | RATERID) + (1 | RATERID:ITEM)
+ (1 | RATERID:NCTETID)
+ (1 | RATERID:OBSID:CHAPNUM)
+ (1 | OBSID/CHAPNUM)
+ (1 | OBSID:CHAPNUM:ITEM)
+ (1 | OBSID:ITEM)
+ (1 | NCTETID:ITEM)
+ (1 | RATERID:ITEM:NCTETID)
+ (1 | RATERID:OBSID:ITEM)
+ (1 | RATERID:OBSID), data = df,
control=lmerControl(optimizer="bobyqa"))
\caption{Target Teaching Skill
    ↪ Estimation Model Code}
\label{code:lmefit}
```

BLUP Extraction BLUPs were extracted `lme4::predict(model, re.form = (1 | NCTETID:ITEM))` and `lme4::predict(model, re.form = (1 | OBSID:CHAPNUM:ITEM))` for the respective ν_{ji} and $\nu_{js:l:i}$

Parameter Estimates Parameter estimates for the model are in Table 9.

Table 9: Mixed Effect Estimates for Teacher Skill Model

Parameter	Group	Effect Type	Coefficient	CI Low	CI High
(Intercept)	—	Fixed	0.46	0.33	0.58
Segment	w_k	Random (σ_k^2)	0.03	–	–
	RATERID:OBSID:ITEM	Random	0.09	–	–
	RATERID:ITEM:NCTETID	Random	0.03	–	–
	RATERID:OBS_CHAPS	Random	0.03	–	–
	RATERID:ITEM	Random	0.05	–	–
	RATERID:OBSID	Random	0.04	–	–
	RATERID:NCTETID	Random	0.02	–	–
	OBSID	Random	0.02	–	–
	Residual	Random	0.16	–	–
	NCTETID	Random	0.03	–	–
	RATERID	Random	0.03	–	–
	ITEM	Random	0.32	–	–

G Alternate Methods for Intended Impact Alignment: Refined VAM Residualization Model

G.1 Refining VAM Measurements

For the y-axis, we measure the alignment between transcript ratings and end of year value-added measures. To assess the alignment between Large Language Model (LLM) ratings of teacher skills and student learning gains, we employ teacher value-added measures (VAMs) derived from standardized assessments. However, VAMs operate at the teacher or teacher-year level, while LLM ratings are generated at the finer-grained lesson-segment-by-teacher-skill level.

Naïve aggregation of LLM ratings would be inappropriate due to the sparse and unbalanced nature of the data, as LLM evaluations were only available for short segments of lessons and for a subset of teacher skills. To bridge this gap in granularity and isolate the relationship between LLM ratings and VAMs, we employ a two-stage approach involving a mixed-effects model and subsequent semipartial correlation analysis.

The residuals from this model, which represent the VAMs after accounting for confounding variables, are then used to compute semipartial correlations. We opted for semipartial correlations to ensure that the controls applied to the dependent variable (VAMs) remain constant across all LLMs, facilitating fair comparisons.

To remove confounding variance from the VAM signal, we construct a comprehensive mixed-effects model to partition the variance in teacher VAMs, accounting for 1) variables that would contribute to VAM scores that are unrelated to a teacher's instructional practice and 2) variables that would mediate the relationship between an observer's score and VAM. This model allows us to isolate the residual variance in VAMs specifically attributable to observable instructional skills, along with any remaining measurement error. By residualizing the VAMs, we remove variation attributable to the representativeness of the lesson segment, and remove confounds potentially introduced by our decision to model with all the information (various VAMs and all 25 items), and account for the various programmatic, curricular, student population, and school-year related relationships to

This methodological approach addresses several key challenges:

1. **Granularity mismatch**: By using lesson-segment level residuals, we preserve the fine-grained nature of LLM ratings while relating them to year-level VAMs.

2. **Unbalanced observations**: Our approach accounts for the fact that LLM ratings typically cover only one section of a lesson, with little overlap between teachers observed.

3. **Preservation of meaningful variance**: The random effect for teachers, scaled by the difference between observed and expected performance, retains more variance for teacher observations that are closest to their typical performance.

