Knowledge without Wisdom: Measuring Misalignment between LLMs and Intended Impact

Anonymous ACL submission

Abstract

LLMs increasingly excel on AI benchmarks, but doing so does not guarantee validity for downstream tasks. This study evaluates the performance of leading foundation models (FMs, i.e., generative pre-trained base LLMs) with out-of-distribution (OOD) tasks of the teaching and learning of schoolchildren. Across all FMs, 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 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 50% of the variation in misalignment error is shared across foundation models, 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 both educational applications of foundation models and to understanding limitations of models.

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). These limitations increase when tasks move downstream, away from "correct answers" and toward some *intended impact* in the world.

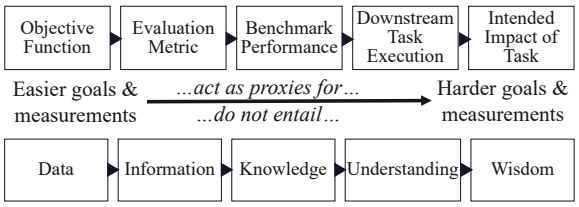


Figure 1: Cascading levels of inference found in LLM development and evaluation and a parallel adaption of Ackoff's progression (Ackoff, 1989; Rowley, 2007)

We use the research of LLMs-in-scientific-discovery studies to illustrate a gap between benchmarks and downstream tasks that is not reserved for 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 evidence (discourse) and where the ultimate objective is delayed: student learning. In this setting, it is easy to build systems that *sound* pedagogically literate while failing to identify teaching practices that actually improve achievement (Bastani et al., 2024; Shein, 2024). Moreover, annotated classroom dis-

course is out-of-distribution (OOD)[1] with respect to the Internet text that dominates LLM pretraining, raising the generalization questions about what is attributable to pretraining or LLM idiosyncrasy: *what shared behaviors do LLMs exhibit when asked to perform an out-of-distribution task?*

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 what schooling ultimately values: student growth. This concern mirrors the scientific-discovery literature, 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 conduct real science (Song et al., 2025; Alampara et al., 2025; Mirza et al., 2025; Kim et al., 2025; Kalai et al., 2025). Similarly, in education, we observe in this study that models may reproduce the *language* of effective pedagogy without tracking the features of instruction that causally support learning. We suspect that there are many other domains with congruent gaps.

Standard AI benchmarks, 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). Without rigorous evaluation against meaningful outcomes, we risk deploying technologies that are not only ineffective but potentially harmful to student learning (Bastani et al., 2024; Shein, 2024).

We use two external criteria that, to our knowledge, has only been linked by one other 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.Methodologically, we treat both as alignment targets: alignment to the *downstream task* (replicating expert ratings of teaching practice) and alignment to the *intended impact* (predicting which classrooms produce greater learning gains). The latter is considered the "gold standard" for measuring impact on student learning.

The study provides several important contributions to current study of applied LLMs. First, 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). Second, it directly quantifies a novel gap between LLM execution of downstream tasks and the *intended impact* (see section 5.2). Third, 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, it decomposes 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.

Across analyses, a consistent picture emerges that parallels recent results in LLM-enabled scientific discovery: models can converge on confident, mutually reinforcing judgments that are poorly tethered to the underlying target. In classrooms, this means that "knowledge" of pedagogical concepts does not reliably translate into the "wisdom" to discern what is relevant to human student learning.

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 FMs 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 inde-

[1] Information regarding children and their schooling is protected. This study uses the only publicly available anonymized dataset with these data of which we are aware.

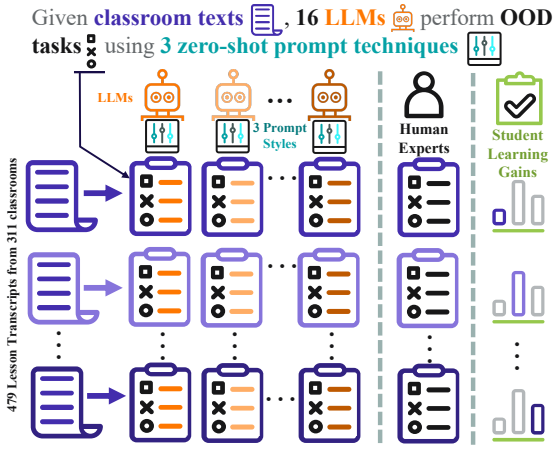


Figure 2: **Task Data and Experimental Design**. Each LLM is provided classroom transcripts. Using several prompting techniques for each model, LLMs place an ordinal rating on the quality on an aspect of teaching and learning. This is done across seven distinct tasks. We evaluate for alignment of FM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

pendence between the ratings. Our core alignment analysis evaluates pairwise comparisons. For any two lessons, A and B, if a model rates lesson A as being of higher quality than lesson B on a given task, we assess whether expert human ratings or student learning data also place A over B. 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 allowing for authentic baseline comparisons.

3 Methods

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 FMs, 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 FM 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. 8 and 9. Additional details are in Appendix C.

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, 1938; 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 also allows us to establish

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

meaningful baselines, such as using a teacher's years of experience or with a prior VAM as the ratings (see Fig. 4.) The ultimate magnitudes of the baselines are consistent with the literature (McCaffrey et al., 2009; Kane and Staiger, 2008, 2012).

3.3 Measuring Sources of Error with Variance Decomposition

In our final line of inquiry, we quantitatively analyze failure modes. To understand the sources of the observed misalignment, we partition the total variance in the prediction errors into components attributable to different facets of variation in the experimental design. This allows us to quantify how much of the misalignment is "controllable" by a developer (i.e., due to the choice of model or prompt) versus how much is systemic or idiosyncratic to the task and classroom context.

