Principled Design of Interpretable Automated Scoring for Large-Scale Educational Assessments

Yunsung Kim, **Mike Hardy**, **Joseph Tey**, **Candace Thille***, and **Chris Piech***

Stanford University
yunsung@stanford.edu

Abstract

AI-driven automated scoring systems offer scalable and efficient means of evaluating complex student-generated responses. Yet, despite increasing demand for transparency and interpretability, the field has yet to develop a widely accepted solution for interpretable automated scoring to be used in large-scale real-world assessments. This work takes a principled approach to address this challenge. We analyze the needs and potential benefits of interpretable automated scoring for various assessment stakeholders and develop four principles of interpretability – **F**aithfulness, **G**roundedness, **T**raceability, and **I**nterchangeability (**FGTI**) – targeted at those needs. To illustrate the feasibility of implementing these principles, we develop the ANALYTICSCORE framework for short answer scoring as a baseline reference framework for future research. ANALYTICSCORE operates by (1) extracting explicitly identifiable elements of the responses, (2) featurizing each response into human-interpretable values using LLMs, and (3) applying an intuitive ordinal logistic regression model for scoring. In terms of scoring accuracy, ANALYTICSCORE outperforms many uninterpretable scoring methods, and is within only 0.06 QWK of the uninterpretable SOTA on average across 10 items from the ASAP-SAS dataset. By comparing against human annotators conducting the same featurization task, we further demonstrate that the featurization behavior of ANALYTICSCORE aligns well with that of humans.

1 Introduction

Accurate and credible assessment of knowledge and skills forms the basis for effective decision making in a variety of educational contexts, from student learning and instructional design to program development and policy making (Berman et al., 2019). When the set of knowledge and skills to be gauged involves complex, open-ended problem-solving and communication abilities, AI-driven *automated scoring systems* can offer rapid, accessible, and scalable alternatives to the otherwise labor-intensive and costly process of training and deploying human scorers (Foltz et al., 2020). Automated scoring systems have been increasingly adopted across various assessment contexts over the past several decades. Today's scoring algorithms achieve acceptable levels of scoring accuracy in various areas of human learning (Whitmer and Beiting-Parrish, 2023, 2024).

Despite progress, automated scoring of open-ended responses has yet to reliably obtain generalizable scoring accuracy across diverse scoring contexts. Even when automated scoring meets acceptable levels of scoring accuracy, errors or biases inherent in the scoring algorithm can profoundly harm student learning and equity, policy evaluation, and public trust (Berman et al., 2019, Pellegrino, 2022). For these reasons, improving transparency and interpretability in automated scoring has now

* Equal Advising

become a moral imperative, not a mere technical preference (Holmes et al., 2022, Khosravi et al., 2022, Memarian and Doleck, 2023, Schlippe et al., 2022). Yet, in spite of the growing research on interpretable and explainable AI as well as its applications specifically within educational assessment, interpretable automated scoring remains mostly confined to academic research with limited adoption in large-scale, real-world assessment (Institute of Education Statistics, 2023, Whitmer and Beiting-Parrish, 2023, 2024).

In this paper, we take a principled approach towards building a practical interpretable automated scoring solution for large-scale assessments. An effective interpretability solution begins by identifying the diverse needs of each stakeholder in understanding the system's decisions, and by grounding the development of interpretable AI systems in those needs (Bhatt et al., 2020, Páez, 2019, Preece et al., 2018). Research on explainable automated scoring, on the other hand, has largely ignored this need-finding process. As we observe later in this paper (Section 2), this neglect has often led to several claimed interpretability solutions that fail to address the diverse and nuanced interpretability needs of the human actors in educational assessment.

We identify the needs and benefits of model explanations for various large-scale assessment stakeholder groups consisting of test takers, assessment developers, and test users (Section 2.1). Targeted at those needs, we develop the principles of **faithful**, **grounded**, **traceable**, and **interchangeable** model interpretations for AI-driven automated scoring (Section 2.2).

We further illustrate the feasibility of implementing these principles in practice and establish a concrete baseline for future work (Section 3). **ANALYTICSCORE** is the first interpretable automated short-answer scoring framework to embody our principles. It operates by extracting explicitly identifiable elements from unannotated response texts and featurizing each response into human-interpretable values based on those elements. These features are input to an intuitive ordinal logistic regression module for scoring.

We measure the performance of ANALYTICSCORE on a real-world response dataset by measuring (1) scoring accuracy and (2) alignment of featurization behaviors with human judgments (Sections 4 and 5). ANALYTICSCORE outperforms many uninterpretable scoring methods, and is within only 0.06 QWK of the uninterpretable SOTA on average across 10 items from the ASAP-SAS dataset. The featurization behavior of ANALYTICSCORE also aligns well with humans (0.90, 0.72, 0.81 QWK across assessment areas). Our findings indicate strong potential for implementing accurate and well-aligned interpretability solutions that meet the real needs of assessment stakeholders.

