

# A Principled Design of Educational AI Benchmarks

November 2026

**Prepared by:**
**SRI**
Leigh Ann DeLyser
Arif Rachmattulah

**November 2026**

**Acknowledgments:**
This work was supported by SRI International through its internal research and development (IRAD) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of SRI International. We thank Satabdi Basu and Nonye Alozie, who reviewed and provided thoughtful feedback on the earlier version of this work. We also thank Jessica Mislevy and Todd Grindal, who supported this work.

**Suggested citation:**
DeLyser, L. A., & Rachmatullah, A. (2026). *A principled design of educational AI benchmarks*. SRI.

# Contents

List of Tables and Figures ..... vi
Executive Summary ..... vii
Generative AI Benchmarks for Education ..... 1
Three Complementary Forms of Evidence ..... 2
What an AI Benchmark Is ..... 3
Why Benchmark Design Matters ..... 5
The Need for Educational and Domain Expertise ..... 7
Principled Design of Benchmarks ..... 9
Principled Assessment Design as a Conceptual Foundation ..... 11
1. Clearly Defined Assessment Targets ..... 13
2. Statements of Intended Score Interpretations and Uses ..... 14
3. Models of Cognition, Learning, or Performance ..... 14
4. Aligned Measurement Models and Reporting Scales ..... 14
5. Purposeful Assessment Activities ..... 14
Adapting PAD to Benchmark Design and Evaluation ..... 15
A Framework for Principled Benchmark Design ..... 17
1. Domain Scoping ..... 19
2. Domain Analysis ..... 21
3. Task and Construct Design ..... 22
4. Dataset and Scenario Development ..... 23
5. Metrics and Evaluation Design ..... 24
6. Implementation and Testing ..... 25
7. Validation, Documentation, and Release ..... 26
8. Sustain ..... 27
From Framework to Practice ..... 28
Applications of Principled Benchmark Design ..... 29
Use Case 1: First-Grade Science Assessment-Task Generation ..... 29
1. Domain Scoping: Purpose, Users, and Boundaries ..... 29
2. Domain Analysis: Defining Content Quality and Assessment Targets ..... 30
3. Task and Construct Design: Claims and Evidence Models ..... 32

4. Dataset and Scenario Development: Task-Model Variables and Pilot Conditions ..... 33
5. Metrics and Evaluation Design: Scoring and Reporting ..... 35
6. Implementation and Testing: Executable Implementation in Kaggle Benchmarks ..... 36
7. Validation, Documentation, and Release: Preliminary Evidence and Limitations ..... 38
8. Sustain: Future Maintenance and Revision ..... 38
Use Case 2: Information Extraction and PII Redaction in IEP Documents ..... 39
1. Domain Scoping: Purpose, Users, and Boundaries ..... 39
2. Domain Analysis: Defining Extraction and Redaction Performance ..... 40
3. Task and Construct Design: Claims and Evidence Model ..... 41
4. Dataset and Scenario Development: Document and Task Model ..... 42
5. Metrics and Evaluation Design: Deterministic Scoring and Reporting ..... 43
6. Executable Implementation in Kaggle Benchmarks ..... 45
7. Validation, Documentation, and Release: Preliminary Evidence and Limitations ..... 47
8. Suatain: Future Maintenance and Revision ..... 48
Use Case 3: AI-Based Evaluation of Science Lesson Plans ..... 49
1. Domain Scoping: Purpose, Users, and Boundaries ..... 49
2. Domain Analysis: Defining KISR Performance ..... 50
3. Task and Construct Design: Claims and Evidence Models ..... 51
4. Dataset and Scenario Development: Lesson Plans and Task-Model Conditions ..... 53
5. Metrics and Evaluation Design: Agreement With Expert Ratings ..... 55
6. Implementation and Testing: Executable Kaggle Benchmark ..... 56
7. Validation, Documentation, and Release: Preliminary Evidence and Limitations ..... 58
8. Sustain: Future Maintenance and Revision ..... 59
References ..... 60
Appendix A: Documentation Scaffold ..... 64
1 Domain Scoping ..... 64
Benchmark purpose and intended uses ..... 64
Users and settings ..... 64
Boundaries of the evaluation ..... 64
2 Domain Analysis ..... 64
Research and source materials ..... 64
Strong and weak performance ..... 64
Gaps in the evidence base ..... 64

3 Task and Construct Design ..... 65
Claims and observable evidence ..... 65
Task specifications ..... 65
Reference answers and labels ..... 65
Evaluation by an AI judge ..... 65
4 Dataset and Scenario Development ..... 65
Materials and example selection ..... 65
Scenario variation ..... 65
Interaction conditions ..... 65
Review of scenario coverage ..... 65
5 Metrics and Evaluation Design ..... 66
Scoring rules ..... 66
Weights and combined scores ..... 66
Results available to users ..... 66
6 Implementation and Testing ..... 66
Evaluation procedure ..... 66
Tests and observed behavior ..... 66
Problems and revisions ..... 66
7 Validation, Documentation, and Release ..... 66
Review and validity evidence ..... 66
Supported interpretations and limitations ..... 66
Access and reuse ..... 66
8 Sustain ..... 67
Version history and comparability ..... 67
Maintenance and future review ..... 67

# List of Tables and Figures

**Table** **Page**

Table 1. Similarities and Differences Between Educational Assessment and AI Benchmarking.. 11
Table 2. Illustration of Construct Granularity....13
Table 3. Domain-Scoping Specification (Use Case 1)....30
Table 4. Domain Analysis Specification (Use Case 1)....31
Table 5. Evidence Models for the Content Subconstructs (Use Case 1) .... 32
Table 6. Task-Model Variables and Levels (Use Case 1).... 34
Table 7. Domain-scoping specification (Use Case 2).... 39
Table 8. Domain-analysis specification (Use Case 2)....40
Table 9. Evidence Models (Use Case 2)....41
Table 10. Task-model variables and possible levels (Use Case 2) .... 42
Table 11. Domain-scoping specification (Use Case 3) .... 49
Table 12. Domain-analysis specification (Use Case 3) .... 50
Table 13. Evidence models for the selected KISR subconstructs (Use Case 3) ....51
Table 14. Task-model variables and possible levels (Use Case 3) .... 54

**Figure** **Page**

Figure 1. Types of Evidence .... 3
Figure 2. SuperGLUE Benchmark Example Questions (A. Wang et al., 2020) .... 4
Figure 3. Shared and Use-Specific Benchmarking of an AI Model Across Educational Contexts . 6
Figure 4. A Framework for the Principled Design of Benchmarks....19
Figure 5. Preliminary Kaggle Benchmarks Leaderboard for the First-Grade Science Assessment-Task Generation Pilot .... 36
Figure 6. Expanded Criterion-Level Results for the Scientific Accuracy Benchmark Task .... 37
Figure 7. Preliminary Kaggle Benchmarks leaderboard for the IEP information-extraction and student-name-redaction pilot.... 45
Figure 8. Expanded assertion-level results for an IEP extraction or student-name-redaction task. The interface displays the deterministic criteria that passed or failed for the evaluated model.... 47
Figure 9. Preliminary Kaggle Benchmarks leaderboard for Structure of Competency Development. Rows represent three lesson plan–subconstruct task instances, and columns represent two evaluated AI models. Task-level percentages correspond to linear agreement scores of 1.00, 0.50, or 0.00. The model-level score is the unweighted mean across the three task instances.... 56
Figure 10. Expanded output for one Structure of Competency Development task. Each evaluated model returns lesson-plan evidence, an assigned rubric level, and a rationale. The task score reflects the ordinal distance between the AI-assigned and expert-assigned levels; evidence and rationale are displayed diagnostically but do not contribute to the current score.... 58

# Executive Summary

Generative artificial intelligence (AI) holds significant promise for education. AI-enabled tools are being developed to support lesson and assessment design, provide feedback and tutoring, analyze educational data, and reduce administrative burden. At the same time, decisions by major school districts to restrict the use of AI expose a deep level of uncertainty and lack of trust in AI. We cannot blame decision-makers for this uncertainty. Schools have been given access to chatbots and AI-enabled products with promises of accelerating learning and improving educational work, often without rigorous evidence of what AI can do well and where it still needs improvement.

Without rigorous evaluation, it is difficult to know whether AI performs better or worse than an experienced educator, whether it produces a subpar lesson or assessment, or whether it responds appropriately when interacting with a student. A spot inspection in which an answer looks "surprisingly good" is not enough. Sometimes poor-quality AI output is difficult to recognize, and education cannot afford a reduction in the quality of lessons, assessments, instructional materials, grading, or tutoring interactions because we accepted the promise of AI without verifying its performance.

AI in educational settings is often evaluated through one of two approaches: based on the resulting impact on learners after they engage with the tool or product, or based on descriptive studies of the interactions between AI and its users. In the first, AI is treated as a black box within an educational product, practice, or intervention. Researchers examine student or teacher behavior, learning outcomes, or attitudinal or engagement outcomes to determine whether the practice or intervention involving AI had an effect. Such studies remain essential for understanding educational impact, but they may not explain whether an observed effect, or lack of effect, reflects the AI system's performance, its implementation, or how users interacted with it. This limitation is especially important for generative AI because its responses can vary based on the model, context, user input, prompts, system guidance, and other conditions.

In the second approach, teams of researchers review transcripts of human–AI interactions to understand how the collaboration is working and what types of questions and responses are being generated. This human-intensive work can produce rich evidence about interaction processes, but it is expensive, time-consuming, and difficult to scale. The method of assessing all of the interactions cannot easily keep pace with the rapid influx of new models, products, and tools entering educational settings, even with AI supporting researchers in the analysis.

Well-designed educational AI benchmarks provide a needed third form of evidence: evidence about performance and capabilities demonstrated by the AI system itself. A benchmark is similar to a student assessment: It is designed to measure specific capabilities across a range of settings, inputs, and contexts. Just as a well-designed assessment has a clear model of what is

being measured and how a student would demonstrate knowledge or skill, an educational AI benchmark should define the capability being assessed, how an AI system would demonstrate that capability, and the conditions under which performance may vary.

Benchmarks provide repeatable evaluation protocols that can be administered across multiple models and products. Although high-quality benchmarks require substantial upfront investment in domain modeling, task development, expert input, scoring, and validation, they can make evaluations more scalable than relying on teams of human researchers to review every output. The benchmark's tasks and scoring procedures must themselves be validated to determine whether they operate consistently across models and contexts and whether the resulting scores support their intended interpretations.

**SRI and the Development of Benchmarking**

Defense Advanced Research Projects Agency's (DARPA) common task framework was one of the first widely used structured approaches in which researchers addressed shared problems using common datasets, predetermined measures, and open leaderboards. SRI participated in these early efforts, helping to demonstrate how standardized benchmarks and open competitions could accelerate technology development. This history also illustrates the connection between benchmarking and improving technology quality during development, grounded in a specific purpose.

Most benchmarks for AI are currently designed by technologists—computer scientists and AI specialists. The Stanford Institute for Human-Centered Artificial Intelligence has highlighted the need to engage domain experts because benchmarks developed without them may not reflect the needs of applied fields. Education has spent decades studying assessment design. Principled Assessment Design (PAD) offers structured approaches for carefully defining the content and skills to be measured, identifying the evidence that would demonstrate them, and designing and creating tasks and the scoring rubrics that elicit that evidence and contribute to understanding whether the resulting scores support their intended interpretations and uses.

This white paper is written for education researchers and experts who have an interest in developing robust measures of AI. The paper has three related aims. First, it argues that benchmarking should become an important additional component of educational AI research and design by providing evidence about AI system performance under specified conditions. Second, it proposes a preliminary framework, grounded in PAD, for developing educational AI benchmarks. Third, it illustrates the application of the framework through three use cases situated in educational settings, involving assessment generation, information extraction and redaction, and rubric-based evaluation of lesson plans.

The proposed framework asks benchmark designers to:

- Define the skill, knowledge, behavior, or capability they are trying to assess. A question such as "Can AI write a good lesson plan?" is not specific enough until "good" is clearly defined and connected to evidence that can be observed in the AI's response.

- Create tasks, similar to questions on a test, that allow the AI to demonstrate the targeted capability across relevant inputs and contexts.
- Define what strong and weak performance looks like and how responses will be scored.
- Test the benchmark itself to determine whether it works as intended and produces reliable scores that support valid interpretations.
- Document what capability was assessed and under what conditions, how the benchmark was administered, and how its results can and cannot be used.
- Revisit and refine the benchmark as models, products, and educational uses change.

Designing a high-quality benchmark is closer to designing the SAT than a teacher writing a quiz for a class the following week. The required rigor depends on the claims and decisions the benchmark will support, but collecting a set of plausible tasks is not enough for a benchmark to be broadly applicable and produce evidence regarding AI's performance.

In this paper, we also share three use cases to illustrate how the proposed framework can be applied to familiar uses of AI in education: generating first-grade science assessments for two specific standards, extracting and redacting information from Individualized Education Program (IEP) documents, and evaluating science lesson plans using a pedagogical content knowledge framework. These examples span instructional, administrative, and evaluative uses of AI for which products and informal practices already exist, but where little research describes overall model performance on the specific tasks. They also offer insight into considerations of granularity of benchmark design and useful cases for discussion regarding the narrow or broad scope for the design of benchmarks and associated tasks.

Educational AI benchmarks explicitly assess the performance or capability demonstrated by an AI system under specified conditions—not its impact on the people or institutions using it. Other research is needed to understand effects on student learning, educator practice, educational relationships, and institutional systems. Benchmarks help unbox what the AI is doing, how well it is performing, and how those actions and performance levels may contribute to findings about human use. A high score should never become a blanket certification that a model or product is safe or good. Trust depends on transparency: A well-defined benchmark must make clear exactly what was assessed and how the results can be interpreted and used.

This framework is preliminary and requires continued testing and refinement. However, it is grounded in mature fields and ready for researchers to begin using. We hope that education researchers will consider benchmark design, development, and use as part of research involving AI-enabled interventions; that product developers will collaborate with researchers to design benchmarks and compare products rigorously with foundational models; that decision-makers will look for benchmark evidence when assessing product performance and potential harms; and that the broader AI community will recognize Principled Benchmark Design as valuable beyond education.



# Generative AI Benchmarks for Education

Generative AI is being infused into student- and teacher-facing products at an unprecedented pace. In the 2026 paper *Future-Proofing Human Flourishing: The Case for a Learning Sciences Benchmark for AI*, the task force refers to AI as an "arrival technology" (p. 3; Reich & Dukes, 2025):

> *"Arrival technologies are defined by their ability to bypass traditional gatekeepers. They are a class of innovations – in this case, generative AI – that do not simply 'pass through' the economy as tools but instead 'arrive' as a permanent, foundational layer of human infrastructure." – The Great Shift: Education's Evolution in the Era of Arrival Technology, Erin Mote, 2026*

AI is not arriving in classrooms and educational institutions in its final form. The underlying models continue to evolve, and each new evolution updates how users experience the AI features of products. The AI Model Release Tracker from AI Flash Report (2026) estimates that, on average, new AI models arrive every two days. Major labs such as OpenAI, Google, and Anthropic release new generations of models every six to 15 months while making smaller adjustments every four to six weeks. These models are reaching users in educational settings

faster than the community can test their performance, let alone measure their impact on students and student learning.

As a result, there is a mismatch between the pace of AI development and the pace of traditional education research. Educational technology companies may cite studies of products using generative AI models that are now generations old as evidence of impact.

## Three Complementary Forms of Evidence

Education research commonly generates two forms of evidence about AI-enabled products and practices. The first is **impact evidence**. In these studies, AI is treated as a black box within a practice, product, or intervention. Researchers look at student or teacher behavior, educational outcomes, or attitudinal or engagement outcomes to understand whether a practice, product, or intervention has affected learning or workflows. This approach remains essential for understanding the impact. However, generative AI is not deterministic, and when it is treated as a black box, the findings may not explain whether an observed effect, or lack of effect, reflects the performance of the AI system, the way the product was implemented, or how students and educators interacted with it.