For every observation, the misalignment error, $\hat{y}$, is the residual from a simple linear model predicting student learning gains (VAM) from the FM's rating. This error term represents the degree to which an FM's rating fails to align with that teacher's measured impact on student learning. We then structurally decompose this error, using Generalizability Theory framework (Brennan, 2001), implemented as a random effects model for which the sources of variation from the items I, FM M, prompt P, and transcript segment C are fully-crossed: $I \times M \times P \times C$. $\hat{y}_{cimp}$, for a given segment c, instructional task (item) i, foundation model m, and prompt p, as the sum of these effects:

$$\hat{y}_{cimp} = \mu + \alpha_c + \beta_i + \gamma_m + \delta_p + (\alpha\beta)_{ci} + (\alpha\gamma)_{cm} + (\beta\gamma)_{im} + \cdots + \varepsilon_{cimp} \quad (1)$$

Each term in the model represents a source of variance including interactions. By calculating the proportion of total variance accounted for by each component (e.g., $\hat{\sigma}^2_M/\hat{\sigma}^2_{total}$), we can determine the relative contribution of each factor to the overall misalignment between FM ratings and student learning outcomes. Appendix B contains a variance diagram (Fig. 7), a table of component variances estimated, formula, and code.

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 the few educational datasets that contains measures of teaching practice, authentic classroom artifacts, and student learning outcomes at scale (VAMs). Our analysis utilizes three key components from this rich resource.

4.1 Classrooms and Downstream Observation Instruments

Our primary input for the FMs 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 benchmark for FM alignment.[3]

4.2 Value-Added Measures (VAMs) of 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

[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

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 (see Appendix C.2, Kane et al., 2015) for each teacher-year corresponding to the observed lesson estimate of a teacher's contribution to student learning. Crucially, this evaluation assumes 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 FMs.

5 Results and Discussion

5.1 The convergent bias of foundation models

A primary result is the striking *behavioral homogeneity* of FMs when evaluating OOD 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, **FM-FM agreement is consistently higher than FM-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 good.

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 *more negatively* associated with learning outcomes.

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 yields 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 the construct that predicts learning gains.

5.3 Ensembling exacerbates misalignment

A natural response to noisy model behavior is ensembling. We evaluate two conceptually opposite approaches: (i) a *pedagogy-expertise-weighted*

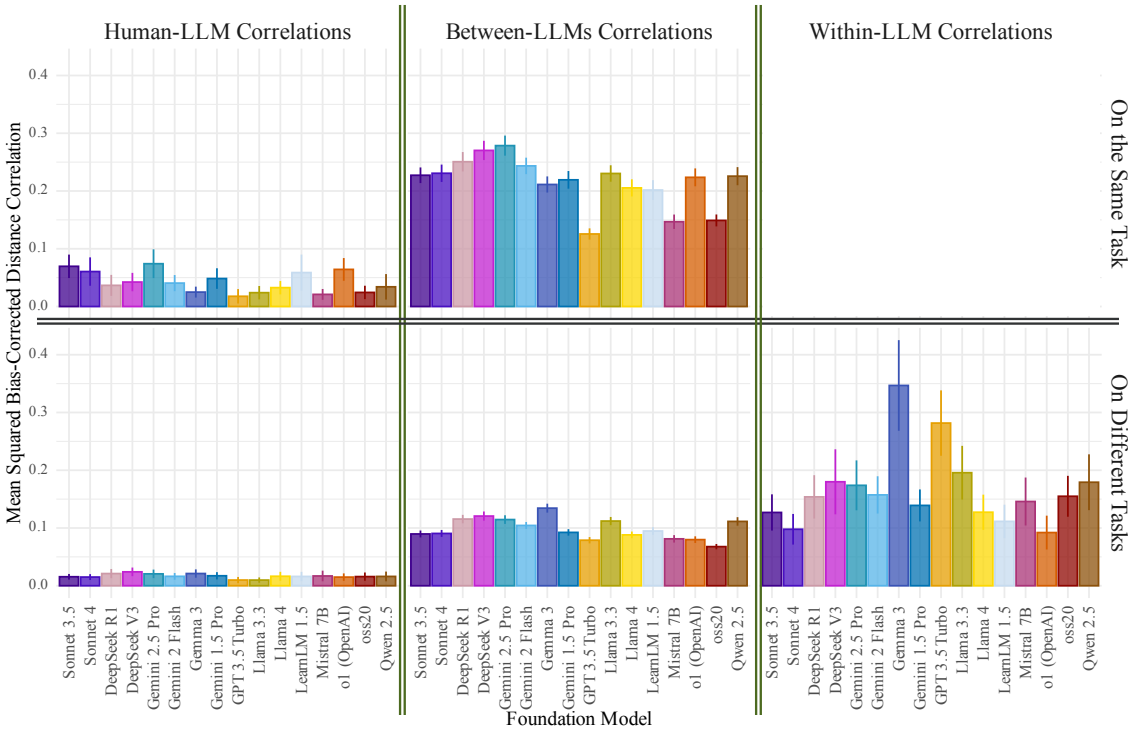


Figure 3: **Mean Inter-task Bias Corrected Squared Distance Correlations** dCor_n^2 : between LLMs and human raters across different evaluation tasks. **Top row: Same-task Correlation** Mean inter-rater distance correlations across transcripts for the same task and (**bottom row: different task correlation**) for different tasks using the same transcript. (**left: correlations with humans**) Mean inter-rater distance correlations with expert human raters, (**center: correlations with other FMs**) 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. 8 and 9

ensemble, where model votes are weighted by pedagogical benchmark performance, and (ii) a *unanimous-vote* ensemble, where we only score classrooms when multiple models agree.