Automated Scoring and Interpretable AI As AI-driven automated scoring systems became more complex and opaque, researchers have increasingly noted the need to enhance the transparency of these systems through model explanations (Bauer and Zapata-Rivera, 2020, Bennett and Zhang, 2015, Schlippe et al., 2022). Several approaches have been proposed to address interpretable automated scoring, and we discuss them in detail in Section 2.2 in connection with our four principles. Despite growing research interests, interpretable automated scoring still lacks practical adoption and meaningful field use. The 2023 NAEP Math Automated Scoring Challenge[1] for open-ended math responses organized by the US National Center for Education Statistics (NCES) found that none of the submissions met the criteria for "interpretability" despite several methods achieving near-human scoring accuracy (Institute of Education Statistics, 2023, Whitmer and Beiting-Parrish, 2024). This gap highlights the need for human-centered explainability solutions driven by assessment stakeholders' real needs.

[1] https://github.com/NAEP-AS-Challenge/math-prediction

2 Building the Principles of Interpretable Automated Scoring

Insights derived from scoring support various stakeholders throughout the overall assessment process. Below we analyze three main stakeholder groups in large-scale assessment – test takers, assessment developers, and test users (AERA et al., 2014, Berman et al., 2019). Each stakeholder group's distinct roles and priorities uniquely shape how interpretable automated scoring can address their specific needs and improve their assessment experience.

2.1 Explainability Needs and Potential Benefits

Test Takers The needs and benefits of interpretable scoring vary depending on the assessment type: summative or formative. Most large-scale assessments are summative assessments, which are assessments *of* learning that support evaluating learner achievement, assigning grades, or determining proficiency levels (Harlen, 2005). Because these assessments often drive high-stakes decisions, test takers need to trust the fairness and justifiability of scoring decisions (Williamson et al., 2012). Provided that the scoring algorithm implements sound scoring logic, allowing test takers or their representatives to examine traceable explanations for scoring decisions can foster trust (Bauer and Zapata-Rivera, 2020, Ferrara and Qunbar, 2022). These explanations can also support a streamlined quality control process by facilitating the identification and correction of errors, improving the overall integrity of the assessment (see Bennett and Zhang (2015) and Ferrara and Qunbar (2022)).

Formative assessments are assessments *for* learning, intended to guide and improve learner performance through frequent practice, progress monitoring, and skill diagnosis (Black and Wiliam, 1998, Wiliam, 2011). In this context, the function of automated scoring is primarily to provide timely, effective and actionable feedback to support learner learning (Bennett, 2006, DiCerbo et al., 2020). Effective feedback should help learners understand the discrepancy between their work and a desired outcome (Schwartz et al., 2016). A step-by-step explanation of the features observed in a learner's work, coupled with human understandable descriptions of how those features were processed can be used to provide such elaborative feedback.

Assessment Developers Scoring algorithms should reliably identify evidence of the constructs (target knowledge, skills, and abilities) measured by the task (Bejar et al., 2016). Understanding the types of evidence that an automated scoring algorithm reliably detects also informs other key aspects of assessment design, such as construct selection and task design (Bennett and Bejar, 1998). Model explanations can facilitate this understanding by transparently revealing the features used by the scoring algorithm and its intermediate reasoning steps. Explanations can also help determine which parts of the algorithm can be reused, avoiding the costly and time-consuming process of training a new scoring algorithm for each new task (see DiCerbo et al. (2020)).

Model explanations also yield specific insights into areas where the scoring model can be improved and how. Scoring models often need to be tuned for various reasons. For instance, models trained on data may reflect biases related to response strategies specific to student groups (Ferrara and Qunbar, 2022, Rupp, 2018). Scoring models may also become less stable over time as the test-taker population and/or scoring criteria change (Bejar et al., 2016). Transparent inspection of model decisions helps identify problematic model elements, enabling targeted data collection and modified training objectives to improve the model.

Test Users Test users, including professionals who select and administer tests, educators, administrators, and policymakers, depend on score reliability and interpretation validity to make

system-level decisions or instructional differentiation. Their reliance on the integrity and validity of scores to drive decisions is significant (AERA et al., 2014). Model explanations provide concrete evidence to validate the choice of the scoring model[2]. This includes understanding whether the extracted features and scoring logic fully capture the rubric and the construct definition, and whether the internal structure of the automated scores align with the construct of interest (Bennett and Zhang, 2015).

2.2 The FGTI Interpretability Principles

We develop four foundational interpretability principles – **F**aithful, **G**rounded, **T**raceable, and **I**nterchangeable (**FGTI**) – targeting the needs and benefits of large-scale assessment stakeholders from Section 2.1. Our first foundational principle is that explanations should be *faithful* (Jacovi and Goldberg, 2020). Faithfulness is an important requirement in many high-stakes applications of interpretable AI (Rudin, 2019). Similar expectations extend to assessments, and all of the needs and benefits outlined in Section 2.1 depend crucially on faithfulness.

Principle 1 (Faithful)**.** *Explanations of scoring decisions should accurately reflect the computational mechanism behind the scoring model's prediction.*

A notable example of **un**faithful scoring explanations are texts produced by prompting LLMs to generate an explanation (e.g., Lee et al. (2024) and Li et al. (2025)). Stepwise reasoning verbalized by LLM through prompting strategies such as chain-of-thought (Wei et al., 2022) are not explanations of their internal computation (Sarkar, 2024) and often fail to reflect the model's true reasoning behavior (Arcuschin et al., 2025, Turpin et al., 2023). Moreover, LLMs are highly sensitive to superficial changes in prompts and input text, frequently exhibiting inconsistent judgments (Wang et al., 2024). Therefore, even when LLMs achieve high scoring accuracy, prompting them for explanations cannot reliably address the stakeholder needs identified in Section 2.1.