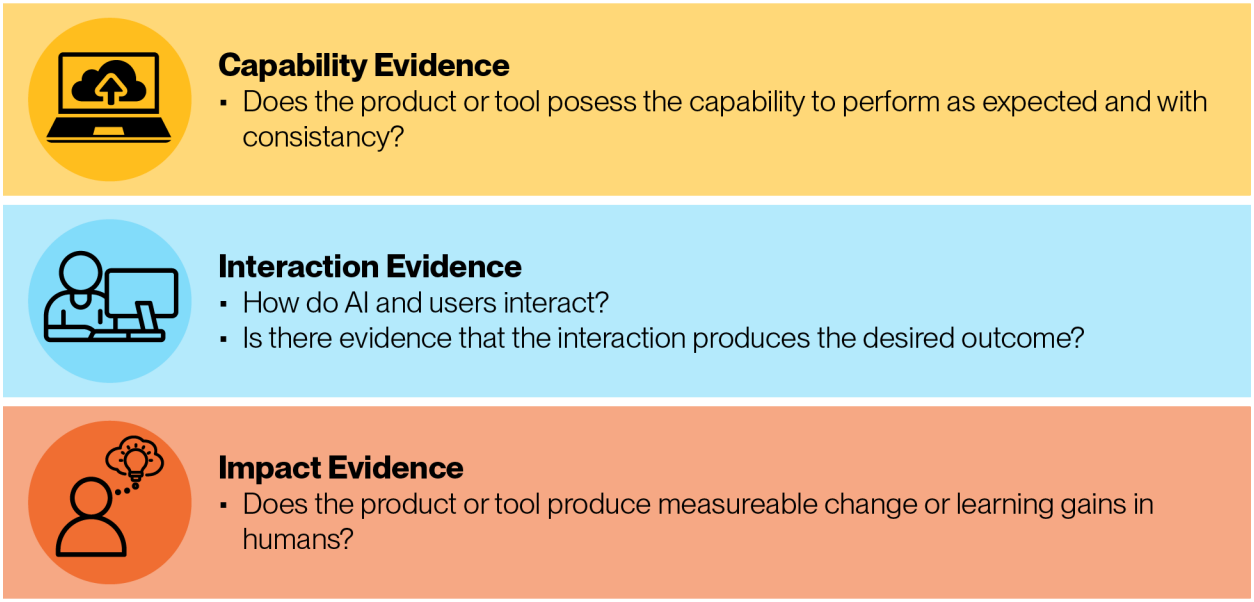
This limitation is particularly important for generative AI. Unlike worksheets, textbooks, and educational systems with deterministic or tightly controlled outputs, generative AI systems can produce different responses depending on the context, model version, user input, prompts, system guidance, and other conditions. Its outputs are also not always well controlled, as demonstrated by examples of students prompting their way outside of a product's intended boundaries (Humble, 2026; Udeshi et al., 2026). A study can therefore evaluate the impact of a product without fully documenting or evaluating the quality and variability of the AI-generated responses experienced by its users.

The second form is **interaction evidence**, often generated through close examination of human–AI interactions. Teams of researchers examine transcripts to understand how a human–AI collaboration is working and what kinds of questions and responses are being generated (Dang et al., 2025; Gomez et al., 2025). This method can produce rich evidence about processes that cannot be captured through outcome measures alone, but it is expensive; takes too much time to keep up with the onslaught of new models, products, and tools available for use in educational settings; and may not observe the full range of interactions, including off-task behavior between users and the AI systems.

Benchmarking offers a complementary third form of evidence: **capability evidence**. Educational AI benchmarks assess what an AI model, system, or product demonstrates under specified task and evaluation conditions. They do not replace research on student and educator outcomes or close qualitative examination of human–AI interactions. Instead, benchmarks provide evidence about the performance and capability of the AI itself. They can help researchers unbox what the AI is doing, how it is performing, and whether that performance

may be a correlating factor in findings about the effects of an intervention on students or educators. Figure 1 presents a summary of these three types of evidence in about AI-enabled products and practices.

***Figure 1. Types of Evidence***




## What an AI Benchmark Is

In the broader field of AI and machine learning, automated benchmarks are a standard mechanism for assessing models and systems. Raji et al. (2021) define a benchmark and its components—tasks and metrics—as:

> *a particular combination of dataset or sets of datasets (at least test data, sometimes also training data), and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods. The task is a particular specification of a problem, as represented in the dataset. A metric is a way to summarize system performance over some set or sets of tasks as a single number or score. The metric provides a means of counting success and failure at the level of individual system outputs and summarizing those counts over the full dataset.*

A well-designed benchmark is similar to a well-designed assessment. Benchmarks are meant to measure specific capabilities across a range of settings, inputs, and contexts. For example, Figure 2 shows sample questions from the SuperGLUE benchmark, where AI models are presented with a passage and a question to answer. SuperGLUE is used to measure performance for general-purpose language understanding (A. Wang et al., 2020). A benchmark should have a

clear model of what is being assessed, how the AI would demonstrate capability in that particular skill or area of knowledge, and the circumstances under which performance may vary.

### *Figure 2. SuperGLUE Benchmark Example Questions (A. Wang et al., 2020)*

| | |
|---|---|
| **BoolQ** | **Passage:** *Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.*<br>**Question:** *is barq's root beer a pepsi product* **Answer:** `No` |
| **CB** | **Text:** *B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?*<br>**Hypothesis:** *they are setting a trend* **Entailment:** `Unknown` |
| **COPA** | **Premise:** *My body cast a shadow over the grass.* **Question:** *What's the CAUSE for this?*<br>**Alternative 1:** *The sun was rising.* **Alternative 2:** *The grass was cut.*<br>**Correct Alternative:** `1` |
| **MultiRC** | **Paragraph:** *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*<br>**Question:** *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (`T`), *No* (`F`), *Yes* (`T`), *No, she didn't recover* (`F`), *Yes, she was at Susan's party* (`T`) |

Well-designed benchmarks also provide repeatable evaluation protocols that can be administered across multiple models and products. Repeatability applies to the evaluation protocol (the tasks, inputs, conditions, scoring procedures, and reporting conventions), not necessarily to the generated responses, which may vary across runs. Although rigorous benchmarks require substantial initial investment in domain modeling, task development, expert annotation, scoring, and validation, they can make repeated comparisons more scalable than relying on teams of researchers to review every output manually.

Benchmarks can help product developers compare a product with other products or with foundational models prompted to perform similar activities. Such comparisons can help decision-makers determine whether a product performs well enough for its intended use and whether it provides value beyond a free or generally available model. For researchers, benchmark evidence can document how the AI component performed during an intervention and provide information that may help explain observed outcomes. Researchers and decision-makers should be made aware of the differences in benchmark results between the model version included in a study and the version currently being used, particularly when those differences are related to the product's intended use. Product developers can use changes in the benchmark performance to identify areas requiring further scrutiny and determine whether a model update has affected the theoretical underpinnings of the solution being delivered.

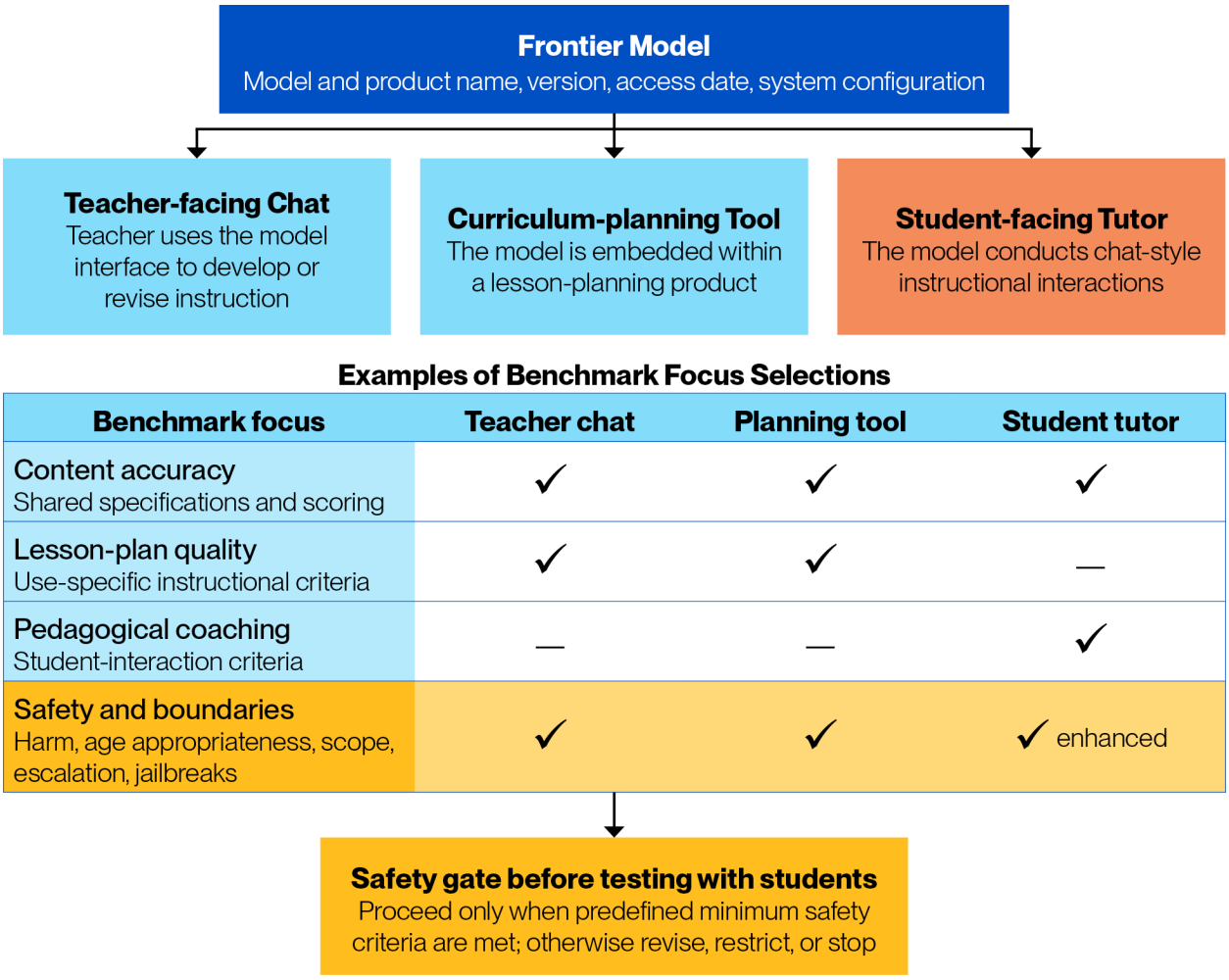
# Why Benchmark Design Matters

The nondeterministic nature of generative AI makes careful benchmark design especially important. The same prompt given to the same model may generate different responses depending on the context of the conversation, the user, or even the day and time. Small changes to a model, product, prompt, or task may also result in different or undesirable behavior. Evaluation cannot rely on a single observation in which an answer looks "surprisingly good" or "good enough." It must test performance across multiple scenarios and clearly define and validate the capabilities, constructs, and assessment mechanisms being used.

The mere existence of a benchmark does not ensure that its results support accurate or useful conclusions. Without careful construct and task design, a benchmark can overestimate or underestimate AI performance. For example, a benchmark may represent only the easiest portion of capability, omit context in which performance is likely to vary, or use a scoring procedure that rewards surface features rather than the intended construct. At a minimum, inaccurate conclusions about AI capability can have downstream effects on student or educator outcomes. At worst, poorly designed benchmarks can communicate trust and safety in a particular AI system or AI-enabled tool while model outputs contribute to harms to students or other users (e.g., Maples et al., 2024; Shimgekar et al., 2026). At a moment when education needs to rebuild trust in technology, well-designed benchmarks provide an important independent verification step that complements other forms of data and storytelling.

Basic safety benchmarking should occur before any AI-enabled product is tested with students, who are a vulnerable population. This evaluation should examine whether the product produces harmful or age-inappropriate responses, goes off topic, or provides guidance in situations that should instead be addressed by mental health or medical professionals. It should also examine whether students can prompt the system outside of its intended boundaries. As developments in agentic AI continue to demonstrate that models can be over indexed toward task completion rather than safety (Slutzky et al., 2026), the field must be able to assess not only pedagogical and content accuracy but also the safety of educational users (Australian Education Research Organisation, 2026; Yang et al., 2026). Figure 3 shows an illustration of shared and use-specific benchmarking of an AI model across educational contexts.

### *Figure 3. Shared and Use-Specific Benchmarking of an AI Model Across Educational Contexts*




*Note.* The same documented model snapshot may be used through different products and practices. Some capabilities, such as content accuracy and safety, should be evaluated across uses, whereas other capabilities require use-specific benchmarks. Benchmark findings are bounded by the model version, product configuration, use context, tasks, and evaluation conditions.

More rigorous capability benchmarking should be used whenever foundational models are integrated into educational products. For product developers and decision-makers, this evidence can help determine whether a product is worth its cost and whether a foundational model can perform the same activity better and safely without the added cost. For researchers, benchmark evidence provides important information about how the AI performed during an intervention and how that performance may relate to observed student or educator outcomes.

Models and the products that rely on them should be benchmarked again when an update is made to the model or to the tool built on the model. High-quality benchmarks should be easy to rerun and hosted on platforms that make them available to the community. Researchers should use and cite relevant benchmarks in studies of AI-enabled interventions, including the date and

model or product version benchmarked, so that the benchmark evidence directly corresponds to the intervention being assessed.

## The Need for Educational and Domain Expertise

Education as a sector must invest in developing and maintaining automated benchmarks with carefully designed tasks for the rapid testing and evaluation of generative AI in educational settings. This requires financial and infrastructure investment as well as investment in human capacity, shared norms, and changes to standard research practice, shared documentation, and standard reporting practices.

Most benchmark developers are currently computer scientists or AI specialists. Guidance from the Stanford Institute for Human-Centered Artificial Intelligence recommends engaging domain experts in benchmark design because benchmarks developed without them may not meet the needs of applied fields (Reuel et al., 2024b). Education researchers and educators have spent decades building expertise in breaking complex capabilities into discrete, measurable components. Deep domain knowledge is needed to define capabilities that matter for a particular educational use, identify evidence that would demonstrate those capabilities, design tasks that elicit that evidence, and determine how performance on those tasks should be scored and interpreted.

Technical partners and tools such as Kaggle (n.d.) can provide the structure and examples needed to turn these task models into executable benchmarks that are visible, easily run across a wide variety of models, and measurable using large datasets for verification. These tools make implementation more accessible to researchers outside computer science and AI specialties. Education researchers and learning scientists do not need to wait to be invited into benchmark development as domain experts by technologists. They can lead benchmark development from their own areas of expertise while collaborating with technical partners on implementation and scaling.

Education researchers and learning scientists do not need to wait to be invited into benchmark development as domain experts by technologists. They can lead benchmark development from their own areas of expertise while collaborating with technical partners on implementation and scaling

Benchmarking is therefore not only a technical expertise; it is a form of assessment design—work with which education researchers already have considerable experience. Yet benchmark scalability does not, by itself, make benchmark results trustworthy. The conclusions drawn from a benchmark depend on how its target capabilities are defined, which tasks and contexts are represented, how outputs are scored, and whether the resulting scores support their intended interpretations and uses. These are fundamentally assessment-design questions.

This white paper responds to these questions in three ways. First, it argues for benchmarking as a complementary but important component of educational AI research alongside impact studies and analyses of human–AI interaction. Second, it adapts principles from Principled Assessment Design into a preliminary framework for developing educational AI benchmarks. Third, it illustrates the application of the framework through three use cases involving assessment generation, information extraction and redaction, and rubric-based evaluation of lesson plans.

The paper is written for researchers with an interest in assessing the capabilities of generative AI within educational settings. By aligning benchmarking with a familiar framework of assessment design, we hope to inspire education researchers to leverage expertise to design benchmarks relevant to the application of generative AI in educational settings. This paper is meant to bridge the expertise gap between AI evaluators and education researchers in order to enable broader collaboration in the design of benchmarks for educational settings.

The following sections connect decades of research on Principled Assessment Design to emerging guidance for benchmark development and offer a preliminary framework that education researchers can begin using and refining. The framework is intended to help researchers explore the benchmarking domain, design and evaluate benchmark tasks with rigor, and integrate benchmarking as a component of high-quality education research involving AI-enabled interventions.

# Principled Design of Benchmarks

*On Assessment: measure what you value instead of valuing only what you can measure.*

—Andy Hargreaves

Hargreaves's observation is especially important for the design of AI benchmarks. Careful consideration of the domain and construct is necessary if benchmarks are to assess the capabilities and behaviors that education values, rather than only those that are easiest to measure computationally. At the same time, the community may come to value what benchmarks measure. Developers frequently use benchmarks as targets when training and adjusting models, meaning that benchmark design can influence not only how model performance is described but also which capabilities developers prioritize improving. Researchers must attend to this interplay as benchmarks become a complementary method for evaluating AI in education.

