

A HUMAN-FRIENDLY GUIDE TO

IMPACTER's Machine Learning Model

A Resource For Our Partners

Purpose

This guide helps school and district leaders understand how IMPACTER's scoring system works, why it's trustworthy, and what makes it different from traditional assessment approaches.

**Technical terms marked with an asterisk are defined in the comprehensive glossary at the end of this document.*

Contents

- 03 — What You Need to Know
- 04 — Why You Can Trust ML Scoring
- 06 — How the Scoring System Works
- 07 — Key Points to Remember
- 08 — Glossary of Technical Terms
- 09 — Frequently Asked Questions

ABOUT THE AUTHOR

Impacter Pathway designs technology-enabled assessment systems that capture authentic evidence of how students learn, grow, and apply future-ready skills. Powered by machine learning, natural language processing, and evidence-based rubrics, Impacter's platform reimagines assessment as a window into thinking, reflection, and human development. Working alongside states, districts, and research partners, the company leads innovation in measuring the mindsets that prepare students for the future. Its mission is simple yet urgent: to ensure every learner is seen as a developing thinker and engaged citizen.

What You Need to Know

The Core Principle: Humans Lead, Technology Scales

IMPACTER's scoring system doesn't replace educators—it follows their judgment at a scale that would otherwise be impossible. Here's what makes it work:



Trained educators score tens of thousands of real student responses.

These human-scored examples become the foundation—called the [training data](#)*—that teaches the [machine learning \(ML\)](#)* system how to recognize strong evidence of skills like curiosity, compassion, and grit in authentic student language.



The technology learns patterns from these expert scorers.

It studies how trained educators apply research-based rubrics to diverse student voices, learning to recognize what strong attribute demonstration looks like across different ways of expressing ideas.



Human reviewers continuously check the system's work.

We regularly pull new student responses, have educators rescore them, and verify the technology still matches human judgment. If we detect any drift in either the model or the human scorers, we [retrain](#)* to ensure consistency.



Each scoring model is locked once validated.

This is called **model versioning** or using [frozen models](#)*—the same quality of response always receives the same score—no hidden changes, no moving targets, and no advantage for students who complete assessments later in the window.

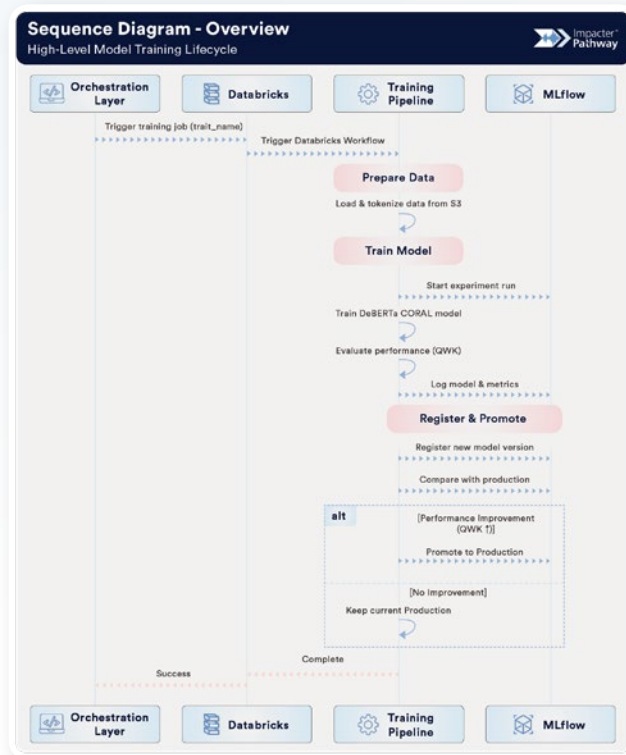
Why You Can Trust ML Scoring

1

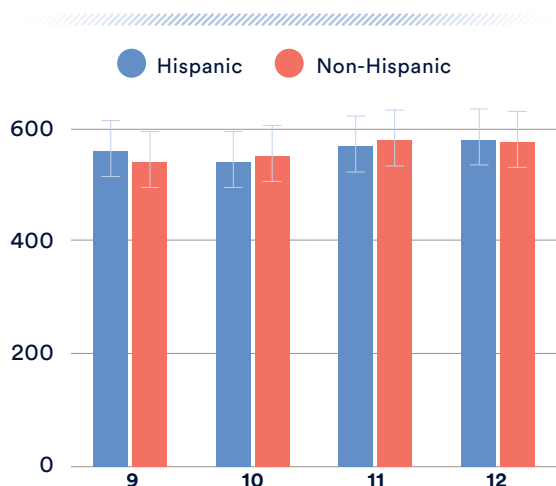
The Technology Matches Trained Educators