Next, the model should use meaningful features that are explicitly linked to each student's work and rely only on those features for the downstream computation.

Principle 2 (Grounded)**.** *Initial features computed by the scoring model should represent human-understandable, explicitly identifiable elements of student work and item task.*

Regardless of the routine used to derive those features, these feature values should possess meaning that is understandable to humans and be explicitly based both in the student work and item task. For instance, cosine similarity of sentence embeddings used as an input feature (e.g., Condor and Pardos (2024)) is less human-understandable than discrete features whose values are associated with clear, verbalizable meaning. Having features that are grounded addresses the need to scrutinize the features of the scoring engines.

How should the model process these features to ultimately produce a final score? Scoring is inherently an evidentiary reasoning process, where elements of the student response and item tasks serve as evidence to support the inference about student knowledge and skills that the score represents (DiCerbo et al., 2020, Mislevy, 2020). Stakeholders need to be able to inspect and interact with the internal structure of the scoring model to ensure soundness, construct-relevance, fairness, and model understanding (Section 2.1). To meet this need, the model's evidentiary reasoning process must be decomposable into clear, sequential steps that a human could reliably execute and

[2]More examples of validity arguments on the use of automated scoring can be found in (Bennett and Zhang, 2015, Table 7.7).

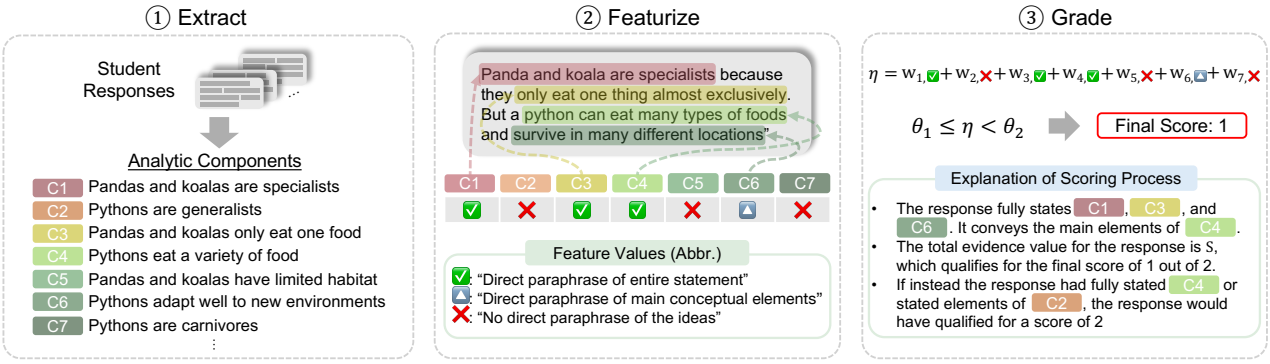

Figure 1: Schematic of the ANALYTICSCORE framework. The example question is: **"Explain how pandas in China and koalas in Australia are similar, and how they both are different from pythons."**

possibly intervene. Our next 2 principles state that the scoring model should be conducive to this decomposition and intervention:

Principle 3 (Traceable)**.** *The scoring model should consist of subroutines that each represent a specific, well-defined evidentiary reasoning step on clearly specified inputs.*

Principle 4 (Interchangeable)**.** *A human should be able to act interchangeably on each of the reasoning subroutines.*

Not all intermediate representations calculated by the model need to be understandable by human, but the reasoning subroutines should collectively account for the entire scoring logic. Moreover, humans should be able to act interchangeably with each decomposed module and replace the module outputs with human-generated results if deemed necessary.

Many proposed interpretability approaches are not grounded, traceable, or interchangeable in the sense described above. These include, for instance, calculating feature importance values (Asazuma et al., 2023, Kumar and Boulanger, 2020, 2021, Schlippe et al., 2022), displaying feature attribution maps (Li et al., 2025, Schlippe et al., 2022), or presenting confidence metrics for scoring decisions (Conijn et al., 2023). This limits the capacity to thoroughly inspect the model's features and internal structure, which is critical to meeting the needs of the stakeholders.

3 A Principled Framework for Interpretable Automated Scoring

How would the FGTI principles be implemented in practice? To illustrate the feasibility of implementing these principles and to set a baseline for future research, we present ANALYTICSCORE as a reference framework in the domain of *short-answer scoring*. In this setting, students write a short 1-5 sentence answer in response to an assessment item which is scored with an emphasis on content correctness and demonstrated reasoning (Leacock and Chodorow, 2003, Shermis, 2015). The scoring model has access to a training set of student response texts paired with human-annotated scores $(r_1, s_1), ..., (r_n, s_n)$ and possibly additional unannotated responses $\{r_{n+1}, ..., r_m\}$. The goal at inference time is to predict the score s for a new response r.

ANALYTICSCORE (Figure 1) is a 3-phase, LLM-based framework grounded in our four principles of interpretable automated scoring. Phase 1 identifies explicitly grounded *analytic components* to be used. Phase 2 catalogs, or *featurizes*, the presence of these components in student responses. Phase 3

uses the features to compute a score. Phases 1 and 2 depend only on the response texts without any annotations. Human score labels are only used with Phase 3.