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 Permanent artifacts of autoregression

We next ask whether practitioners can fix misalignment through the two most accessible levers: selecting a different FM (M) and altering the zero-shot prompt (P). A variance decomposition of prediction error shows these levers explain only a small fraction of the total variance: the main effect of model choice accounts for 4.8%, prompt choice for 1.0%, and their interaction for 1.4%. This result complements the correlation evidence in Section 5.1. If misalignment were driven by a few poor models or poorly chosen prompts, we would expect much larger explainable variance from M and P. Instead, the small contributions imply that **misalignment is largely systemic**: it reflects common artifacts of current FM training and generalization rather than idiosyncratic implementation choices. In practical terms, prompt engineering and model shopping are unlikely to make classroom-evaluation systems safe or useful when judged against learning outcomes.

Summing all variance components involving a developer's choice (any term with 'M' or 'P'), we find they collectively account for only half (50%) of the total variance in misalignment. The major-

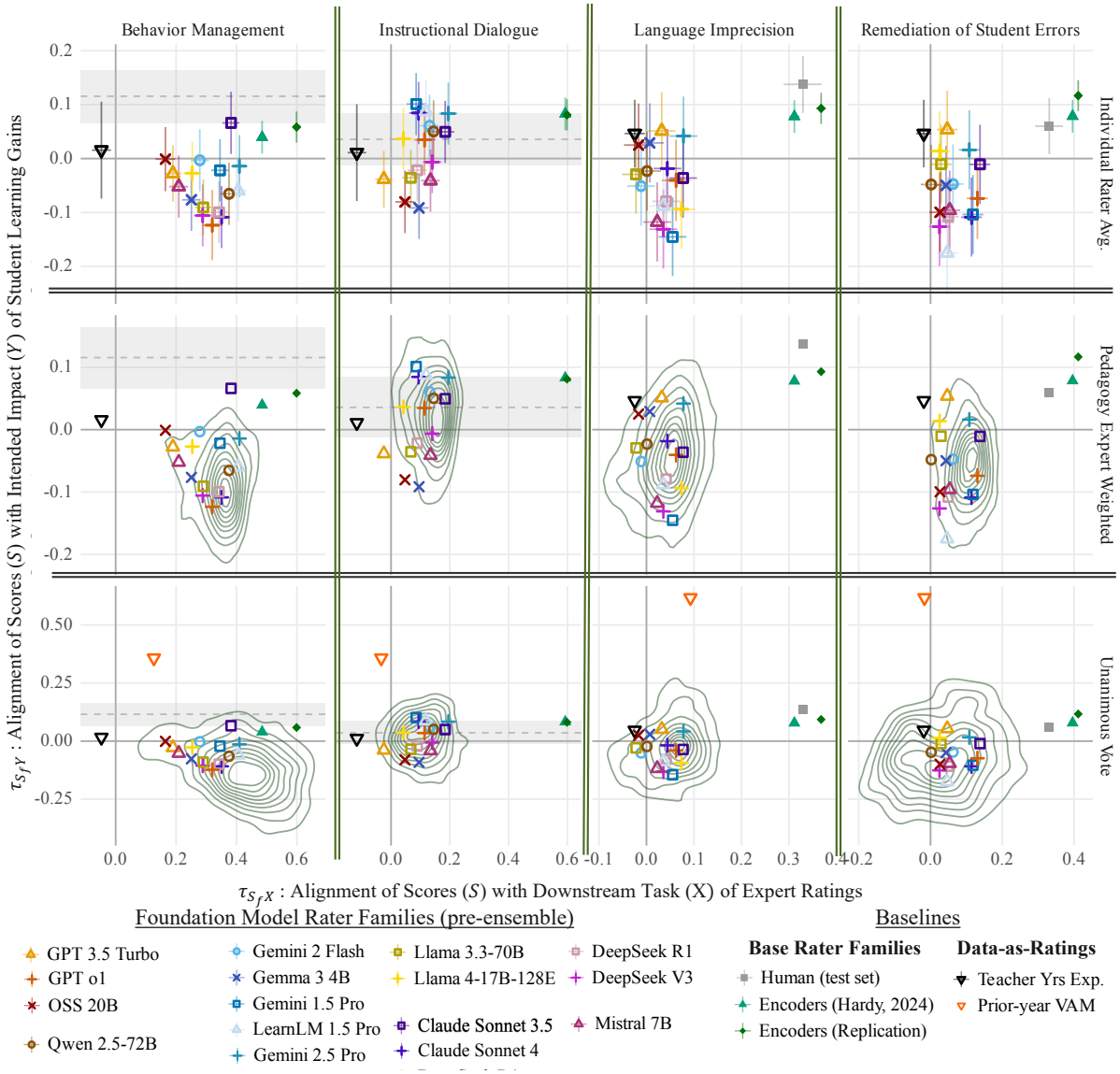


Figure 4: **(Mis)alignment with Downstream Task (Teaching) and Intended Impact (Learning)**: The **x-axes** measure the alignment of scores (S_f) from each FM, ensemble, and baseline f with expert human ratings on downstream tasks (X) on the quality of teaching skills in a given lesson: $\tau_{S_f X}$. 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)** for these models, we only consider for measurement of alignment observations where all three models are in agreement. 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. **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 FMs, and they merely serve as a baseline for possible alignment based on transcripts. The **baseline of Teacher Experience** represent pairwise comparisons that always put a higher value for more experienced teachers. 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.