We measure how well our scoring system aligns with trained human scorers using the same statistical methods used in large-scale state assessments—like [Quadratic Weighted Kappa \(QWK\)](#)^{*}, which evaluates both accuracy and the severity of scoring differences. Our results consistently exceed typical [inter-rater reliability](#)^{*} between human scorers—not because the system is smarter than educators, but because it maintains perfect consistency. It's never affected by fatigue, order effects, or subjective drift that can influence even well-trained human raters.

What this means: If a trained teacher would score a student response as a 3, our system almost always scores it as a 3 or very close.



RESPONSE LENGTH	FEMALE SMD	MALE SMD	BIAS INTERPRETATION
Very Short	0.039	0.037	Negligible
Medium	0.027	0.026	Negligible
Long	0.022	0.025	Negligible



2

We Run Rigorous Fairness Checks

We verify that the system treats similar responses the same way, regardless of who said them. This includes checking across response lengths (what we call [structural fairness](#)^{*}), demonstrated skill levels, and—when districts provide demographic data—across student groups using [Standardized Mean Difference \(SMD\)](#)^{*} analysis.

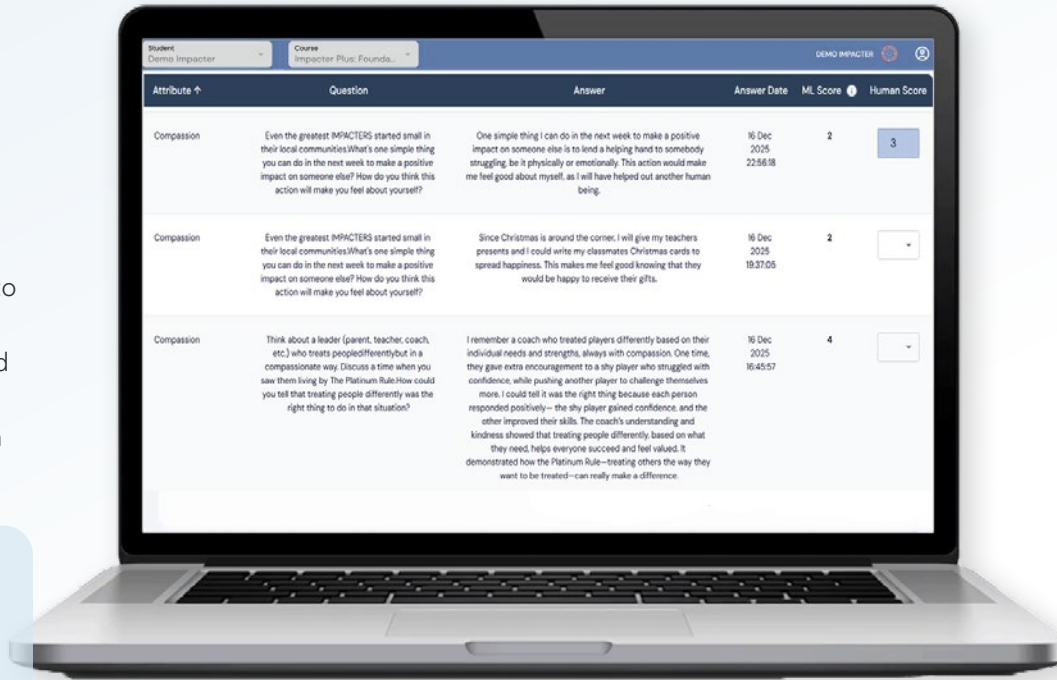
What this means: No student or student group receives an advantage or disadvantage when providing the same-quality response.