A principled benchmark is not an ad hoc collection of prompts, datasets, and scores. It is built through an explicit chain of reasoning connecting intended uses and claims, targeted capabilities, observable evidence, tasks, scoring procedures, and score interpretations. Principled Assessment Design (PAD; Ferrara et al., 2017; Nichols et al., 2017) provides the broader principles guiding this design process. Within the family of PAD approaches, we draw selectively on Evidence-Centered Design (ECD; Mislevy & Haertel, 2006) because its claim–evidence–task logic provides a practical method for translating an intended interpretation into observable evidence and tasks designed to elicit that evidence. We adapt this logic to AI benchmarking by defining what an AI system is expected to demonstrate, what observable features of its output or behavior would constitute evidence of that capability, and what tasks would provide opportunities for the system to produce that evidence.

In this context, a benchmark task is a standardized opportunity for an AI system to demonstrate a targeted capability. It specifies the input or context provided to the system, the instructions or prompt, the expected output, and the conditions under which the output is generated. A benchmark generally includes multiple task instances so that performance can be examined across relevant variations in content, format, context, or difficulty. Designers should specify the task features intended to elicit observable evidence of the targeted capability. In ECD, these specifications are formalized in a task model, which distinguishes features that must remain consistent from those that may vary across task instances. Benchmark designers should also document why particular tasks and variations were selected, how the resulting outputs will be scored, and what evidence the scores are intended to provide. The benchmark's boundaries must make clear what it assesses and what it does not.

**SRI's History in Task Specification**

In the late 1960s, SRI researchers created STRIPS (Stanford Research Institute Problem Solver), which established a standardized language for task definition, including initial states, goals, preconditions, and actions. STRIPS was one of the earliest foundations for computational agents to be evaluated on tasks. This research was one of the early approaches to measuring algorithmic performance, a field that later developed into benchmarking.

These design decisions create an assessment argument for the benchmark. Claims about an AI system's capabilities are supported by scores; those scores are based on evidence identified in the system's outputs or behaviors; and those outputs or behaviors are elicited through tasks aligned with the targeted construct. The benchmark-design process therefore produces both standardized tasks that present an AI system with specified inputs, instructions, and conditions and scoring procedures for evaluating the resulting outputs. Making this chain explicit helps explain why a system received a particular score, what interpretation that score supports, and which assumptions must be reviewed during design and tested empirically after implementation.

The design of educational assessments offers a useful foundation for this work. Educational assessment design does not begin with writing questions. It first considers the purpose of the assessment, what is to be measured, the ways students can demonstrate their knowledge and skills, and how the resulting evidence will be scored and interpreted. Rigorous approaches to construct definition, task design, scoring, and evaluation and validation have been developed and studied in educational assessment for decades. The design of educational AI benchmarks should employ similar rigor.

The analogy between student assessment and AI benchmarking is useful, but it is not exact (see Table 1). Students often demonstrate knowledge and skills in ways that align with learning trajectories or are coherent with the instruction they have received. AI systems do not necessarily develop capabilities through comparable trajectories. Models and products change through retraining, fine-tuning, changes in data, prompting, system instructions, retrieval mechanisms, and other technical adjustments. A task that is difficult for one model version may become easy in the next, while performance on other tasks may decline. Benchmark designers therefore cannot assume that AI capabilities develop in a coherent sequence comparable to human learning.

***Table 1. Similarities and Differences Between Educational Assessment and AI Benchmarking***

| Design Feature | Educational Assessment | Educational AI Benchmark |
|---|---|---|
| Object of evaluation | Student knowledge, skills, or performance | AI system capabilities or behaviors |
| Opportunity to demonstrate performance | Assessment item or activity | Benchmark task containing inputs, instructions, and conditions |
| Observed evidence | Student responses or work products | AI outputs, actions, or interaction patterns |
| Evaluation mechanism | Answer key, rubric, or measurement model | Reference value, deterministic rule, rubric, judge, or other evaluator |
| Result | Evidence supporting interpretations of student performance | Evidence supporting interpretations of AI system performance under specified conditions |
| Important source of variation | Items, occasions, examinee conditions, and raters | Prompts, task instances, runs, contexts, evaluators, and model versions |
| Developmental assumption | May be informed by a model of learning or developmental progression | Does not assume a human-like learning trajectory |

Despite these differences, benchmark designers can draw on the same principles used to define what is being assessed, identify observable evidence, design tasks, create scoring procedures, and bound the interpretations supported by the results. The central question for this white paper is therefore not whether assessment design principles apply to AI benchmarking without modification, but how those principles can be adapted to the distinctive characteristics of AI systems:

> *In what ways can principles from rigorous educational assessment design be adapted to support the design, implementation, and evaluation of educational AI benchmarks?*

The following subsection describes PAD as the conceptual foundation for the proposed framework and explains how selected operational components from ECD are adapted to educational AI benchmark design.

# Principled Assessment Design as a Conceptual Foundation

The need for a principled approach to assessment design became especially visible during the expansion of standards-based accountability in the United States during the late 1990s and early 2000s. During this period, increases in scores on some high-stakes state assessments were not always accompanied by comparable gains on external measures such as the National Assessment of Educational Progress (NAEP; Kober et al., 2008; Koretz & Barron, 1998). Such discrepancies raised questions about whether score increases reflected broader student proficiency or knowledge and strategies specific to particular tests. They also drew attention to the test development process itself. Assessment development could be distributed across groups

responsible for standards, curriculum and instruction, item development, scoring, and reporting without an explicit framework connecting their decisions. This fragmentation could produce assessments for which it was unclear what knowledge or skills were being measured, why particular tasks were selected, what evidence student responses provided, and how the resulting scores should be interpreted.

Without a systematic connection among intended constructs, assessment tasks, observable evidence, and score interpretations, assessments are vulnerable to construct-irrelevant variance, where performance is influenced by skills unrelated to the intended construct, and to construct underrepresentation, in which the assessment fails to capture the breadth or depth of the targeted knowledge and skills. These challenges motivated the development of PAD frameworks that make the reasoning underlying assessment development explicit and connect assessment purposes, claims, evidence, tasks, and score interpretations.

PAD refers to a family of systematic, theory-driven approaches to assessment development rather than a single methodology (Ferrara et al., 2017; Nichols et al., 2017). These approaches include Cognitive Design Systems (Embretson, 1998), Evidence-Centered Design (Mislevy & Haertel, 2006), Principled Design for Efficacy (Nichols et al., 2014), and Assessment Engineering (Luecht, 2013). Although these approaches differ in their implementation, they share a common philosophy that meaningful assessment requires explicit alignment between what is intended to be measured, the evidence collected, the assessment tasks used to elicit that evidence, and the interpretations made from assessment results.

PAD is closely aligned with the National Research Council's (2001) assessment triangle, which emphasizes the interdependence of cognition, observation, and interpretation in assessment design. Collectively, PAD approaches seek to ensure that assessment decisions are guided by explicit theories of learning and performance, systematically translated into assessment tasks and scoring procedures, and supported by evidence for the intended interpretations and uses of assessment results.

Ferrara et al. (2017) identify five foundational design elements that characterize PAD: (1) clearly defined assessment targets; (2) statements of intended score interpretations and uses; (3) models of cognition, learning, or performance; (4) aligned measurement models and reporting scales; and (5) purposeful assessment activities designed to elicit evidence aligned with the intended interpretations. These five elements provide a coherent design process in which assessment targets, evidence, measurement, and interpretation are intentionally aligned rather than developed independently.

Underlying the five elements is the overarching objective of constructing a validity argument supporting the intended interpretations and uses of assessment scores. Consistent with Kane's (2013) argument-based approach, validity is not treated as a property of an assessment or score by itself. Instead, validation involves accumulating evidence supporting the interpretations, claims, and decisions that assessment developers intend to make from assessment results.

> Validity reasoning therefore begins during assessment design, when targets, evidence, tasks, scoring, and intended interpretations are aligned, and continues through empirical evaluation after implementation.

The following subsections describe each of the five foundational elements of PAD. These elements subsequently serve as the conceptual foundation for our benchmark guidance framework, which adapts principles of PAD to the development, implementation, and evaluation of educational AI benchmarks.

## 1. Clearly Defined Assessment Targets

The first element of PAD is the development of clearly defined assessment targets. Assessment targets specify the knowledge, skills, and abilities (KSAs) that an assessment is intended to measure and establish the foundation for subsequent design decisions. They may be derived from construct definitions, models of cognition or learning, or existing content standards such as those used in state assessments. Defining assessment targets explicitly encourages the use of theory and empirical evidence to characterize how learners develop, demonstrate, and apply the targeted knowledge and skills. Once established, the targets guide the design of assessment tasks, scoring procedures, reporting practices, and other components of the assessment system.

Assessment targets can be specified at different levels of granularity (see Table 2). A broad target may represent an integrated domain but require a larger and more varied collection of tasks to represent that domain adequately. More fine-grained targets can support specific and diagnostically useful claims, but they may represent only a limited portion of the broader domain. The appropriate grain size therefore depends on the assessment's purpose, context, intended interpretations, and available task sample. This same trade-off later applies to AI benchmark design: Broader capabilities may better approximate complex real-world performance, whereas narrower sub-capabilities can provide more precise information about the sources of system success or failure.

***Table 2. Illustration of Construct Granularity***

| Level of Granularity | Human Assessment Example | Educational AI Benchmark Example |
|---|---|---|
| Broad construct | Computational thinking | Generate a high-quality science assessment |
| Component or capability | Algorithmic thinking | Standards alignment |
| Fine-grained target | Construct a sequence of instructions satisfying specified constraints | Generate an assessment task that elicits evidence of a specified performance expectation |
| Observable evidence | Student produces a correct and efficient sequence | Generated task requires students to demonstrate the targeted knowledge or practice |

## 2. Statements of Intended Score Interpretations and Uses

The second element of PAD is the development of explicit statements describing the intended interpretations and uses of assessment scores. These statements articulate the conclusions that assessment developers intend stakeholders to draw from assessment results and the decisions those results are intended to support. Intended score interpretations should be developed concurrently with assessment targets because they provide direction for all subsequent design, development, and implementation decisions. Traditionally, these interpretations are operationalized through mechanisms such as proficiency-level descriptors, which define what examinees at different performance levels are expected to know and be able to demonstrate.

## 3. Models of Cognition, Learning, or Performance

The third element of PAD is the development of explicit models of cognition, learning, or performance that explain how students or examinees acquire, organize, and demonstrate competence within the targeted domain. Models of cognition describe how students represent and apply knowledge to solve problems. Models of learning describe developmental progressions through which competence develops over time. Models of performance specify how competence is manifested through observable behaviors or work products. These models provide the theoretical basis for determining what evidence an assessment should collect and what inferences can reasonably be drawn from observed performance. Within ECD, for example, claims about an examinee's competence are represented in the student or proficiency model, while the evidence model specifies how observable features of the examinee's performance provide evidence for those claims. The task model then specifies the situations and features of tasks designed to elicit that evidence.

## 4. Aligned Measurement Models and Reporting Scales

The fourth foundational element of PAD focuses on selecting measurement models and reporting scales that are aligned with the assessment targets, intended interpretations, and underlying cognitive or performance models. Measurement models provide the inferential link between observed assessment responses and the conclusions that assessment developers intend to make about student or examinee competence. Depending on the assessment purpose, the models may be based on classical test theory, item response theory, Bayesian measurement approaches, diagnostic classification models, or other psychometric frameworks. Reporting scales translate measurement results into forms that are meaningful and useful for the intended users, whether through scale scores, proficiency levels, classifications, or other reporting frameworks.

## 5. Purposeful Assessment Activities

The fifth element of PAD is the purposeful design of assessment activities that elicit evidence aligned with the assessment targets and intended score interpretations. Assessment tasks

provide opportunities for students to demonstrate the targeted KSAs while minimizing sources of construct-irrelevant variance. Task features—including context, response format, cognitive complexity, required knowledge and skills, and supporting materials—are deliberately selected so that the resulting performances provide the evidence needed to support the assessment's intended interpretations.

Task designers must also consider additional or prerequisite KSAs that a task may require but that are not part of the primary assessment target. Within ECD, these are sometimes described as additional KSAs, or AKSAs. For example, a task intended to assess scientific reasoning may also place demands on reading comprehension. When such demands are not part of the intended construct, designers should minimize their influence or provide sufficient support so that performance more directly reflects the targeted KSA.

Tasks specifications can distinguish between required features and variable features. Required features ensure that each task provides an appropriate opportunity to elicit the targeted evidence. Variable features may be deliberately changed across tasks, such as context, complexity, reading load, cultural setting, programming language, or response format. Varying these features allows designers to represent the target domain and examine whether performance changes across relevant conditions. Within ECD, these specifications are formalized in the task model.

The same logic applies to AI benchmark tasks. A benchmark task may unintentionally require capabilities unrelated to the intended target—for example, processing a complex document format when the intended capability is scientific reasoning. Benchmark designers should identify these additional demands and decide whether to minimize them, provide support for them, or include them explicitly within the benchmark claim. Similarly, benchmark task models should distinguish features that must remain consistent from contextual features that can be varied to examine performance across relevant conditions.

## Adapting PAD to Benchmark Design and Evaluation

PAD informs AI benchmark design through the same general chain of reasoning used in PAD: intended uses and claims, targeted performance, observable evidence, tasks, scoring, and score interpretation. Aforementioned, within the broader family of PAD approaches, we draw selectively on ECD because its claim–evidence–task logic provides a practical structure for translating an intended interpretation into observable evidence and tasks designed to elicit that evidence.

This adaptation is functional rather than literal. The framework does not assume that AI systems possess or develop human knowledge, skills, and abilities or that their capabilities follow human learning trajectories. Instead, it emphasizes a performance model: an explicit account of what an AI system must demonstrate, how that performance should become observable in its outputs or behavior, and under which conditions the evidence will be elicited

and evaluated. The appropriate level of granularity—from a broad, integrated capability to a more specific sub-capability—depends on the benchmark's intended uses and claims.

A benchmark score should describe an observed demonstration of performance rather than treat capability as a stable property of a model or product. Claims should therefore be bounded by the prompts, data, task sample, scoring procedures, model or product version, and evaluation conditions included in the benchmark. A score may provide evidence that a system demonstrated specified capabilities under those conditions, but it does not establish that the system will demonstrate them consistently across all prompts, contexts, users, future runs, or model versions.

The proposed framework supports an initial argument-based validity rationale by requiring explicit alignment between intended uses and claims, targeted capabilities, observable evidence, tasks, scoring procedures, and score interpretations. Design-stage alignment does not, by itself, establish the validity or reliability of benchmark scores. Empirical evidence is still needed to examine consistency, generalizability, fairness, evaluator performance, and relationships with external criteria. The following section operationalizes these principles through the components of the proposed Principled Benchmark Design framework.

# A Framework for Principled Benchmark Design

The design of benchmarks has similar components to the design of educational assessments. Although benchmarking has been used in practice since the 1980s, the rise of generative AI has created a surge in the need for benchmarks, a broader variety of the capabilities that benchmarks assess, and a broader set of domains, including education, where content-specific benchmarks measure AI capabilities for common uses within the field.

A benchmark is a purposeful collection of tasks, such as individual questions on an assessment, where the designer of the benchmark has selected the items to measure a specific competency (Bean et al., 2025; Y. Wang et al., 2024). Tasks can be used (and reused) across multiple benchmarks where the capabilities being assessed share contructs that contribute to the broader performance of the model. Tasks are often combined in different ways to evaluate similar but slightly different capabilities of a model or program being tested. For example, a benchmark designed to test a model's capabilities to tutor a student in math, and another benchmark designed to test a model's capabilities to tutor a student in literacy, may share tasks related to the model's ability to stay on task with the student.

Benchmarks can be written to test models or to assess the behavior and capabilities of a specific program or software. Evaluating a model with a benchmark is about more than just creating the tasks, it is also about designing a collection of tasks which combine to create a meaningful score on the benchmark. The scoring design process for a benchmark is similar to creating a scoring guide in order to apply it to students' assessment responses. Like an assessment scoring guide, a benchmark's scoring procedures translate performance across tasks into interpretable evidence about how well an AI model or educational software system demonstrates the targeted capability under the conditions tested.

In this framework, we discuss creating tasks, compiling tasks into purposeful benchmarks, and evaluating the benchmark results to determine the AI model's capabilities as well as the benchmark's quality, reliability, validity, and appropriateness for its intended use.

Principled Benchmark Design offers a structured approach to aligning goals, task and benchmark design, and evaluation to the specific capabilities to be assessed. Just as a well-designed collection of questions or assessment items does not automatically guarantee the validity and reliability of the assessment, a collection of carefully designed tasks with construct validity does not necessarily guarantee a meaningful benchmark. Therefore, in this framework we connect PAD approaches associated with the design and evaluation of assessments to the steps to design, implement, and evaluate the benchmark itself before relying on the benchmark's assessment of the AI's capabilities.