ity of the error is attributable to factors outside of specific FMs and prompts and how pretraining shared among all FMs collective responds to a transcript. The implication is profound: if this shared bias is a structural feature inherited from pre-training, it is unlikely to be resolved simply by scaling up current architectures or training on more of the same data. The challenge of predicting and steering these inherent biases remains an open and significant problem (Jones et al., 2025). Our results suggest that, for the domain of classroom instruction, this shared bias is not neutral; it is systematically misaligned with the goals of effective pedagogy. These correlated errors across diverse providers mirrors similar findings in scientific discovery, concluding that "frontier models are approaching a performance plateau likely imposed by similar pre-training data distributions rather than distinct architectural limitations." (Song et al., 2025). This reinforces our central conclusion: the gap between FM judgments and the goals of education of children is not a surface-level issue to be solved with system prompting, but rather a deep, structural problem demanding a more fundamental rethinking of how these models are built, trained, and evaluated for educational applications.

6 Conclusion

Our findings reveal a fundamental and systemic misalignment between the evaluative judgments of foundation models and the real-world outcomes of classroom instruction. Across a suite of leading models and prompting techniques, we observed three critical patterns: (1) a strong behavioral convergence among models, suggesting a shared underlying bias; (2) a disconnect between alignment with expert human ratings and a perilous misalignment with student learning gains; and (3) a failure of common ensemble methods to mitigate, and in some cases, an exacerbation of this misalignment. Here, we interpret these findings and discuss their profound implications for the use of AI in classroom education (see 6 Ethical Considerations).

In this OOD, high-stakes domain, foundation models exhibit strong internal consensus yet weak—and often negative—association with the intended impact metric. The pattern closely parallels recent findings in LLM-assisted scientific discovery: systems can accumulate and deploy vast amounts of domain-relevant language while failing to reliably support the downstream goals that language is meant to serve. We demonstrate that the shared architectural and data lineage of today's FMs leads them to converge on a view of teaching that is not only disconnected from expert human judgment but is, on average, negatively correlated with student learning. Furthermore, we show that common strategies for improving AI performance, such as ensembling and expert weighting, can perversely amplify this misalignment.

It is important to note that our findings do not prove that LLMs could never fulfill instructional evaluation tasks. Rather, they show that current simple prompting strategies are insufficient; the models' vast pretraining has imparted biases and tendencies that are not conducive to helpfully contextualizing real-world classroom discourse. Out-of-distribution texts, such as authentic transcripts from elementary school classrooms, appear beyond the scope of what these models can reliably interpret using conventional prompts or chain-of-thought instructions. Our data show that LLM judgments on transcripts, on average, tend to remain invariant to differences in prompting–and even differences in instructional tasks. The findings place the burden on those using FMs to build classroom tools to demonstrate that they have adequately addressed known shortcomings.

As training data and model size increase, more "emergent" capabilities may be identified (Diaz and Madaio, 2024; Ruan et al., 2024; Wei et al., 2022; Schaeffer et al., 2023) with performance tracked across AI benchmarks (Liang et al., 2023; Spangher et al., 2024). However, for some tasks, increased pretraining can worsen performance on downstream tasks if there is misalignment between the pretraining data and downstream tasks (Isik et al., 2025). Although it is unclear how much pretraining would be sufficient, the protected educational data of children will not be readily available for such pursuits.

The challenge is not one of individual model failure but of a systemic bias rooted in current autoregressive training regimes. Moving forward, the development and deployment of AI in education must shift its focus from performance on proxy benchmarks to a rigorous, evidence-based evaluation of its alignment with the ultimate goal of improving student outcomes. Without this fundamental shift, we risk deploying technologies that, despite their sophistication, are systemically misaligned with the very purpose of education.

Limitations

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, 2024, 2025a). The test set used (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 Appendix E, 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.

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.

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 of school-age children, is effectively non-existent on the Internet and therefore absent in the pretraining data of large LLMs.[4] Studies about the qualit of teaching and learning are expensive (Grissom et al., 2013; Liu and Cohen, 2021; Jurenka et al., 2024), with only two major studies with 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 that the call to create more datasets found in many LLM-as-scientist studies (Song et al., 2025; Alampara et al., 2025; Mirza et al., 2025)

Ethical considerations

6.1 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

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

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. Illustrating these continua, 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, remediation 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 (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 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 edu-

cational 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 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 learning K12 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 idea generation that could boost the performance of top experts. 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. The remainder of these Ethical Considerations for using LLMs in classroom contexts will be continued in Appendix D.

Acknowledgments

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. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19815.

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.

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. Publisher: Nature Publishing Group.

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.

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

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. Publisher: Routledge _eprint: https://doi.org/10.1080/10627197.2017.1309274.

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. 2001. *Generalizability Theory*. Springer, New York, NY.

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

Paul-Christian Bürkner. 2021. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, 100:1–54.

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.

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].

Dorottya Demszky and Heather Hill. 2022. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. Publisher: arXiv Version Number: 1.

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.

Fernando Diaz and Michael Madaio. 2024. Scaling Laws Do Not Scale. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:341–357.

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.

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

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. *arXiv.org*.

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.

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. Publisher: American Educational Research Association.

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. Publisher: Taylor & Francis, Ltd.

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. Publisher: Routledge _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. Publisher: American Educational Research Association.

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].

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. 2025. Scaling Laws for Downstream Task Performance in Machine Translation. *arXiv preprint*. ArXiv:2402.04177 [cs].

Erik Jones, Meg Tong, Jesse Mu, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma. 2025. Forecasting Rare Language Model Behaviors. *arXiv preprint*. ArXiv:2502.16797 [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. Publisher: Routledge _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.