3.1 Phase 1: Extracting Analytic Components.

With the response texts from the training set (and optionally assessment content), ANALYTICSCORE first extracts a set of **analytic components**, which are explicitly identifiable elements of student responses described in Principle 2. In this work, we consider a specific type of components which are representative, atomic units of explicit statements, arguments, or claims as in Figure 1.

$$[c_1, ..., c_k] = \text{Extract}(r_1, ..., r_m),$$

Component extraction is implemented using an LLM with the prompts shown in Figure 2. Having too many analytic components could diminish the interchangeability (Principle 4) of the overall framework by exploding the number of features used in scoring (Lipton, 2018), so we limit to generating 15 components per request.

3.2 Phase 2: Featurizing Responses

Once the analytic components have been identified, student responses are featurized according to the presence of these components $c_1, ..., c_m$ in each response r. This step uses a labeling function $f(r; c)$ whose outputs are associated with human-understandable meaning (Principles 2 & 3). The exact label definitions used can be selected using natural language. In this work, we explore the following general purpose labeling function for $f(r; c)$:

$$f(r; c) = \begin{cases} 2, \text{ if } r \text{ contains direct paraphrase of } c \\ 1, \text{ if } r \text{ contains partial paraphrase of } c \\ 0, \text{ if } r \text{ does not contain paraphrase of } c \end{cases} \quad (1)$$

We implement $f(r; c)$ using a Chain-of-Thought (Wei et al., 2022) prompting template shown in Figure 2[3]. Inspired by the self-consistency decoding strategy for LLMs (Wang et al., 2022), we apply the first-to-three aggregation rule to consider the possibly diverse interpretation of the labeling criteria when selecting the final output. Easily interpretable one-hot encodings of each $f(r; c_m)$ are then concatenated to produce a $3m$-dimensional binary featurization of r:

$$\text{Featurize}(r) = \text{OneHot}(f(r; c_1)) \,\|\, \cdots \,\|\, \text{OneHot}(f(r; c_k))$$

Distilling LLM Featurizer into Open Source Using proprietary LLMs for featurization can quickly become too expensive in large-scale assessment settings, especially with many analytic components. To avoid the linearly growing cost of featurization, we supervised fine-tuned a small open-source model using a subsample of (r, c) pairs, where r is a response from the training set and c is an analytic component from Phase 1. More specifically, we randomly sampled 10k pairs across all 10 items, calculated the featurization labels on these samples using o4-mini, and collected the full LLM model requests and outputs generated during this process which aligned with the aggregated final decision. This dataset was used to fine-tune Llama-3.1-8b-instruct with QLoRA (Dettmers et al., 2023).

[3]Note that we CoT is used solely as a prompting technique, and the generated "thoughts" are explicitly discarded.

Component Extractor Prompt

Students were asked the following question: ```[QUESTION PROMPT]```

Here are several examples of student responses to the question:

Student Response: [RESPONSE TEXT] x 1000

Please tell me 15 short, simple, and representative statement, claims, or arguments that are common across many student responses and that distinguish these student responses from others. Each of these "features" should identify an actual statement or claim made in the student responses, not a general description of it.

When identifying statements or claims, adhere to the following guidelines:
- Use the specific wording mentioned in the student response when it is crucial to distinguishing between responses.
- Ensure that each statement or claim is **atomic and isolated**. This means that the statements and claims should **not combine** multiple ideas, contexts, or supporting details. For example, instead of:
  * "[Statement A], because [Statement B]", or
  * "[Statement A], and [Statement B]", or
  * "[Statement A], so [Statement B]" etc.,

"[Statement A]" and "[Statement B]" should be listed separately.

Featurizer Prompt

Students were asked the following question: ```[QUESTION PROMPT]```

Here is a response from a student: ```[RESPONSE]```

Here is a statement: ```[ANALYTIC COMPONENT]```

Choose among the following:
- A: The text contains a *direct* paraphrase of the given statement using clearly synonymous wording.
- B: The text does not contain a direct paraphrase of the entire statement, but it contains direct paraphrases of the main parts of the statement.
- C: The text may potentially imply the given statement, but it does not contain a direct paraphrase of the ideas in the statement.

Limit your answer to 100 words, and format your answer in the following python dictionary format:
{
  "Explanation": "[Why the text disqualifies for other options]",
  "Answer": "A"/"B"/"C"
}

Figure 2: Prompts used in ANALYTICSCORE

3.3 Phase 3: Logically Traceable Scoring

Based on the featurized responses, a traceable and interchangeable model (Principle 3 and 4) is selected and trained using the labeled response pairs $(r_1, s_1), ..., (r_n, s_n)$. Given the nature of the score categories, we employ the Immediate-Threshold variant of Ordinal Logistic Regression (Pedregosa et al., 2017, Rennie and Srebro, 2005) as our scoring module. Combined with the one-hot encoding featurization from Phase 2, the resulting algorithm calculates the sum of weights for each component and feature label: $\eta = \sum_{i=1}^{c} w_{i,f(r,c_i)}$, where w are the trained weights. Scores are determined by comparing η to a set of learned thresholds θ_j; the predicted score corresponds to the ordinal category j for which $\theta_j \leq \eta < \theta_{j+1}$. η can be understood as "evidence values" used for scoring.

