Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1385

Leveraging Large Language Models to Assess Short Text Responses

Jacob M. Rubin
Vanderbilt University

Jason A. Grissom
Vanderbilt University

Educational practitioners and researchers often score short, unstructured text for the presence or strength of domain-specific constructs. Manual scoring, however, faces limitations, including time- and labor-intensiveness. Large language models (LLMs) offer an automated alternative to manual scoring, yet questions remain regarding LLM implementation and performance when scoring text requires domain-specific knowledge. Drawing from two assessments of aspiring principals' teacher-hiring capacities, this study demonstrates a four-stage workflow for implementing LLM-generated scoring of open-ended text while evaluating six LLMs across three prompting methods. Models with higher performance on language comprehension benchmarks and more detailed prompting methods reduced scoring variability and demonstrated strong alignment to trained human scorers. Further, we highlight key design considerations, including how many LLM scoring iterations are necessary, how many entries must be scored manually for precise estimates of consistency, and checks for algorithmic bias.

VERSION: January 2026

Suggested citation: Rubin, Jacob M., and Jason A. Grissom. (2026). Leveraging Large Language Models to Assess Short Text Responses. (EdWorkingPaper: 26-1385). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/07w3-by46

Leveraging Large Language Models (LLMs) to Assess Short Text Responses

Jacob M. Rubin

Jason A. Grissom

Vanderbilt University

Abstract

Educational practitioners and researchers often score short, unstructured text for the presence or strength of domain-specific constructs. Manual scoring, however, faces limitations, including time- and labor-intensiveness. Large language models (LLMs) offer an automated alternative to manual scoring, yet questions remain regarding LLM implementation and performance when scoring text requires domain-specific knowledge. Drawing from two assessments of aspiring principals' teacher-hiring capacities, this study demonstrates a four-stage workflow for implementing LLM-generated scoring of open-ended text while evaluating six LLMs across three prompting methods. Models with higher performance on language comprehension benchmarks and more detailed prompting methods reduced scoring variability and demonstrated strong alignment to trained human scorers. Further, we highlight key design considerations including the number of scoring iterations, manual scoring for precise estimates of consistency, and validity checks for algorithmic bias.

Introduction

In education, scoring short, unstructured text using domain-specific knowledge is commonplace for both practitioners and researchers. A high school social studies teacher, for example, might use a rubric to grade short-answer questions about the American Revolution to assess how well students integrate evidence from multiple sources. A school principal might review teacher lesson plans to evaluate whether plans align with grade-level curriculum standards. A researcher may code transcribed interview data to explore teachers' sensemaking about new instructional materials. However, manually scoring text or artifactual evidence for domain-specific content presents challenges. It becomes more time-intensive and financially expensive as the text length or the number of responses to evaluate increases. It requires scorers to have substantive knowledge, requiring costly training in some settings (Krippendorff, 2018). It can be (too) cognitively demanding even when few constructs are assessed in each text (Saldana, 2021). Cognitive demand also grows with the number of texts, which, aside from causing scoring fatigue, may lead to "drift" in scoring (Leckie & Baird, 2011). Maintaining scoring consistency requires continuous recalibration, often impractical for practitioners and time-consuming for researchers.

Large language models (LLMs) offer a potential solution to the challenges of manual human coding. LLMs operate as next-word prediction systems, estimating the probability of a given text segment (referred to as a token) occurring given the preceding text. In the context of scoring, LLMs may predict an appropriate score for a text (e.g., 1 to 5) by drawing on patterns learned from training on large datasets. LLMs may allow scaling manual scoring efforts while reducing cost, cognitive overload, and inconsistency (Gilardi et al., 2023; Liu et al., 2025; Parker et al., 2024). Recent advances have markedly improved their ability to perform a range of text analysis tasks, including content labeling, document summarization, and language generation (Xing et al., 2025; Ziems et al., 2024). As a result, researchers are rapidly adopting LLMs to streamline text analysis tasks such as analyzing end-of-course survey comments (Parker et al., 2024) and evaluating feedback given to pre-service teachers

(Avitabile et al., 2025; R. Wang & Demszky, 2023).

Recent studies find that LLMs may complete text analysis tasks like data labeling and classification with accuracy comparable to or exceeding trained human scorers and crowdsourcing platforms like Amazon Mechanical Turk (Gilardi et al., 2023). However, these findings are highly dependent on LLM implementation decisions and measurement contexts. Researchers have demonstrated that variation in model type and version (Mellon et al., 2024), prompting method and level of detail (Brown et al., 2020; Gilardi et al., 2023; Kirsten et al., 2024; Liu et al., 2025; Parker et al., 2024; Ruckdeschel, 2025), and properties of the constructs being measured (Atreja et al., 2024; Kirsten et al., 2024; Liu et al., 2025) can influence the reliability of LLM-generated scores.

In other words, while LLMs can scale scoring efforts, their performance and reliability can vary with implementation decisions and measurement contexts. Maximizing LLM scoring usefulness requires establishing systematic workflows that evaluate LLM-generated scores' variation and reliability across model type, prompting method, and construct properties. Education research has few resources for setting up such workflows to ensure LLM scoring can be used productively and assessed responsibly.

This article lays out such a workflow for generating and evaluating LLM scores across model types, prompting methods, and construct domains. To illustrate it, we analyze two forms of open text from a study of aspiring principals' capacities for teacher hiring: (1) written text from a short-response item and (2) transcribed response text from a structured interview question. Each was scored by multiple human scorers using rubrics designed to measure different facets of subjects' approaches to hiring. We then replicated rubric-based coding with six widely available LLMs. In this process, we asked: (1) To what extent can LLMs be considered feasible and reliable tools for scoring domain-specific content from short text data? (2) What design considerations should practitioners and researchers attend to when implementing LLM-generated scoring? In answering these questions, we structured a workflow that others in the field employing LLMs might follow in other work. We also

highlighted key considerations for practitioners and researchers implementing LLMs to score open-ended assessments beyond this empirical case.

The remainder of this article unfolds as follows. First, we describe the research context and data for this study. Second, we lay out the workflow we used—and, we argue, that others can use—to generate LLM scores and analyze their outputs. Throughout our discussion of this process, we describe its application to the case study of aspiring school leaders, including what we found regarding the comparison of human and LLM scoring. Third, we summarize key design considerations for researchers implementing LLM scoring that emerged from our analysis. We conclude with implications, limitations, and ideas for future research.

Data

Research Setting

This research is part of a multi-year study examining the leadership capacities of aspiring school principals. For the broader study, researchers recruited aspiring principals in Tennessee[1] to take a series of diagnostics to measure pre-service skills across domains associated with effective early-career school leadership (Grissom et al., 2021). These diagnostics were administered via a series of surveys that included a mix of closed- and open-ended response items. Surveys included brief job performance tasks designed to simulate a scenario the leader might encounter on-the-job, such as conducting mock teacher observations and then providing feedback. Such performance tasks are increasingly common in leadership assessments (e.g., Orr and Hollingworth, 2018). In addition to survey-based diagnostics, participants took part in a structured 30-minute interview with a member of the research team. Interviews were conducted via Zoom and transcribed. We make use of data from

[1]Aspiring leaders were recruited to participate via two channels. First, researchers reached out to district superintendents asking them to identify potential future principals in their districts. Second, aspiring leaders self-nominated through the Tennessee Educator Survey, a statewide survey of teachers and school administrators conducted by the Tennessee Department of Education and Tennessee Education Research Alliance each spring. In both cases, researchers followed up with potential participants to describe the study and obtain their consent. Participants who completed all diagnostics received a small honorarium in recognition of their time.

200 aspiring principals from 54 school districts and charter networks, representing 36% of all Tennessee districts.[2] Aspiring principals in the sample were, on average, 44 years old with 15 years of experience in Tennessee public schools. Roughly 87% were White, 9% Black, and 68% female. Thirty-three percent were located in urban districts, 34% in suburban districts, and 33% in town/rural districts. For more information regarding participant characteristics, see Appendix A.

Case: Measuring Aspiring Leaders' Capacities for Teacher Hiring

From the broader study, we focus on responses to two data elements designed to elicit aspiring leaders' approaches to teacher hiring. Teacher hiring is a key area of principals' human capital management responsibilities (Cannata et al., 2017; Donaldson, 2013) and one new principals are increasingly likely to encounter as hiring becomes less centralized at the district level (Engel et al., 2018). It is thus worthwhile to understand soon-to-be principals' expertise and conceptualization of leaders' hiring roles. First is a survey short-response item:

> You are considering two candidates for an open teaching position. Candidate 1 has an exceptional record of student achievement growth at their prior school but has received complaints from parents for being too strict with student discipline. Candidate 2 is well-liked by parents and students at their prior school but has a weaker record of student achievement growth. What do you do?

By varying the candidates' relational and instructional skills, the item aims to distinguish what aspiring principals might trade off when hiring teachers and how they use evidence and experience to support their decision. We obtained written responses to this prompt from 188 aspiring principals. Second are responses to a structured interview question:

> A key goal for every school is to hire effective teachers for every vacancy that comes up. How do you go about meeting that goal?

This question aims to elicit leaders' decision-making processes and strategies for hiring teach-

[2]This dataset is considerably smaller than those used in recent studies of LLM-based scoring, which typically involve thousands of text responses (Kim et al., 2025; Parker et al., 2024). In this sense, we demonstrate the efficacy of LLM scoring using a dataset size that is more reasonable for an educational leadership setting.

ers. We analyzed responses from 124 aspiring leaders to this interview prompt.[3]

Beyond their substantive relevance, responses to these items have technical relevance for assessing the feasibility of LLM scoring. The prompts generate open-ended responses of varying length both across and within items. Answers to the short-response item, by design, were briefer, with an average of 49 words (SD = 24), compared to the interview responses, which averaged 253 words (SD = 126). Also, given that scoring rubrics for these data tasks share several items, we can examine how LLMs respond to the same criteria across contexts and responses of varying length. Further, as described later, responses were scored on a series of rubric dimensions that vary by measurement scale, frequency, and complexity. Such variation extends generalizability towards other real-world scoring contexts. Sample responses from both response types appear in Appendix B.

A Procedure for Generating and Analyzing LLM Scores

Exploring the viability of LLM scoring of our two text types led us to a four-stage workflow that this section describes and illustrates. We adapted our approach from Anglin et al.'s (2025) and Halterman and Keith's (2025) frameworks for LLM measurement with rubrics. Our workflow consists of the following stages: (1) prepare rubrics for human scorers and LLMs, (2) split the dataset and test LLMs' basic capabilities, (3) conduct LLM scoring with multiple models and prompting methods, and (4) evaluate LLMs' *compliance* with the rubric, *variation* in scoring, *consistency* with trained human scorers, and self-reported *uncertainty*. The rest of this section describes each stage.

Stage 1: Prepare Rubrics for Human Scorers and LLMs

We began by constructing detailed rubrics, grounded in prior literature and refined through multiple rounds of human review, to guide human scorers and LLMs. During this

[3]The interview typically was the last diagnostic task participants took part in for the study. Only a subset of participants made it all the way to the end of the assessment set, which is why the sample size for the interview sample is lower.

process, researchers established common definitions for constructs that may differ from an LLM's pretraining data (Ruckdeschel, 2025). For instance, even basic terms such as "leadership" carry a distinct meaning in educational contexts. Establishing content-specific rubrics ensures that scoring reflects the context's conceptualization of constructs rather than generic language model priors.

Given we are interested in the same fundamental question (how aspiring principals approach teacher hiring) for both the survey short-response and interview, rubrics share the same overarching constructs. Rubrics were defined by three sets of items:

1. **Reasoning**. For the survey short-response rubric, a single item asked scorers to "Rate how strong and convincing the reasoning is" on a 5-point Likert scale. For the interview rubric, reasoning was operationalized into four distinct items: intensity of recruiting strategies, engagement with evidence, collaborative decision-making, and responsiveness to local labor conditions. Each was scored on a 4-point Likert scale.

2. **Candidate Preference**. For the survey short-response, rubric items asked scorers to "Rate the strength of preference for Candidate 1/Candidate 2/an alternative response" on a 3-point Likert scale. These items only appeared on the survey short-response rubric since respondents were provided predefined options. Options were mutually exclusive; a scorer could only indicate preference for one option.

3. **Hiring Priorities**. For both the survey short-response and interview rubrics, hiring priorities items asked whether respondents appeared to prioritize which, if any, of the following six factors: *cultural responsiveness*, *parent/community engagement*, *academic achievement*, *candidate experience/expertise*, *evaluation*, and/or *school culture fit*. The development of these factors was informed by prior theory and exploratory analysis of data from other assessments in the broader study. See Appendix C for more information on the factor development process and full definitions.

Beyond capturing different theoretical constructs of how aspiring principals approached hir-

ing, rubric items also test distinct tasks relevant to LLM scoring. The set of *hiring priorities* items represent a binary classification task, in which the scorer determines whether a given construct is present (1) or absent (0). Scoring *candidate preference* items constitutes a multiclass classification task, in which the scorer selects a single preference among a mutually exclusive set of options (candidate 1, candidate 2, or an alternative response). Assessing *reasoning* represents an ordinal classification task, in which scorers rate responses along a ranked but discrete 4- or 5-point Likert scale. Aligning with Halterman and Keith (2025), the finalized rubrics that human scorers and LLMs accessed included a label, instructions, and examples for each item. In addition, we created step-by-step rationales illustrating why a given score would be assigned. Depending on the prompt design, the LLM could receive the full rubric entry (i.e., label, instructions, examples, rationales), or a simplified version containing only instructions and/or examples. Prompt design and testing is further discussed in Stage 3. Full rubrics are provided in Appendix C.