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. Publisher: American Educational Research Association.

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].

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].

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].

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. Publisher: American Educational Research Association.

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. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12061.

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 Publisher: American Psychological Association.

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. Publisher: Nature Publishing Group.

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].

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. Publisher: American Educational Research Association.

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.

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. Publisher: SAGE Publications Ltd.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational Scaling Laws and the Predictability of Language Model Performance. *arXiv preprint*. ArXiv:2405.10938 [cs].

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.

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

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

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.

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. _eprint: https://ila.onlinelibrary.wiley.com/doi/pdf/10.1002/rrq.357.

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].

Lucas Spangher, Tianle Li, William F. Arnold, Nick Masiewicki, Xerxes Dotiwalla, Rama Parusmathi, Peter Grabowski, Eugene Ie, and Dan Gruhl. 2024. Project MPG: towards a generalized performance benchmark for LLM capabilities. *arXiv preprint*. ArXiv:2410.22368 [cs].

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. Publisher: [Wiley, International Reading Association].

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. Publisher: Nature Publishing Group.

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. Publisher: Institute of Mathematical Statistics.

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].

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

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].

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *arXiv preprint*. ArXiv:2206.07682 [cs].

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.

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. Publisher: Frontiers.

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. Publisher: Nature Publishing Group.

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. Event-place: Copenhagen, Denmark.

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 Observation Instruments

For each of the observation instruments, the abbreviation codes used in this study are listed with the expanded names in Table 1. The distributions of scores across all items for all rater families are in Figure **??**. 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.

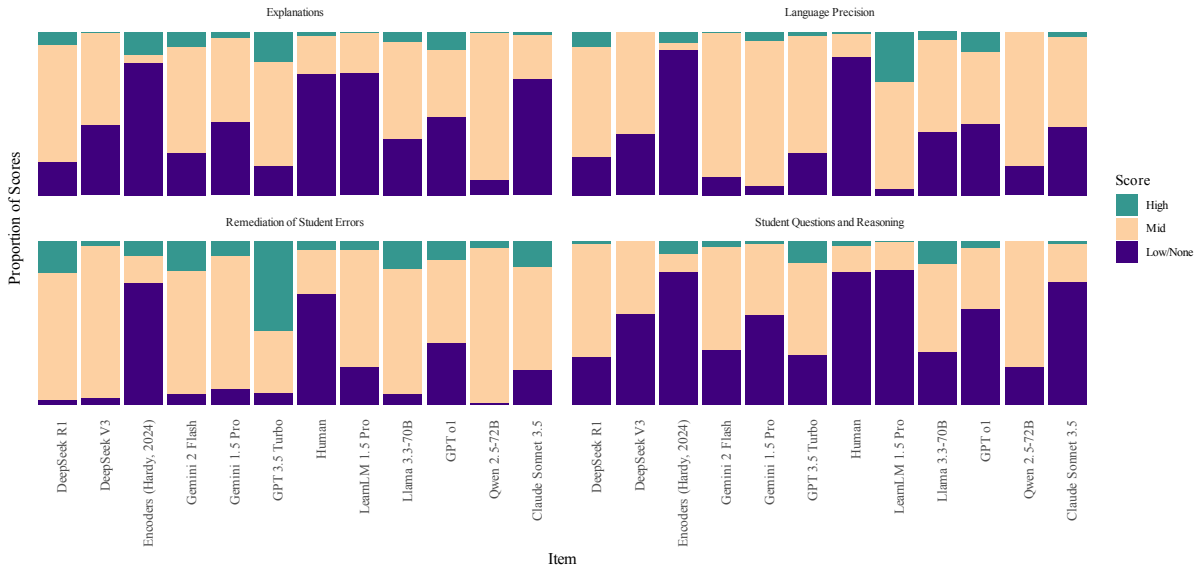


Figure 5: Proportions of rater scores by item.

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, 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 6. 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).

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

A.0.1 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.

A.0.2 Models and Prompts

We selected 13 of the FMs 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). 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 model predictions are publicly available to support future research.[6] for reproducibility. 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.

B Variance Decomposition Model

The variance decomposition model was fit using the `brms` package (Bürkner, 2021) in R with weakly informative default priors. The model converged with all $\hat{R} \leq 1.075$.

Each term in the model represents a source of variance: μ: The grand mean of the misalignment error; $\alpha_c \sim N(0, \sigma_C^2)$: The random effect of classroom c; $\beta_i \sim N(0, \sigma_I^2)$: The random effect of instructional item i; $\gamma_m \sim N(0, \sigma_M^2)$: The random effect of model m; $\delta_p \sim N(0, \sigma_P^2)$: The random effect of prompt p; Subsequent terms, e.g., $(\alpha\beta)_{ci} \sim N(0, \sigma_{CI}^2)$, represent the two-way, three-way, and four-way interaction effects; $\varepsilon_{cimp} \sim N(0, \sigma_{res}^2)$: The final unexplained residual variance. By calculating the proportion of total variance accounted for by each component (e.g., $\hat{\sigma}_M^2 / \hat{\sigma}_{total}^2$), we can determine the relative contribution of each factor to the overall misalignment between FM ratings and student learning outcomes. This is illustrated in Figure 7.

[6] https://drive.google.com/file/d/1fP7xyKasJ4Ui6di-SlY3TLdNQCeg59Tr/view?usp=sharing

C Measures and Metrics

C.1 Distance Correlation

To assess the extent to which LLMs respond similarly on tasks in the presence of out-of-distribution classroom text, we employ model-item distance correlation tests, using the bias corrected squared distance correlation dCor_n^2. The distance correlation is robust to ties, supports discrete values, and can capture non-linear dependencies (Szekely and Rizzo, 2014; Székely et al., 2007).

