Interpretability from the Ground Up: Stakeholder-Centric Design of Automated Scoring in Educational Assessments

Yunsung Kim* Mike Hardy Joseph Tey Candace Thille Chris Piech

*yunsung@stanford.edu, Stanford University

ACL 2026 SAN DIEGO JULY 2–7

Background

- **Motivation**: Automated scoring (AS) is rapid and scalable, but stakeholder trust is crucial for adoption in high-stakes educational assessments. Interpretable AS helps foster that trust.
- **Research-to-Practice Gap**: Despite demand, interpretable AS remains limited in practice.
- **Illustrative Example**: 2023 NAEP Math Automated Scoring Challenge winners achieved human-like scoring accuracy, but *none of the* submissions met the interpretability criteria.
- **Underlying Issue**: We attribute this gap to the lack of a stakeholder need-finding process.

Key Takeaways

Design of interpretable AI should be grounded in **stakeholder needs**.

We derive the **"FGTI" principles** of **interpretable automated scoring** by examining the needs of the **assessment stakeholders**.

AnalyticScore embodies the FGTI. It is accurate and human-aligned.

Contributions

Problem Formulation. Shift from treating automated scoring as an isolated prediction task to situating it within the broader stakeholder-centered assessment lifecycle.

Stakeholder-Centric Interpretability Principles. The FGTI principles are derived from the stakeholder needs and benefits.

Method. AnalyticScore is a reference framework implementing FGTI

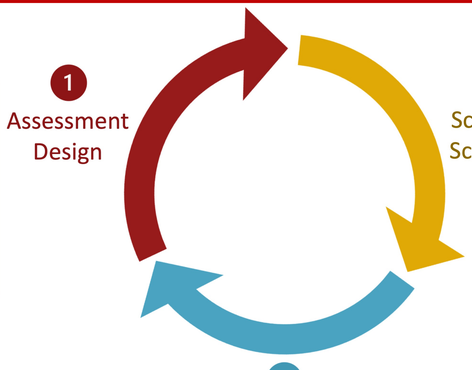
Evaluations. AnalyticScore achieves high scoring accuracy and aligns well with human judgment

Stakeholder Needs & Benefits of Interpretable Automated Scoring

Ensure construct alignment
Verify that the scorer can detect evidence of the target constructs measured by the task

Inform assessment design
Design target constructs, rubrics, and tasks based on what the scorer can reliably detect

Select and reuse scorer components
Identify scorer components that can be reused to reduce training cost. (DiCerbo et al. (2020))



Build trust in fair, justifiable scoring
Make individual scores and the overall scoring process auditable, traceable, and defensible

Human-in-the-loop quality control
Help reviewers efficiently identify and correct scoring errors to maintain assessment integrity

Build confidence in score use
Helps users make score-based instructional, evaluative, or system-wide decisions.

Trace the sources of problematic scoring
Identify errors, bias, or construct mismatch across items and test-taker population and trace these problems to specific model components

Refine and Improve scorer and tasks
Tune and improve scorer components, scoring criteria, rubric, and tasks.

Key Desiderata

- Accurately inspect model decisions
- Identify features & intermediate reasoning steps
- Make necessary adjustments to model decisions

Interpretable automated scoring is **not just about scoring**.

It should be designed to support the **full assessment lifecycle**.

The FGTI Principles of Interpretable Automated Scoring

(Transparent inspection) | (Which features are extracted) | (How the features are processed)

Principle 1. (Faithful)

*"Explanation reflects the **actual computational mechanism**"*

Violating Examples

- Text produced by prompting LLMs to generate an explanation[3]

Principle 2. (Grounded)

*"**Features** represent **understandable & identifiable elements** of response"*

Violating Examples

- Cosine similarity of sentence embedding

Principle 3. (Traceable)

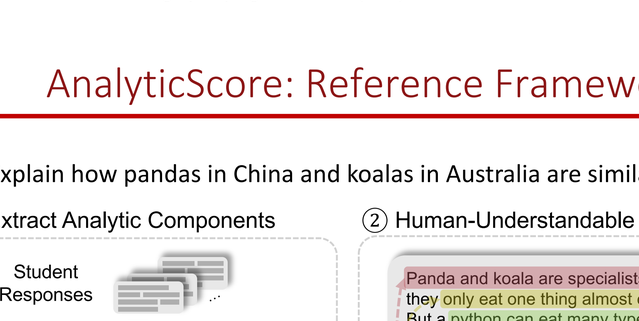
*"Scorer consists of **subroutines** that represent a **specific evidentiary reasoning step** on well-defined inputs."*

Principle 4. (Interchangeable)

*"**A human** should be able to **act interchangeably** on these subroutines."*

Violating Examples

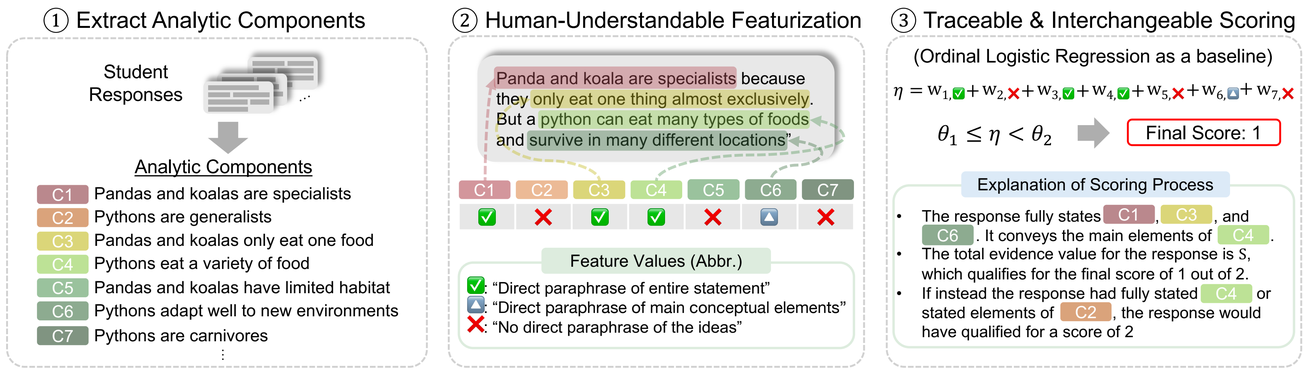
- Feature importance values
- Feature attribution maps
- Attention weights & saliency maps
- Confidence metrics
- Circuit tracing