3.4 Analysis of ANALYTICSCORE

An example of ANALYTICSCORE's model explanation is shown in the right panel of Figure 1. By demonstrating human-understandable features of the response (Principle 2) and the exact decision process (Principle 3), the explanation transparently and faithfully reveals the actual scoring mechanism used (Principle 1). If, based on the explanation, the model is suspected to have made an error (e.g., C6 should be a check, not a triangle), a human inspector can modify the featurization and rerun the scoring algorithm (Principle 4), which is also how the "if instead..." explanation is generated.

The structure of ANALYTICSCORE's scoring model is akin to Concept Bottleneck Models (Koh et al., 2020, Yang et al., 2023) in that we enforce a layer of intermediate representations with human-understandable "concepts." Our framework ensures that the intermediate features have human-understandable values that are associated with explicitly identifiable elements (Principle 2), as opposed to characteristics that are inferred from the response.

	Token Len.	Train	Valid	Test	Assessment Area
Q1	47.5 ± 22.2	1,341	331	557	Science
Q2	59.2 ± 22.6	1,024	254	426	
Q3	47.9 ± 14.6	1,445	363	406	Reading (Informational Text)
Q4	40.3 ± 15.5	1,308	349	295	
Q5	25.1 ± 21.5	1,459	336	598	Science
Q6	23.8 ± 22.6	1,418	379	599	
Q7	41.3 ± 25.1	1,432	367	599	Reading (Literature)
Q8	53.0 ± 32.6	1,446	353	599	
Q9	49.7 ± 36.3	1,453	345	599	Reading (Informational Text)
Q10	41.1 ± 28.5	1,314	326	546	Science

Table 1: ASAP-SAS dataset detail by item.

4 Evaluating AnalyticScore

Having introduced AnalyticScore and discussed its interpretability, we now evaluate its scoring performance and how its featurization aligns with human judgments on a real-world response scoring dataset.

Dataset The ASAP-SAS dataset (Shermis, 2015)[4] is the largest publicly available dataset with short answer responses from schoolchildren for 10 different open-ended exam questions. Human raters double-scored and assigned a single number to each student response using a 3 or 4 point rubric. The assessment area for each question, as well as the sample sizes and response lengths are reported in Table 1. We use the original test set and split the public training set into training and validation sets with a 8:2 ratio.

AnalyticScore Implementation Details For each assessment item, we used `GPT-4.1` as the base LLM and extracted 15 analytic components except for Q7. This item uses a two-part scoring scheme to separately assess a character trait identified from the reading and its supporting evidence. We extracted 15 analytic components from each part, totaling 30 components. For the featurizer, we experimented with GPT-4.1-mini and Llama-3.1-8B-Instruct as our base LLM, each with temperature set to 0.7 and 1.0. We distilled the Llama featurizer for 2 epochs using a batch size of 4 and learning rate of 1e-4. All model calls were made through the official OpenAI API. Fine-tuning was conducted on an Ubuntu 20.04 machine with 2 RTX A6000 GPUs (49Gb memory), 16 AMD EPYC 9224 24-Core Processors, and 250Gb of CPU RAM.

4.1 Scoring Accuracy Experiment

We measured scoring accuracy in terms of quadratic weighted kappa (QWK) against the model scores in the test set, following the convention of the automated scoring literature (Institute of Education Statistics, 2023, Shermis, 2015).

The following baseline models were compared :

[4] https://www.kaggle.com/competitions/asap-sas/data

Few-Shot Prompting: We few-shot prompt **GPT-4.1** with 10 randomly selected responses from each score category, including a rubric for the score categories.

Supervised Fine-tuned LLM: The following LLM-based classifiers were fine-tuned on the response-score pairs: **BERT** (Devlin et al., 2019), **DeBERTa** (He et al., 2020), **Llama-3.1-8b**, and **Llama-3.1-8b-Instruct** (Grattafiori et al., 2024). We also fine-tune **Llama-3.1-8B-Instruct** with a rubric of the score categories added to the input.

Automated Scorer Baselines: Methods included are: **AutoSAS** (Kumar et al., 2019), **AsRRN** (Li et al., 2023), and **NAM** (Condor and Pardos, 2024).

The only baseline method that has aspects of interpretability is NAM. This method requires hand-crafting a specific form of rubric describing the key phrases and concepts to be used by the response. Using sentence embeddings with n-gram matching as its features, this method implements a logistic regression score classifier. To implement this baseline, we replace the rubrics with the analytic components extracted by our ANALYTICSCORE.

4.2 Featurization Alignment Experiment

The feature labeling task described in Figure 2 was designed to produce human-*understandable* features (Principle 2). But how well does the LLM's featurization behavior align with how humans actually understand this task? Even more fundamentally, how well do humans themselves agree in their understanding of this task?

To answer these questions, we sampled 50 (response, analytic component) pairs for each of the 3 assessment areas. To ensure balanced representation, the sample included a balanced number of pairs from each of the three score categories, as initially determined by the GPT-4.1-mini featurizer. We then asked 7 human annotators to conduct the labeling task on these samples. The human annotators consisted of five volunteers from an R1 university and two of the study's authors. None of the annotators had prior exposure to any of the LLM's featurization outputs. All annotators had advanced academic training (PhD-level) and teaching experience, five of whom have been instructors at the primary, secondary, and/or post-secondary level.

The annotators received an oral presentation of the purpose of the study along with links to 3 Qualtrics forms to be filled out, one in each assessment area. The form reiterated the study's purpose, explained the task, and presented 50 items to annotate, each containing the context of the assessment item and the same featurizer prompt shown in Figure 2. The overall process took each annotator between 2.5 to 3.5 hours.