We quantify inter-rater agreement using bias-corrected squared distance correlation dCor_n^2 (Székely et al., 2007), which captures both linear and nonlinear dependence while controlling for finite-sample bias. For random variables U and V with finite first moments, the population distance covariance is defined as:

$$\begin{aligned}\mathrm{dCov}^2(U,V) &= \mathbb{E}[\|U-U'\|\|V-V'\|] \\ &\quad + \mathbb{E}[\|U-U'\|]\mathbb{E}[\|V-V'\|] \\ &\quad - 2\mathbb{E}[\|U-U'\| \cdot \|V-V''\|]\end{aligned}$$

where U', U'' are independent copies of U and V', V'' are independent copies of V. The squared distance correlation is then:

$$\mathrm{dCor}^2(U,V) = \frac{\mathrm{dCov}^2(U,V)}{\sqrt{\mathrm{dCov}^2(U,U) \cdot \mathrm{dCov}^2(V,V)}} \quad (2)$$

with dCor_n^2 denoting the bias-corrected sample estimator, which is constructed similarly, except using the population covariance construction.

C.2 Disattenuated Stacked Correlations for Underlying VAM

By stacking both VAM measures (STA and ALT in Eq. 3) (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).

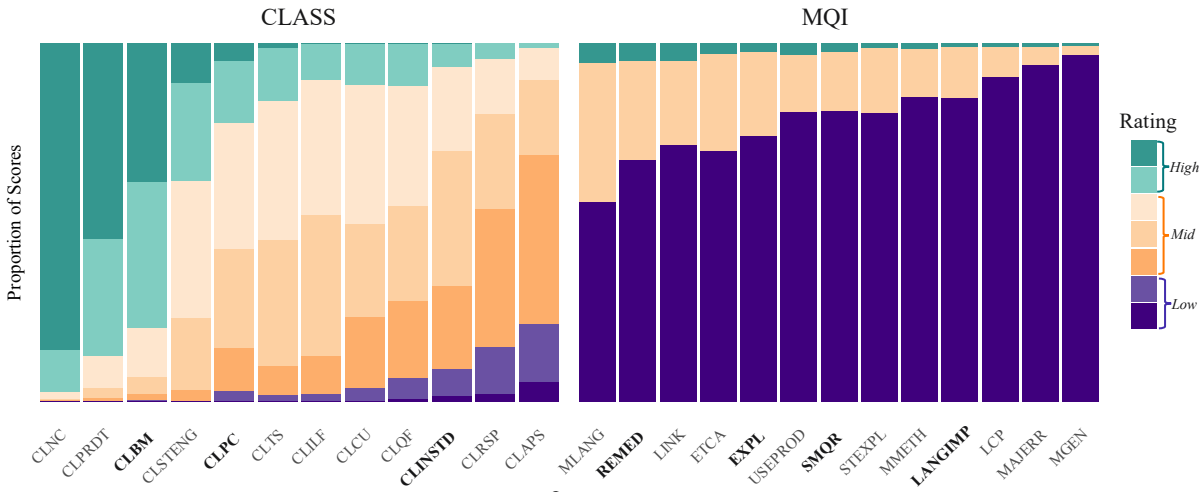
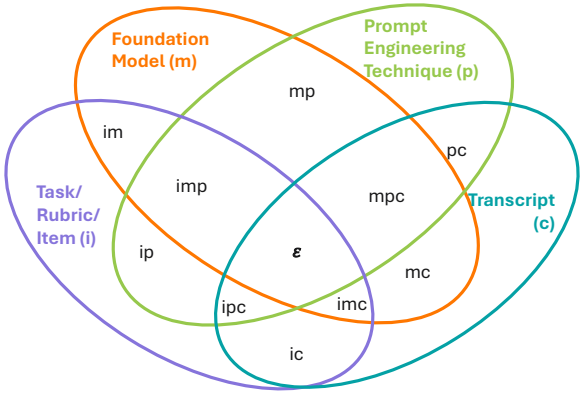


Figure 6: Proportions of human rater scores by item.



I×M×P×C Bayesian Variance Design

```
brms(formula = residVAM ~
(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=25)
```

component	var	pct
I	1.1512	34.7847
CxI	0.2789	8.4257
M	0.1585	4.7891
MxI	0.1405	4.2462
MxCxI	0.3895	11.7674
MxC	0.0296	0.8957
PxM	0.0466	1.4088
PxMxI	0.1541	4.6560
PxMxC	0.1368	4.1336
C	0.2176	6.5748
P	0.0333	1.0050
PxI	0.0056	0.1693
PxCxI	0.0811	2.4493
PxC	0.0001	0.0040
residual	0.4862	14.6903

Figure 7: **Bayesian Variance Decomposition**: (*left*) fully crossed facet diagram for sources of variance. (*middle*) corresponding `brms` code listing. (*right*) table of estimated variance components and their percentage from the total.