Stage 2: Split the Dataset and Test LLMs' Basic Capabilities

Following general guidance for supervised machine learning methods, we split the dataset of responses into *training* and *testing* subsets. Splitting the data ensures that models do not see the full dataset prior to scoring, potentially biasing score estimates. Like Anglin et al. (2025), we allocated 25% of each dataset for training.

The training subset was used for two purposes. First was designing and refining prompt text. We checked whether LLMs can follow instructions (e.g., "What are the labels of the rubric items to score?") and evaluated sensitivity to slight changes in the prompt text. We also tested different approaches for dividing responses into smaller sections to comply with model input limits. Second, the training set was used to determine a minimum number of scoring iterations that balances precision of score estimates with capturing sufficient variation in model scores (we discuss this process more later). By running the same text through LLMs multiple times, we seek to capture the internal reliability of each LLMs' scoring and generate

a distribution of scores. Once the prompt text and number of scoring iterations was finalized, the testing dataset was used in Stages 3 and 4 to conduct and evaluate LLM scoring.[4]

Stage 3: Conduct LLM Scoring with Multiple Models and Prompting Methods

To understand variability in LLM-generated scores, we scored both texts using six LLMs: OpenAI's GPT-5, GPT-4o, and GPT-4.1 mini; Anthropic's Claude Sonnet 4 and Claude 3.5 Haiku; and Amazon's Nova Lite. Models were selected by their variety in cost, speed, and performance on widely adopted language comprehension and complex reasoning benchmarks (see Table 1). To address privacy concerns associated with proprietary models, all text scoring was conducted through a private instance of each LLM offered by the authors' institution.

We scored each assessment using three prompting methods, reflecting research findings that prompting strategies can influence model behavior (Atreja et al., 2024): zero-shot, few-shot, and chain-of-thought. *Zero-shot* prompting refers to the case in which the LLM is provided only a rubric with label definitions and instructions. *Few-shot* refers to the case in which the LLM is additionally provided with examples to illustrate appropriate coding. *Chain-of-thought* (CoT) prompting augments the few-shot case with step-by-step rationales for each example. Several studies demonstrate the viability of zero-shot prompting for some cases (e.g., Kirsten et al., 2024; Parker et al., 2024), while others show that providing LLMs with additional annotated examples can improve model performance with more complex constructs (Brown et al., 2020; Liu et al., 2025).

[4]Notably, Anglin et al. (2025) define a third development (dev) subset used for identifying the best-performing combination of prompting method and sample size prior to scoring the testing subset. This study does not define a development set for several reasons. Development splits generally are used in machine learning for tuning model hyperparameters, an omitted step when using off-the-shelf and proprietary LLMs. While other research does investigate the supervised fine-tuning of open-source model weights (e.g., Halterman and Keith, 2025; Kim et al., 2025) this approach is generally inaccessible to the majority of researchers and practitioners. In addition, the goal of our work is to identify general trends of high-performing combinations of model type and prompting method, rather than a single "best" combination. In this sense, our work is model- and prompt-agnostic, enabling researchers and practitioners to adopt this workflow within their own available model and computational constraints.

Stage 4: Evaluate LLM Scoring: Compliance, Variation, Consistency, and Uncertainty

We evaluated model performance across four dimensions: compliance, variation, consistency, and uncertainty. First, we investigated LLM *compliance* by checking whether outputs included valid responses and complete rubric scores. Second, we assessed *variation* by analyzing score distributions across models and prompting methods. Third, we tested LLM-generated scores' *consistency* with scores generated by human scorers. Finally, we documented *uncertainty* in the models' assignment of scores across constructs. We describe and illustrate each criterion below in the context of our two teacher hiring assessments.

Compliance

To investigate LLMs' compliance with scoring instructions, we documented the degree to which a given model output scores as expected (e.g., within the correct item scale bounds) and was free of "hallucinations," defined as model outputs that directly contradict the prompt or real-world knowledge (e.g., assigning scores to responses that do not exist) (Zhang et al., 2025). Specifically, we calculated the percentage of LLM scoring responses that (1) were correctly formatted, (2) only scored text responses that existed in the data (without generating extra hallucinated content), and (3) scored each item within their respective scale bounds.

Across the three compliance checks for both the survey short-response and interview response, average model performance ranged from 93% to 100% (see Appendix D). LLMs produced correctly formatted scores in 96% of survey short-response scoring attempts and 99% of interview scoring attempts. Only one instance of hallucinating a non-existent response occurred across scoring responses, suggesting strong comprehension of the response data. For the item scale-bounds compliance check, average model performance ranged from 88% to 100%, with a median compliance rate of 100%.

Although no single model consistently outperformed the others, GPT-5, GPT-4o, Claude

Sonnet 4, and Claude 3.5 Haiku models met all compliance checks in approximately 99% of their scoring outputs. The two poorer-performing models, Amazon Nova Lite and GPT-4.1 mini models, still met all requirements in approximately 93% and 97% of their scoring outputs, respectively. Overall, strong compliance suggested that off-the-shelf LLMs reliably followed detailed rubric instructions, supporting their use for scaling scoring efforts.

Variation

To examine variation in LLM scores, we calculated means and standard deviations across model type and prompting method. We compared these statistics to those calculated for human-assigned scores under the assumption that if LLMs can substitute for human scorers, score distributions will be similar.

In Tables 2–3, we present LLM-generated score means and standard deviations from the survey short-response item and interview question by model type, prompting method, and rubric item. With respect to model type, we found that the LLMs that performed better on the language comprehension and complex reasoning benchmarks summarized in Table 1 (namely, GPT-5, GPT-4o, and Claude Sonnet 4) tended to produce scores that more closely mimic the human average. Normalized by item scale, GPT-4o's and GPT-5's mean scores differed from the human average across rubric items by only 4% and 5% of the total scale range, respectively. For a 5-point Likert scale item, these percentages were equivalent to a difference of 0.16 and 0.20 scale points from the human mean. Such LLMs also generated more precise scores, with standard deviations (normalized by item scale) lowest for GPT-5, Claude Sonnet 4, and GPT-4o. These patterns appeared similar for rubric items across and within rubric categories.

Prompting methods that included examples (few-shot and CoT) also yielded scores that more closely aligned with the human average. We observed that moving from a zero-shot approach to a few-shot or CoT approach reduced scores' distance from the human average by roughly 3% of the total scale range, averaged across rubric item categories. The CoT

prompting method, however, offered minimal improvement beyond the few-shot approach. Standard deviations were similar across prompting methods.

Consistency

Key to the feasibility of substituting LLM for human scoring is that LLMs produce scores that are consistent with those provided by human scorers. We measured this consistency via intra-class correlations (ICCs), which represent the degree of variance that is explained by differences in the scores themselves relative to differences between the scorers or random noise. For each rubric item, we compute ICCs (1) between each model–prompt combination mean score and the human mean score, and (2) between human scorers themselves. Higher ICC values reflect stronger inter-rater reliability. For instance, an ICC of 0.79 indicates that 79% of the total variance in scores is explained by true response differences. Comparable ICCs between a model–prompt combination and the human mean, and those observed among human scorers themselves, suggests that the LLM aligns closely with the human scoring norm with regards to inter-rater reliability.[5]

Notably, other commonly used inter-rater reliability (e.g., Cohen's Kappa) and predictive analytics (e.g., accuracy, precision, F1 score) metrics assume discrete categories. These metrics assume perfect agreement and therefore poorly assess continuous variables such as average rubric ratings. By contrast, ICCs account for both agreement and score distance, making them more suitable for this context (Shrout & Fleiss, 1979).

In Table 4, we present intra-class correlations (ICCs) from the survey short-response scores. Results for the interview response (see Appendix E) followed the same patterns.

[5]We calculated ICCs using a two-way random-effects model for single measures, denoted as ICC(2,1). This approach assumes that each response was evaluated by every scorer and that scorers represented a random sample from a larger population of potential scorers (Koo & Li, 2016). The ICC(2,1) was computed as shown in Equation (1):

$$ICC(2,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{1}$$

where MS_R was the mean square for responses, MS_E was the mean square for error, MS_C was the mean square for scorers, k was the number of scorers, and n was the number of responses.

We found that inter-rater reliability varied systematically by rubric category, model type, and prompting method. For rubric category, agreement between LLMs and the human average was stronger for rubric categories with more concrete criteria (e.g., *candidate preference*, *hiring priorities*) versus more subjective constructs (e.g., *reasoning*). For model type, LLMs with better established language comprehension and complex reasoning benchmark capabilities exhibited stronger alignment to human scoring. Specifically, Claude Sonnet 4, GPT-5, and GPT-4o often showed greater alignment with the human mean scores than human scorers did with one another (indicated by the bold values in Table 4). Lastly, prompting methods that included examples (few-shot and CoT) improved inter-rater reliability between LLMs and the human mean. On average, moving from a zero-shot to a few-shot prompt design was associated with an ICC increase of 0.07, and moving from a zero-shot to a CoT prompt design was associated with an ICC increase of 0.10. Notably, these increases were smaller for LLMs with the highest reasoning capabilities. For example, Claude Sonnet 4 and GPT-5 showed minimal gains (0.00–0.03 and 0.03–0.06, respectively). Prompting detail improved inter-rater reliability most for LLMs that performed poorest on the language comprehension and complex reasoning benchmarks (e.g., Amazon Nova Lite).

Uncertainty

To evaluate the degree of *uncertainty* in LLM scoring, we followed guidance from Li et al. (2023) by examining models' self-evaluation and entropy. We elicited self-evaluation by appending the following instruction to the prompt: "Please give a confidence score on a scale of 0 to 1 for each predicted score." In other words, models were directly asked to report their scoring confidence.

We computed entropy, defined as a measure of the impurity in a dataset (Shannon, 1948), from the relative frequencies of each predicted score among all possible options. For instance, if the model always gives the same score across scoring iterations, there is no uncertainty and entropy is zero. If the model scores many different answer choices with roughly equal

frequency, uncertainty is high and entropy approaches one.[6] Whereas reviewing variation in LLM scoring provides information about the spread of scores (especially useful for ordinal Likert-scale items), entropy measures provide information about the consistency of scores.

We found a moderately positive correlation ($r = 0.37$) between LLMs' self-reported confidence scores and the LLM-generated score means across rubric items. In other words, models tended to express greater confidence when assigning higher scores. One possible explanation is that LLMs exhibit tendencies towards "social sycophancy" (Cheng et al., 2025), the inclination to provide more positive and agreeable responses. Alternatively, this pattern may simply reflect poor separation in the confidence scores themselves, with 91% (survey short-response) and 87% (interview) of scores falling in the range of 0.6–1.0.[7]

Finally, we calculated entropy scores normalized by item scale for the survey short-response (see Appendix F). For instance, a normalized entropy score of 0.30 indicates that 30% of the maximum possible uncertainty (or inconsistency) is present in the scoring distribution. Aligned with findings from the variation analysis, more concrete rubric item categories and LLMs that performed higher on language comprehension and complex reasoning benchmarks produced more consistent scores. Similarly, entropy scores varied little by prompting method, indicating that model type is more influential than prompt design for scoring consistency.

[6] Given c possible answer choices, an entropy score u_i was calculated via Equation 2 as:

$$u_i = -\sum_{j=1}^{c} p_j * ln(p_j) \tag{2}$$

where p_j represented the proportion of scoring iterations that a model assigned the jth scoring choice.

[7] Further, we found a weakly negative correlation (r = -0.14) between LLMs' self-reported confidence scores and the LLM-generated score standard deviations across rubric items. This inverse relationship reveals that as models report being more confident about score assignment, their variability of scoring decreases. This pattern appeared to be strongest for LLMs that performed higher on language comprehension and complex reasoning benchmarks (r_{GPT-5} = -0.46, $r_{ClaudeSonnet4}$ = -0.26), and did not differ across prompting methods. This pattern may suggest that these LLMs, when confident, are also more consistent in scoring.

Design and Evaluation Considerations

Through this empirical demonstration of our LLM-based scoring workflow, we found that models with stronger performance on language comprehension and complex reasoning benchmarks produced scores that closely reflected trained human scorers and were more precise than lower-performing models. Additionally, scores from prompting methods that incorporated examples produced scores that closely reflected human scorers, although precision did not vary much by prompting method. Next, we discuss key design and evaluation considerations that emerged during this process.

How many LLM scoring iterations are necessary to ensure scoring precision?

Given the non-deterministic nature of LLMs, LLM-generated scores exhibit some degree of randomness. Running multiple iterations of LLM scoring begins to address this concern, capturing variation in scoring and producing score distributions. To determine an appropriate number of scoring iterations to run for the testing dataset while minimizing computational costliness, we first ran 50 scoring iterations for each model–rubric item combination in the training dataset and calculated the mean score for each (e.g., the mean score for the *reasoning* item produced by GPT-5 was 3.3). We then generated random subsamples of scores ranging in size from 1 to 50. For each subsample size, we computed the root mean squared error (RMSE) relative to the full sample mean, with lower values indicating that a given subsample closely resembles the mean.

Figure 1a illustrates that, for the survey short-response item, RMSE decreased sharply within the first few subsample sizes before plateauing. In other words, as the number of iterations increased, the stability of model scores increased quickly to a point of diminishing returns. The "elbow" of each model's curve is indicated by the point farthest from a normalized straight line connecting the endpoints of the curve (Satopaa et al., 2011). For our particular case, a conservative estimate of approximately 10 scoring iterations was sufficient to obtain a stable estimate of scoring behavior while limiting computational cost. As a re-

sult, we ran 10 scoring iterations for the testing dataset in later analysis. Similar patterns emerged for the interview item.

How many human-scored responses are necessary to achieve a precise estimate of human-model consistency?