Aggregate Human label was generated by majority voting (ties resolved randomly). We calculated inter-rater reliability among human labelers (Krippendorff's α) and alignment between each LLM featurizer and aggregate human labels (QWK and class-wise F1). We report the 95% bootstrap CI of each metric, reweighting the sampling probability to account for the initial balanced sampling of score categories.

5 Experiment Results

5.1 Scoring Accuracy Results

Table 2 shows the results of the scoring accuracy experiments. Across items and within each assessment area, ANALYTICSCORE outperforms* several automated scoring baselines on average

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	All Avg.	Sci Avg.	R(Inf) Avg.	R(Lit) Avg.
Human	0.95	0.93	0.77	0.75	0.95	0.93	0.96	0.86	0.84	0.87	0.88±0.02	0.93±0.01	0.79±0.03	0.91±0.05
ANALYTICGRADE														
w/ GPT-4.1-mini (*)	0.80	**0.86**	0.64	0.59	0.79	0.78	0.61	0.59	0.80	0.68	0.72±0.03	0.78±0.03	0.68±0.06	0.60±0.01
w/ Llama-3.1-8b (*)	0.57	0.57	0.59	0.56	0.69	0.47	0.52	0.45	0.74	0.60	0.58±0.03	0.58±0.04	0.63±0.05	0.48±0.04
+ Distillation (*)	0.80	0.82	0.68	0.59	0.81	0.76	0.62	0.59	0.78	0.64	0.71±0.03	0.77±0.03	0.68±0.06	0.60±0.01
Fewshot														
GPT-4.1	0.69	0.65	0.61	0.65	0.72	0.61	0.34	0.57	0.76	0.69	0.63±0.04	0.67±0.02	0.68±0.04	0.45±0.12
Supervised LLM														
BERT	0.80	0.80	0.70	0.70	0.80	0.81	0.69	**0.68**	***0.84***	0.71	0.75±0.02	0.79±0.02	0.74±0.05	**0.69±0.01**
DeBERTa	0.85	**0.86**	0.66	0.70	0.81	0.83	0.71	0.64	0.79	0.71	0.76±0.03	0.81±0.03	0.72±0.04	0.67±0.04
Llama-3.1-8b Inst.	0.84	0.73	**0.72**	0.71	0.82	0.81	0.71	0.66	0.82	0.75	0.76±0.02	0.79±0.02	0.75±0.03	**0.69±0.02**
w/ rubric	**0.87**	0.80	0.68	***0.77***	**0.85**	0.80	**0.72**	0.65	***0.84***	**0.79**	**0.78±0.02**	**0.82±0.02**	**0.76±0.05**	0.68±0.04
Llama-3.1-8b	0.83	0.75	0.70	***0.77***	0.82	**0.84**	0.68	0.65	0.82	0.74	0.76±0.02	0.80±0.02	**0.76±0.03**	0.67±0.02
Baseline														
AutoSAS	0.68	0.47	0.57	0.61	0.50	0.54	0.37	0.44	0.77	0.68	0.56±0.04	0.57±0.04	0.65±0.06	0.41±0.04
ASRRN	0.60	0.43	0.57	0.60	0.61	0.64	0.59	0.51	0.71	0.66	0.59±0.02	0.59±0.04	0.63±0.04	0.55±0.04
NAM (*)	0.63	0.62	0.43	0.35	0.72	0.63	0.42	0.38	0.76	0.62	0.56±0.05	0.64±0.02	0.52±0.13	0.40±0.02

Table 2: Test-time Quadratic Weighted Kappa (QWK) of scoring models per item, along with average per assessment area. **Best**, second-best, and ***at human-level*** performance scores are marked respectively. **Sci.**: Science (Q1,2,5,6). **R(Inf)**: Reading(Informational Text) (Q3,4,9). **R(Lit)**: Reading(Literature) (Q7,8). **(*)** are methods that are considered interpretable.

and, given its interpretability, achieves reasonable performance compared to state-of-the-art black-box models. Except for the untuned Llama featurizer, each ANALYTICSCORE variant outperforms* the few-shot prompting and automated scoring baselines. Compared to the best-performing models in each assessment area, these three ANALYTICSCORE models are, on average, within 0.06 QWK over all items, 0.04 QWK for Science, 0.08 QWK for Reading (Informational Text), and 0.09 QWK for Reading (Literature) items.

Also noticeable is the striking improvement* in the performance of the Llama featurizer post-distillation, with an average increase of 0.13 QWK. The distilled Llama featurizer performs comparably to both variants of GPT-4.1 mini. Increase in average QWK is most notable for Science items (+0.19), followed by Reading (Literature) (+0.12) and Reading (Informational Text) (+0.05).

5.2 Featurization Alignment Results

Table 4 displays the Krippendorff's α[5] measured among the human raters in conducting the featurization task from Section 3.2. For all assessment areas, we observe $0.667 \leq \alpha < 0.8$, which fall into a range of acceptable inter-rater reliability (Krippendorff, 2018). We interpret this as a good level of rater agreement on the featurization process as defined in this work and acknowledge that there is still potential to refine and improve the task further.