Relatedly, in January 2026, the National Institute of Standards and Technology (NIST, 2026) published *Practices for Automated Benchmark Evaluations of Language Models* offering a three-step process for benchmark evaluation. While the paper offers design insights, it states that "consistent practices to support the validity and reproducibility of such evaluations are only beginning to emerge" (p. 6). NIST's three steps are:

- Defining model evaluation objectives and selecting benchmarks
- Implementing and running model evaluations
- Analyzing and reporting results

Although the steps are simple enough in their description, the careful design, implementation choices, and analysis can significantly impact not only the quality of the benchmark itself, but also the community's ability to use the benchmark evaluation data to inform adoption, development, and trust of AI models and educational software.

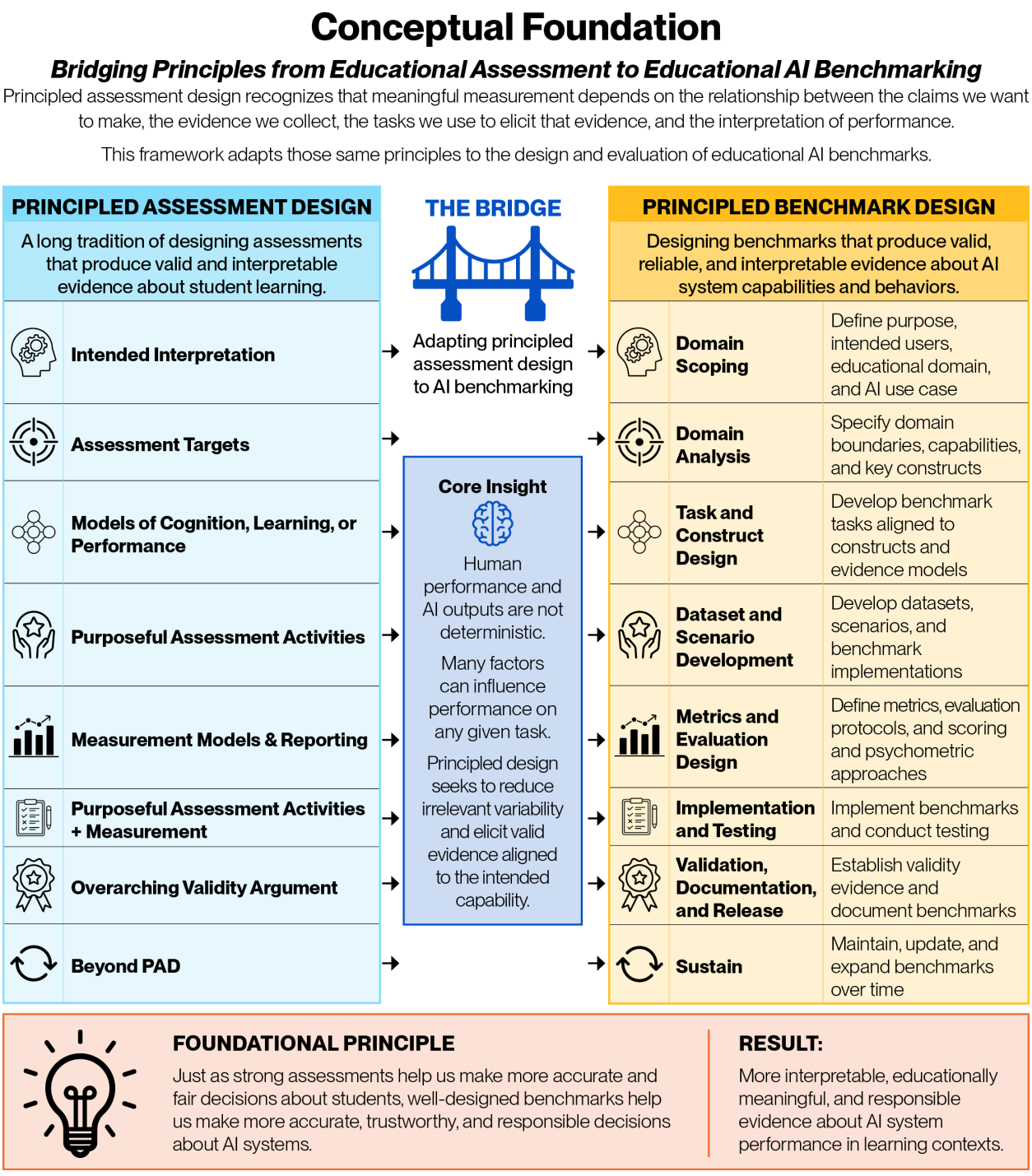
In this section, we complement NIST's guidance with domain-specific examples for education and with guidance for defining a benchmark's purpose, creating the evaluation protocol, and measuring initial results and connecting the guidance to PAD for familiar frameworks for education researchers.

Next, we describe our proposed framework for the design of educational AI benchmarks. The key to the overall creation of tasks, and ultimately benchmarks, is the benchmark's intended claims and uses as well as clear definitions of the capabilities and constructs to be measured. Figure 4 shows the connection between PAD and our proposed framework. In a later section (Applications of Principled Benchmark Design), we apply our proposed benchmark design framework to three common educational use cases. These examples focus on:

- Generating a first-grade science assessment—a bounded content generation task
- Extracting and redacting sensitive information in Individualized Education Program (IEP) documents
- Evaluating quality of lesson plans for teachers

Each use case highlights a different level of complexity in the design and evaluation of benchmarks and illustrates the components of our proposed framework and their alignment to PAD as an iterative approach to the task and benchmark design. For guidance, in the framework components for each use case, we pose questions to the benchmark designer to prompt reflection, potential revision, and alignment, as well as suggest opportunities to evaluate or check the task and benchmark design. These examples are preliminary in their design and meant to illustrate a process as opposed to represent the deep analysis and design to produce completed benchmarks. Future work will need include robust benchmark design and analysis outside the scope of this white paper.

*Figure 4. A Framework for the Principled Design of Benchmarks*




# 1. Domain Scoping

The first component of Principled Benchmark Design is domain scoping: outlining what the benchmark is intended to measure and what is outside its scope. Benchmark design should be aligned with a clearly defined capability, characteristic, or the scope it covers (Reuel et al.,

2024a). These boundaries connect to PAD's emphasis on intended interpretations and uses to establish the direction for the detailed construct definitions and subsequent task design.

For example, a benchmark evaluating AI's ability to generate a first-grade assessment for Earth and Space science might focus specifically on the ability of AI to produce an assessment aligned with standards. Or the benchmark might be broader to include whether there was variability in the type of question formats used, the reading level of the questions (and whether they match the expected grade reading level), or other factors. Whether the assessment to be designed is a quiz, a unit test, or a high-stakes standardized assessment will also inform the scope of the domain evaluated by the benchmark.

Clear boundaries help designers identify the range of prompts, desired output functions, and exemplars that benchmark tasks should include. They can also help constrain claims about the capabilities and contexts the benchmark examines, addressing concerns about benchmarks being used to support claims of general capability that their tasks and data cannot substantiate (Raji et al., 2021).

Scoping requires consideration of both the users of the AI and the users of the benchmark. The intended users of the AI help benchmark designers anticipate variability in user interactions for the AI application within the educational setting and expectations for how the AI should respond to the user. For example, a student using an AI chatbot in a social studies product may have a larger variability to their interactions, including some that are off task, than those of an educator using AI to support lesson planning. Combining the profile of the AI's users with the boundaries of the benchmark domain helps in later phases of the benchmark design process by enabling alignment in task design.

The intended users of the benchmark also inform what evidence will be useful and how results may be interpreted. Researchers may use benchmarks to complement other methodologies, examining AI performance to help explain its effects on users. Developers may use the same benchmarks to compare their products with others, while the broader education community may use benchmark results to understand what AI does and assess the evidence for its reliability on the tasks that make up the benchmark. Across these uses, benchmark designers should define the scope of the benchmark to make explicit the connection between observed performance and the capability being assessed so that users can understand how to interpret the benchmark scores (Jo & Wilson, 2025).

Domain expertise, relevant research, and practitioner input should inform the definition of these boundaries, helping benchmark designers identify important variables and elements of instructional practice, learning, or administrative tasks that the AI is intended to support. This approach is consistent with the emphasis on scoping and co-design in benchmark development guidance (Reuel et al., 2024a). The resulting scope provides a foundation for the next step, where we define the domain and deconstruct the broader activity into component concepts that an expert would consider when evaluating an AI response.

# 2. Domain Analysis

Domain analysis establishes the foundation for defining what high-quality performance looks like within the benchmark scope and is an important step in defining the theoretical underpinning of the benchmark's design. This component connects to PAD's clearly defined assessment targets by drawing on theory and empirical evidence, where available, to identify the knowledge, sub-capabilities, and behaviors relevant to the intended domain. Education research often distinguishes novice and expert performance on the activities that AI is being asked to replicate or support. Designers should examine how researchers have characterized these differences and the criteria used to evaluate resulting work.

For example, the Science Assessments for Language-Diverse Elementary Education (SALDEE) project has focused on developing and validating Next Generation Science Standards (NGSS)-aligned formative assessments for linguistically diverse first-grade students (e.g., Rachmatullah et al., 2026b). The SALDEE assessment development work and the project team's substantive expertise provided a starting point for identifying dimensions of assessment task quality relevant to AI-generated assessments. These dimensions included Content, Assessment Design, Fairness and Accessibility, and Implementation Feasibility. In Use Case 1 (see Applications of Principled Benchmark Design), we chose to focus benchmark design on Content as the main construct, which includes identifying Scientific Accuracy, Standards Alignment, and Standards Coverage as the sub-construct capabilities for the use case. The other dimensions of Fairness and Accessibility and Implementation Feasibility remained outside the scope for this use case. This example illustrates how domain analysis draws from a broader evidence base to identify relevant criteria within the established boundaries.

The boundaries established during domain scoping guide which parts of an evidence base are relevant and which fall outside the benchmark's intended scope. Within those boundaries, expert human performance can provide a reference for the quality expected of the AI. Markers of nonexpert or lower-quality outputs are also useful because they help identify explicit variables to examine during the next phase of task and construct design, given that generative AI models are trained to behave as humans. Existing rubrics, criteria, and exemplars from previous literature can inform the definition of the capabilities being assessed and make the relationship between observed performance and those capabilities explicit (Jo & Wilson, 2025).

Documentation should explain the inclusion and exclusion of relevant research, criteria, and examples, making the reasoning available for future replication of benchmark task design and for publication or sharing. Where the evidence base is insufficient, designers should identify those gaps and document the resulting risks to the validity of the benchmark's intended interpretations, just as researchers would specify threats to validity in other studies. At the end of this step, designers should create a documented account of expected performance, indicators of lower-quality outputs, and limitations in the supporting evidence to inform the task and

construct design. These documented examples will form the basis for task design and support the eventual construction of scoring guidelines in later steps.

# 3. Task and Construct Design

Task and construct design connects the constructs identified through domain analysis to explicit claims and observable evidence in the AI system's output. This component draws on PAD's alignment of intended claims with evidence of performance and ECD's specification of how observable features support those claims (Ferrara et al., 2016; Mislevy & Haertel, 2006). Benchmark designers specify the tasks the AI system should perform, the capabilities it should demonstrate, and the features of its output that would support or contradict the intended claim. Prompts, setup, and other information provided to the system are addressed in Component 4: Dataset and Scenario Development, while scoring scales and aggregation are addressed in Component 5: Metrics and Evaluation Design.

The Appendix A provides a template to scaffold the connection between constructs and observable evidence. Researchers should continually ask whether they are defining construct-evidence pairs at the smallest appropriate level of granularity so that nuance and detail are represented in the benchmark's evaluation. For example, evaluating the scientific accuracy of a first-grade assessment requires criteria that make specific inaccuracies observable, beyond a general judgment that the assessment appears correct. As AI improves, these distinctions will become increasingly important for identifying where its outputs remain incorrect or fall short of the expected quality. AI often produces outputs that look "good enough." When the system supports a learner or novice teacher, rigorous criteria are especially important because the user may be less able to recognize or interrogate subtle errors.

Designers must also consider what reference data are available to evaluate the output. When using a preexisting labeled dataset, they should first establish that its labels rigorously align with the intended constructs. For example, in Use Case 2 (see Applications of Principled Benchmark Design), a dataset containing simulated student IEPs and the accompanying de-identified documents could be used to assess AI's capability of redacting identifiable information. The task would present the full IEP and compare the model's output with the correctly de-identified file. By using a prelabeled dataset, the task can evaluate the AI system's performance against curated reference labels. If the labels capture only part of the intended construct to be measured, designers should narrow the scope and corresponding claims or develop complementary tasks that cover the remaining dimensions. The availability of data should not substitute for an explicit account of the capability being evaluated (Jo & Wilson, 2025).

When suitable labeled reference data are unavailable, one approach is to use a separate large language model (LLM) to evaluate generated outputs against defined criteria. Throughout this framework, the AI system being evaluated is the model or product performing the benchmark task; the judge LLM will refer to the model applying evaluation criteria to that system's output.

The judge LLM must receive explicit guidance or training on the constructs and observable evidence it is expected to assess, just as a human rater would need to be trained to apply a codebook or scoring rubric. The ability of the judge LLM to apply those criteria accurately and consistently must itself be evaluated, including comparison with expert human judgment. Defining the criteria and evidence needed to assess the judge LLM at this stage establishes requirements for subsequent evaluation design and testing.

# 4. Dataset and Scenario Development

Dataset and scenario development translates the task specifications into the materials, prompts, and interaction conditions presented to the AI system being evaluated. This component connects to PAD's purposeful assessment activities and ECD's task model, which specifies the situations and task features used to elicit relevant evidence (Ferrara et al., 2016; Mislevy & Haertel, 2006). The user profiles and use cases established during scoping guide what should vary and what should remain constant when constructing prompts and different tasks. Designers should deliberately consider which sources of variation are relevant to the intended capability and which should be controlled (Jo & Wilson, 2025).

For example, different educators may vary in how they ask AI to generate a science quiz for their class. Some educators may ask for items about a topic such as first-grade space science while not explicitly specifying the standards the assessment is meant to assess, the grade level, students' backgrounds, or instructional context. Other educators may be more specific. These are important variabilities to consider because benchmarks should measure capabilities across a variety of user inputs if the capability being tested involves human collaboration. Scenarios should represent meaningful variation in how intended users engage with the AI system, allowing the benchmark to examine how that variation affects the quality of the system's output. The first-grade science assessment generation example in the Use Case 1 illustrates how task-model variables can organize these decisions.

Scenarios may allow the AI system to ask clarifying questions, make explicit assumptions, or generate a response from the information available. Unless the interaction process itself is the capability being measured, the benchmark should evaluate the final product against the defined criteria. Where scenarios include multiple turns, designers should specify how follow-up information is supplied and when the interaction ends so that the conditions leading to the final product are documented and reproducible.

Datasets should draw on a combination of authentic materials or interactions and expert-authored or simulated examples, consistent with recommendations to include both naturally occurring and expert-created examples in benchmark development (Bowman & Dahl, 2021). The selection and breadth of use cases must account for the resources, including financial resources, available to develop the benchmark. Documentation should make clear what materials were used, how examples were selected or created, and what variability they represent

so that benchmark users can understand how the intended claims are supported and where coverage remains limited.

Collaborators, expert reviewers, and community input can also help designers assess whether they have covered important variability. AI can also support brainstorming about edge cases, nontraditional use cases, and plausible ways in which users might approach similar tasks. Such suggestions require review against domain knowledge and evidence of actual use; generated descriptions of past use should not be treated as empirical evidence. Together, these sources can help designers broaden scenario coverage while maintaining alignment with the benchmark's scope and intended claims.

At the end of this step, designers should have documented not only the tasks the model should perform in the benchmark, but also the scenarios under which those tasks would occur in the real world. Outputs of this step include what prompts would be used to elicit the task, what variability exists in the users or software where the model could be asked to complete a task, and what data would be useful to evaluate the model's response to the task.

# 5. Metrics and Evaluation Design

Metrics and evaluation design specify how observable evidence is translated into scores that help users understand the capability demonstrated by the AI system being evaluated. This component connects to PAD's emphasis on measurement models and reporting scales aligned with assessment targets and intended interpretations (Ferrara et al., 2016). The scoring approach should follow from the domain scope, domain analysis, task specifications, and scenario details. As in human assessment, the choice of binary correct/incorrect judgments, rubric-based ratings, or more-complex evaluation methods depends on which approach best captures meaningful differences in performance and supports a useful benchmark score. When rigorous comparisons among AI systems are intended, designers should plan for consistently labeled reference data and enough items in the dataset or collection of benchmark tasks, where possible, so that sufficient statistical power can be achieved and the benchmark is able to detect meaningful performance differences (Bowman & Dahl, 2021).