Reference
[3] Turpin, Miles, et al. "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting." NeurIPS 36 (2023): 74952-74965.

AnalyticScore: Reference Framework for the FGTI Principles

Q: Explain how pandas in China and koalas in Australia are similar, and how they both are different from pythons.



Analytic Components: "Explicitly identifiable elements of student responses" (Groundedness)

- Can be reviewed and modified by assessment developers during assessment design & refinement
- For constructed response scoring, analytic components are propositions

Distilling to Open Source. Supervised Fine-Tuning on 10,000 featurization examples from a larger LLM featurizer to train an open-source featurizer to avoid linearly growing cost.

Evaluation Dataset

ASAP-SAS Dataset: Largest public constructed-response scoring dataset with complex response structures. 10 items, 3 assessment areas.

Evaluation of AnalyticScore

Scoring Accuracy Experiments

- Despite limited architectural flexibilities, AnalyticScore outperforms several baseline scorers including proprietary LLMs and achieves accuracy close to black-box SOTA.
- Distillation substantially improves accuracy to the level of proprietary LLM featurizers.

	All Avg.	Sci Avg.	R(Inf) Avg.	R(Lit) Avg.
Human Scorers	0.88±0.02	0.93±0.01	0.79±0.03	0.91±0.05
AnalyticScore				
w/ GPT-4.1-mini*	0.72±0.03	0.78±0.03	0.68±0.06	0.60±0.01
w/ Llama-3.1-8b Inst*	0.58±0.03	0.58±0.04	0.63±0.05	0.48±0.04
+ Distillation*	0.71±0.03	0.77±0.03	0.68±0.06	0.60±0.01
Few-Shot				
GPT-4.1	0.63±0.04	0.67±0.02	0.68±0.04	0.45±0.12
Supervised LLM				
BERT	0.75±0.02	0.79±0.02	0.74±0.05	**0.69±0.01**
DeBERTa	0.76±0.03	0.81±0.03	0.72±0.04	0.67±0.04
Llama-3.1-8b Inst.	0.76±0.02	0.79±0.02	0.75±0.03	**0.69±0.02**
w/ rubric	**0.78±0.02**	**0.82±0.02**	**0.76±0.05**	0.68±0.04
Llama-3.1-8b	0.76±0.02	0.80±0.02	**0.76±0.03**	0.67±0.02
Baseline				
AutoSAS	0.56±0.04	0.57±0.04	0.65±0.06	0.41±0.04
ASRRN	0.59±0.02	0.59±0.04	0.63±0.04	0.55±0.04
NAM*	0.56±0.05	0.64±0.02	0.52±0.13	0.40±0.02

Table 1. Scoring accuracy measured by QWK against gold standard scoring labels

Featurization Alignment Experiments

7 annotators independently performed AnalyticScore's featurization.

Question 1: Is AnalyticScore's featurization task clear enough to yield human agreement?

- Human raters agree well on the featurization task (Table 2)

Question 2: Does AnalyticScore's featurization align with human judgment?

- Distilled featurizer achieves high agreement with aggregate human features
- Further study should reduce the ambiguity of the "partial paraphrase" category

Assessment Area	Krippendorff's α
Science	(0.718, 0.723)
Reading (Informational Text)	(0.696, 0.700)
Reading (Literature)	(0.672, 0.680)

Table 2. Inter-rater reliability among human raters on AnalyticScore's featurization task

Assessment Area	Featurizer Model	QWK	Label-wise F1 ✅	F1 🔺	F1 ❌
Science	GPT-4.1-mini	(0.89, 0.89)	(0.83, 0.84)	(0.20, 0.21)	(0.96, 0.96)
	o4-mini	(0.94, 0.95)	(0.93, 0.93)	(0.49, 0.51)	(0.98, 0.98)
	Distilled	(0.90, 0.90)	(0.89, 0.89)	(0.20, 0.22)	(0.97, 0.97)
Reading (Informational Text)	GPT-4.1-mini	(0.71, 0.72)	(0.68, 0.69)	(0.54, 0.55)	(0.87, 0.88)
	o4-mini	(0.81, 0.82)	(0.73, 0.74)	(0.69, 0.70)	(0.94, 0.94)
	Distilled	(0.72, 0.73)	(0.61, 0.62)	(0.24, 0.26)	(0.91, 0.91)
Reading (Literature)	GPT-4.1-mini	(0.52, 0.54)	(0.44, 0.46)	(0.67, 0.69)	(0.95, 0.95)
	o4-mini	(0.49, 0.51)	(0.49, 0.51)	(0.10, 0.12)	(0.92, 0.92)
	Distilled	(0.81, 0.82)	(0.86, 0.87)	(0.17, 0.19)	(0.92, 0.92)

Table 3. LLM featurizers vs aggregate human features