A key consideration for LLM scoring is determining the fraction of responses for humans to manually score to ensure a precise estimate of inter-rater reliability with LLM-generated scores while minimizing human effort. In other words, how many responses need to be human-scored to assess with confidence whether LLM-generated scores have strong agreement with human scores?[8]

To investigate, we calculated inter-rater reliability for various subsample sizes of the testing data, beginning with the minimum subsample size needed to calculate stable variance estimates for intraclass correlations (ICCs) (Shrout & Fleiss, 1979). For each subsample size, we drew 50 random samples of that size from the testing data. Then, for each sample, we calculated the intraclass correlation coefficient (ICC) between the human scores and LLM-generated scores. ICCs were averaged across the 50 samples to obtain a mean ICC for that subsample size; mean ICCs were plotted to identify the subsample size for which ICC estimates plateaued.

Figure 1b illustrates that, for the survey short-response rubric categories, subsample sizes ranging from 6% to 15% yielded ICC estimates within 10% of the full-sample ICC value, and subsamples ranging from 8% to 30% yielded within 5%. These findings indicate that only a small fraction of human-scored responses is needed for convergence to a relatively precise estimate of human–model agreement. Similar patterns followed for the interview rubric categories, with subsamples ranging from 6% to 26% for a within-10% value of the

[8]To be clear, this is not a training question but rather an evaluation question. Rather than answer "How many human-scored responses should we train the model with to obtain stronger human-model agreement?", we answer "How many human-scored responses are necessary to obtain accurate estimates of human-model agreement?" With a larger human-scored subsample size, we are not improving inter-rater reliability, but rather generating a more precise estimate of inter-rater reliability.

full-sample ICC and 7% to 35% for a within-5% value (see Appendix G).

How might researchers qualitatively investigate LLM-generated scores?

Adapting guidance from Li et al. (2023), we recommend leveraging scoring variation and uncertainty to choose which LLM-generated scores to qualitatively investigate. In Stage 4 of our workflow, we quantified uncertainty by prompting models to self-report confidence scores and calculating entropy scores. Consistent with Li et al. (2023), we favor using entropy scores to guide re-scoring efforts given that confidence scores poorly separated the data, with 91% and 87% of confidence scores falling between 0.6 and 1.0 for the survey short-response and interview item, respectively. Alongside entropy, we propose that standard deviations from LLM-generated scores can also help determine which responses to manually rescore. Researchers may, contingent on their available time and labor, begin by selecting the n instances with the greatest entropy (to assess consistency) and standard deviation (to assess variability).

How might researchers investigate potential algorithmic bias in LLM-generated scores?

An ongoing concern in LLM-generated scoring is algorithmic bias—that models trained on data containing social biases may replicate or even magnify such biases (Chouldechova & Roth, 2018). In this study, we make no claim that LLM-generated scoring is objective or bias-free. Rather, we examine whether LLM-based scores exacerbate existing social biases relative to the "business-as-usual" case of human scoring.

Following Baker et al. (2023), who recommend that "there are greater benefits to fairness if demographic variables are used to validate fairness rather than as predictors within models," we used participant demographic data in a validation exercise to determine if LLMs differentially score respondents by race/ethnicity and sex, over and above differences in human scoring (p. 22). For each rubric item, we regressed scores on model type (human mean

scores was the reference group), demographic variables of interest (White/Non-White and male/female), and their interaction. We found that, across items, interaction coefficients were generally small and not statistically significant, suggesting that LLMs did not differentially score text by race/ethnicity or sex (relative to humans). Full interaction coefficients appear in Appendix H.

Discussion

This article contributes to a growing body of literature that examines the viability and implementation practices of LLMs to score text data produced from educational assessments. Our analyses suggested that LLM-generated scoring may be appropriate for scaling human scoring efforts, with human-model inter-rater reliability comparable to that between human scorers, though results varied by rubric construct, model type, and prompting methods. LLMs that (1) score constructs with concrete criteria, (2) perform higher on language comprehension and complex reasoning benchmarks, and (3) are provided detailed prompting methods produced scores that represented the human norm and exhibited precise distributions while increasing scale and avoiding cognitive overload. Findings align with recent work that suggests, while LLMs are generally promising tools, measurement context matters (Liu et al., 2025; Mellon et al., 2024; Parker et al., 2024). Further, this work illustrates the role of design and evaluation decisions for researchers and practitioners considering LLM-generated scoring to scale human efforts.

Notably, we did not set out to answer questions regarding the validity of LLM-generated scores for broader research contexts. Evaluation of validity requires building evidence about the prompts, rubrics, and scoring processes—both human scoring and LLM scoring—that are beyond the scope of this article (Haertel, 2013; Kane, 2006). Instead, we took on the more modest task of investigating whether LLM scoring could be a feasible replacement for human scoring in this context. That is, are there LLMs that could essentially replicate the "business-as-usual" approach of human scoring, without consideration of the validity of the

(human-scored) measures themselves? The answer appears to be yes.

Implications and Limitations

These findings can inform policy and practice in several key ways. First, we provided an empirical demonstration of scaling LLM-generated scoring in an education research context. Adapting Anglin et al.'s (2025) and Halterman and Keith's (2025) frameworks for codebook-LLM measurement, we outlined a replicable approach for generating and evaluating the reliability of LLM scores across model types, prompting methods, and construct domains. We publish an example workflow with rubrics and prompts on Github for researchers to adapt to their own studies. Next, we identified key considerations for practitioners and researchers when producing LLM-generated scores. Considerations included examining trade-offs in model type and prompting method, design details such as the number of scoring iterations to run and the proportion of responses to be manually scored, and evaluation details such as determining which scores to investigate and examining potential embedded algorithmic bias.

Finally, while this work points to using LLM-generated scores on other assessments that produce large volumes of short-response texts (e.g., observational feedback), we caution against hastily adopting LLM-generated scoring across any context. Given the demonstrated variability in scoring by construct domain, we note that task and model details matter. Further, although we demonstrated that LLM-generated scores do not appear to exacerbate potential social biases when scoring compared to humans, they still may replicate such biases at the same rates. LLM-generated scores should not be seen as a shield of objectivity to hide behind, especially in assessments with high-stakes decisions.

Our analysis faces several limitations. We looked only at the viability of scoring for off-the-shelf LLMs rather than machine learning models specifically developed for our assessments. Other researchers have demonstrated that human-model consistency may also improve with assessment-specific models or fine-tuned open-source LLMs (Halterman &

Keith, 2025; Kim et al., 2025; Mozer & Miratrix, 2025). We restricted our analysis to off-the-shelf models given that (1) most proprietary models are not open-source, limiting users' access to fine-tune model weights, and (2) researchers and practitioners are unlikely to build and fine-tune models due to technical and resource constraints.

There are also limitations of the models themselves. The use and choice of proprietary models (e.g., OpenAI's ChatGPT) is associated with trade-offs in input/output text length limits, cost, and speed, as well as privacy and interpretability concerns. This work is made possible because of institutional access to private model instances, essential when dealing with sensitive data.

We also see several opportunities to extend this work. Future research should continue focusing on the use of LLM-generated scores for downstream inference. Recent approaches that combine LLM-generated scores with a small number of gold-standard human scores may help correct for scoring bias and invalid confidence intervals while achieving statistical properties of consistency (Egami et al., 2023; Mozer & Miratrix, 2025). In addition, further work should explore the assumption that human scoring represents a definitive gold standard. Human-generated "gold" or "ground truth" labels are not error-free (Hardy, 2024). More work is required to understand what it means when human and model scores converge or diverge, given these relative judgments of reliability.

Acknowledgements

This research was funded by the Bill and Crissy Haslam Foundation. We thank Jim Soland for helpful feedback on earlier versions of this work. Macie Fitzgerald, Kurt Urban, and Isabella Yao provided excellent research assistance. We also thank Kaitlin Binsted and Francisco Santelli for their contributions to this research.

References

Anglin, K. L., Bertrand, A., Gottlieb, J., & Elefante, J. (2025). Scaling up with integrity: Valid and efficient narrative policy framework analyses using large language models. *Policy Studies Journal*. https://doi.org/10.1111/psj.70045

ArtificialAnalysis.ai. (n.d.). *Comparison of AI models across intelligence, performance, price — artificial analysis*. https://artificialanalysis.ai/models

Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., & Hemphill, L. (2024). Prompt design matters for computational social science tasks but in unpredictable ways. *arXiv preprint arXiv:2406.11980*. https://doi.org/10.48550/arXiv.2406.11980

Avitabile, A., Bartanen, B., & Kwok, A. (2025). Using large language models to analyze preservice teacher feedback and reflections during clinical teaching. EdWorkingPaper No. 25-1203. *Annenberg Institute for School Reform at Brown University*.

Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining*, *15*(2), 22–52.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cannata, M., Rubin, M., Goldring, E., Grissom, J. A., Neumerski, C. M., Drake, T. A., & Schuermann, P. (2017). Using teacher effectiveness data for information-rich hiring. *Educational Administration Quarterly*, *53*(2), 180–222. https://doi.org/10.1177/0013161X16681629

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., & Jurafsky, D. (2025). ELEPHANT: Measuring and understanding social sycophancy in LLMs. *arXiv preprint arXiv:2505.13995*. https://doi.org/10.48550/arXiv.2505.13995

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*. https://doi.org/10.48550/arXiv.1810.08810

Donaldson, M. L. (2013). Principals' approaches to cultivating teacher effectiveness: Constraints and opportunities in hiring, assigning, evaluating, and developing teachers. *Educational Administration Quarterly*, *49*(5), 838–882. https://doi.org/10.1177/0013161X13485961

Egami, N., Hinck, M., Stewart, B., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, *36*, 68589–68601.

Engel, M., Cannata, M., & Curran, F. C. (2018). Principal influence in teacher hiring: Documenting decentralization over time. *Journal of Educational Administration*, *56*(3), 277–296. https://doi.org/10.1108/JEA-05-2017-0061

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30). https://doi.org/10.1073/pnas.2305016120

Grissom, J. A., Egalite, A. J., & Lindsay, C. A. (2021). How principals affect students and schools. *Wallace Foundation*, *2*(1), 30–41.

Haertel, E. H. (2013). Reliability and Validity of Inferences about Teachers Based on Student Scores. William H. Angoff Memorial Lecture Series. *Educational Testing Service*.

Halterman, A., & Keith, K. A. (2025). Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *arXiv preprint arXiv:2407.10747*. https://doi.org/10.48550/arXiv.2407.10747

Hardy, M. (2024). "All that Glitters": Approaches to evaluations with unreliable model and human annotations. *arXiv preprint arXiv:2411.15634*. https://doi.org/10.48550/arXiv.2411.15634

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64).

Kim, Y., Mozer, R., Al-Adeimi, S., & Miratrix, L. (2025). ChatGPT vs. machine learning: Assessing the efficacy and accuracy of large language models for automated essay scoring. EdWorkingPaper No. no. 25-1335. *Annenberg Institute for School Reform at Brown University*.

Kirsten, E., Buckmann, A., Mhaidli, A., & Becker, S. (2024). Decoding complexity: Exploring human-ai concordance in qualitative coding. *arXiv preprint arXiv:2403.06607*. https://doi.org/10.48550/arXiv.2403.06607

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE Publications.

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Li, M., Shi, T., Ziems, C., Kan, M.-Y., Chen, N., Liu, Z., & Yang, D. (2023). CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1487–1505. https://doi.org/10.18653/v1/2023.emnlp-main.92

Liu, X., Zambrano, A. F., Baker, R. S., Barany, A., Ocumpaugh, J., Zhang, J., Pankiewicz, M., Nasiar, N., & Wei, Z. (2025). Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics*, *12*(1), 169–185.

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? using language models to code open-text

social survey responses at scale. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241231468

Mozer, R., & Miratrix, L. (2025). More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials. *The Annals of Applied Statistics*, *19*(1), 440–464. https://doi.org/10.1214/24-AOAS1967

Orr, M. T., & Hollingworth, L. (2018). How performance assessment for leaders (PAL) influences preparation program quality and effectiveness. *School Leadership & Management*, *38*(5), 496–517. https://doi.org/10.1080/13632434.2018.1439464

Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00414-0

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., ... Hendrycks, D. (2025). Humanity's last exam. *arXiv preprint arXiv:2501.14249*. https://doi.org/10.48550/arXiv.2501.14249

Ruckdeschel, M. (2025). Just read the codebook! make use of quality codebooks in zero-shot classification of multilabel frame datasets. *Proceedings of the 31st International Conference on Computational Linguistics*, 6317–6337. https://aclanthology.org/2025.coling-main.422/

Saldana, J. (2021). The coding manual for qualitative researchers.

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, 166–171.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Wang, R., & Demszky, D. (2023). Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 626–667. https://doi.org/10.18653/v1/2023.bea-1.53

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024). MMLU-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, *37*, 95266–95290. https://doi.org/10.52202/079017-3018

Xing, W., Nixon, N., Crossley, S., Denny, P., Lan, A., Stamper, J., & Yu, Z. (2025). The use of large language models in education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00457-x

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Shi, S., et al. (2025). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics*, 1–46.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Tables and Figures

Table 1: LLM Characteristic Comparison

Model	Version Date	Token Output Length	Output Cost	Output Speed	MMLU-Pro	HLE
GPT-5	Aug. 25	128k tokens	$10.00 / 1M tokens	114 tokens/second	0.82	0.06
GPT-4o	Nov. 24	128k tokens	$15.00 / 1M tokens	238 tokens/second	0.80	0.05
GPT-4.1 mini	Apr. 25	1.00M tokens	$1.60 / 1M tokens	73 tokens/second	0.78	0.05
Claude Sonnet 4	May 25	1.00M tokens	$15.00 / 1M tokens	76 tokens/second	0.84	0.10
Claude 3.5 Haiku	Oct. 24	200k tokens	$4.00 / 1M tokens	49 tokens/second	0.63	0.04
Amazon Nova Lite	Dec. 24	300k tokens	$0.24 / 1M tokens	155 tokens/second	0.59	0.05

Notes: Comparison of LLMs on output length, output cost, output speed, and two intelligence benchmarks (as of November, 2025) (ArtificialAnalysis.ai, n.d.). Intelligence benchmarks employed to assess language comprehension and complex reasoning are the Massive Multitask Language Understanding Pro (MMLU-Pro) (Y. Wang et al., 2024) and Humanity's Last Exam (HLE) (Phan et al., 2025).