4. **Ordinal nature of ratings**: The use of Kendall correlations acknowledges the ordinal nature of most teaching quality ratings and focuses on directional alignment rather than precise numeric agreement.

5. **Multiple sources of VAMs**: By incorporating multiple types of VA measures, we account for differences in assessment season relative to classroom observations.

G.2 Value-added Measures

Two complementary value-added measures are used at two levels of aggregation: VAMs calculated from formal state standardized assessments and another from study-specific, low-stakes and informal assessments, aggregated at the teacher-year level and across years. Although imperfect, together the VAMs provide a more robust measure with the underlying teacher value-add. The residualization model is estimated using all information from available VAM types. Only VAM residuals corresponding to the same teacher-year as observed lesson transcript are used for the subsequent semi-partial correlations.

By stacking VAMs from both the state and the study-specific exams, we can correlate rater and LLM scores against the underlying teacher value-add as measured through these instrument. We can then use the correlation of these VAMs to correct for rater correlations against this underlying value-add.

G.2.1 Controls for Confounding

The analysis further controls for numerous confounding factors, including district, season of observation, grade level, class composition, teacher experience, class size, and subject-specific expertise, to isolate the effect of instructional quality on student outcomes.

G.2.2 Controls for Divergence from Typical Teacher Skill

To address the issue of differing levels of granularity, we leverage Best Linear Unbiased Predictors (BLUPs) extracted from the fully crossed random effects model described in the previous section (and explicitly defined in Eq, 18) to isolate the signal most relevant to interpreting the relationships between ratings of instructional quality and student learning.

Specifically, we estimate two BLUPs with this model. The first, $\hat{\nu}_{jsli}$, was used in the previous section and estimates the skill level displayed by a teacher during a specific segment of a lesson. The second, $\hat{\nu}_{ji}$, estimates the latent ability for a specific skill for a teacher across all observations. The difference between these, $\Delta_{si,j} = \hat{\nu}_{js:l:i} - \hat{\nu}_{j:i}$, provides a metric that captures the *representativeness* of a particular lesson segment in relation to the teacher's overall skill profile. This difference is then used as a predictor in a linear mixed-effects model to adjust the VAMs for the representativeness of each lesson segment.

The residualizing model is stylized as follows:

$$
\begin{aligned}
V_{dgs:l:ityvj} = \; & \beta_1 D_d + \beta_2 G_g + \beta_3 (D_d G_g) \\
& + \boldsymbol{\beta_C} \mathbf{C}_{cgSyt} + \boldsymbol{\beta_S} \mathbf{S}_{dSyi} \\
& + \boldsymbol{\beta_T} \mathbf{T}_i + \tau \boldsymbol{M}_{s:l} \\
& + \boldsymbol{\nu}_{dglsv} [D_d \times (G_g \times Y_y + m_{s:l})] \\
& + \boldsymbol{\nu}_{dglsv} [D_d \times (G_g \times Y_y + m_{s:l}) \times v] \\
& + \boldsymbol{\nu}_{dglsv} S_S + \gamma_t \Delta_{si,j} + \epsilon_{dgsyt}
\end{aligned}
$$

where:

- V_{dgsytj} represents the stacked teacher VAM for district d, grade g, school S, year y, and term t, incorporating multiple standardized assessment outcomes. Stacking multiple VAMs increases the reliability of the overall teacher effectiveness measure and mitigates potential biases associated with the season of assessments relative to classroom observations.
- D_d fixed effects systematic differences in student achievement attributable to District- programmatic and curricular policies and practices. The inclusion of the interaction term acknowledges the potential for district-specific effects by grade-level G_g for curricular and programming decisions.
- $\mathbf{C}_{cgSyt}$, $\mathbf{S}_{dSyt}$, and $\mathbf{T}_i$ represent vectors of class-, school-, and teacher-level covariates, respectively. These include prior achievement, student demographics, class size, school size, teacher experience, and measures of teacher knowledge. These covariates control for factors that influence student achievement that would not be observable by raters of instruction or are not directly related to the teacher instructional quality.
- $\mathbf{M}_{s:l}$, vectors of lesson segment (s) within the lesson (l)-specific covariates with respect to the time during the school year, length of class, and the unrepresentativeness of the observation with respect to the teacher's average expected level of performance, $\Delta_{si,j}$.