Next, alignment between each featurizing model with the human ratings using majority vote is shown in Table 3. Most notably, the distilled Llama featurizer achieves substantially high agreement with the aggregate human features across all assessment areas. Other featurizers also achieve high

*$p < 0.05$ for Wilcoxon signed-rank test across all items. Due to small n, no area-specific difference was statistically significant.

[5]α ranges between -1 and 1. 0 indicates chance agreement.

Assessment Area	Featurizer Model	QWK	Label Distribution[6]			Label-wise F1		
			2	1	0	2	1	0
Science	**Human**		15.32%	3.70%	80.98%			
	GPT-4.1-mini	(0.89, 0.89)	7.56%	12.59%	79.85%	(0.83, 0.84)	(0.20, 0.22)	(0.96, 0.96)
	o4-mini	(0.94, 0.95)	14.35%	8.17%	77.47%	(0.93, 0.93)	(0.49, 0.51)	(0.98, 0.98)
	Llama-3.1-8B (Distilled)	(0.90, 0.90)	11.54%	5.27%	83.19%	(0.89, 0.89)	(0.20, 0.22)	(0.97, 0.97)
Reading (Informational Text)	**Human**		11.64%	18.53%	69.83%			
	GPT-4.1-mini	(0.72, 0.72)	9.82%	22.35%	67.83%	(0.68, 0.69)	(0.54, 0.55)	(0.87, 0.88)
	o4-mini	(0.81, 0.81)	18.30%	16.60%	65.10%	(0.73, 0.74)	(0.68, 0.69)	(0.94, 0.94)
	Llama-3.1-8B (Distilled)	(0.72, 0.73)	20.51%	10.16%	69.34%	(0.61, 0.62)	(0.24, 0.26)	(0.91, 0.91)
Reading (Literature)	**Human**		9.48%	6.57%	83.95%			
	GPT-4.1-mini	(0.54, 0.56)	2.60%	13.86%	83.54%	(0.45, 0.47)	(0.67, 0.69)	(0.95, 0.95)
	o4-mini	(0.52, 0.54)	7.22%	6.80%	85.98%	(0.50, 0.52)	(0.12, 0.14)	(0.92, 0.92)
	Llama-3.1-8B (Distilled)	(0.81, 0.81)	7.22%	7.52%	85.26%	(0.83, 0.84)	(0.20, 0.22)	(0.92, 0.92)

Table 3: Alignment between LLM featurizers and Aggregate Human featurization obtained by majority voting for different models and assessment area. QWK and F1 values presented are 95% Bootstrap CI.

agreement in Science and Reading (Information Text) but achieves moderate agreement in Reading (Literature).

Assessment Area	Krippendorff's α
Science	$(0.718, 0.723)$
Reading (Informational Text)	$(0.696, 0.700)$
Reading (Literature)	$(0.669, 0.678)$

Table 4: Inter-rater reliability among human raters for the featurization alignment experiment (95% bootstrap CI).

F1 scores and label distribution for each feature label[6] provide a more detailed insight and reveal areas for further improvement. Notice that the F1 score is exceptionally high (near or above 0.9) for label 0, and moderate-to-high (0.6$\sim$0.93) for label 2, with higher agreement for Science items. Yet, alignment for label 1 is moderate-to-low, ranging from 0.68 down to 0.12. We believe this is due to the relatively ambiguous nature of the label category 1, coupled with the rarity of label 1 in human rating. While LLM featurizers achieve high overall alignment with aggregate human featurization, additional study needs to be done to ensure that the labeling task incurs less ambiguity and that the featurizer models match the natural distribution of human labels.

6 Discussions and Limitations

The principles outlined in Section 2.2 address a specific aspect of "interpretability" in the broader domain of automated scoring. While these principles are foundational to supporting the design of a valid scoring process and fostering trust in the scoring system, simply adhering to the principles

[6]For Aggregate Human and o4-mini, prevalence weighting was used to extrapolate the label distribution from the 50 study samples for comparison with the full distribution of all (r, c) pairs.

does not, on its own, guarantee that the needs and potential benefits of the stakeholders (Section 2.1) will be met.

Educational assessment literature is ripe with guidelines, frameworks, and best practices for ensuring that automated scoring properly serves the stakeholders' needs and produce valid, reliable, and fair scoring results (e.g., Bejar et al. (2016), Bennett and Bejar (1998), Bennett and Zhang (2015), Williamson et al. (2012)). These studies emphasize that evidence for the validity of automated scoring should be collected throughout the assessment process from a variety of evidentiary sources, such as model features, agreement with human raters, treatment of unusual responses, generalizability of score interpretations, population invariance of scores, and impact on teaching and learning (Bennett and Zhang, 2015).

7 Conclusion

AI and education research community has yet to produce a practical interpretability solution for automated scoring in large-scale educational assessments despite a pressing need. To address this challenge, we analyzed the needs and potential benefits of interpretable automated scoring for various assessment stakeholders (students, assessment developers, and test users) and developed 4 foundational principles – **F**aithful, **G**rounded, **T**raceable, and **I**nterchangeable (**FGTI**) – aimed at addressing those needs. We also demonstrated the feasibility of implementing these principles by developing ANALYTICSCORE for short-answer scoring. This framework generates human-interpretable features for each response based on explicitly identifiable elements, and uses an intuitive ordinal logistic regression scorer. On a real-world short-answer scoring dataset, ANALYTICSCORE outperforms many uninterpretable scoring methods, achieves a narrow performance gap relative to the uninterpretable SOTA, and demonstrates featurization behaviors that align with human judgment. Our findings show strong promise for implementing accurate and well-aligned interpretability solutions that address the real needs of assessment stakeholders. We hope our work illuminates exciting new directions in developing practical and effective interpretable automated scoring for large-scale educational assessments.