3

We Continuously Monitor Quality

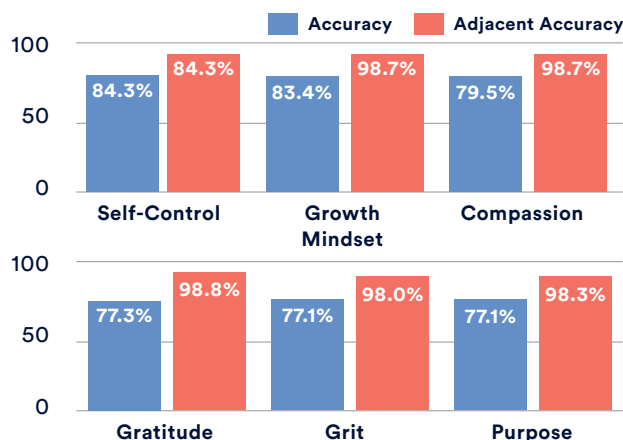
Human reviewers regularly examine new student samples. If there's any drift in scoring patterns, we retrain the system. We also review human scoring to ensure our expert raters remain calibrated to the research-based rubric, so both the technology and our human scorers maintain alignment.

What this means: The system receives regular tune-ups to stay aligned with educator judgment and best practices.



ML Score	Human Score				
	0	1	2	3	4
0	8.20%	2.85%	0.68%	0.27%	0.15%
1	2.45%	14.90%	4.15%	0.90%	0.30%
2	0.62%	3.40%	17.65%	5.25%	1.18%
3	0.20%	1.05%	4.30%	15.15%	3.30%
4	0.12%	0.35%	1.15%	3.18%	10.50%

Accuracy Metrics by Trait



4

We're Built on Assessment Science Standards

We use the same types of validity checks used in established K-12 assessment scoring: Quadratic Weighted Kappa (QWK), agreement rates, and fairness analyses. This isn't experimental—it's built on proven scoring science.

What this means:

IMPACTER's approach follows the same rigorous standards districts already trust in state testing systems.

How the Scoring System Works

Step 1

Building the Foundation with Human Expertise

Everything starts with real educators scoring real student work. We begin with tens of thousands of authentic student responses scored by trained educators who are calibrated on research-based rubrics designed to minimize bias and maximize asset-based evaluation.

These human-scored examples become the “training set” that teaches the machine learning system how educators make scoring decisions.

Step 2

The System Learns by Example

The technology doesn’t read the rubric like a checklist. Instead, it studies the patterns in how trained humans apply it. It learns things like:

- » What strong attribute demonstration looks like in authentic student language
- » What linguistic markers indicate stronger or weaker reflection
- » What low-evidence answers look like across diverse student voices

As it learns, the system organizes student responses in what’s called a [multidimensional vector space](#)*—essentially an invisible map where similar responses cluster together based on their characteristics (tone, complexity, evidence of reasoning, etc.). This helps it recognize patterns in how students express traits like curiosity, purpose, or self-control, and match those patterns to the scores humans have given in the past.

Step 3

Scoring New Student Work

When a student provides a response, the system compares it to the thousands of examples it has already learned from and predicts the score a trained human would give.

Key point: The system is not replacing humans—it is following the scoring patterns humans have established, but doing so at a scale that would be unsustainable and fiscally prohibitive for human graders.

Step 4

Ongoing Human Review

We routinely pull new student responses, have humans rescore them, and check whether the system still matches educator judgment. If we see the system drift, we retrain it. If we see educators drift, we recalibrate with them to ensure scoring continues to reflect the research-based, unbiased rubric.

Step 5

Locking the Model for Consistency

Once a model passes all accuracy and fairness checks, we freeze it—this is called model versioning or frozen models. That means the same answer always gets the same score throughout an assessment window—no surprises, no hidden changes, no moving targets.

Key Points to Remember

The system learns from tens of thousands of educator-scored examples. Humans set the standard.

We regularly recheck the model with new human scoring to keep it aligned.

Unlike humans, the system maintains perfect consistency: it's never affected by fatigue, order effects, or subjectivity that can influence even well-trained human raters.

We run fairness checks to ensure no group is treated differently for the same-quality work.

This is the same approach used in K-12 large-scale assessments, just faster and more scalable.

Glossary of Technical Terms

Machine Learning (ML)

The system that reads student writing (or transcripts of student speech) and predicts a score based on what it learned from tens of thousands of human-scored examples—not from a generic dataset, but from real student voices scored by trained educators.

ML Pipeline

A pipeline is the full scoring workflow for a specific skill or attribute. Think of pipelines as individual “scoring specialists,” each trained to evaluate a different human skill. IMPACTER currently maintains 13 specialized scoring pipelines: the IMPACTER 8 (Curiosity, Perspective-Taking, Purpose, Self-Control, Grit, Growth Mindset, Compassion, Gratitude) and the CASEL 5 (Self-Awareness, Self-Management, Social Awareness, Relationship Skills, Responsible Decision-Making).