  - $\Delta_{si,j}$ represents the difference between the teacher's observed skill level in a specific lesson segment ($\hat{\nu}_{jsli}$) and their

overall mean skill level ($\hat{\nu}_{ji}$), calculated as $\Delta_{si,j} = \hat{\nu}_{jsli} - \hat{\nu}_{ji}$. These Best Linear Unbiased Predictors (BLUPs) are derived from a separate, fully crossed mixed-effects model in Equation 18 and incorporated here to weight each lesson segment according to its representativeness of the teacher's overall skill profile. This approach allows us to preserve the variance contributed by lesson segments that are most indicative of the teacher's typical performance, supporting observation-level inference.

- Random effects, $\boldsymbol{\nu}_{dgltsv}$, are crossed and indexed by district, grade, time during the school year, type of value-added measure $\boldsymbol{\nu}_{dglsv}[D_d \times (G_g \times Y_y + m_{s:l}) \times (1+v) + S_S]$, account for the nested structure of the data and the potential correlations within districts and between different types of VAM (v). Additionally, each school S_S has a random slope to account for variance not attributable to the fixed effects. It is important to not use a school-level random effect since a fixed effect would distort the residuals, for example, in schools where all teachers were good
- The random slopes for each teacher γ_t are with respect to $\Delta_{si,j}$, allows for teacher-specific variation in the relationship between lesson segment representativeness and VAM, preserving the variation for those segments that are more representative of the teacher.
- $\epsilon_{dgs:l:ityvj}$ represents the residual error, which retains the 1) variation in VAM attributable to each lesson segment, scaled by its representativeness vis-a-vis the teacher, whose effects on VAM are lessened as the segment diverges from the teacher's average, 2) measurement error, and 3) any remaining bias from omitted variables.

We represent the residualized VAM values for an individual teacher $\epsilon_{dgs:l:ityvj} = \tilde{V}_{sjv}$, for each unique nested segment-skill-VAM intersection. For use in the semi-partial correlations of the main portion of this study, we use only the subset of the residuals corresponding to the end of year VAMs on the state assessment and alternative assessment (original study) $v \in \{STA, ALT\}_y$, to better align teaching practices to same-year outcomes.

Below is the code listing for the model specification using the `lme4` software in `R`.

```
outcome ~ 0
        + DISTRICT*GRADE ## District
            ↪ fixed effects and grade
            ↪ fixed effect baselines
        + V_CS_ALT_IRT_M_TM1 ## class
            ↪ mean standardized student
            ↪ performance on BOY
            ↪ assessment
        + V_CS_STATE_STD_M_TM1 ## class
            ↪ mean standardized student
            ↪ performance on previous
            ↪ year math assessment
        + V_CS_STATE_STD_E_TM1 ## class
            ↪ mean standardized student
            ↪ performance on previous
            ↪ year English/reading