Table 2: LLM and Human Scoring Descriptives (Survey Short-Response)

Rubric Item	Item Scale	Model Type						Prompting Method			Human Avg
		GPT-5	GPT-4o	GPT-4.1 mini	Claude Sonnet 4	Claude 3.5 Haiku	Amazon Nova Lite	Zero-Shot	Few-Shot	CoT	
Reasoning											
Reasoning	1/5	3.34 (0.37)	3.19 (0.39)	3.38 (0.42)	3.29 (0.35)	3.38 (0.53)	3.12 (0.53)	3.42 (0.43)	3.27 (0.42)	3.16 (0.40)	2.88
Candidate Preference											
Preference for Candidate 1	0/2	1.02 (0.06)	1.11 (0.10)	1.17 (0.11)	1.06 (0.07)	1.16 (0.15)	1.10 (0.25)	1.17 (0.15)	1.08 (0.12)	1.07 (0.11)	1.02
Preference for Candidate 2	0/2	0.28 (0.04)	0.33 (0.08)	0.34 (0.08)	0.28 (0.04)	0.34 (0.11)	0.36 (0.13)	0.37 (0.12)	0.30 (0.06)	0.30 (0.06)	0.27
Alternative response	0/1	0.31 (0.07)	0.29 (0.09)	0.20 (0.11)	0.27 (0.06)	0.20 (0.10)	0.18 (0.14)	0.20 (0.12)	0.26 (0.10)	0.26 (0.07)	0.29
Hiring Priorities											
Cultural Responsiveness	0/1	0.03 (0.02)	0.02 (0.01)	0.05 (0.07)	0.04 (0.02)	0.05 (0.04)	0.03 (0.05)	0.03 (0.03)	0.04 (0.04)	0.04 (0.03)	0.05
Parent/Community Engagement	0/1	0.17 (0.06)	0.12 (0.08)	0.22 (0.17)	0.23 (0.08)	0.21 (0.16)	0.19 (0.18)	0.20 (0.14)	0.18 (0.11)	0.19 (0.11)	0.15
Academic Achievement	0/1	0.46 (0.13)	0.56 (0.08)	0.55 (0.22)	0.58 (0.16)	0.65 (0.22)	0.64 (0.15)	0.62 (0.16)	0.54 (0.16)	0.56 (0.16)	0.41
Candidate Experience/Expertise	0/1	0.02 (0.02)	0.01 (0.01)	0.02 (0.05)	0.04 (0.04)	0.01 (0.02)	0.02 (0.04)	0.02 (0.02)	0.02 (0.03)	0.02 (0.03)	0.06
Evaluation	0/1	0.23 (0.06)	0.16 (0.07)	0.21 (0.16)	0.24 (0.07)	0.28 (0.17)	0.16 (0.12)	0.20 (0.12)	0.21 (0.10)	0.22 (0.10)	0.15
School Culture Fit	0/1	0.47 (0.12)	0.38 (0.21)	0.45 (0.25)	0.48 (0.11)	0.46 (0.23)	0.22 (0.21)	0.47 (0.22)	0.37 (0.17)	0.39 (0.17)	0.34
Avg. Distance from Human Average (normalized)		0.05	0.04	0.07	0.06	0.09	0.07	0.08	0.05	0.05	0.00
Avg. SD (normalized)		0.06	0.07	0.12	0.07	0.12	0.12	0.10	0.09	0.08	–

Notes: N=141. Descriptives are represented as: Mean of item-level response means (Average SD of item-level response SDs). Avg. Distance from Human Avg. (normalized) represents the average distance from the human average score, normalized by item scale and averaged over rubric item categories. Avg. SD (normalized) represents the average standard deviation, normalized by item scale and averaged over rubric item categories.

Table 3: LLM and Human Scoring Descriptives (Interview)

Rubric Item	Item Scale	Model Type						Prompting Method			Human Avg
		GPT-5	GPT-4o	GPT-4.1 mini	Claude Sonnet 4	Claude 3.5 Haiku	Amazon Nova Lite	Zero-Shot	Few-Shot	CoT	
Reasoning											
Intensity of Recruiting Strategies	1/4	2.05 (0.27)	2.01 (0.36)	1.96 (0.38)	1.90 (0.24)	2.65 (0.38)	2.26 (0.50)	2.03 (0.35)	2.18 (0.35)	2.20 (0.37)	2.12
Engaging with Evidence	1/4	2.21 (0.29)	2.33 (0.44)	2.27 (0.48)	2.22 (0.33)	2.48 (0.37)	2.07 (0.54)	1.99 (0.41)	2.35 (0.40)	2.45 (0.41)	2.48
Collaborative Decision-Making	1/4	1.90 (0.23)	2.05 (0.42)	2.12 (0.45)	2.11 (0.32)	2.11 (0.42)	1.81 (0.56)	1.92 (0.41)	2.04 (0.38)	2.09 (0.41)	2.15
Responsiveness to Local Labor Conditions	1/4	2.25 (0.29)	2.14 (0.48)	2.29 (0.56)	2.26 (0.36)	2.50 (0.39)	2.12 (0.55)	2.22 (0.44)	2.33 (0.44)	2.23 (0.43)	2.32
Hiring Priorities											
Cultural Responsiveness	0/1	0.03 (0.02)	0.03 (0.01)	0.05 (0.06)	0.08 (0.04)	0.07 (0.06)	0.02 (0.03)	0.04 (0.04)	0.06 (0.05)	0.04 (0.03)	0.04
Parent/Community Engagement	0/1	0.04 (0.03)	0.04 (0.03)	0.14 (0.09)	0.09 (0.04)	0.09 (0.08)	0.04 (0.05)	0.08 (0.06)	0.07 (0.05)	0.06 (0.05)	0.04
Academic Achievement	0/1	0.14 (0.07)	0.10 (0.07)	0.18 (0.11)	0.28 (0.13)	0.50 (0.25)	0.18 (0.16)	0.25 (0.14)	0.23 (0.13)	0.21 (0.11)	0.22
Candidate Experience/Expertise	0/1	0.14 (0.09)	0.15 (0.15)	0.25 (0.20)	0.27 (0.14)	0.44 (0.21)	0.07 (0.10)	0.27 (0.16)	0.21 (0.16)	0.17 (0.12)	0.13
Evaluation	0/1	0.59 (0.06)	0.41 (0.18)	0.43 (0.18)	0.64 (0.09)	0.61 (0.16)	0.42 (0.22)	0.50 (0.16)	0.52 (0.14)	0.53 (0.15)	0.46
School Culture Fit	0/1	0.59 (0.14)	0.66 (0.25)	0.67 (0.26)	0.67 (0.15)	0.82 (0.18)	0.57 (0.32)	0.64 (0.22)	0.69 (0.22)	0.66 (0.22)	0.55
Avg. Distance from Human Average (normalized)		0.04	0.04	0.05	0.07	0.13	0.05	0.06	0.04	0.03	0.00
Avg. SD (normalized)		0.07	0.12	0.14	0.09	0.13	0.14	0.12	0.11	0.11	-

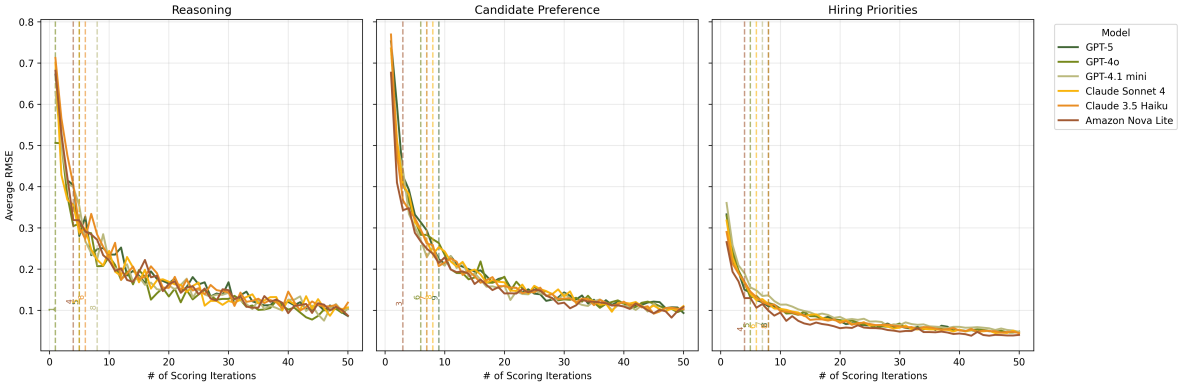
Notes: N=93. Descriptives are represented as: Mean of item-level response means (Average SD of item-level response SDs). Avg. Distance from Human Avg. (normalized) represents the average distance from the human average score, normalized by item scale and averaged over rubric item categories. Avg. SD (normalized) represents the average standard deviation, normalized by item scale and averaged over rubric item categories.

Table 4: Intra-Class Correlations: LLM-Human Score Consistency (Survey Short-Response)

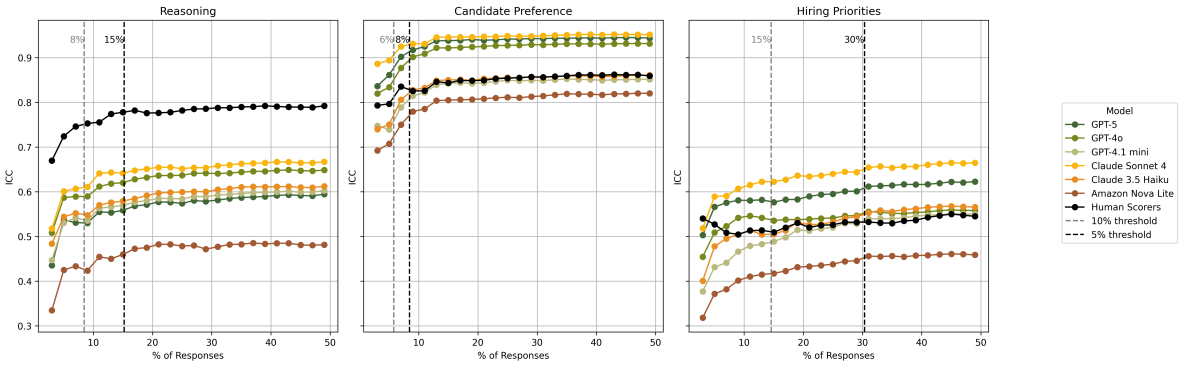
	Reasoning			Candidate Preference			Hiring Priorities		
	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT
GPT-5	0.60	0.55	0.63	**0.95**	**0.92**	**0.96**	**0.59**	**0.67**	**0.65**
GPT-4o	0.52	0.69	0.74	**0.91**	**0.93**	**0.96**	0.48	**0.62**	**0.63**
GPT-4.1 mini	0.50	0.63	0.68	0.76	**0.87**	**0.92**	0.47	**0.60**	**0.62**
Claude Sonnet 4	0.63	0.63	0.74	**0.95**	**0.95**	**0.96**	**0.62**	**0.71**	**0.69**
Claude 3.5 Haiku	0.50	0.66	0.68	0.76	**0.91**	**0.92**	0.51	**0.62**	**0.62**
Amazon Nova Lite	0.40	0.50	0.60	0.73	0.85	**0.89**	0.39	0.51	0.54
Human Scorers		0.79			0.86			0.56	

Notes: N=141. For each model and prompt type, the ICC is calculated between the model's average score over 10 iterations of scoring for a given response rubric item with the average human score for that same response rubric item. Bold values represent ICC scores greater than or equal to the Human Scorers' ICC in that category. CoT stands for chain-of-thought prompting. Following Koo & Li (2016), we suggest that "based on the 95% confident interval of the ICC estimate, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively" (p. 155).

Figure 1: Design Considerations (Survey Short-Response)



(a) Average RMSE vs. # of Scoring Iterations



(b) ICC vs. % of Responses

Panel 1 Notes: N=47 (training set size). RMSE is not exactly 0 when the number of scoring iterations is 50 due to random subsamples being drawn with replacement. Vertical lines represent the point on the normalized curve with the furthest distance from the line y = -x + 1 (kneedle method) (Satopaa et al., 2011)

Panel 2 Notes: N=141 (testing set size). The % of responses does not start at 1% to ensure a high enough number of degrees of freedom to calculate ICCs, confidence intervals, and perform hypothesis testing. The 10% and 5% thresholds indicate at what % of responses the ICC estimates are within 10% and 5% of the full-sample ICC value, respectively. Graphs represent average ICCs across all prompting methods. We demonstrate these same graphs broken down by prompting method in Appendix G, though note no systematic threshold differences by prompting method.

Appendix A: Participant Demographic and District Characteristics

Table A.1: Participant Demographic and District Characteristics (n=200, $n_{dist} = 54$)

	Mean
Years of Experience	15 years
Age	44 years
Female	68%
White	87%
Black	9%
Urban District	33%
Suburban District	34%
Town/Rural District	33%

Appendix B: Sample Responses

B.1 Survey Short-Response

Prompt: You are considering two candidates for an open teaching position. Candidate 1 has an exceptional record of student achievement growth at their prior school but has received complaints from parents for being too strict with student discipline. Candidate 2 is well-liked by parents and students at their prior school but has a weaker record of student achievement growth. What do you do?