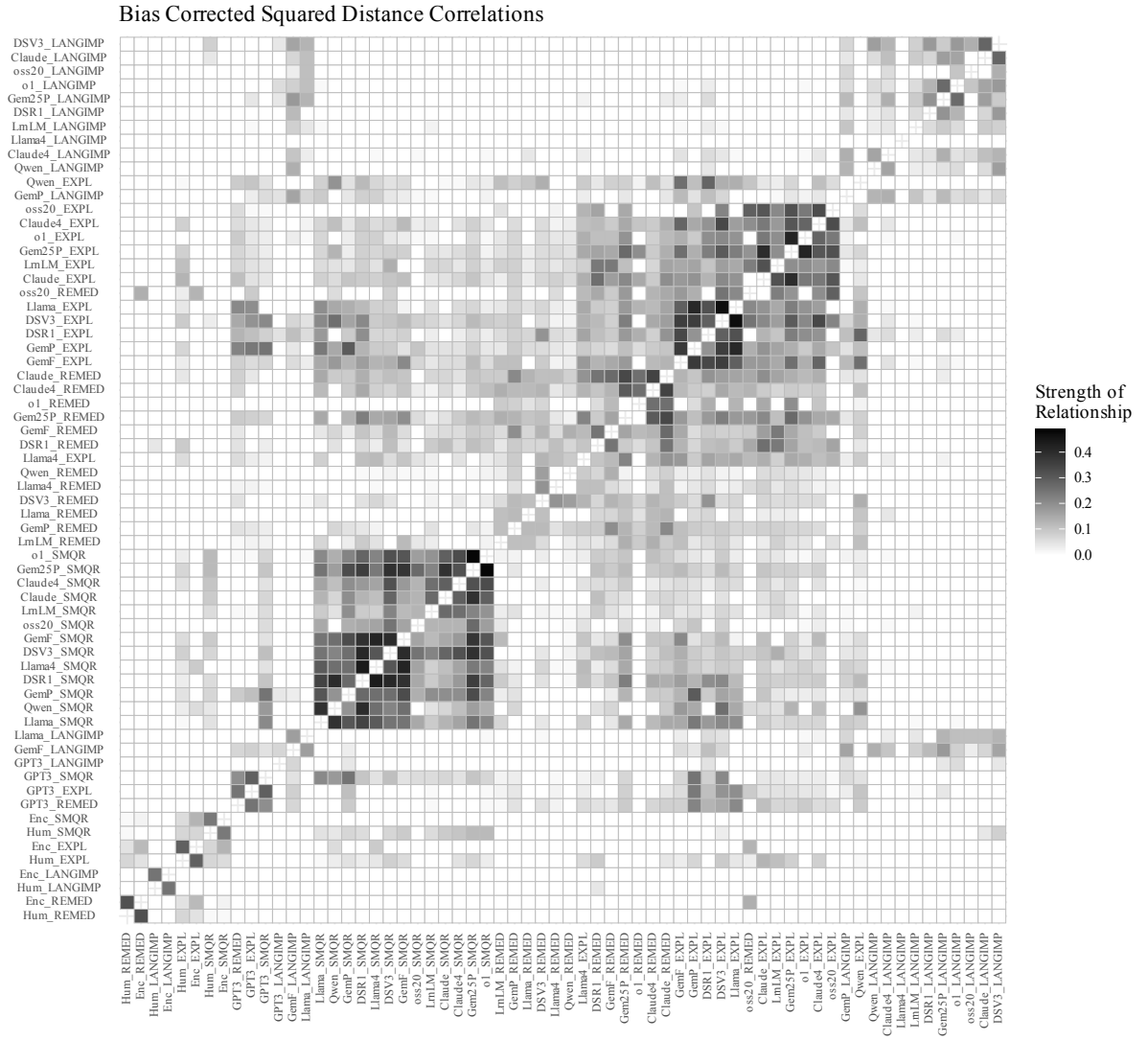


Figure 8: **Between and Within LLM Distance Correlations: MQI**, distance correlations nonparametric measure of dependence between and within rater families across MQI items. Correlation are conducted at the item-transcript level using pairwise-complete observations. Nonsignificant relationships (at $\alpha < 0.05$) are shown as blank after adjusting for family-wise error rate using the Bonferroni correction. Hierarchical clustering is done using complete linkages.