            ↪ assessment
        + V_CCLASS_SIZE ## class size
        + MAXCHAP ## length of class
            ↪ observed
        + V_CS_SPED ## class prop. of
            ↪ SPED students
        + V_CS_LEP  ## class prop. of
            ↪ English Learners
        + V_CS_FRPL ## class prop. of Low
            ↪ -SES students
        + ACCURACY_yr_est ## standardized
            ↪  measure of teacher
            ↪ accuracy in predicting
            ↪ their student's errors
        + KOSM_yr_est ## standardized
            ↪ measure of teacher
            ↪ knowledge of common
            ↪ student misconceptions
        + MATH_KNOWLEDGE_yr_est ##
            ↪ standardized measure of
            ↪ teacher knowledge of
            ↪ mathematics for
            ↪ instruction
        + RepDelta ## difference between
            ↪ the BLUP-estimated scores
            ↪ of this observation and
            ↪ BLUP-estimated teacher
            ↪ average for that
            ↪ observational item
        + TIMING ## season/month during
            ↪ the year of the
            ↪ observation
        + V_SS_FRPL ## school prop. of
            ↪ Low-SES students
        + V_SS_SPED ## school prop. of
            ↪ SPED students
        + V_SS_LEP ## school prop. of
            ↪ English Learners
        + V_SS_STATE_STD_M_TM1 ## school
            ↪ mean standardized student
            ↪ performance on previous
            ↪ year math assessment
        + V_SS_STATE_STD_E_TM1 ## school
            ↪ mean standardized student
            ↪ performance on previous
            ↪ year English/reading
            ↪ assessment
        + V_SSCHOOL_SIZE ## school size
        + (1|DISTRICT:GRADE:SCHOOLYEAR_SP
            ↪ ) ## variations in
            ↪ programming, resources,
            ↪ and capacity in districts
            ↪ by grade and year
```

```
        + (1|ITEM:outvar) ## variation in
            ↪  how each teacher skill
            ↪ may be related to
            ↪ different VAMs
        + (1|DISTRICT:TIMING:outvar) ##
            ↪ variation with respect to
            ↪ the academic calendar and
            ↪ the timing of the
            ↪ assessments used in VAM (
            ↪ capturing district-level
            ↪ variation around test prep
            ↪  or instructional
            ↪ initiatives)
        + (1|DISTRICT:GRADE:SCHOOLYEAR_SP
            ↪ :outvar) ## variation of
            ↪ emphases and curricula for
            ↪  districts during
            ↪ different seasons of the
            ↪ year, with respect to the
            ↪ kind of outcome (capturing
            ↪  district-level variation
            ↪ around test prep or
            ↪ instructional initiatives)
        + (0 + RepDelta|NCTETID) ##
            ↪ variation by teacher,
            ↪ which have a random slope
            ↪ to control the teacher
            ↪ variation with difference
            ↪ between the BLUP-estimated
            ↪  scores of this
            ↪ observation and BLUP-
            ↪ estimated teacher average