- "I would consider the needs of the students for which this teacher is being considered. If there is a need to greater discipline and student achievement, I would select the first candidate. If the students are generally engaged in learning and achievement is satisfactory, I would likely hire the second candidate and provide them with instructional coaching supports to improve student achievement within that class."

- "I would prefer candidate 1. An exceptional record of student achievement is probably more beneficial to students in the long run than being well-liked. I had teachers in school who were strict and weren't necessarily my favorites at the time, however, these educators did more to prepare me for difficult college classes and tasks at work than the ones who were more popular among students and parents. Students need structure and discipline as well as the opportunity to grow academically. Candidate 1 sounds like the best option for students. "

- "Easy. Candidate 2. I have worked with many a Candidate 1. To reduce disparities and be culturally responsive, strict is not always the best path for a teacher."

- "I could go either way and support either decision, but for this purpose, I would select Candidate 2. I feel strongly that teaching methods and effectiveness can be taught and coached, but it is much more difficult to change attitudes, beliefs, and personal qualities. In addition, as Rita Pierson says, 'Kids don't learn from people they don't like.' I have seen this, and I would rather students and families have a teacher that they like because the teacher is more likely to get students to perform at high levels. We could coach instructional methods to get there."

- "I would hire Candidate 1. Student growth and achievement is the purpose of school. Being well-liked is not the main intent. Candidate 1 can be coached in methods of handling student discipline with parental feelings taken into account. Hiring a teacher with a weak record of student achievement is a much more debilitating risk for the school."

B.2 Interview

Prompt: A key goal for every school is to hire effective teachers for every vacancy that comes up. How do you go about meeting that goal?

- "This year we had to hire 7 teachers. Some of it was because of retirement. And so we started, probably about mid-year last year, talking to other principals in the county that have good student teachers, strong student teachers, and finding out that they were doing well, and that's where we found some of ours. We also in our county have a program called Grow-Your-Own, where we have teachers who are able to work as assistants within our schools for 3 years, and they can get a degree through [University A]. And that's where we had 3 of ours. 3 of ours that we hired came from that. So they're getting to work in the schools and teach part of the time and also be an

assistant, and we have found that has been an excellent way for us to find teachers. And they wait until I see how they are in the classroom. But they get to work at our school, too, and that's been excellent because we can see how they fit in our culture of our school. And then some of it is, I know some of ours came from a local job fair and we have those here in our county. One of the things we're lucky here in [County A] is we have [University A] here that has about 180 new graduates every year in education. So we lean on that heavily. And then just word of mouth, if someone knows about someone that's gonna be great. We ask around because it is getting harder and harder to find good candidates. We scour everyone we can find. But those the local college and the Grow-Your-Own program has been great. And then, we just ask everybody around, especially local schools and the college. We have a great relationship with our local college."

- "Well, I would say that that starts with having a list of people in mind that you would call if a position came open. Like you always have to have somebody that's on deck. And I think that as a good principal, you start networking and putting people. Like I always say like I have a list for when I get my own school. And I will definitely be using that, because at that point, my school and my students and my staff are the most important. So if they want to come work with me and do something great, then I will definitely ask them. I also would say that, it's interesting because like I was looking at the survey question that's like, 'how do I decide who is a good teacher?' and a lot of times it has to do with quality questions about how they make instructional decisions and how they feel about kids. If I cannot tell a person does not like kids from their interview, I don't care if they are the most pedagogically sound human being, or if they have content knowledge that would rival, you know, a college professor. I don't want them. Because if they don't like kids, kids are gonna want to be in their room. So there definitely have to be some indicators there. I would even go as far as, like

one of the principles, and I think I would do this as well, would even have kids in a screening interview at first. Like they wouldn't sit in the actual interview. But like when kids are waiting out in the lobby, kids would be like, 'what's your name? Where you from?' And then they would tell the principal, like, 'oh, he was cool.' Or 'he seemed a little nervous,' you know. So anything that I could do to get a read on how they felt about kids and how they fit in the culture of the school. I think it's incredibly difficult to make a hiring decision based on a resume and interview and then letters of recommendation. I think ultimately you can feel like something's a great hire and then it totally blows up in your face. But I think that's not just teaching, I think that's any industry. But I think having a streamlined process that's a fair to everybody. You'll get into a string of that eventually, but ultimately, I have not had hiring decisions be solely up to me ever, so that will be something that I think will be a lot of on the job learning and adjusting as I go throughout the principalship."

- "Yeah, going back to one of the things I said earlier. Relationships are important, but one of the most important things as well is hiring and then retaining high quality teachers. And so thankfully, in the school I'm in, it's a smaller school, so I've been told there's not a lot of turnover here. Now, I do have a couple of teachers scheduled to retire at the end of this year. So, I'm gonna start early, you know, as early as I can, as early as the district will post the position. But even prior to the posting, I'm gonna be keeping my eye open and talking to both our local universities here, [University A] and [University B]. We have a good relationship with them, and I know I have a couple of student teachers that are coming in the spring from [University B]. And so just thinking through that as early as I can, but I think the most important thing, well, obviously, I would typically rather hire a veteran teacher who has already had some experience. But that's not always the case, and students right out of college have to start somewhere, so I'm not opposed to it. But I would look for a veteran teacher.

But even more importantly than that, I want someone who is gonna be all-in. High character, right? We can train the other things. We can teach the specifics of the curriculum. But I want somebody who is willing to be all in, and join in our shared vision of moving kids forward, and if I can find that, the other things can be trained."

- "I mean, if I'm being 100% honest, that's not a realistic goal. We are currently down a social studies teacher and a SpEd teacher. And we have been for the year. One of our math teachers was a bandaid. A person that we pulled off the street, basically. I don't think most schools, and I know that we're a Title I school, but I would say some of the top schools in the district do not have the luxury of looking through 6 applications and saying, 'Oh, this is the one that I will hire.' What you have to do, I think, is make sure you are hiring a culture fit. Now, if you have that opportunity, then great, but the reality is we don't have enough teachers. So I would love to be able to say, 'Hey, you need to come in and and do a lesson, and then we're gonna do 3 interviews for every time we hire.' But the reality is, if you have a license and you fit the culture of our school and you're gonna love on kids, we're gonna hire you. Cause otherwise we're not gonna have anybody."

- "It's tough. I've been very blessed. I've worked in districts and schools where people want to be there. So I've been very blessed in my career, because I know friends and colleagues, and they're in tough schools, and it's hard to keep and retain them. And a lot of districts have a thing where they incentivize for folks to come to some of these difficult schools, or they, you know, districts often have a minimum. You gotta serve in one school a minimum of 3 consecutive years or x amount of years before you can transfer to another school. That's a tough one, because when you're at a school where people want to be there, your applicant pool is much, much better. I think the

challenge in leadership, and I've not had to face this a whole lot in my career, is having a difficult school where you attract the best. And every once in a while, some of those best of the best want to go to those schools, but that's a difficult thing to do. That goes back to your culture piece, right? I mean, if you're in there, and you've got a strong culture, regardless teachers are going to be there. But hiring good teachers, if you do that right, it takes care of itself. I mean, it's one of the best things you can do if you get the right folks. But it's one of the most difficult things to do. The teacher shortage is a real thing, and especially in those critical fields, special education, math. And now states like the state of Tennessee, [State A] did it a few years ago, computer science for all where everybody's gotta do computer science. It's hard to find computer science teachers. You know, those types of things. It's tough. But not settling, doing your homework, doing your background checks with them, talking to former employers. You know, getting on- I don't do social media personally- but when I was hiring in [State A], I always had an assistant principal doing social media checks for us so we could look at that. So I think those are some things that you do when you're hiring."

Appendix C: Full Rubrics and Hiring Priorities Information

Note: While the full rubrics are provided below, the detail of information provided to models varies by prompting method. In zero-shot prompting, models are provided only labels and instructions. In few-shot prompting, models are provided labels, instructions, and examples; however, examples do not contain step-by-step rationales. In chain-of-thought prompting, models are provided labels, instructions, examples, and step-by-step rationales for examples.

C.1 Survey Short-Response - Full Rubric

Rubric Category: Reasoning

Label: Reasoning

Instructions: Rate how strong and convincing the reasoning is on a scale of 1 to 5, with 1 denoting the lowest strength and 5 denoting the highest strength.

Example: "I will choose Candidate x." The respondent does not articulate a plan to decide which candidate they will choose, refer to any data, list goals or expectations for hiring teachers, or define priorities in a candidate. This response scores a 1.

Example: "During the hiring process, I will consider the needs of the school and which teacher meets those needs best but will likely hire candidate x." The respondent discusses a plan to weigh the needs of the school, does not provide any details, but still makes a decision about choosing a candidate. This response scores a 2.

Example: "Being strict is not necessarily bad if the teacher is still following along with school policies. I would hire Candidate x but also make sure to explain our school's behavioral procedures in place before hiring." The respondent identifies a specific weakness of a candidate and details expectations about school policies. This response scores a 3.

Example: "I like that candidate x has strong relationships with school parents and the community. An instructional coach may improve candidate x's student achievement scores, and a mentor teacher may help the new hire adapt to the curriculum expectations." The respondent demonstrates a priority to hire the candidate with strong community relationships, presents two plans to help the teacher improve their record of student achievement, and identifies teaching expectations. This response scores a 4.

Example: "Before hiring a candidate, I would try to get more information in the decision-making process. This includes holding an interview to directly assess the candidates' strengths and weaknesses on student growth and discipline. I would also try to use prior knowledge based on principal evaluations and see which grade and subject levels need filling. With this structure, I will hire the candidate whose answers stand

out and can best fit our school culture." The respondent has a multistep plan to assess the candidates, presents highly detailed logic, highlights several key areas, and provides a solution to the question. This response scores a 5.

Rubric Category: Candidate Preference

Label: Preference for Candidate 1

Instructions: On a scale of 0 to 2, rate the respondent's strength of preference for choosing Candidate 1.

Example: "If I can allocate professional development funds to improve classroom behaviors for Candidate 1, then I will choose candidate 1." The respondent is not completely decisive but still expresses a preference for candidate 1 if they are able to use resources to help candidate 1. This response scores a 1 for Preference for Candidate 1.

Example: "Candidate 1. Strong academic performance is key for student's growth and development." The respondent is decisive and lists candidate 1 as their hiring choice. This response scores a 2 for Preference for Candidate 1.

Label: Preference for Candidate 2

Instructions: On a scale of 0 to 2, rate the respondent's strength of preference for choosing Candidate 2.

Example: "I would like to get more information before choosing, but if I have to choose, then I will select candidate 2." The respondent is not fully certain and would like more information before hiring but still prefers candidate 2. This response scores a 1 for Preference for Candidate 2.

Example: "Candidate 2 will contribute positively to the school climate. An instructional coach can also help candidate 2 improve their academic track record." The respondent compliments candidate 2, plans to dedicate resources to helping candidate 2,

and is decisive. This response scores a 2 for Preference for Candidate 2.

Label: Alternative response

Instructions: Respondent provides an alternative response. Score 1 if yes, 0 if no.

Example: "I would hold another round of interviews and decide based on their responses to my questions. If one candidate seems to best fit the needs of the school, then I will hire that candidate." The respondent does not reveal a preference for either candidate and needs more information before reaching a decision. This response scores a 1 for Alternative response.

Example: "This is a hard decision, but I think I will choose Candidate x". The respondent mentions the decision is difficult but is still decisive and prefers one candidate over another. This response scores a 0 for Alternative response.

Rubric Category: Hiring Priorities

Category Instructions: Multiple items may be scored with a 1 for a given response.

Label: Cultural Responsiveness

Instructions: Score 1 if the response appears to prioritize any of the following: Ensuring racial/ethnic representation among school faculty, Placing teachers in classrooms with students that share racial/ethnic identities, Hiring teachers that engage in culturally responsive instruction. Else score 0.

Positive Example: "I would choose Candidate x because they are committed to creating a safe and inclusive space for all students." The respondent prefers candidates that will foster inclusivity among all students, indicating a priority for cultural responsiveness. This response scores a 1 for Cultural Responsiveness.

Negative Example: "My choice for candidate will depend on how the teacher fits in with the culture of the staff and students more broadly." The respondent does not mention whether teacher or student identity plays a role in hiring a candidate nor culturally responsive instruction. This response scores a 0 for Cultural Responsiveness. However, this response scores a 1 for School Culture Fit.

Label: Parent/Community Engagement

Instructions: Score 1 if the response appears to prioritize any of the following: Ensuring parents/community members are satisfied with their children's teachers, Faculty have connections to the local community. Else score 0.

Positive Example: "I would rather hire Candidate x because of their positive relationships with students and parents. Strictness and academic record matter, but parent and community feedback are also important." The respondent prefers candidates with strong positive community and parent relationships. This response scores a 1 for Parent/Community Engagement.

Negative Example: "Student success and achievement are paramount, and I feel we can coach classroom management." The respondent does not mention whether parents or community members matters in deciding which candidate to choose. This response scores a 0 for Parent/Community Engagement. However, this response scores a 1 for Academic Achievement.

Label: Academic Achievement

Instructions: Score 1 if the response appears to prioritize any of the following: Hiring faculty with consistent records of improving student achievement, Subject-matter expertise, Placing high performing teachers in tested grades / with lower achieving students.

Else score 0.

Positive Example: "I would not choose Candidate x because they do not show the same track record of student growth as Candidate y. We should help students perform the best they can." The respondent values student growth, and chooses a candidate based on their ability to improve academic performance. This response scores a 1 for Academic Achievement.