Designers should determine how scores for individual criteria and tasks are weighted to produce a measurement of each capability. These decisions should reflect what matters most to the intended users of the AI models and to the system's developers, who may use benchmark results to guide performance improvement. For example, weighting scientific accuracy, standards alignment, and standards coverage in the first-grade science benchmark expresses judgments about their relative importance to the quality of the resulting assessment. Designers should document the rationale for those weights and consider how aggregation affects the visibility of strengths and weaknesses. Making the relationship between scores and capabilities explicit is necessary for interpreting what higher performance on a benchmark means (Jo & Wilson, 2025).

Evaluation design also includes specifying the information made available alongside an overall score. Depending on the implementation, platforms such as Kaggle may allow benchmark users to expand scores, inspect outputs from the AI systems being evaluated, and examine other information produced during evaluation. Designers should plan outputs that report performance across tasks and transparently connect those tasks to the capabilities identified through domain analysis and construct design. This approach reflects the argument that benchmarks should help researchers understand where systems work and where they do not, with their usefulness grounded in appropriate scope and content (Raji et al., 2021). These connections help researchers interpret performance, developers identify areas for improvement, and the education community understand the evidence underlying a benchmark score.

This step establishes the scoring rules, aggregation approach, and reporting requirements that guide implementation. Where a judge LLM applies the criteria, its scoring instructions should reflect these specifications. Testing whether the scoring procedures—including judgments made by a judge LLM—work accurately and consistently is addressed in the next component.

# 6. Implementation and Testing

Implementation and testing translates the benchmark design into an executable system and examines whether it performs as intended. As with any computational artifact, a benchmark requires rigorous testing. Encoding the benchmark, running it against AI systems, examining its outputs, and testing it with known inputs provide evidence that the realized benchmark follows the design.

Testing should include simulated responses that mimic outputs from the AI system being evaluated, test datasets with known expected results, and checks of individual scoring components as well as the complete tasks and benchmark. These inputs should represent different levels of quality, common errors, and edge cases so that designers can examine whether the benchmark recognizes the distinctions it was designed to measure. Research on improving agentic benchmarks illustrates the value of reviewing problem descriptions, reference solutions, and tests that cover boundary conditions and common errors (e.g., Zhu et al., 2026). Reference answers and labels also require scrutiny: An implementation that faithfully reproduces flawed reference judgments can still produce misleading scores. Consistently labeled data are therefore an important foundation for trustworthy evaluation (Bowman & Dahl, 2021).

When a judge LLM is used, its ability to evaluate outputs must be tested separately from the performance of the AI system being evaluated. Designers should compare the judge LLM's scores with expert-labeled responses spanning relevant quality levels and error types. Repeated scoring of the same responses can help assess consistency, while comparison with expert judgments examines whether the judge applies the intended criteria appropriately. Acceptance criteria should reflect the benchmark's intended use. Agreement and consistency provide

complementary evidence; a judge that repeatedly assigns the same incorrect score is consistent but does not adequately implement the evaluation criteria.

Unexpected behavior should prompt systematic debugging. Designers can create targeted test inputs, print intermediate results, and trace the execution of the benchmark to determine whether a problem arises from the implementation, the scoring criteria, or the underlying task and construct design. For example, an unexpected score may reflect a calculation error, ambiguity in a criterion, or a task that does not elicit the intended evidence. Each finding calls for revision at the corresponding point in the framework and subsequent testing of the change.

The design, implementation, and testing of benchmarks is an iterative process in pursuit of quality. Testing may require designers to revisit earlier decisions, refine criteria, correct reference data, or revise scenarios. Documenting these changes and the evidence supporting them makes the process transparent and provides a foundation for the subsequent validation, documentation, and release of the benchmark.

# 7. Validation, Documentation, and Release

Validation, documentation, and release bring together the evidence supporting the benchmark's intended interpretations and uses. This component of the framework also makes that evidence available for review and further development. For benchmarks, this requires examining whether the connections between scope, constructs, tasks, observable evidence, and scoring support the claims designers intend users to make.

This white paper focuses on design and emphasizes review approaches that can inform an initial validity argument. Expert review can examine whether tasks adequately represent the intended domain, whether evaluation criteria align with the constructs, and whether important dimensions are missing or unrelated task components influence scores. Review by intended users can also help establish whether the benchmark appears appropriate for its stated purpose and whether the results are understandable. Such judgments contribute evidence about content coverage, construct alignment, and apparent appropriateness; they do not independently establish construct validity. Claims should remain bounded by what the benchmark's tasks, data, and evaluation conditions can support (Raji et al., 2021). In assessment design, the assessment items should not follow the curriculum, but both the curriculum and its assessments should be aligned to the learning goals. Similarly, a benchmark should not follow from a specific products or educational AI software's design, but from the capability the AI model is contributing to the software's performance or user interactions.

The extent and methods of validation will vary with the benchmark's intended use and available resources. Resource constraints should inform a thoughtful selection of validation activities, with explicit documentation of what was examined, what evidence was obtained, and what remains uncertain. Those resource constraints should not lead designers to omit consideration of validity. This documentation enables future researchers and designers to build additional

evidence over time. Further research could explore psychometric methods for educational AI benchmarks. Readers pursuing technical analysis should also consult emerging applications of methods such as item response theory to benchmark tasks, considering whether their assumptions fit the intended evaluation (Botter, 2026; P. Li et al., 2025; Y. Li et al., 2024; Zhou et al., 2026).

Release documentation should consolidate the design rationale, intended claims and uses, review and testing evidence, known limitations, and instructions for use. It should distinguish supported interpretations from those that still require investigation and make the relationships between tasks, capabilities, and scores transparent. This emphasis on documentation supports scrutiny and informed use of benchmark results (Reuel et al., 2024a). Appendix A provides a documentation outline, organized similarly to a structure abstract and paper outline, to help novice benchmark designers describe their decisions and supporting evidence.

Release makes the benchmark and its evidence available for use, examination, and continued improvement. Clear documentation allows others to understand the conditions under which results were obtained, reproduce the evaluation where possible, and contribute evidence that strengthens or revises the benchmark's intended claims.

# 8. Sustain

Sustaining a benchmark requires continued attention to its usefulness as AI systems and educational practices evolve. Just as initial benchmark design is iterative, ongoing development should respond to community input, model advancements, and evidence that tasks no longer detect meaningful nuances in AI performance. Sustaining a benchmark therefore extends beyond hosting its code to revisiting the design decisions and evidence supporting its intended uses. This emphasis reflects the value of benchmarks as tools for understanding where systems perform well and where limitations remain (Raji et al., 2021).

The rigorous attention to documentation throughout the Principled Benchmark Design process and enables participation beyond the original designer. Clear accounts of scope, constructs, tasks, scenarios, scoring, and supporting evidence allow others to remix, adapt, and extend a benchmark while understanding the reasoning behind its design. Revisions should prompt review of the affected components and renewed testing, with additional evidence of validity gathered where changes alter the intended interpretations or use.

As with other software, versioning systems are important tools for documenting change over time. Records should identify what changed, when, and why; preserve earlier versions where feasible; and make clear which version produced a reported result. Changes to tasks, datasets, or scoring may affect score comparability, so users should not assume that results from different versions are interchangeable. As educational AI benchmarking matures as a field, community practices aligned with platforms hosting concentrations of benchmarks will help shape how revisions and adaptations are made, documented, and shared.

Maintenance also depends on available resources. When original designers can no longer sustain a benchmark, documenting its maintenance status and enabling others to continue the work can preserve its value. If a benchmark no longer supports useful interpretations, retirement from active use may be appropriate while its documentation and historical results are retained for reference. These practices allow the community to build on prior work while making the status and limitations of available benchmarks visible.

## From Framework to Practice

Principled Benchmark Design connects the intended uses of a benchmark to the constructs, tasks, scenarios, and evaluation criteria needed to produce meaningful evidence about AI performance. Drawing on decades of educational assessment research, the framework makes the reasoning behind these decisions explicit and available for review. Its components form an iterative process: Implementation and testing may reveal a need to refine constructs or scenarios, while community use and model advancements may prompt further revision. Throughout this process, documentation helps users understand what the benchmark results support, where uncertainty remains, and how future work can strengthen the evidence.

The use cases in the next section illustrate how these framework components work together in three educational applications: generating first-grade science assessments, redacting personally identifiable information in IEP documents, and evaluating science lesson plans. Each use case draws on existing research or development work to establish a bounded purpose, identify relevant and authentic constructs, connect claims to observable evidence, and implement tasks and scoring procedures. They show how the framework can guide different choices of reference data, evaluation methods, and task conditions while maintaining alignment with the intended use. They also offer practical starting points for benchmark designers to adapt, test, and refine within their own domains. These use cases were developed as preliminary illustrative examples, not as exhaustive or fully detailed benchmarks for community use.

# Applications of Principled Benchmark Design

To illustrate the process of Principled Benchmark Design, we offer the following three use cases. These use cases progress from an example using a judge LLM to examples using dataset-driven evaluation. They were developed to illustrate the framework described in the previous section and decisions made during the design process.

## Use Case 1: First-Grade Science Assessment-Task Generation

This use case provides a preliminary example of applying the eight components of the Principled Benchmark Design framework. The subheadings below correspond directly to the framework components presented in the previous section. Components 1 through 6 are illustrated through the design and executable implementation of the pilot benchmark. Component 7 is addressed only through preliminary review and documentation, and Component 8 was outside the scope of the use case pilot. We explicitly identify these limitations because the use case illustrates the Principled Benchmark Design process rather than presenting a fully validated and sustained benchmark. The design yielded an implemented series of tasks and tested in the Kaggle platform.

### 1. Domain Scoping: Purpose, Users, and Boundaries

Formative and summative assessment design is a challenge for emerging educators and is often described as a task for AI to support teachers (Black & Wiliam, 1998; Webb & Galamba, 2026). A June 2025 poll by Gallup indicated that 25% of educators are using AI at least monthly to write assessments and 33% of educators are using AI at least monthly to create worksheets (Walton Family Foundation, 2025). Recent work has compared AI-generated and human-generated assessment items in postsecondary and medical-education settings (Ahmed et al., 2025; Isley et al., 2026). Findings have been mixed and appear to depend on the domain, intended learners, and evaluation criteria. In one postsecondary study, AI-generated questions performed comparably to human-generated questions on some measures and exhibited lower difficulty but higher discrimination (Isley et al., 2026). In a separate medical-education study, however, reviewers rejected more than 30% of AI-generated questions because they exceeded expected student knowledge, were confusing, or contained other flaws (Ahmed et al., 2025). These findings are not directly contradictory because they concern different populations and evaluation criteria.

Domain Scoping established the benchmark's educational context, intended users and uses, target capability, and boundaries. The pilot focused on a bounded generation capability rather than the overall quality of an AI-generated formative assessment. For this use case, we asked, *How well do general-purpose generative AI models generate scientifically accurate and NGSS-aligned first-grade Earth and Space science assessment-task sets intended to support*

*formative use?* The intended benchmark users included assessment and curriculum developers, classroom teachers, AI developers, researchers, and education decision-makers. The pilot evaluated only the Content dimension of the generated assessment tasks: Scientific Accuracy, Standards Alignment, and Standards Coverage. It did not evaluate assessment design, diagnostic scoring guidance, fairness and accessibility, implementation feasibility, instructional effectiveness, or effects on student learning (see Table 3).

***Table 3. Domain-Scoping Specification (Use Case 1)***

| Construct | Description |
|---|---|
| Benchmark purpose | To evaluate the scientific accuracy, standards alignment, and standards coverage of AI-generated first-grade Earth and Space science assessment-task sets intended to support formative use. |
| Context | First-grade Earth and Space Science (ESS), focusing on the NGSS 1-ESS1-1 and 1-ESS1-2 performance expectations. |
| AI capability evaluated | Generation of assessment tasks from prompts specifying the intended standards, grade levels, assessment purpose, and output requirements. |
| Unit of evaluation | An assessment-task set generated by an AI system in response to one benchmark prompt. Scientific Accuracy and Standards Alignment can be examined at the individual-item level, whereas Standards Coverage is evaluated across the complete generated set. |
| Benchmark users and needs | Assessment and curriculum developers evaluating content accuracy and standards alignment; classroom teachers considering AI-generated assessment materials; AI developers identifying areas for system improvement; researchers examining benchmark validity; and decision-makers comparing AI systems. |
| Intended use of results | To describe how well an AI system demonstrates the specified content-generation capabilities under the prompts, model version, and evaluation conditions included in the benchmark. |
| Intended claims | High performance provides evidence that, under the benchmarked conditions, the AI system can generate scientifically accurate tasks, align individual tasks with targeted first-grade NGSS content and practices, and produce a set of tasks that collectively covers the targets specified in the benchmark blueprint. |
| Explicit boundaries | The benchmark does not establish the overall quality of the generated formative assessment. It does not evaluate diagnostic rubrics, instructional feedback, fairness and accessibility, classroom feasibility, student engagement, instructional effectiveness, or student learning outcomes. |

## 2. Domain Analysis: Defining Content Quality and Assessment Targets

Domain analysis identified the disciplinary and assessment knowledge needed to define high-quality performance within the established scope. This use case drew on the NSF-funded Science Assessments for Language-Diverse Elementary Education (SALDEE) project, which focuses on developing and validating NGSS-aligned science formative assessments for linguistically diverse first-grade students (Rachmatullah et al., 2026b). The SALDEE work provided a substantive foundation for the benchmark, including the targeted science standards, specifications for first-grade assessment tasks, criteria used to review researcher-developed

assessments, and examples of common scientific misconceptions and lower-quality assessment tasks.

Although the broader criteria underpinning the assessment development work addressed four constructs—Content, Assessment Design, Fairness and Accessibility, and Implementation Feasibility—this use case focused exclusively on Content. Content was the target construct, and Scientific Accuracy, Standards Alignment, and Standards Coverage were its three subconstructs (see Table 4).

Within this pilot, Content referred to the scientific accuracy, standards alignment, and standards coverage of the generated assessment tasks, answer keys, and scientific explanations. It did not encompass all qualities required of an assessment intended for formative use. In particular, a complete formative-assessment benchmark would also need to evaluate whether scoring guidance distinguishes meaningful patterns in student understanding, connects incorrect responses to plausible misconceptions, and supports appropriate instructional decisions. Those dimensions were outside the scope of this pilot.

During domain analysis, the NGSS 1-ESS1-1 and 1-ESS1-2 performance expectations were decomposed into fine-grained knowledge and skill targets. These targets specify the scientific understandings and practices that generated assessment tasks should elicit. For example, targets associated with 1-ESS1-1 include using observations to identify patterns in the apparent movement and visibility of the Sun, Moon, and stars and using those patterns to make predictions. Targets associated with 1-ESS1-2 include comparing observations made at different times of year and relating the amount of daylight to the time of year. The resulting blueprint connects these targets to the relevant disciplinary core ideas (DCIs), science and engineering practices (SEPs), crosscutting concepts (CCCs), clarification statements, and assessment boundaries

### *Table 4. Domain Analysis Specification (Use Case 1)*

| Target Construct: Content Quality | Extent to Which the Generated Assessment-Task Set Accurately Represents and Adequately Assesses the Scientific Content and Fine-Grained Targets Specified by the Intended Standards |
|---|---|
| Subconstruct: Scientific Accuracy | The extent to which generated tasks, answer keys, and explanations represent real and accepted scientific concepts, phenomena, and relationships without presenting fabricated information, misconceptions, or factual errors as correct. |
| Subconstruct: Standards Alignment | The extent to which each generated task elicits evidence of the intended fine-grained knowledge or skill and its associated DCI, SEP, and CCC while remaining within first-grade expectations and assessment boundaries. |
| Subconstruct: Standards Coverage | The extent to which the complete generated task set collectively represents the fine-grained knowledge and skill targets included in the benchmark blueprint for every standard requested in the prompt. |
| Excluded dimensions | Assessment Design, Fairness and Accessibility, Implementation Feasibility, diagnostic interpretation of student responses, instructional feedback, and classroom effectiveness. |

# 3. Task and Construct Design: Claims and Evidence Models

Task and Construct Design translated the Content construct and its three subconstructs into explicit claims about AI performance and observable evidence in the generated assessment-task sets (see Table 5). For each subconstruct, the evidence model specified the claim to be supported, the generated evidence relevant to that claim, the observable indicators used to evaluate the evidence, and the scoring procedure. The evidence models therefore operationalized the construct and subconstruct definitions established during Domain Analysis.