            ↪ for a given teacher skill,
            ↪  preserving more of the
            ↪ teacher effect in the
            ↪ residual where the
            ↪ observation is more
            ↪ similar to the teacher's
            ↪ average.
```

G.3 Residualized VAM model parameters and results

Figure 12, which follows the same format and notation as Figure 4 and the following tables report out the results from the alternative estimation methods. Note the similarity in distribution shapes when estimated using the more sophisticated noise-controlling technique to simple alignment of just tau from the main body.

Table 10: Measures of fit for Residualization

Parameter	Fit
R2 (conditional)	0.26
R2 (marginal)	0.14
Sigma	0.12

H AI Use

AI assistants (Gemini 2.5) was used during final revision to polish writing. We have mixed feelings about its ability to do the task well. We tried Claude for camera-ready polishing, which we mostly had to unpolish.

Table 11: Mean VAM-alignment estimates across models after noise control, including three additional items from replication study.

Item	Estimate (CI)	p.value
All	-0.156 (-0.205, -0.107)	0
CLBM	-0.356 (-0.426, -0.287)	0
CLINSTD	-0.139 (-0.21, -0.069)	0
LANGIMP	-0.181 (-0.23, -0.131)	0
REMED	-0.233 (-0.282, -0.183)	0
CLPC	-0.141 (-0.212, -0.071)	0
EXPL	-0.212 (-0.261, -0.162)	0
SMQR	-0.124 (-0.174, -0.074)	0

Parameter	Coefficient	CI	t	p
DISTRICT [11]	0.41	[0.15, 0.67]	3.09	0.00
DISTRICT [12]	0.20	[-0.06, 0.46]	1.54	0.12
DISTRICT [13]	0.60	[0.34, 0.86]	4.52	0.00
DISTRICT [14]	0.14	[-0.12, 0.4]	1.07	0.29
GRADE	-0.10	[-0.16, -0.04]	-3.47	0.00
V CS ALT IRT M TM1	0.14	[0.14, 0.14]	319.98	0.00
V CS STATE STD M TM1	-0.02	[-0.02, -0.02]	-34.29	0.00
V CS STATE STD E TM1	-0.11	[-0.11, -0.11]	-194.07	0.00
V CCLASS SIZE	0.00	[0, 0]	9.34	0.00
MAXCHAP	0.00	[0, 0]	66.59	0.00
V CS SPED	0.01	[0.01, 0.01]	11.58	0.00
V CS LEP	0.03	[0.03, 0.03]	39.37	0.00
V CS FRPL	0.03	[0.03, 0.03]	27.82	0.00
ACCURACY yr est	0.01	[0, 0.01]	21.93	0.00
KOSM yr est	0.00	[0, 0]	-12.57	0.00
MATH KNOWLEDGE yr est	0.01	[0.01, 0.01]	88.31	0.00
samprep	0.00	[-0.01, 0]	-0.80	0.43
TIMING [WINTER]	0.02	[0.01, 0.04]	2.51	0.01
TIMING [SPRING]	0.03	[0.01, 0.05]	3.30	0.00
TIMING [FALL]	0.04	[0.02, 0.05]	3.91	0.00
V SS FRPL	0.03	[0.03, 0.04]	25.00	0.00
V SS SPED	-0.16	[-0.16, -0.15]	-90.17	0.00
V SS LEP	-0.02	[-0.02, -0.02]	-15.78	0.00
V SS STATE STD M TM1	-0.04	[-0.04, -0.04]	-49.53	0.00
V SS STATE STD E TM1	0.05	[0.05, 0.05]	69.43	0.00
V SSCHOOL SIZE	0.00	[0, 0]	-114.14	0.00
DISTRICT [12] × GRADE	0.05	[-0.03, 0.13]	1.13	0.26
DISTRICT [13] × GRADE	-0.03	[-0.11, 0.05]	-0.70	0.48
DISTRICT [14] × GRADE	0.06	[-0.02, 0.14]	1.36	0.17
sd(NCTETID)	0.04		NA	NA
sd(DISTRICT:GRADE:SCHOOLYEAR_SP:outvar)	0.02		NA	NA
sd(DISTRICT:TIMING:outvar)	0.03		NA	NA
sd(DISTRICT:GRADE:SCHOOLYEAR_SP)	0.03		NA	NA
sd(Residual)	0.12		NA	NA

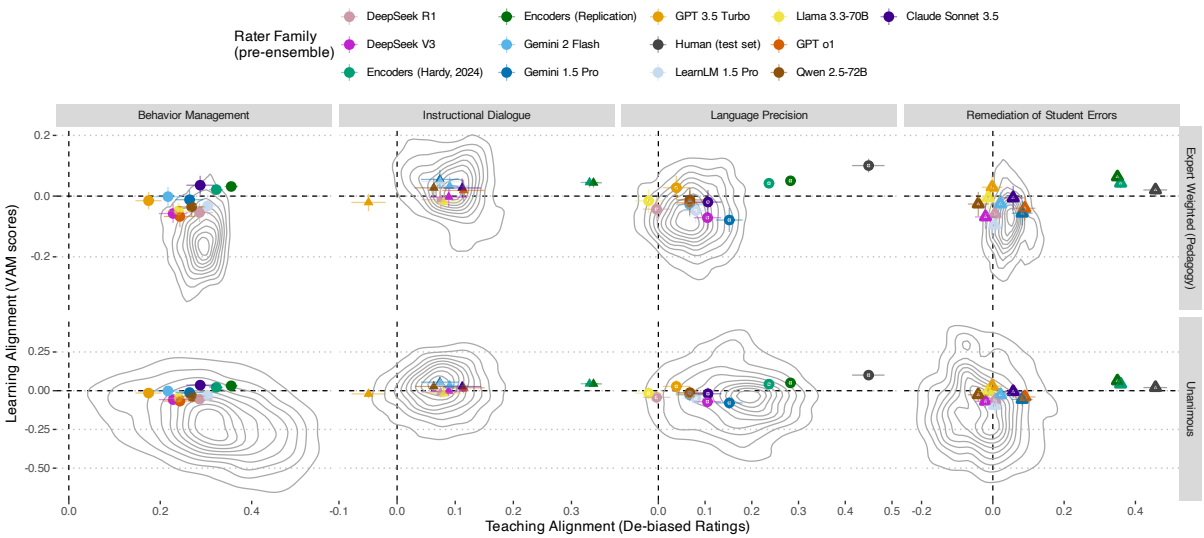


Figure 12: see the caption of Figure 4 for full description of details. The difference is that the axes have undergone the noise controlling transformations of this appendix prior to plotting.