References

AERA, APA, and NCME. The standards for educational and psychological testing. 2014.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Yuya Asazuma, Hiroaki Funayama, Yuichiroh Matsubayashi, Tomoya Mizumoto, Paul Reisert, and Kentaro Inui. Take no shortcuts! stick to the rubric: A method for building trustworthy short answer scoring models. In *International Conference on Higher Education Learning Methodologies and Technologies Online*, pages 337–358. Springer, 2023.

Malcolm I Bauer and Diego Zapata-Rivera. Cognitive foundations of automated scoring. In *Handbook of automated scoring*, pages 13–28. Chapman and Hall/CRC, 2020.

Isaac I Bejar, Robert J Mislevy, and Mo Zhang. Automated scoring with validity in mind. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, pages 226–246, 2016.

Randy Elliot Bennett. Moving the field forward: Some thoughts on validity and automated scoring. *Automated scoring of complex tasks in computer-based testing*, pages 403–412, 2006.

Randy Elliot Bennett and Isaac I Bejar. Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17(4):9–17, 1998.

Randy Elliot Bennett and Mo Zhang. Validity and automated scoring. In *Technology and testing*, pages 142–173. Routledge, 2015.

Amy I Berman, Michael J Feuer, and James W Pellegrino. What use is educational assessment?, 2019.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020.

Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74, 1998.

Aubrey Condor and Zachary Pardos. Explainable automatic grading with neural additive models. In *International Conference on Artificial Intelligence in Education*, pages 18–31. Springer, 2024.

Rianne Conijn, Patricia Kahr, and Chris CP Snijders. The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, 10(1):37–53, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Kristen DiCerbo, Emily Lai, and Ventura Matthew. Assessment design with automated scoring in mind. In *Handbook of Automated Scoring*, pages 29–48. Chapman and Hall/CRC, 2020.

Steve Ferrara and Saed Qunbar. Validity arguments for ai-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3):288–313, 2022.

Peter W Foltz, Duanli Yan, and André A Rupp. The past, present, and future of automated scoring. In *Handbook of Automated Scoring*, pages 1–10. Chapman and Hall/CRC, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wynne Harlen. Teachers' summative practices and assessment for learning–tensions and synergies. *Curriculum Journal*, 16(2):207–223, 2005.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pages 1–23, 2022.

Institute of Education Statistics. Math autoscoring is finally here—let's tap its potential for improving student performance. https://ies.ed.gov/learn/blog/math-autoscoring-finally-here-lets-tap-its-potential-improving-student-performance, Oct 2023. [Accessed: Feb 21. 2025].

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020.

Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, 3:100074, 2022.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

Vivekanandan Kumar and David Boulanger. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA, 2020.

Vivekanandan S Kumar and David Boulanger. Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584, 2021.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. Get it scored using autosas—an automated system for scoring short answers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9662–9669, 2019.

Claudia Leacock and Martin Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213, 2024.

Jiazheng Li, Artem Bobrov, David West, Cesare Aloisi, and Yulan He. An automated explainable educational assessment system built on llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29658–29660, 2025.

Zhaohui Li, Susan Lloyd, Matthew Beckman, and Rebecca J Passonneau. Answer-state recurrent relational network (asrrn) for constructed response assessment and feedback grouping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3879–3891, 2023.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Bahar Memarian and Tenzin Doleck. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5:100152, 2023.

Robert J Mislevy. An evidentiary-reasoning perspective on automated scoring: Commentary on part i. In *Handbook of Automated Scoring*, pages 151–168. Chapman and Hall/CRC, 2020.

Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29 (3):441–459, 2019.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18(55):1–35, 2017.

James W. Pellegrino. *A Learning Sciences Perspective on the Design and Use of Assessment in Education*, page 238–258. Cambridge Handbooks in Psychology. Cambridge University Press, 2022.

Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*, 2018.

Jason DM Rennie and Nathan Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1, pages 1–6. AAAI Press, Menlo Park, CA, 2005.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

André A Rupp. Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, 31(3):191–214, 2018.

Advait Sarkar. Large language models cannot explain themselves. *arXiv preprint arXiv:2405.04382*, 2024.

Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, and Paul Libbrecht. Explainability in automatic short answer grading. In *International conference on artificial intelligence in education technology*, pages 69–87. Springer, 2022.

Daniel L Schwartz, Jessica M Tsang, and Kristen P Blair. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*. WW Norton & Company, 2016.

Mark D Shermis. Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment*, 20(1):46–65, 2015.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

John Whitmer and Magdalen Beiting-Parrish. Results of naep math item automated scoring data challenge & comparison between reading & math challenges. 2023.

John Whitmer and Magdalen Beiting-Parrish. Lessons learned about transparency, fairness, and explainability from two automated scoring challenges. In *AI for Education: Bridging Innovation and Responsibility*, 2024.

Dylan Wiliam. *Embedded formative assessment*. Solution tree press, 2011.

David M Williamson, Xiaoming Xi, and F Jay Breyer. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13, 2012.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19187–19197, 2023.