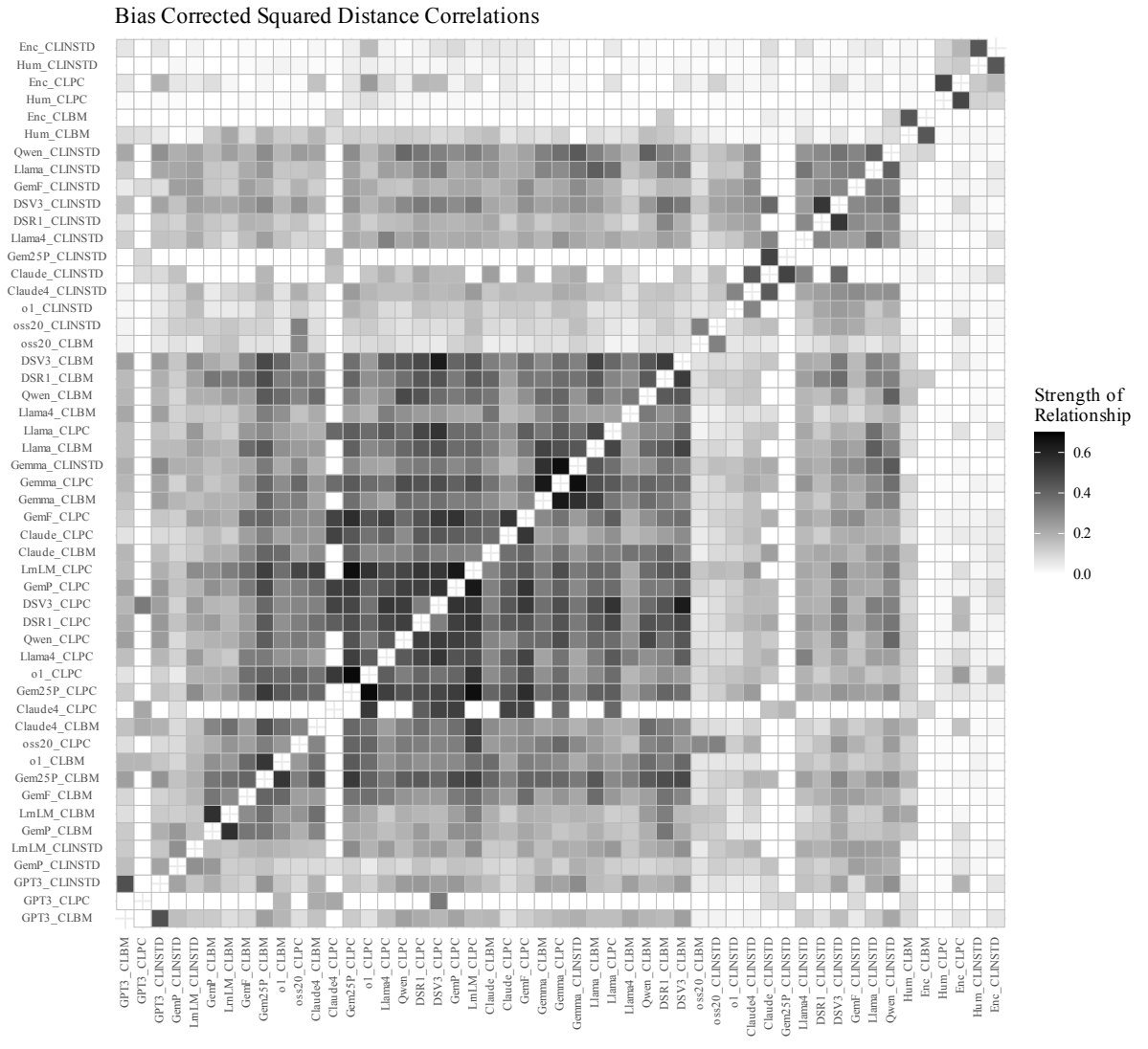


Figure 9: **Between and Within LLM Distance Correlations: CLASS**, distance correlations nonparametric measure of dependence between and within rater families across CLASS items. Correlation are conducted at the item-transcript level using pairwise-complete observations. Nonsignificant relationships (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.

Table 1: 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 identified by Blazar et al. in Appendix 2.b of the original study (Kane et al., 2015)

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.

$$\tau_{S_j,Y} = \frac{\tau_{S_j\tilde{Y}}}{\sqrt{\tau_{\tilde{Y}_{STA},\tilde{Y}_{ALT}}}} \tag{3}$$

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 additionally utilizing Greiner's equality,

$$\sin\left(\frac{\pi}{2}E[\tau_A]\right)$$

(Greiner, 1909) when performing this transformation.

C.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, 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:

$$S_{\text{weighted},tj} = \frac{\sum_{k=1}^{K} w_k \cdot S_{f_k,tj}}{\sum_{k=1}^{K} w_k} \tag{4}$$

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.

D Extended Ethical Considerations

D.1 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 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 not built for a handful of specific teachers with hyperlocalized expertise. In their study on evaluating the quality of instruction, for raters being exposed to unfamiliar teachers, Ho and Kane found that school administrators were more discerning of differences in teaching quality and produced ratings that were 12% and 25% more reliable compared to other teachers with similar and different teaching certifications, respectively to the instruction observed (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, nor 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). 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, the assumed pedagogical capacity lost from removing the teacher from parts of the design, 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.

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 the maximize needed subject expertise for the assumptions of the solution across the target audiences and contexts of generalization of the downstream task. 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.

E Alternate Methods for Measuring Alignment

E.1 Teacher Skill Model BLUPs

The teacher skill model described in Section **??** 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:

$$\begin{aligned} \mathrm{S}_{rjsli} = \mu + \nu_i + \nu_r + \nu_j + \nu_{l:i} + \nu_{s:l:i} + \\ \nu_{ji} + \nu_{jl:i} + \nu_{js:l:i} + \nu_{ri} + \nu_{rl:i} + \nu_{rs:l:i} \\ + \nu_{rji} + \nu_{rjl:i} + \nu_{rj} + \epsilon_{rjs:l:i} \end{aligned} \quad (5)$$

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. [7]
- 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

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

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.

E.1.1 Code Listing

Below is the code used for the estimation of the model defined by Equation 5 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.

```
lmer(SCORE ~ (1 | NCTETID)  + (1 | ITEM)
+ (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"))
```

Listing 1: Target Teaching Skill Estimation Model Code

E.1.2 Parameter Estimates

Parameter estimates for the model are in Table 2.

E.2 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.

F Refined VAM Residualization Model

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 rep-

Table 2: Random 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	–	–

Item	Estimate (CI)	N	p.value
All	-0.156 (-0.205, -0.107)	6299.86	0
CLBM	-0.356 (-0.426, -0.287)	3059.80	0
CLINSTD	-0.139 (-0.21, -0.069)	3049.52	0
CLPC	-0.141 (-0.212, -0.071)	3080.57	0
EXPL	-0.212 (-0.261, -0.162)	6216.73	0
LANGIMP	-0.181 (-0.23, -0.131)	6065.87	0
REMED	-0.233 (-0.282, -0.183)	6230.79	0
SMQR	-0.124 (-0.174, -0.074)	6218.39	0

resentativeness 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.

F.1 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.

F.1.1 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.

F.1.2 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, 5) 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 5 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`.

Listing 2: Estimation Model Specification

```
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.
```

F.2 Residualized VAM model parameters

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

G 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.

Table 4: Measures of fit for Residualization

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