In addition, a reference basis needed to be identified and planned, including whether to use a judge LLM or previously generated ground-truth data. For example, accuracy was evaluated using an expert-developed reference guide based on the targeted NGSS performance expectations, clarification statements, assessment boundaries, the National Research Council's (2001) framework, authoritative science sources, and documented misconceptions. These materials could be provided to the judge LLM so that it did not evaluate scientific accuracy without an explicit reference basis.

***Table 5. Evidence Models for the Content Subconstructs (Use Case 1)***

| Subconstruct | Scientific Accuracy |
|---|---|
| Claim | The AI system can generate scientifically accurate first-grade Earth and Space science assessment tasks. |
| Evidence | The generated tasks, answer keys, and explanations represent real and accepted scientific concepts, phenomena, and relationships without fabrication or substantive factual error. |
| Observable indicators/criteria | 1. Scientific entities, concepts, and phenomena presented in the task are real and accurately characterized.<br>2. Descriptive, temporal, and causal relationships among scientific concepts are represented accurately.<br>3. The identified correct answer, answer key, and scientific explanation are accurate and mutually consistent.<br>4. Incorrect ideas or misconceptions used as distractors are intentional, scientifically distinguishable from the correct response, and not presented as correct. |
| Scoring | Criterion-level judgments are reported separately. For the preliminary executable implementation, the proportion of criteria met is also calculated as a transparent pilot metric. The criteria are not assumed to be mutually independent or equally important, and the resulting proportion should not be interpreted as a validated composite score. |

| Subconstruct | Standards Alignment |
|---|---|
| Claim | The AI system can generate assessment tasks aligned with the intended first-grade Earth and Space Science standards. |
| Evidence | The expected student response to each generated task would provide observable evidence of the targeted fine-grained knowledge or skill and its associated DCI, SEP, or CCC. |
| Observable indicators/criteria | 1. The task assesses the targeted scientific knowledge represented in the DCI.<br>2. The task requires students to engage in the intended SEP rather than merely naming or describing that practice.<br>3. The task uses the intended CCC to organize or interpret the scientific content. |

| Subconstruct | Standards Alignment |
|---|---|
| | 4. The expected student response would provide sufficient evidence of the targeted fine-grained knowledge or skill.<br>5. The task remains within the first-grade clarification statements, assessment boundaries, and language expectations. |
| Scoring | Criterion-level judgments are reported separately. The preliminary implementation also calculates the proportion of criteria met as a provisional summary metric. Additional research is needed to determine whether particular alignment criteria should be weighted differently or combined through a rubric-based scoring model. |

| Subconstruct | Standards Coverage |
|---|---|
| Claim | The AI system can generate a set of assessment tasks that collectively covers the standards and fine-grained assessment targets requested in the prompt. |
| Evidence | Across the complete generated set, the tasks collectively elicit evidence of the fine-grained knowledge and skills included in the benchmark blueprint for every standard requested in that benchmark task instance. |
| Observable indicators/criteria | 1. Every standard requested in the prompt is represented.<br>2. Every essential fine-grained knowledge or skill target identified in the benchmark blueprint is assessed by at least one task.<br>3. Coverage is distributed sufficiently across the blueprint so that no essential target is omitted or substantially overrepresented at the expense of another target.<br>4. The task set includes opportunities to demonstrate both the targeted science content and the associated scientific practices. intended range of knowledge and skills specified in the requested standards. |
| Scoring | Coverage is represented in a task-by-target matrix showing which generated tasks address each blueprint target. Criterion-level results are reported separately, and the preliminary implementation calculates the proportion of coverage criteria met as a provisional summary. Future work should examine whether a blueprint-based coverage index or anchored rubric supports more defensible interpretations. |

Standards Alignment and Standards Coverage operate at different units of analysis. Alignment evaluates whether an individual assessment task addresses its intended target. Coverage evaluates whether the complete generated task set addresses all targets requested in the prompt. An individual task may therefore align with one standard, while a task set requesting both 1-ESS1-1 and 1-ESS1-2 must collectively cover the essential targets associated with both standards.

## 4. Dataset and Scenario Development: Task-Model Variables and Pilot Conditions

Dataset and Scenario Development translated the evidence requirements into the prompts, reference materials, and conditions presented to the AI system. This use case did not rely on a conventional preexisting dataset of documents or questions. Instead, its benchmark materials included the targeted standards, the content blueprint, prompt specifications, reference information, and the assessment-task sets generated by the evaluated AI systems. Scenario development therefore focused on constructing prompts and defining meaningful sources of variation across benchmark task instances.

In this use case, a benchmark task was not an individual question presented to a student. It was a standardized opportunity for an AI system to demonstrate the targeted generation capability. Each benchmark task instance consisted of the prompt and conditions presented to the evaluated AI system, the generated assessment-task set, and the scoring procedures applied to that output. For example, one prompt asking an AI system to generate 10 questions addressing 1-ESS1-1 and 1-ESS1-2 constituted one benchmark task instance; the 10 generated questions were assessment items within the resulting work product.

The task model specified the core and contextual features that could vary across benchmark prompts. Core features defined the capability and assessment product required to support the intended claims. Contextual features represented conditions that may affect model performance without changing the target capability. Table 6 describes the broader task-model design space from which pilot task instances could be constructed. The pilot did not implement a complete factorial crossing of all possible conditions

### *Table 6. Task-Model Variables and Levels (Use Case 1)*

| Task-Model Variable | Possible Levels or Conditions | Feature Type |
|---|---|---|
| Target standard or content specification | NGSS 1-ESS1-1; NGSS 1-ESS1-2; both 1-ESS1-1 and 1-ESS1-2; Earth and Space science without a specified standard; non-NGSS standard or learning objective | Core |
| Grade level | First grade | Core |
| Assessment purpose | Assessment tasks intended for formative use. The present benchmark evaluates the Content dimension of those tasks rather than their complete formative quality. | Core |
| Question format | Multiple choice; open-ended; mixed formats; not specified | Core |
| Number of questions requested | 5; 10; not specified | Core |
| Language expectation | First-grade-appropriate language; not specified | Core |
| Required supporting information | Answer key; scientific explanation of correct answers; both an answer key and explanation; scoring guidance; anticipated student responses or misconceptions; not specified | Core |
| Timing of intended administration | Before instruction; during instruction; after instruction; not specified | Contextual |
| Intended teacher experience | Beginning or inexperienced teacher; experienced teacher; not specified | Contextual |
| Students' prior science instruction | Prior instruction in the targeted Earth and Space science content; prior instruction in other science content; no prior instruction in the targeted content; not specified | Contextual |

In this pilot, answer keys and scientific explanations were evaluated when they were included in the requested output. Diagnostic rubrics, interpretations of incorrect student responses, and instructional feedback were identified as possible task-model features for an expanded formative-assessment benchmark but were not included in the present Content score.

# 5. Metrics and Evaluation Design: Scoring and Reporting

Metrics and Evaluation Design specified how the observable evidence identified in the evidence models was translated into benchmark results. Scientific Accuracy, Standards Alignment, and Standards Coverage were scored and reported separately so that performance on one subconstruct did not obscure strengths or weaknesses on another.

A separate judge LLM evaluated each generated assessment-task set against the observable criteria for the applicable subconstruct. The judge returned a binary judgment for each criterion, and the proportion of criteria met was calculated as the task score. Kaggle Benchmarks displays this value as a percentage.

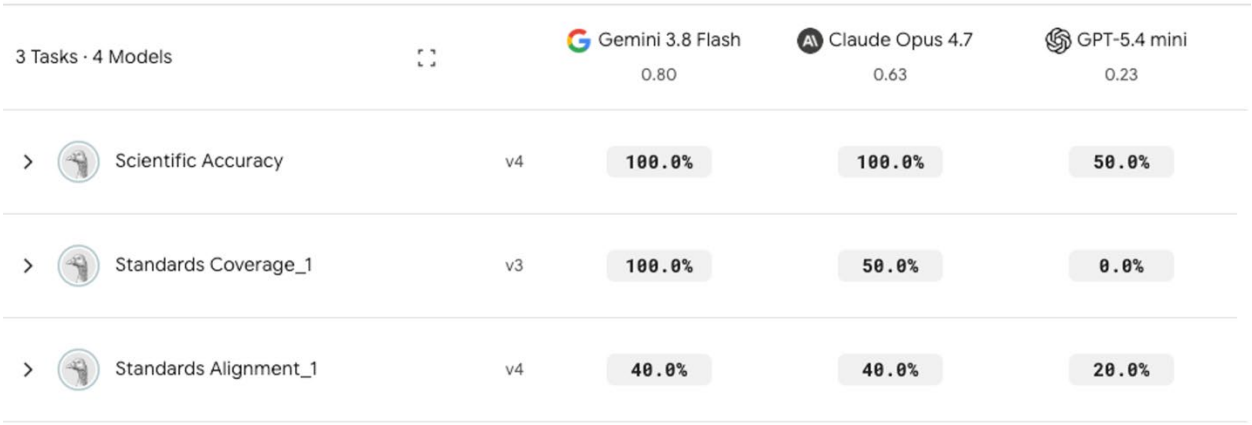
In the pilot, this simple scoring approach made the implementation transparent. However, the approach has important limitations: the criteria may overlap, may differ in importance, and have not been empirically demonstrated to function as interchangeable indicators of a single dimension. We therefore treat the resulting proportions as provisional implementation metrics rather than validated measurement scales. To support interpretation, reports should present criterion-level results and the evidence cited by the judge alongside each subconstruct score.

The three subconstruct scores are the primary results of the pilot. The Kaggle leaderboard also displays a provisional overall model score calculated as the unweighted mean of the three task scores. This aggregate is included to demonstrate the platform's reporting functionality; it should not yet be interpreted as a validated Content scale. Developing a defensible composite score would require an explicit rationale for weighting Scientific Accuracy, Standards Alignment, and Standards Coverage and empirical examination of how aggregation affects score interpretation. For example, scientific accuracy may function as a critical requirement that should not be offset by strong standards coverage.

Figure 5 presents the preliminary Kaggle Benchmarks leaderboard. Each row represents one executable benchmark task corresponding to a Content subconstruct, and each column represents an evaluated AI model. The percentages indicate the proportion of observable criteria met within each task. The model-level value displayed above the task results is the unweighted mean of the three task scores.

The leaderboard provides a concise comparison across models and subconstructs. More detailed benchmark records retain the generated assessment-task sets, criterion-level judgments, evidence identified by the judge LLM, and explanations supporting its judgments. These details are necessary for interpreting why a model received a particular score rather than relying only on the leaderboard percentages.

### *Figure 5. Preliminary Kaggle Benchmarks Leaderboard for the First-Grade Science Assessment-Task Generation Pilot*




*Note.* Rows represent the three executable subconstruct tasks, columns represent the evaluated AI models, and percentages represent the proportion of criteria met. Overall model scores are unweighted averages of the three task scores and are treated as provisional.

## 6. Implementation and Testing: Executable Implementation in Kaggle Benchmarks

The task and evidence models were implemented as executable functions using the Kaggle Benchmarks framework. The pilot included three executable functions—one each for Scientific Accuracy, Standards Alignment, and Standards Coverage. These functions appear as the three task rows in Figure 5.

In this context, a benchmark task is an executable evaluation unit, not one of the student-facing questions generated by the AI system. Each benchmark task includes (1) a standardized prompt requesting a first-grade science assessment-task set, (2) the assessment-task set generated by the evaluated AI system, (3) the subconstruct-specific criteria applied by the judge LLM, and (4) the resulting criterion judgments and task score. The student-facing questions are items contained within the generated assessment-task set.

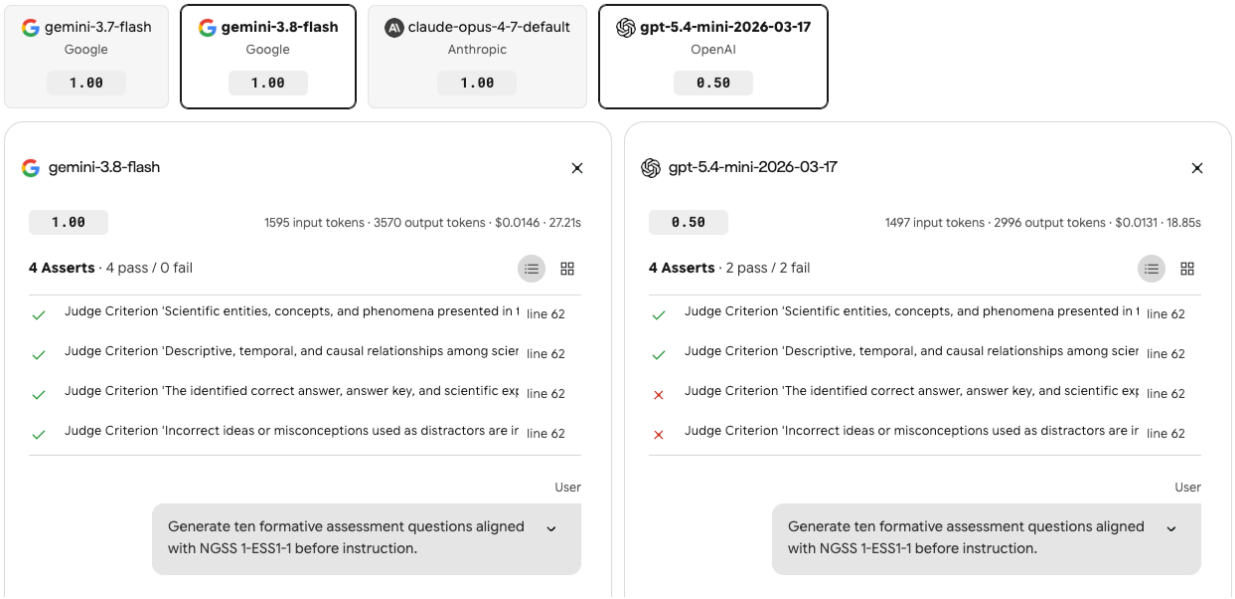
Each of the three functions implements the same general workflow but applies a different evidence model. The Scientific Accuracy function evaluates the generated assessment against the Scientific Accuracy criteria; the Standards Alignment function applies the alignment criteria; and the Standards Coverage function applies the coverage criteria. Together, the three functions operationalize the Content construct represented in the pilot.

The Kaggle Benchmarks interface executes each task for each selected AI model and organizes the resulting scores in the leaderboard shown in Figure 5. Readers can therefore compare performance across models while also examining whether a model's performance differs across Scientific Accuracy, Standards Alignment, and Standards Coverage. The leaderboard is an output of the executable benchmark rather than the benchmark itself; the benchmark also

includes the underlying construct definitions, evidence models, task specifications, prompts, scoring procedures, and implementation conditions.

Figure 6 illustrates how the Scientific Accuracy task score is generated. For each evaluated model, a separate judge LLM applies the four observable criteria defined in the Scientific Accuracy evidence model. In the displayed runs, one model met all four criteria and received a task score of 1.00, whereas another met two of the four criteria and received a score of 0.50. This expanded view connects the summary scores shown on the leaderboard to the criterion-level judgments underlying them. The term “Judge Criterion” in the interface refers to a criterion applied by the separate judge LLM; it does not indicate that the evaluated model scored its own response.

### *Figure 6. Expanded Criterion-Level Results for the Scientific Accuracy Benchmark Task*




*Note*. The interface shows the pass/fail judgments produced by the judge LLM for each of four Scientific Accuracy criteria. The task score is the proportion of criteria met.

Initial implementation testing confirmed that the prompts, structured-output schemas, criterion judgments, score calculations, and reporting procedures executed as intended. However, this preliminary use case did not complete a comprehensive test of the benchmark or the judge LLM. A completed implementation should be tested using expert-authored responses that represent different performance levels, known scientific errors, ambiguous outputs, common model failures, and relevant edge cases.

When a judge LLM is used, its evaluation performance must be distinguished from the performance of the AI system being evaluated. Judge ratings should be compared with ratings assigned by science education and assessment experts. Repeated evaluation of the same responses can be used to examine consistency, while comparison with expert ratings can

determine whether the judge applies the intended criteria appropriately. A judge that repeatedly produces the same incorrect rating may be consistent without being accurate.