Negative Example: "I would look for candidates that can meet the needs of the school's grade levels and subject material." The respondent does not mention prioritizing candidates with a history of student achievement or growth. This response scores a 0 for Academic Achievement. However, this response scores a 1 for Candidate Experience/Expertise.

Label: Candidate Experience/Expertise

Instructions: Score 1 if the response appears to prioritize any of the following: Years of teaching experience, Hiring/Placing teachers in grades/subjects where they have experience. Else score 0.

Positive Example: "It is important to consider what subject and expertise I need for my school. I would choose a candidate x based on if they have done that work in the past." The respondent prefers candidates with past experience and meeting the subject needs of the school. This response scores a 1 for Candidate Experience/Expertise.

Negative Example: "I will choose the candidate who manages the classroom in a culturally relevant manner. Teachers who are strict disciplinarians can be good or bad, but it's important to show my teachers restorative justice practices too." The respondent does not discuss the candidates' experience or specific grade and subject expertise when making their decision. This response scores a 0 for Candidate Experience/Expertise. However, this response scores a 1 for Cultural Responsiveness.

Label: Evaluation

Instructions: Score 1 if the response appears to prioritize any of the following: References from prior employers, Evaluation scores in prior positions, Classroom observations including sample lessons, Responses to hiring interview questions. Else score 0.

Positive Example: "I would ask the candidates questions about how they discipline students and improve achievement. It would also be helpful to talk to Candidate x's current principal about past performance." The respondent prioritizes asking the candidates questions and interviewing them. The respondent also wants to know more information from prior leaders such as the candidate's current principal. This response scores a 1 for Evaluation.

Negative Example: "If candidate x has a lot of parent complaints, that would make me hesitant to hire them." The respondent does not mention interviewing the candidate, reviewing their past evaluation scores from leaders, or observing how the candidate teaches in the classroom. This response scores a 0 for Evaluation. However, this response scores a 1 for Parent/Community Engagement.

Label: School Culture Fit

Instructions: Score 1 if the response appears to prioritize any of the following: Fit with the school culture / existing procedures (including discipline), Cultural fit and relationships with students and staff. Else score 0.

Positive Example: "Candidate x is liked by their school leaders, and it is important to know how the teacher will work within the school climate. So, I will choose Candidate x." The respondent mentions the importance of a candidate to be liked and fit in the school culture and climate. This response scores a 1 for School Culture Fit.

Negative Example: "I would hold another round of interviews and speak with each teacher about how they handle behavior and academic policies." The respondent does not mention the school culture or community or how the candidates will work with other teachers. This response scores a 0 for School Culture Fit. However, this response scores a 1 for Evaluation.

C.2 Interview - Full Rubric

Rubric Category: Reasoning

Label: Intensity of Recruiting Strategies

Instructions: On a scale of 1 to 4, with 1 denoting the lowest strength and 4 denoting the highest strength, rate the degree to which the respondent is actively involved in multiple recruiting strategies, with some strategies tailored to local contexts (e.g., job fairs, Grow-Your-Own (GYO) programs, university partnerships). Only score 1 through 4.

Example: "Within our district, I keep a good relationship with X and Y universities in our county, so whenever there's open positions, I can reach out early to a few places and see if a student teacher is available". The respondent is actively involved in multiple recruiting strategies, with some strategies tailored to local contexts. This response scores a 4.

Example: "Our school participates in a local job fair and we attract a few teachers through that every year. It also helps when the HR team uses social media to reach out to candidates". The respondent incorporates non-traditional or multiple recruiting strategies, and some strategies are tailored to their local context. This response scores a 3.

Example: "When we have openings in the summer, I usually start by looking through the postings to see who's applied...". The respondent mentions a traditional or passive method of recruiting (e.g., job postings, word-of-mouth) with no adaptation to local

context. This response scores a 2.

Example: "Hiring is tricky with teacher shortages in our county, but when I review applicants I try to figure out...". The respondent does not mention recruitment during the hiring process. This response scores a 1.

Label: Engaging with Evidence

Instructions: On a scale of 1 to 4, with 1 denoting the lowest strength and 4 denoting the highest strength, rate the degree to which the respondent incorporates multiple data sources (e.g., sample lessons, student achievement data, references, interviews) and explains how evidence informs their hiring strategy. Only score 1 through 4.

Example: "Data can indicate if someone is an effective teacher, but it's also important to see if they will fit the culture and needs of the school. When I interview or do a sample lesson, I try to figure out if you're a team player with the students and the staff, because that will usually let me know if you enjoy teaching and will be effective". The respondent incorporates multiple data sources and explains how evidence informs their hiring strategy. This response scores a 4.

Example: "Hiring can be tough right now. Usually I reach out to recommenders and try to ask good interview questions". The respondent uses evidence (e.g., references, resumes, interviews) during the hiring process but does not explain how evidence informs hiring strategy or decision-making. This response scores a 3.

Example: "One of the strategies I have seen is getting a feel for if you'll be a good fit and you meet the needs of the school". The respondent references only on impressions or resumes during the hiring process. This response scores a 2.

Example: "Well this year we are going to try and start the hiring process early, and we can give good supports to all teachers even if they are not immediately qualified for the whole job". The respondent does not mention evidence-based evaluation when making

hiring decisions. This response scores a 1.

Label: Collaborative Decision-Making

Instructions: On a scale of 1 to 4, with 1 denoting the lowest strength and 4 denoting the highest strength, rate the degree to which the respondent seeks out current staff, students, or professional networks during the hiring process (e.g., interview input, content expertise, recruitment connections). Only score 1 through 4.

Example: "I always try to bring in a subject teacher in our interview process, and we can bounce off each other's questions. So they can judge if the teacher knows their content and then I also talk to former employers to see if they will be a good culture fit here". The respondent seeks out current staff, students, or professional networks at multiple points during the hiring process. This response scores a 4.

Example: "So I have a great relationship with X university, and every summer I talk to them about student teachers that can be a great fit here". The respondent regularly seeks out current staff, students, or professional networks at some point during the hiring processes. This response scores a 3.

Example: "With vacancies, I might reach out to other principals in the area and see if they know someone who's looking for a job". The respondent may seek out input from current staff, students, or professional networks, but has no structured or consistent strategy (e.g., word-of-mouth, inconsistently talking to others). This response scores a 2.

Example: "I usually try to see in an interview if they know the content and figure out if they would fit with the school culture and the students". The respondent does not mention involving current staff, students, or professional networks in the hiring process. This response scores a 1.

Label: Responsiveness to Local Labor Conditions

Instructions: On a scale of 1 to 4, with 1 denoting the lowest strength and 4 denoting the highest strength, rate the degree to which the respondent identifies local teacher labor market conditions (including supply, shortages, hard-to-staff positions, and/or competition) and implements targeted strategies for recruitment and hiring, and/or describes additional supports for new/underqualified hires. Only score 1 through 4.

Example: "If you asked me this years ago, I would have told you about doing multiple interviews and a sample lesson. But lately there are so few teachers, so instead I try to make sure if I hire someone underqualified, we provide coaching and mentor teachers to get them in a good place". The respondent identifies labor market conditions and implements additional supports for new/underqualified hires. This response scores a 4.

Example: "Right now we have 4 vacancies and even though our county might be able to pay a little more than the surrounding ones, so sometimes I just need to act as a substitute teacher and that's how it is". The respondent identifies labor market conditions and describes reactive measures. This response scores a 3.

Example: "It has been difficult with the teacher shortage. But in an ideal world, I would have 5 candidates and we would do a sample lesson...". The respondent identifies labor market conditions with no reference to recruitment/hiring strategies. This response scores a 2.

Example: "When it comes to recruiting, I always try to start early and try to show off what our high school can bring to the table...". The respondent does not mention local labor market conditions. This response scores a 1.

Rubric Category: Hiring Priorities

Category Instructions: Score either a 1 or a 0. Multiple items may be scored with a

1 for a given response.

Label: Cultural Responsiveness

Instructions: Score 1 if the response appears to prioritize any of the following: Ensuring racial/ethnic representation among school faculty, Placing teachers in classrooms with students that share racial/ethnic identities, Hiring teachers that engage in culturally responsive instruction. Else score 0.

Positive Example: "And you want to see if a teacher would be a good fit with the population of the students". The respondent prioritizes teacher representation among student populations. This response scores a 1 for Cultural Responsiveness.

Negative Example: "So with the lack of qualified candidates right now, we really try to look for fit, if they will work well with the kids and other teachers?". The respondent prioritizes culture fit, not cultural responsiveness. This response scores a 0 for Cultural Responsiveness, however it scores a 1 for School Culture Fit.

Label: Parent/Community Engagement

Instructions: Score 1 if the response appears to prioritize any of the following: Ensuring parents/community members are satisfied with their children's teachers, Faculty have connections to the local community. Else score 0.

Positive Example: "We want to make sure teachers are willing to talk to our parents and community if we ask". The respondent prioritizes teachers' abilities to satisfy parents. This response scores a 1 for Parent/Community Engagement.

Negative Example: "So students could also benefit from teacher diversity like if there's a lot of Spanish speaking students have Spanish speaking teachers". The respondent prioritizes hiring for racial/ethnic representation, not relationships with parents. This response scores a 0 for Parent/Community Engagement, however it scores a 1 for Cul-

tural Responsiveness.

Label: Academic Achievement

Instructions: Score 1 if the response appears to prioritize any of the following: Hiring faculty with consistent records of improving student achievement, Subject-matter expertise, Placing high performing teachers in tested grades / with lower achieving students. Else score 0.

Positive Example: "And so, we can pull evaluation data to see if they'll be an effective teacher". The respondent prioritizes evaluation/achievement data when hiring. This response scores a 1 for Academic Achievement.

Negative Example: "There is a really strong parent community here, so they'll know if a teacher is unqualified". The respondent prioritizes whether a teacher will satisfy parents. This response scores a 0 for Academic Achievement, however it scores a 1 for Parent/Community Engagement.

Label: Candidate Experience/Expertise

Instructions: Score 1 if the response appears to prioritize any of the following: Years of teaching experience, Hiring/Placing teachers in grades/subjects where they have experience. Else score 0.

Positive Example: "Obviously I would prefer to hire a veteran over a new teacher, but sometimes your only candidates are fresh out of college". The respondent discusses prioritizing experienced teachers. This response scores a 1 for Candidate Experience/Expertise.

Negative Example: "You always want to try to get a 5 teacher, so you want to look at the data piece". The respondent references achievement scores and does not

prioritize Candidate Experience/Expertise. This response scores a 0 for Candidate Experience/Expertise, however it scores a 1 for Academic Achievement.

Label: Evaluation

Instructions: Score 1 if the response appears to prioritize any of the following: References from prior employers, Evaluation scores in prior positions, Classroom observations including sample lessons, Responses to hiring interview questions. Else score 0.

Positive Example: "For our interview process, I always try to ask the same questions and avoid any bias". The respondent discusses the importance of evaluating candidates. This response scores a 1 for Evaluation.

Negative Example: "You know, I want to know if you know the content for the position and the grade, otherwise we won't hire you". The respondent prioritizes hiring teachers in grades/subjects with experience. This response scores a 0 for Evaluation, however it scores a 1 for Candidate Experience/Expertise.

Label: School Culture Fit

Instructions: Score 1 if the response appears to prioritize any of the following: Fit with the school culture / existing procedures (including discipline), Cultural fit and relationships with students and staff. Else score 0.

Positive Example: "Your data may look good, but I would rather find people who fit with the school environment and want to be here". The respondent prioritizes whether a teacher will fit with school culture instead of achievement data. This response scores a 1 for School Culture Fit.

Negative Example: "I also think the candidate should do a model lesson or some sort of classroom observation". The respondent prioritizes performance tasks, not school

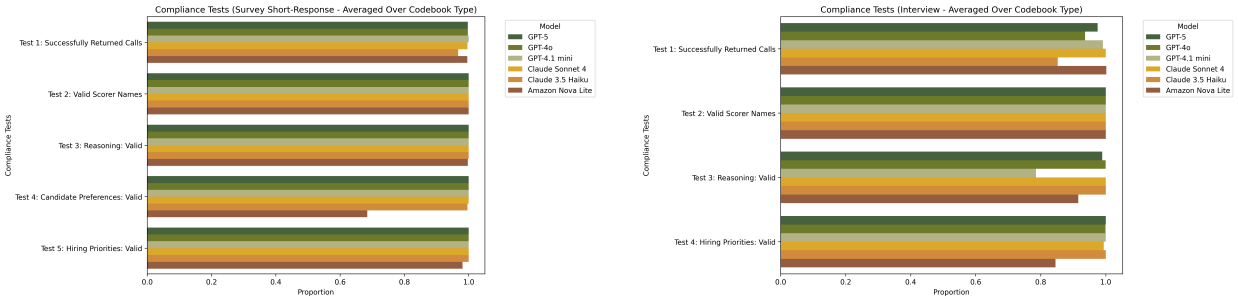
culture fit. This response scores a 0 for School Culture Fit, however it scores a 1 for Evaluation.

C.3 Development of "Hiring Priorities" Rubric Category

To develop the *hiring priorities* rubric category, we drew from an additional survey that participants completed that included several sets of items related to teacher hiring, assignment, and retention. Specifically, survey items asked aspiring principals to rate the importance of twelve priorities when hiring teachers, nine priorities related to hiring fit, and ten priorities for assigning teachers to classes (items measured via 5-point Likert scales). Item development was rooted in prior literature (Donaldson, 2013; Grissom et al., 2021) and intended to cover a wide variety of considerations, including cultural responsiveness (e.g., "Ensuring racial and ethnic representation among students in each class") and academic achievement (e.g., "Hiring faculty with consistent records of improving student achievement"). The six *hiring priorities* factors were derived from predicted factor scores generated through factor analysis of participants' responses.