Training Data

A large collection of real student responses scored by expert humans. These examples “teach” the machine learning system how humans make scoring decisions.

Retraining

A routine process where we pull new student responses, have humans rescore them, and use the results to keep the system aligned with human judgment.

Multidimensional Vector Space

Inside the model, each student response is turned into a kind of “location” in a very large, invisible map. Each direction on this map represents something about the response (tone, complexity, relevance, reasoning, etc.). Similar responses end up close together; very different responses land far apart. This “map” is how the model organizes and compares student responses so it can score them consistently.

Accuracy

How often the system and a trained human give the same score.

Adjacent Accuracy

How often the system is either exact or within one point of a human scorer. Human scorers typically agree within one point, so this is a meaningful measure of consistency.

QWK (Quadratic Weighted Kappa)

A statistical measure that shows, in a single number between 0 and 1, how closely the system matches human scorers. It tells us how close the scoring is to human scoring while giving extra penalties when the system is off by a lot. Higher numbers = better. The industry standard is 0.70; IMPACTER typically achieves 0.90 or higher.

Here’s how it works:

- » If the system and a human give the same score → that’s good
- » If the system is off by 1 point → that’s okay (humans disagree like this too)
- » If the system is off by 3 points → QWK treats this as a much bigger problem

This is why QWK is more meaningful than simple accuracy. It understands how far off the model is, not just whether it’s right or wrong.

Inter-Rater Reliability

A measure of how often two trained human scorers agree when they independently score the same student work. In assessment science, high inter-rater reliability (typically 0.70-0.80 QWK) means scorers are applying the rubric consistently. IMPACTER's system consistently exceeds 0.90 QWK—meaning it agrees with human scorers more reliably than human scorers typically agree with each other.

Fairness / SMD (Standardized Mean Difference)

A way to check whether different groups of students receive similar scores when they write similar-quality work. Small numbers mean no meaningful difference.

Structural Fairness

Checks that the system isn't accidentally favoring things like long answers versus short ones.

Model Versioning (Frozen Models)

Once a model passes all checks, we “freeze” it. That means the same response quality always gets the same score throughout an assessment window, ensuring fairness and consistency.

Frequently Asked Questions

How do you know the scores are accurate?

We compare the system's scores to trained human scorers using industry-standard statistical measures. Our agreement rates consistently match or exceed typical human-to-human agreement in large-scale assessment scoring.

How do you make sure the system isn't biased?

We run comprehensive fairness checks to confirm that similar-quality responses receive similar scores across groups. If districts provide demographic data, we run additional subgroup analyses to verify equitable treatment.

Does the system use a rubric?

Not directly—and that's by design. It learns how humans apply the rubric by studying tens of thousands of real scored examples. This makes it flexible enough to understand diverse student voices and expression styles, not just formulaic answers that “game” a rubric.

Does the system change on its own?

No. Once validated, each model version is locked to ensure consistency throughout an assessment window. Any updates happen between assessment cycles, not during them, and only after rigorous validation.

Do humans ever review what the system scores?

Yes. We routinely pull new student responses and have educators rescore them to ensure the model stays aligned. We also review the human scores themselves to check for any bias or drift, so both the system and our human graders remain calibrated to the same research-based, unbiased rubric.

What happens if a student writes something the system hasn't seen before?

The system doesn't just memorize examples—it learns the underlying patterns of how students demonstrate skills. This means it can recognize quality evidence even in creative or unexpected responses. However, if a response is genuinely ambiguous or unclear, human reviewers flag it for additional review.

How does the system handle different writing styles and abilities?

The system learns from responses across the full range of student writing abilities and styles. It's trained to recognize evidence of skills like curiosity or compassion in both sophisticated and simple language. The focus is on the presence and quality of the attribute being demonstrated, not on vocabulary level or sentence complexity. This makes the assessment more equitable for students with varying language proficiency.

Can the model score in languages other than English?

Yes. IMPACTER supports student responses in 90+ languages. Students can submit reflections in their primary language, and the system evaluates the quality of their thinking and evidence of skills—not their English proficiency. This multilingual capability ensures equitable assessment for English Learners and students from diverse linguistic backgrounds.