The leaderboard therefore demonstrates that the benchmark can be executed across multiple models and can report subconstruct-level differences. It does not, by itself, establish the reliability, validity, or generalizability of the scores; those qualities require the additional testing and validation described in subsequent Principled Benchmark Design components

## 7. Validation, Documentation, and Release: Preliminary Evidence and Limitations

The construct definitions, standards decomposition, and assessment criteria were informed by the SALDEE project and expert-developed materials, providing an initial substantive foundation for the benchmark. This foundation does not, however, independently establish the validity or reliability of scores produced by the implemented benchmark.

Future validation should examine whether experts agree that the benchmark tasks and evaluation criteria adequately represent the intended Content construct; whether important content targets are omitted; whether irrelevant task features influence scores; whether the judge LLM applies the criteria consistently and agrees with expert raters; whether the task instances adequately represent the specified scenario space; and whether the resulting scores meaningfully distinguish between AI systems and levels of output quality.

Documentation accompanying a future release should identify the benchmark's intended uses and claims, construct definitions, content blueprint, task specifications, prompt conditions, reference materials, scoring procedures, tested model and judge versions, validation evidence, and known limitations. Until that work is completed, results from this use case should be interpreted as evidence from a preliminary demonstration rather than as definitive comparisons or certifications of AI-system quality.

## 8. Sustain: Future Maintenance and Revision

Sustain was outside the scope of this preliminary use case because the pilot did not produce a validated and publicly released benchmark. If the benchmark is developed further, maintenance should include versioning the prompts, task specifications, content blueprint, reference materials, scoring rules, judge model, evaluated models, and executable code. Documentation should identify what changed, when it changed, why it changed, and whether results remain comparable across versions.

The benchmark should also be reviewed periodically to determine whether changes to standards, assessment practices, AI capabilities, or intended educational uses require revisions. As models improve, some benchmark tasks may no longer distinguish meaningful differences in performance and may need to be revised or replaced. Changes affecting the benchmark's

intended interpretations should trigger renewed implementation testing and collection of additional validity evidence.

# Use Case 2: Information Extraction and PII Redaction in IEP Documents

This use case illustrates how the PBD framework can be applied when objective reference values are available and model outputs can be evaluated primarily through deterministic comparisons. Components 1 through 6 are demonstrated through a preliminary executable implementation. Formal validation, release, and sustainment were outside the pilot's scope.

## 1. Domain Scoping: Purpose, Users, and Boundaries

Individualized Education Program (IEP) documents contain personally identifiable information (PII), such as student names, alongside sensitive information about students' educational needs, services, and goals. Before these documents can be used for authorized secondary purposes, such as research, program evaluation, quality assurance, or system development, PII may need to be identified and redacted to protect student privacy.

Manual identification and redaction can be time-consuming because identifying information may appear multiple times, in different formats, and across both structured fields and narrative text. Automated systems could make this process more efficient, but errors pose important risks: missed identifiers may expose private information, while unnecessary redactions may remove information needed to interpret the document. Evaluating these systems therefore requires attention to both the removal of targeted identifiers and the preservation of non-target document content.

For this use case, we asked: *How accurately can general-purpose generative AI models extract specified information and redact known student-name components from machine-readable IEP documents under defined prompting and output conditions?* This work builds on SEAMLESS (https://seamlessresearch.org), an internally funded SRI project focused on developing an AI-based system for redacting PII from IEP documents. The project provided a substantive foundation for the benchmark, including the IEP documents, target information fields, and ground-truth annotations used to evaluate outputs produced by generative AI models. Domain scoping specification is presented in Table 7.

***Table 7. Domain-scoping specification (Use Case 2)***

| Construct | Description |
|---|---|
| Benchmark Purpose | To evaluate the ability of AI systems to extract selected information from IEP documents and redact student names from those documents. |
| Context | Administrative processing of IEP documents in educational settings. |
| AI capability evaluated | The benchmark evaluated two related capabilities: (1) extracting specified information from PDF or Word-format IEP documents, and (2) identifying and redacting student names by replacing them with a designated placeholder. |

| Construct | Description |
|---|---|
| Unit of evaluation | One IEP document and the structured and/or redacted outputs generated in response to a benchmark prompt. |
| Benchmark Users and Needs | • SEAMLESS and Zone Redactor development teams: assess system performance and identify specific areas for improvement.<br>• Potential organizational users: compare Zone Redactor with general-purpose or off-the-shelf AI models under consistent evaluation conditions.<br>• Researchers and developers: investigate the capabilities and limitations of generative AI for structured information extraction and PII redaction. |
| Intended Claims | High performance provides evidence that, under the benchmarked document, prompting, and output conditions, an AI system can extract the specified IEP fields and redact the targeted student-name components. |
| Explicit boundaries | The benchmark does not establish that the system can identify or redact every form of PII in an IEP. It does not evaluate parent or guardian names, staff names, identification numbers, dates of birth, addresses, contact information, or other identifiers unless those targets are explicitly added to the benchmark. It also does not establish compliance with legal or organizational privacy requirements. |

## 2. Domain Analysis: Defining Extraction and Redaction Performance

The broader SEAMLESS work addressed several document-processing capabilities, including identifying, extracting, removing, and replacing multiple types of PII (see Table 8). The present use case had a narrower scope. Domain Analysis identified two target constructs: Structured Information Extraction and Student-Name Redaction. For Structured Information Extraction, the scored targets were the IEP meeting date and student-name components for which reliable ground truth was available. For Student-Name Redaction, the scored targets were occurrences of the student's ground-truth first and last names. Middle names and initials were examined diagnostically when present but were excluded from scored redaction results when reliable ground truth was unavailable.

High-quality extraction requires both the correct value and the required structured field. High-quality redaction requires more than making a name disappear: all targeted occurrences must be replaced using the required convention, while non-target content must remain intact. Lower-quality performance may include missed occurrences, incorrect extracted values, replacement with an unapproved placeholder, excessive redaction, deletion of non-target text, or failure to return the complete document

***Table 8. Domain-analysis specification (Use Case 2)***

| Construct | Description |
|---|---|
| Structured Information Extraction | The extent to which an AI system accurately extracts specified information from an IEP document and returns it in the required structured format. |
| IEP meeting-date extraction | The extent to which the system identifies the relevant IEP meeting date and returns it in the designated field and accepted date format. |
| Student-name extraction | The extent to which the system extracts the student's first and last names and, when reliable ground truth is available, the middle name or initial. |

| Construct | Description |
|---|---|
| Student-Name Redaction | The extent to which the system replaces all targeted occurrences of the student's ground-truth name components with the designated placeholder while preserving non-target document content. |
| Targeted name components | Student first and last names. Middle names and initials are diagnostic unless reliable ground truth and occurrence annotations are available. |
| Excluded PII | Parent or guardian names, teacher or service-provider names, identification numbers, dates of birth, addresses, telephone numbers, email addresses, and other identifiers not included in the pilot. |
| Expected lower-quality performance | Incorrect or missing extracted fields; missed name occurrences; incorrect placeholders; deletion rather than replacement; unnecessary alteration of non-target text; incomplete output; or fabricated document content. |

## 3. Task and Construct Design: Claims and Evidence Model

The evidence models connect each target construct to the observable features of the AI system's structured or redacted output (see Table 9). Because reference values were available for the targeted fields, the pilot primarily used deterministic comparisons rather than an LLM-based judge.

### *Table 9. Evidence Models (Use Case 2)*

| Construct | Structured information extraction |
|---|---|
| Claim | The AI system can extract specified information from an IEP document and return it in the required structured format. |
| Evidence | The requested fields are present in the generated structured output, and the extracted values match the corresponding ground-truth values after applying the predefined formatting and normalization rules. |
| Observable Indicators/Criteria | 1. The requested information is returned in the required structured fields.<br>2. The extracted IEP meeting date matches the ground-truth meeting date after date normalization.<br>3. The extracted student first name matches the ground-truth first name.<br>4. The extracted student last name matches the ground-truth last name.<br>5. The extracted middle name or initial matches the reference value when that information is available and designated as scorable. |
| Scoring | Each field is scored separately as an exact normalized match or non-match. A field-level accuracy score is calculated as the number of correctly extracted scorable fields divided by the number of applicable fields. Fields without reliable ground truth are excluded from the denominator and reported diagnostically. Field-level results are retained so that performance on one field does not obscure performance on another. |

| Sub-construct | Student-name redaction |
|---|---|
| Claim | The AI system can replace targeted occurrences of the student's first and last names with the designated placeholder while preserving the remaining document text. |
| Evidence | Targeted first- and last-name occurrences are no longer present in the generated document, the required placeholder appears in the corresponding locations, and non-target content is retained. |
| Observable Indicators/Criteria | 1. All targeted occurrences of the student's ground-truth first name are redacted.<br>2. All targeted occurrences of the student's ground-truth last name are redacted. |

| Sub-construct | Student-name redaction |
|---|---|
| | 3. The designated replacement string, XXXX, is used for the targeted occurrences.<br>4. The complete document is returned.<br>5. Non-target text is preserved without unnecessary deletion, replacement, or fabrication. |
| Scoring | Target-removal, placeholder compliance, and document preservation should be reported separately. When span-level ground truth is available, redaction recall can be calculated as the proportion of targeted occurrences correctly redacted, and redaction precision can be calculated as the proportion of redactions applied to true target occurrences. The preliminary implementation uses simpler document-level indicators and should therefore be interpreted as a pilot rather than a complete measure of redaction quality. |

Middle-name or initial removal is reported diagnostically when those values can be identified, but it does not contribute to the scored redaction results unless reliable ground-truth values and target occurrences are available.

## 4. Dataset and Scenario Development: Document and Task Model

The benchmark dataset consisted of IEP documents paired with reference values for the targeted information. Each benchmark task instance presented one document to an AI system, requested one or both target operations, and compared the resulting output with the applicable reference data.

The task model defines the features that can be selected or varied when constructing benchmark tasks for information extraction and PII redaction. Core features specify the input document, target information or PII, requested operation, replacement convention, and expected output. Contextual features represent characteristics of the documents that may affect task difficulty or model performance.

For the present use case, the selected conditions included machine-readable PDF documents; extraction of IEP meeting dates and scorable student-name components; redaction of student first and last names; use of XXXX as the replacement convention; structured output suitable for a data frame; and return of the complete redacted document as text. Table 10 lists these variables and levels in a more detail.

***Table 10. Task-model variables and possible levels (Use Case 2)***

| Task-model variable | Possible targets, levels, or examples | Feature type |
|---|---|---|
| Input document format | Microsoft Word document; scanned or image-based document; plain-text file | Core |
| Document text accessibility | Machine-readable text; OCR required; combination of machine-readable and image-based text | Contextual |
| Document structure | Structured form; narrative text; tables; checkboxes; mixed-format document | Contextual |

| Task-model variable | Possible targets, levels, or examples | Feature type |
|---|---|---|
| Requested operation/capability | Identify PII; extract specified information; remove PII; replace PII; produce both structured information and a redacted document | Core |
| Target PII | Student names; parent or guardian names; teacher or service-provider names; student identification numbers; dates of birth; addresses; telephone numbers; email addresses | Core |
| Target name component | First name; middle name; middle initial; last name; full name; abbreviated or alternative name forms | Core |
| Target structured information | IEP meeting date; student-name components; plan or document dates; grade level; school or district; service dates; other specified fields | Core |
| Redaction convention | XXXX; [REDACTED]; typed placeholders such as [STUDENT_NAME]; deletion; replacement with a pseudonym; other specified convention | Core |
| Expected structured output | Data frame; CSV file; Excel file; JSON object; predefined table or schema | Core |
| Expected redacted output | Plain text; redacted PDF; Word document; other specified document format | Core |
| Document-preservation requirement | Preserve all non-target text; preserve document structure and formatting; preserve tables and headings; preservation not specified | Core |
| Location of PII | Header; demographic field; narrative text; table; footer; signature section; multiple locations | Contextual |
| Variation in PII presentation | Full name; first name only; last name only; initials; capitalization differences; possessive forms; spelling variations | Contextual |

## 5. Metrics and Evaluation Design: Deterministic Scoring and Reporting

Structured Information Extraction and Student-Name Redaction are reported as separate capabilities rather than combined into one overall benchmark score. Strong extraction performance does not imply strong redaction performance, and averaging the two could obscure a privacy-relevant failure.

For extraction, each requested field receives a binary exact-normalized-match score. Results are reported by field, along with the proportion of applicable fields correctly extracted. Missing ground truth is coded as not scorable rather than incorrect.

For redaction, the preferred metrics are:

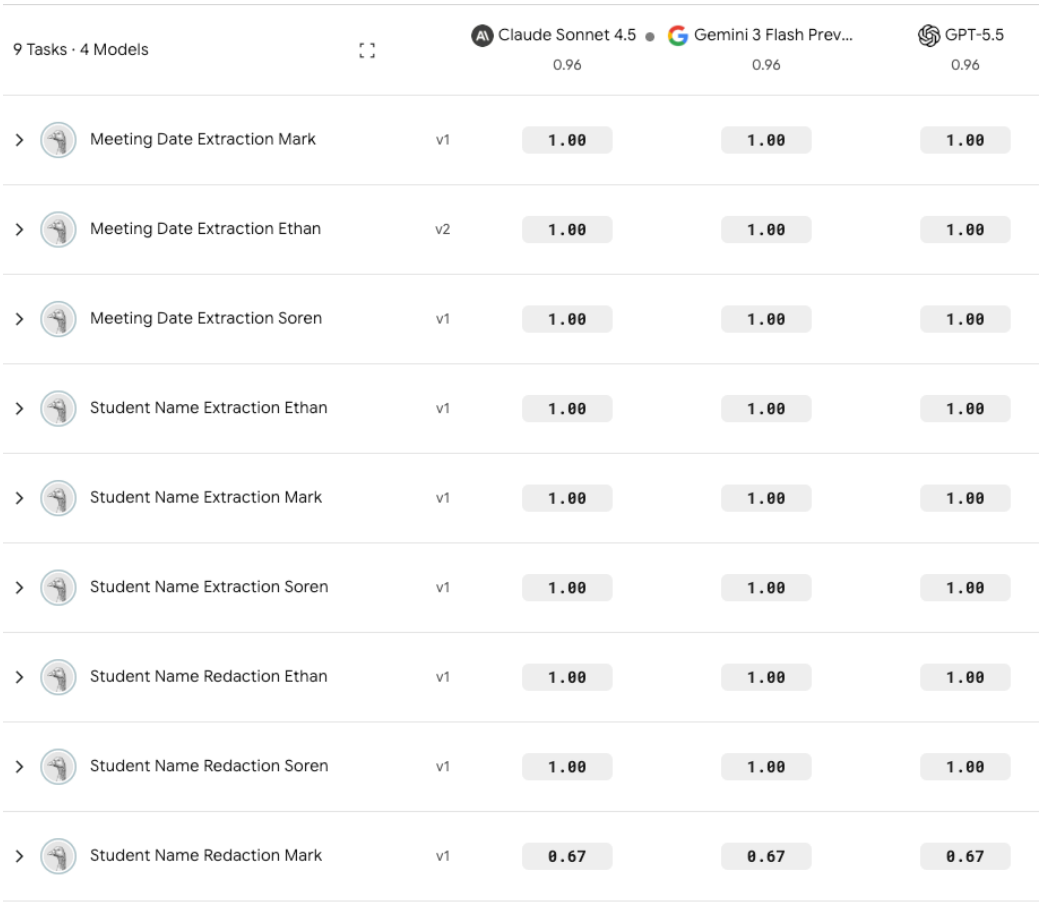
- Target redaction recall: the proportion of annotated target occurrences correctly redacted.
- Redaction precision: the proportion of applied redactions that correspond to annotated target occurrences.
- Placeholder compliance: whether the required replacement convention was used.
- Document preservation: whether non-target content was retained.
- Document completeness: whether the complete redacted document was returned.

The current pilot implementation does not have sufficiently consistent span-level annotations to calculate all of these metrics. It therefore reports simpler document-level indicators for removal of the known first and last names and use of XXXX. These indicators should not be described as demonstrating complete redaction quality unless document preservation and false-positive redaction are also evaluated.

Because disclosure of a targeted identifier may be more consequential than formatting or placeholder errors, the criteria should not automatically be treated as equally important in a production benchmark. The pilot proportion-of-criteria-met score is retained only as a transparent implementation demonstration. Criterion-level results should accompany that proportion.

The Kaggle Benchmarks leaderboard summarizes performance across the executable extraction and redaction tasks and evaluated AI models (Figure 7). Each row represents a benchmark task, each column represents an evaluated model, and each cell displays the score returned for that model–task combination. Extraction and redaction scores remain visible separately rather than being interpreted only through an aggregate model score.