Appendix D: Compliance

Figure D.1: Compliance Proportions by Model, Compliance Test



(1) Survey Short-Response (2) Interview

Notes: $N_{survey} = 47$. $N_{interview} = 31$. Compliance proportions are determined by the number of responses that meet a given test's criteria divided by the total number of responses.

Appendix E: Interview Consistency

Table E.1: Intra-Class Correlations: LLM-Human Score Consistency (Interview)

	Reasoning			Hiring Priorities		
	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT
GPT-5	0.80	**0.87**	**0.87**	**0.64**	**0.70**	**0.73**
GPT-4o	0.64	0.76	0.81	0.59	**0.62**	0.60
GPT-4.1 mini	0.74	0.80	0.83	0.55	0.57	0.60
Claude Sonnet 4	0.76	0.84	0.85	0.53	0.58	**0.63**
Claude 3.5 Haiku	0.59	0.67	0.65	0.36	0.53	0.57
Amazon Nova Lite	0.44	0.55	0.60	0.57	0.59	0.55
Human Scorers		0.86			0.62	

N=93. For each model and prompt type, the ICC is calculated between the model's average score over 10 iterations of scoring for a given response rubric item with the average human score for that same response rubric item. Bold values represent ICC scores greater than or equal to the Human Scorers' ICC in that category. CoT stands for chain-of-thought prompting.

Appendix F: Interview Uncertainty

Table F.1: Entropy Scores (Normalized) (Survey Short-Response)

	Reasoning			Candidate Preference			Hiring Priorities		
	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT
GPT-5	0.30	0.28	0.27	0.10	0.09	0.04	0.14	0.10	0.10
GPT-4o	0.29	0.31	0.28	0.16	0.10	0.09	0.14	0.14	0.13
GPT-4.1 mini	0.33	0.34	0.33	0.17	0.13	0.10	0.35	0.23	0.22
Claude Sonnet 4	0.30	0.26	0.25	0.09	0.08	0.05	0.14	0.13	0.12
Claude 3.5 Haiku	0.36	0.33	0.30	0.21	0.14	0.11	0.27	0.23	0.22
Amazon Nova Lite	0.40	0.42	0.42	0.24	0.20	0.21	0.18	0.23	0.23

N=141. Each cell represents the mean entropy of item-level score distributions for a given model and prompt type across rubric categories. Lower entropy values indicate greater scoring consistency across iterations and alignment with humans. CoT stands for chain-of-thought prompting.

Table F.2: Entropy Scores (Normalized) (Interview)

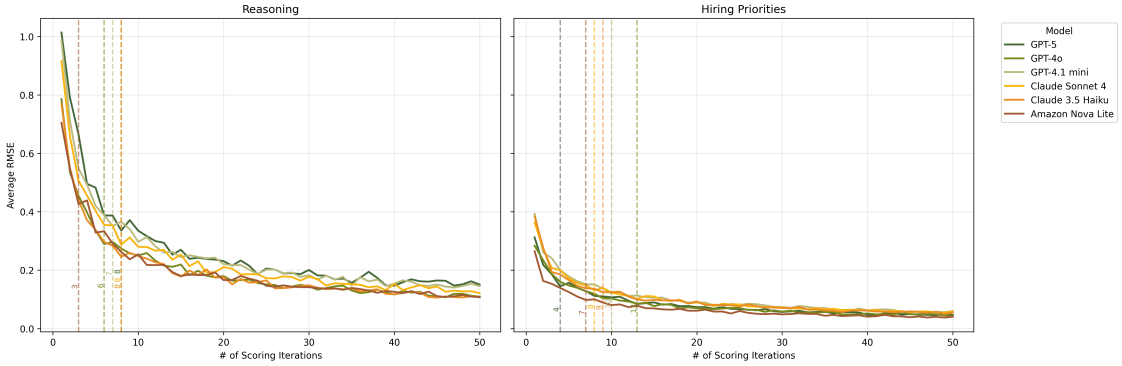
	Reasoning			Hiring Priorities		
	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT
GPT-5	0.20	0.20	0.20	0.13	0.11	0.11
GPT-4o	0.35	0.33	0.31	0.19	0.22	0.19
GPT-4.1 mini	0.36	0.34	0.37	0.29	0.25	0.23
Claude Sonnet 4	0.23	0.26	0.22	0.18	0.19	0.14
Claude 3.5 Haiku	0.30	0.28	0.31	0.28	0.26	0.26
Amazon Nova Lite	0.42	0.42	0.43	0.25	0.26	0.26

N=93. Each cell represents the mean entropy of item-level score distributions for a given model and prompt type across rubric categories. Lower entropy values indicate greater scoring consistency across iterations and alignment with humans. CoT stands for chain-of-thought prompting.

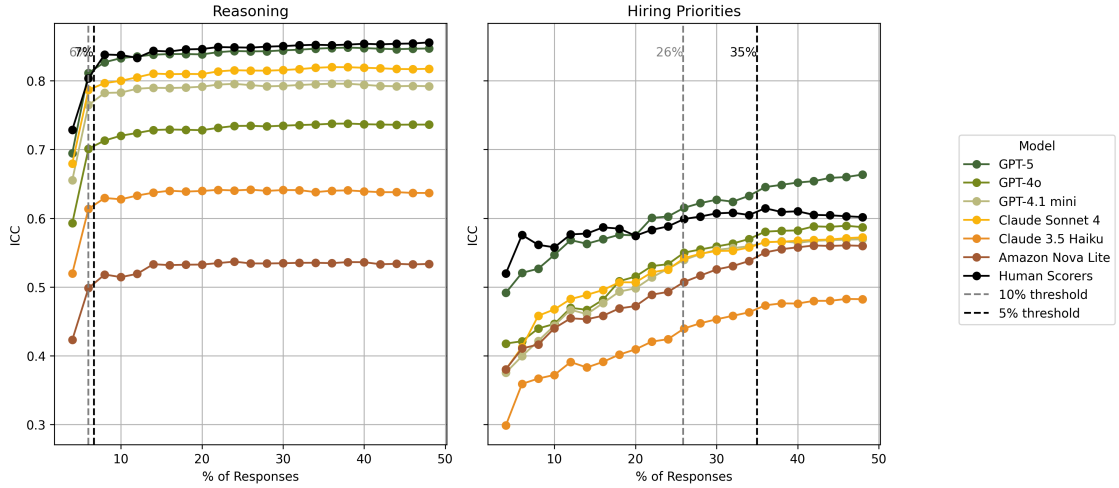
Appendix G: Design and Evaluation Considerations

Figure G.1: Design Considerations (Interview)

Panel 1: Average RMSE vs. # of Scoring Iterations



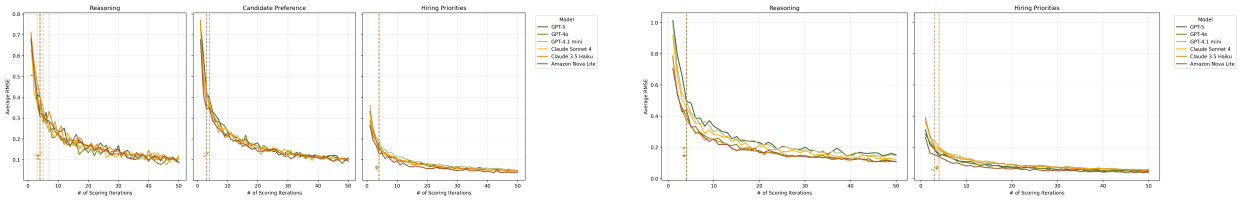
Panel 2: ICC vs. % of Responses



Panel 1 Notes: N=31 (training set size). RMSE is not exactly 0 when the number of scoring iterations is 50 due to random subsamples being generated with replacement. Subsamples are generated with replacement given that not all models' 50 scoring iterations returned correctly formatted scores. Vertical lines represent the point on the normalized curve with the furthest distance from the line y = -x + 1 (kneedle method) (Satopaa et al., 2011)

Panel 2 Notes: N=93 (testing set size). The % of responses does not start at 1% in order to ensure a high enough number of degrees of freedom to calculate ICCs, confidence intervals, and perform hypothesis testing. The 10% and 5% thresholds indicate at what % of responses the ICC estimates are within 10% and 5% of the full-sample ICC value, respectively. Graphs represent average ICCs across all prompting methods. We demonstrate these same graphs broken down by prompting method in Appendix G, though note no systematic threshold differences by prompting method.

Figure G.2: RMSE vs. # of Scoring Iterations - Second Derivative Threshold

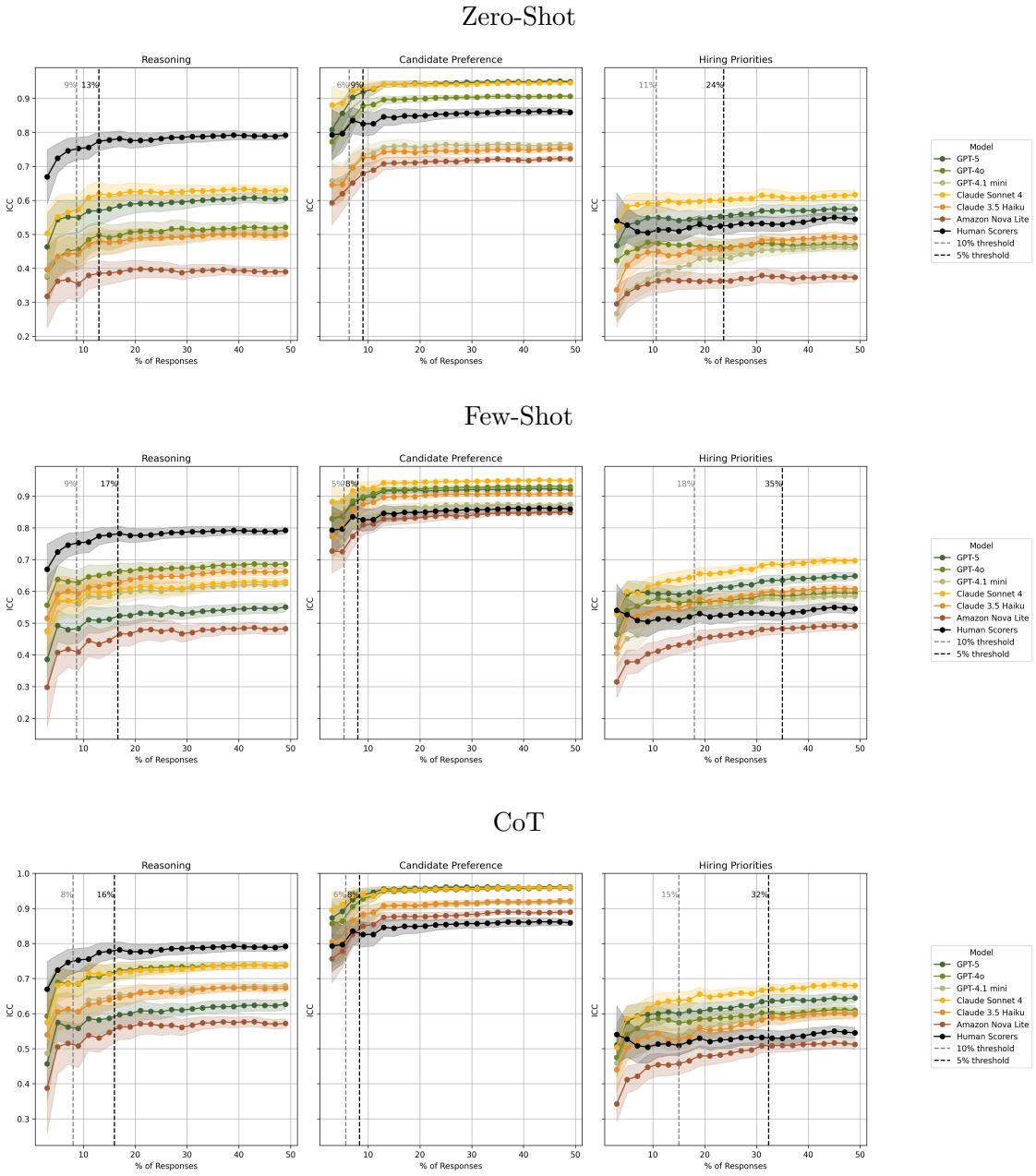


(1) Survey Short-Response (n=141)

(2) Interview (n=93)

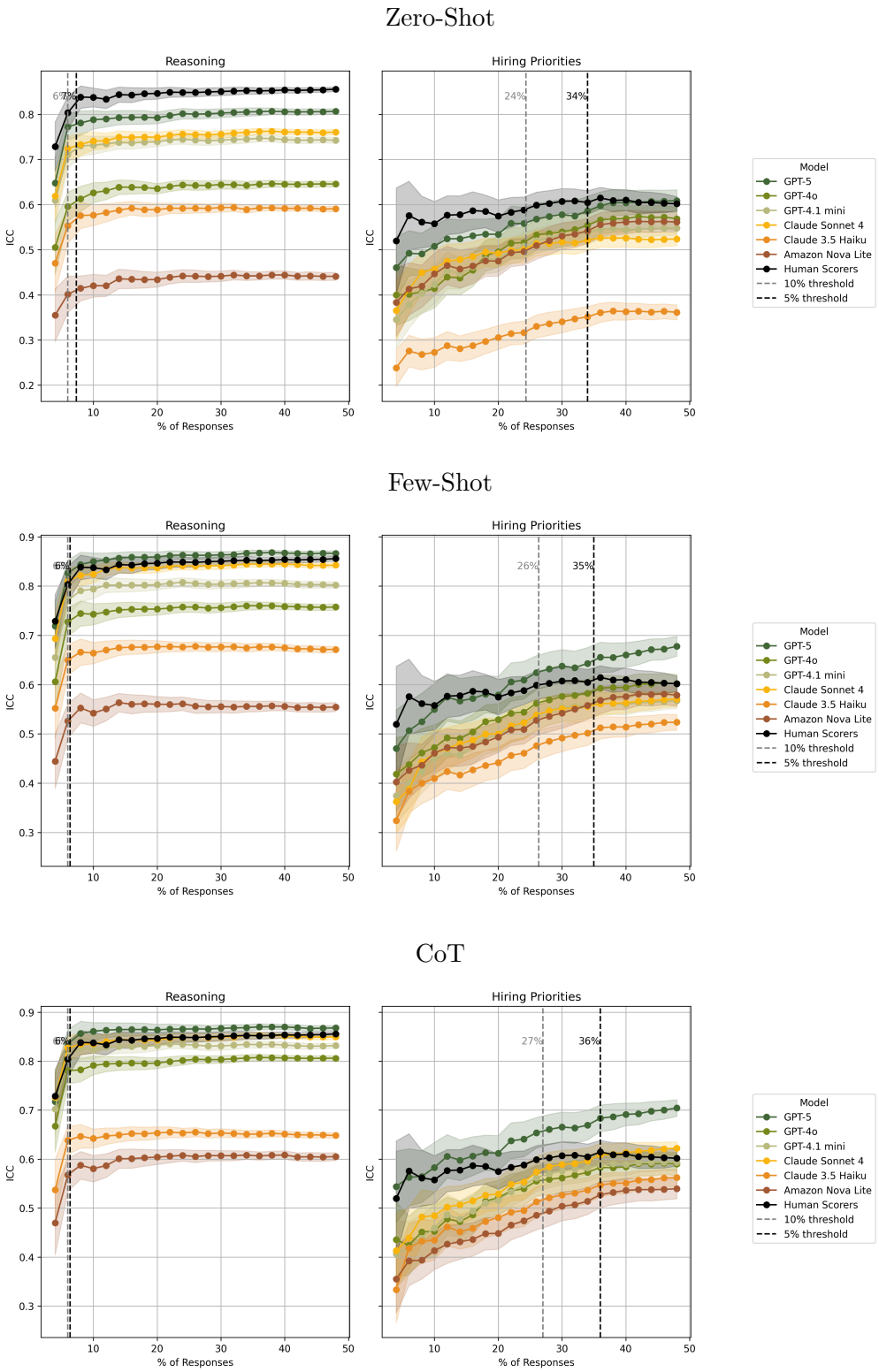
Notes: RMSE is not exactly 0 when the number of scoring iterations is 50 due to random subsamples being generated with replacement. Subsamples are generated with replacement given that not all models' 50 scoring iterations returned correctly formatted scores. Vertical lines represent the point of maximum curvature of the graphs, as determined by smoothing curves and calculating local second derivatives.

Figure G.3: ICC vs. % of Responses (by Prompting Method) (Survey Short-Response)



Notes: N=141. The % of responses does not start at 1% in order to ensure a high enough number of degrees of freedom to calculate ICCs, the confidence intervals for ICCs, and perform hypothesis testing. The 10% and 5% thresholds indicate at what % of responses the ICC estimates are within 10% and 5% of the full-sample ICC value, respectively.

Figure G.4: ICC vs. % of Responses (by Prompting Method) (Interview)



Notes: N=93. The % of responses does not start at 1% in order to ensure a high enough number of degrees of freedom to calculate ICCs, the confidence intervals for ICCs, and perform hypothesis testing. The 10% and 5% thresholds indicate at what % of responses the ICC estimates are within 10% and 5% of the full-sample ICC value, respectively.

Appendix H: Validation Exercise of Algorithmic Bias

Table H.1: Survey - Interaction Coefficients

		Reasoning	Candidate Preference				Hiring Priorities				
		Reasoning	Pref. for Candidate 1	Pref. for Candidate 2	Alternative Response	Cultural Responsiveness	Parent/Community Engagement	Academic Achievement	Candidate Experience/Expertise	Evaluation	School Culture Fit
Race/Ethnicity	Overall	0.02 (0.23)	0.03 (0.05)	-0.03 (0.06)	0.41 (0.37)	-0.26 (0.43)	0.25 (0.45)	0.99** (0.46)	0.33 (0.59)	0.31 (0.45)	0.90* (0.47)
	GPT-5	0.11 (0.25)	0.01 (0.05)	-0.01 (0.04)	0.40 (0.39)	-0.20 (0.41)	0.01 (0.43)	1.25*** (0.49)	0.29 (0.83)	0.25 (0.43)	0.57 (0.51)
	GPT-4o	-0.00 (0.23)	0.03 (0.05)	0.00 (0.05)	0.66* (0.39)	0.42 (0.35)	0.27 (0.54)	1.14** (0.51)	0.25 (0.84)	0.46 (0.44)	1.25** (0.61)
	GPT-4.1 mini	-0.01 (0.23)	0.06 (0.08)	-0.03 (0.06)	0.21 (0.39)	-0.65 (0.58)	0.93 (0.57)	1.26* (0.72)	0.63 (0.83)	0.58 (0.52)	0.68 (0.48)
	Claude Sonnet 4	0.08 (0.23)	-0.03 (0.05)	-0.08 (0.07)	0.54 (0.42)	0.17 (0.50)	-0.54** (0.27)	0.78 (0.54)	0.30 (0.39)	0.35 (0.49)	0.95 (0.63)
	Claude 3.5 Haiku	-0.00 (0.22)	0.03 (0.08)	-0.06 (0.08)	0.23 (0.31)	-0.60 (0.38)	0.18 (0.65)	1.07 (1.10)	-0.32 (0.83)	0.30 (0.62)	1.06* (0.63)
	Amazon Nova Lite	-0.05 (0.24)	0.07 (0.08)	-0.02 (0.09)	0.40 (0.46)	-0.06 (0.64)	0.70 (0.64)	0.47 (0.56)	0.35 (0.85)	0.15 (0.48)	1.09* (0.60)
Sex	Overall	-0.08 (0.12)	0.02 (0.05)	-0.05 (0.04)	0.18 (0.18)	0.43 (0.46)	0.35 (0.29)	0.29 (0.27)	-1.43 (1.02)	0.20 (0.29)	0.52** (0.26)
	GPT-5	-0.05 (0.13)	0.05 (0.05)	-0.04 (0.04)	0.25 (0.19)	0.81 (0.59)	0.12 (0.28)	0.40 (0.28)	0.08 (0.34)	0.09 (0.29)	0.68** (0.30)
	GPT-4o	-0.08 (0.12)	0.04 (0.05)	-0.04 (0.04)	-0.01 (0.19)	0.19 (0.45)	0.48 (0.32)	0.27 (0.36)	-0.35 (1.49)	0.23 (0.29)	0.70** (0.31)
	GPT-4.1 mini	-0.08 (0.13)	-0.00 (0.06)	-0.05 (0.05)	0.32 (0.22)	0.34 (0.58)	0.45 (0.33)	0.04 (0.38)	-1.62 (1.15)	0.08 (0.38)	0.42 (0.30)
	Claude Sonnet 4	-0.08 (0.12)	0.03 (0.05)	-0.05 (0.03)	-0.02 (0.18)	0.86** (0.39)	0.10 (0.29)	0.40 (0.30)	-1.8* (1.06)	0.10 (0.31)	0.63** (0.29)
	Claude 3.5 Haiku	-0.12 (0.13)	-0.01 (0.05)	-0.04 (0.04)	0.18 (0.20)	0.16 (0.50)	0.20 (0.36)	0.39 (0.49)	-0.56 (1.29)	0.35 (0.39)	0.48 (0.32)
	Amazon Nova Lite	-0.05 (0.14)	0.02 (0.06)	-0.08* (0.05)	0.32 (0.25)	0.52 (0.63)	0.63 (0.39)	0.33 (0.42)	-1.75* (1.06)	0.34 (0.31)	0.29 (0.30)

Notes: For each rubric item, we regress item score on model type (human average is the reference point for comparison), the demographic characteristic of interest (we run regressions separately for binary indicators of White/Non-White and Male/Female), and the interaction between the two. For the Race/Ethnicity regressions, the reference group is White. For the Sex regressions, the reference group is Male. The 0-1 rubric items were run with logistic regression. Interaction terms reveal whether models score responses differentially by race/sex, above and beyond how the human average differentially scored responses. Given the interaction coefficients are small and not statistically significant, this provides evidence that models and humans do not differ in how they score responses by race/ethnicity and sex. This indicates that models may be replicating any existing human biases, but not exacerbating such biases.

Table H.2: Interview - Interaction Coefficients

		Reasoning					Hiring Priorities				
		Engaging with Evidence	Intensity of Recruiting Strategies	Collaborative Decision-Making	Responsiveness to Local Labor Conditions	Cultural Responsiveness	Parent/Community Engagement	Academic Achievement	Candidate Experience/Expertise	Evaluation	School Culture Fit
Race/Ethnicity	Overall	-0.28* (0.16)	0.19 (0.19)	0.11 (0.23)	0.33** (0.16)	-0.12 (0.60)	0 (omitted)	-0.25 (0.43)	-0.93 (0.63)	0.84* (0.49)	-0.38 (0.67)
	GPT-5	-0.05 (0.16)	0.10 (0.10)	0.12 (0.22)	0.07 (0.11)	0.51 (0.63)	-0.12 (0.71)	-0.59** (0.29)	-1.43** (0.57)	0.84 (0.58)	-0.47 (0.54)
	GPT-4o	-0.38** (0.18)	0.19 (0.20)	0.21 (0.25)	0.43** (0.20)	0.72** (0.30)	0 (omitted)	-0.07 (0.48)	-1.11 (0.71)	0.57 (0.39)	-0.52 (1.02)
	GPT-4.1 mini	-0.39** (0.17)	0.13 (0.20)	0.03 (0.29)	0.32* (0.16)	-0.91 (0.60)	-1.42 (0.92)	-0.02 (0.39)	-0.94 (0.76)	0.88 (0.55)	-0.45 (1.03)
	Claude Sonnet 4	-0.18 (0.15)	0.13 (0.13)	0.12 (0.24)	0.14 (0.12)	0.07 (0.78)	0 (omitted)	-0.43 (0.49)	-0.86 (0.82)	1.12 (0.82)	0.25 (0.86)
	Claude 3.5 Haiku	-0.43** (0.20)	0.26 (0.29)	0.06 (0.24)	0.34* (0.19)	-0.00 (0.88)	-0.66 (0.86)	-0.36 (0.70)	-0.64 (0.81)	0.85 (0.81)	-1.46 (1.11)
	Amazon Nova Lite	-0.25 (0.21)	0.35 (0.29)	0.12 (0.27)	0.68*** (0.25)	-0.43 (0.65)	0.19 (0.90)	-0.05 (0.65)	-1.20** (0.52)	1.02* (0.62)	-0.24 (0.98)
Sex	Overall	-0.02 (0.16)	0.04 (0.16)	0.03 (0.14)	-0.05 (0.16)	0 (omitted)	0.19 (0.87)	0.92*** (0.35)	-0.96* (0.55)	0.44 (0.29)	-0.35 (0.45)
	GPT-5	0.04 (0.15)	0.07 (0.13)	-0.06 (0.16)	-0.00 (0.15)	0.64 (0.85)	0.38 (0.86)	0.50 (0.36)	-1.04* (0.53)	0.83*** (0.32)	-0.37 (0.52)
	GPT-4o	-0.08 (0.18)	0.09 (0.17)	0.05 (0.15)	-0.10 (0.18)	1.29 (0.87)	0.30 (0.93)	0.60 (0.42)	-1.12* (0.60)	0.10 (0.32)	-0.38 (0.70)
	GPT-4.1 mini	-0.02 (0.15)	0.04 (0.16)	0.10 (0.16)	-0.07 (0.17)	-0.05 (0.54)	-0.14 (0.96)	0.69* (0.38)	-0.86 (0.68)	0.39 (0.35)	-0.74 (0.78)
	Claude Sonnet 4	-0.07 (0.15)	0.09 (0.15)	-0.10 (0.15)	-0.01 (0.17)	0 (omitted)	0.77 (0.88)	1.20*** (0.41)	-0.92 (0.60)	0.13 (0.39)	0.07 (0.55)
	Claude 3.5 Haiku	-0.04 (0.20)	-0.03 (0.22)	0.04 (0.17)	-0.11 (0.18)	-0.21 (0.53)	-0.22 (1.00)	1.22** (0.48)	-1.31** (0.64)	0.37 (0.42)	0 (omitted)
	Amazon Nova Lite	0.04 (0.21)	0.01 (0.23)	0.15 (0.17)	-0.01 (0.24)	1.40 (1.19)	0.56 (0.96)	1.10** (0.46)	-0.89 (0.58)	0.72** (0.32)	-0.07 (0.62)

Notes: For each rubric item, we regress item score on model type (human average is the reference point for comparison), the demographic characteristic of interest (we run regressions separately for binary indicators of White/Non-White and Male/Female), and the interaction between the two. For the Race/Ethnicity regressions, the reference group is White. For the Sex regressions, the reference group is Male. The 0-1 rubric items were run with logistic regression. Interaction terms reveal whether models score responses differentially by race/sex, above and beyond how the human average differentially scored responses. Given the interaction coefficients are small and not statistically significant, this provides evidence that models and humans do not differ in how they score responses by race/ethnicity and sex. This indicates that models may be replicating any existing human biases, but not exacerbating such biases.