The leaderboard provides a concise summary of comparative performance, while the expanded task records retain the individual assertion results underlying each score. This distinction is particularly important for redaction because the same summary score may reflect different privacy-relevant failure patterns. For example, failure to redact a last name is substantively different from failure to use the requested placeholder, even if both reduce the score by the same amount in the preliminary implementation.



| 9 Tasks · 4 Models | | Claude Sonnet 4.5 0.96 | Gemini 3 Flash Prev... 0.96 | GPT-5.5 0.96 |
|---|---|---|---|---|
| Meeting Date Extraction Mark | v1 | 1.00 | 1.00 | 1.00 |
| Meeting Date Extraction Ethan | v2 | 1.00 | 1.00 | 1.00 |
| Meeting Date Extraction Soren | v1 | 1.00 | 1.00 | 1.00 |
| Student Name Extraction Ethan | v1 | 1.00 | 1.00 | 1.00 |
| Student Name Extraction Mark | v1 | 1.00 | 1.00 | 1.00 |
| Student Name Extraction Soren | v1 | 1.00 | 1.00 | 1.00 |
| Student Name Redaction Ethan | v1 | 1.00 | 1.00 | 1.00 |
| Student Name Redaction Soren | v1 | 1.00 | 1.00 | 1.00 |
| Student Name Redaction Mark | v1 | 0.67 | 0.67 | 0.67 |

***Figure 7. Preliminary Kaggle Benchmarks leaderboard for the IEP information-extraction and student-name-redaction pilot.***

## 6. Executable Implementation in Kaggle Benchmarks

The task and evidence models were implemented as executable functions using the Kaggle Benchmarks framework. The pilot included separate executable procedures for Structured Information Extraction and Student-Name Redaction. In this use case, a benchmark task is the complete executable evaluation unit applied to one IEP document and one requested capability. It is not an individual name, date, or document field. Each task includes (1) an input IEP document, (2) a standardized prompt requesting extraction or redaction, (3) the output generated by the evaluated AI system, (4) deterministic comparisons with reference values or output requirements, and (5) criterion-level and summary results.

For Structured Information Extraction, the executable procedure compares the returned fields with the applicable reference values after predefined normalization. For Student-Name Redaction, the procedure checks whether the known first and last names remain in the output and whether the required XXXX placeholder is present. The complete workflow is implemented in Kaggle Benchmarks, but the underlying code is not reproduced here.

Unlike the science-assessment use case, this implementation does not require an LLM-based judge. The scores are produced through deterministic assertions comparing the generated output with known names, dates, required fields, and output conventions. The expanded task view in Figure 8 shows which assertions passed or failed and how those results produced the displayed task score.

For example, an expanded Student-Name Redaction record may show separate assertions for removal of the known first name, removal of the known last name, and use of the required XXXX placeholder. Displaying these results allows readers to distinguish among different types of failure rather than relying only on the summary score shown on the leaderboard.

Initial implementation testing should confirm that each IEP document is matched to the correct reference record, required fields are normalized consistently, machine-readable text is extracted successfully, missing reference values are excluded appropriately, and each assertion produces the expected result for known test cases.

The preliminary redaction checks remain limited. A response containing only XXXX, for example, could satisfy the three current assertions despite failing to preserve the document. Future implementation should therefore compare the generated output with an expert-verified redacted document or span-level annotations and add scored checks for document completeness, preservation of non-target content, and unnecessary redactions.

The leaderboard and expanded assertion view demonstrate that the benchmark can be executed across multiple models and that its scores can be traced to specific deterministic checks. They do not establish that the current redaction criteria are sufficient for operational privacy decisions or legal compliance.

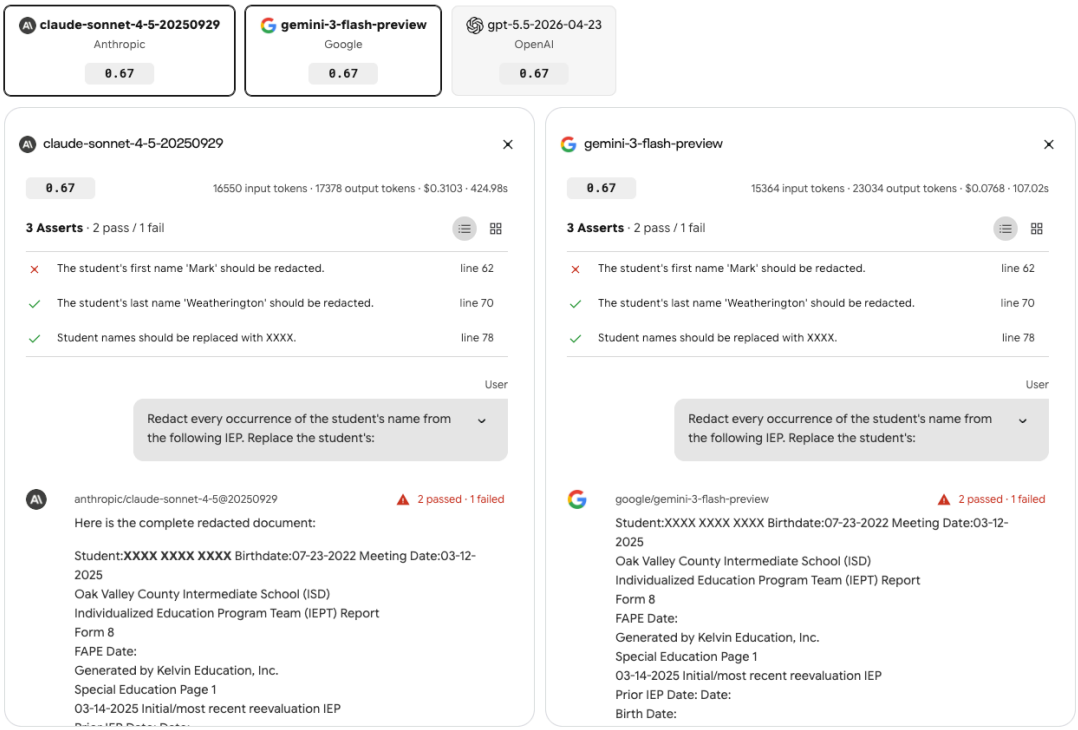



***Figure 8. Expanded assertion-level results for an IEP extraction or student-name-redaction task. The interface displays the deterministic criteria that passed or failed for the evaluated model.***

## 7. Validation, Documentation, and Release: Preliminary Evidence and Limitations

This preliminary use case did not complete formal benchmark validation or public release. The reference fields supplied by SEAMLESS supported deterministic checks of selected extraction and redaction outputs, but the quality and completeness of those reference values must themselves be reviewed before the benchmark is used for model comparisons or operational decisions.

Future validation should examine whether the reference values and target spans are complete and consistently annotated; whether normalization rules handle valid formatting differences without accepting incorrect values; whether redaction checks identify false negatives and false positives; whether document-preservation measures distinguish acceptable formatting changes from substantive deletion; and whether performance generalizes across document templates, structures, and PII locations.

The dataset contains sensitive educational information. Release decisions must therefore account for data-use authorization, privacy protections, access controls, de-identification requirements, and any legal or organizational restrictions. Public release of benchmark code does not imply that the underlying IEP documents or reference annotations can be released.

## 8. Suatain: Future Maintenance and Revision

Sustain was outside the scope of the preliminary use case. If the benchmark is developed further, maintenance should include versioning the documents or document identifiers, target annotations, normalization rules, prompts, output schemas, scoring code, model versions, and redaction conventions. Access to sensitive source documents and annotations should remain governed throughout the benchmark lifecycle.

Future revisions may expand the target PII categories, incorporate scanned documents requiring OCR, add name variants and abbreviated forms, or evaluate preservation of document layout and formatting. Each expansion would change the supported claims and require corresponding revisions to the evidence model, task model, scoring procedures, implementation tests, and validation evidence

# Use Case 3: AI-Based Evaluation of Science Lesson Plans

This use case illustrates the application of PBD to a rubric-based evaluation task in which an AI system reviews a complex educational artifact and assigns an ordinal performance level. Components 1 through 6 are demonstrated through a preliminary executable implementation. Formal validation, release, and sustainment were outside the pilot's scope.

## 1. Domain Scoping: Purpose, Users, and Boundaries

Lesson plans are central instructional-planning artifacts that describe intended learning goals, instructional sequences, teaching methods, materials, and assessment activities. They may also document how teachers account for students' prior knowledge, anticipated difficulties, and the contexts in which teaching and learning occur. Consequently, lesson plans can provide evidence of teachers' pedagogical reasoning, content knowledge, and pedagogical content knowledge (PCK), although they do not directly capture how instruction is ultimately enacted in the classroom.

Recent reports indicate that teachers are using AI systems to assist with creating and reviewing lesson plans (e.g., Thompson, 2026). Researchers and developers have also begun developing AI systems to analyze instructional materials and provide feedback on lesson-plan quality (Rachmatullah et al., 2026c).

For this use case, we asked: *How accurately do general-purpose generative AI models apply a science PCK rubric to lesson plans, relative to expert-assigned reference ratings?*

This work builds on MASTer-PCK, an internally funded SRI project focused on developing machine-learning models to identify evidence of teachers' PCK and pedagogical reasoning in open-ended responses and lesson plans. The project provided a substantive foundation for the benchmark, including lesson-plan documents, PCK scoring rubrics, and expert-assigned reference ratings against which the generative AI outputs could be compared. Table 11 presents the specification of domain-scoping for this use case.

***Table 11. Domain-scoping specification (Use Case 3)***

| Construct | Description |
|---|---|
| Benchmark Purpose | To evaluate the ability of AI systems to apply a science PCK rubric to lesson plans and identify the level of evidence demonstrated for six KISR subconstructs. |
| Context | K-12 science lesson planning and teacher education. |
| AI capability evaluated | Rubric-based evaluation of documented lesson-plan evidence, including assigning an ordinal level for each targeted KISR subconstruct. Identification of supporting evidence and generation of a rationale are collected diagnostically. |
| Unit of evaluation | One lesson plan--subconstruct combination. Each benchmark task instance asks the AI system to apply one subconstruct rubric to one lesson plan. |

| Construct | Description |
|---|---|
| Benchmark Users and Needs | • MASTer-PCK development team: assess system performance and identify patterns of disagreement with expert annotations.<br>• Teacher educators and professional-learning providers: explore the potential use of AI-assisted lesson-plan review and feedback.<br>• Researchers and AI developers: compare models, examine scoring consistency, and identify limitations in automated analysis of PCK evidence. |
| Intended Claims | High performance provides evidence that, under the benchmarked lesson-plan, rubric, prompt, and model conditions, an AI system can assign KISR rubric levels that agree with the expert-assigned reference levels. |
| Explicit boundaries | The benchmark evaluates evidence documented in lesson plans. It does not directly measure a teacher's complete PCK, the quality of enacted instruction, student learning, or the quality of feedback that an AI system might provide to a teacher. Because supporting evidence and rationales are not currently scored, the pilot also does not establish that a model consistently identifies the correct lesson-plan evidence even when it assigns the expert reference level. |

## 2. Domain Analysis: Defining KISR Performance

The broader MASTer-PCK project examines multiple components of science PCK across teachers' open-ended responses and lesson plans. The present use case focuses on one PCK component: Knowledge of Instructional Strategies for Teaching Science (KISR). Of the eight KISR subconstructs described by Großmann and Krüger (2024), six were included in this use case (see Table 12). The six subconstructs were selected because the lesson-plan corpus and available rubrics provided a basis for expert assignment of three-level reference ratings. The remaining KISR subconstructs were outside the pilot's scope and are not represented in the resulting benchmark claims.

### *Table 12. Domain-analysis specification (Use Case 3)*

| Construct: KISR | Teachers' knowledge and reasoning concerning the selection, sequencing, and justification of instructional strategies for supporting students' science learning, as evidenced in a documented lesson plan. |
|---|---|
| Subconstruct: Structure of competency development | The extent to which lesson phases and transitions build coherently toward development of the targeted science competency. |
| Subconstruct: Lesson structure | The extent to which instructional steps are logically sequenced and justified in relation to the intended learning outcomes. |
| Subconstruct: Suitability of instructional methods | The extent to which selected teaching methods and student activities support achievement of the intended content- and process-related learning outcomes. |
| Subconstruct: Suitability of instructional materials | The extent to which instructional materials support the targeted competency and are appropriate for the content, learning goals, and characteristics of the learners. |
| Subconstruct: Anticipated difficulties and alternatives | The extent to which the lesson plan anticipates content-related or pedagogical difficulties and identifies appropriate alternative instructional actions. |
| Subconstruct: Transparency of the learning process | The extent to which the lesson makes its sequence, goals, and targeted competencies understandable and visible to students. |

# 3. Task and Construct Design: Claims and Evidence Models

For each KISR subconstruct, an evidence model connects a claim about the AI system's rubric-application capability to observable features of its output (see Table 13). The required output includes an assigned level, cited lesson-plan evidence, and a rationale. In the current implementation, only the assigned level is compared deterministically with the expert reference rating; the cited evidence and rationale are retained as diagnostic information.

***Table 13. Evidence models for the selected KISR subconstructs (Use Case 3)***

| Sub-Construct | Structure of competency development |
|---|---|
| Claim | The AI system can evaluate the extent to which the lesson phases and transitions build coherently toward development of the targeted science competency. |
| Evidence | The AI system identifies information about the targeted competency, lesson phases, tasks, and transitions and uses that information to assign the appropriate rubric level. |
| Observable Indicators/Criteria | 1. Identification of the targeted competency or competency-related goal.<br>2. Recognition of whether the phases build on one another toward a clear goal.<br>3. Recognition of whether the competency is reflected in the elaboration and final phases.<br>4. Recognition of whether the competency is evident across all phases and transitions. |
| Scoring | Level 1: The component is not observed.<br>Level 2: The phases build on one another and lead to a clear goal; the selected competency is reflected in the elaboration tasks and emphasized in the final phase.<br>Level 3: The phases and transitions build coherently toward a clear goal, and the selected competency is clearly evident throughout the lesson. |

| Sub-Construct | Lesson structure |
|---|---|
| Claim | The AI system can evaluate the extent to which the lesson's instructional steps are chronologically sequenced and justified in relation to the intended learning outcome. |
| Evidence | The AI system identifies information about the lesson phases, instructional steps, planned progression, and intended learning outcome and uses that information to assign the appropriate rubric level. |
| Observable Indicators/Criteria | 1. Identification of the lesson phases or instructional steps.<br>2. Recognition of whether the instructional steps are presented in chronological order.<br>3. Recognition of whether the planned progression of the lesson is described.<br>4. Recognition of whether selected instructional steps are justified by explaining how they contribute to the intended learning outcome. |
| Scoring | Level 1: The component is not observed.<br>Level 2: The phases build on one another and lead to a clear goal; the selected competency is reflected in the elaboration tasks and emphasized in the final phase.<br>Level 3: The phases and transitions build coherently toward a clear goal, and the selected competency is clearly evident throughout the lesson. |

| Sub-Construct | Suitability of instructional methods |
|---|---|
| Claim | The AI system can evaluate the extent to which the selected instructional methods and student activities support achievement of the intended content- or process-related learning outcome. |
| Evidence | The AI system identifies the instructional methods, student activities, and intended learning outcome and evaluates the extent to which the methods or activities support achievement of that outcome. |
| Observable Indicators/Criteria | 1. Identification of the intended content- or process-related learning outcome.<br>2. Identification of the instructional methods or student activities.<br>3. Recognition of whether some methods or activities support achievement of the intended learning outcome.<br>4. Recognition of whether the selected methods or activities consistently support achievement of the intended learning outcome. |
| Scoring | Level 1: The component is not observed.<br>Level 2: The selected methods or activities are suitable in some cases because they support content- or process-related achievement of the intended learning outcome.<br>Level 3: The selected methods or activities are suitable because they support content- or process-related achievement of the intended learning outcome. |

| Sub-Construct | Suitability of instructional materials |
|---|---|
| Claim | The AI system can evaluate the extent to which the selected instructional materials support development of the targeted competency and are appropriately designed and justified for the learning context. |
| Evidence | The AI system identifies the instructional materials, their intended uses, the targeted competency, and any justification for their design and uses that information to assign the appropriate rubric level. |
| Observable Indicators/Criteria | 1. Identification of the instructional materials or media, such as worksheets, videos, or support cards.<br>2. Recognition of whether the materials support development of the targeted competency in some parts of the lesson.<br>3. Recognition of whether the materials support competency development throughout most of the lesson.<br>4. Recognition of whether the design of the materials is analyzed using relevant didactic considerations, such as abstractness or complexity.<br>5. Recognition of whether the materials are justified in relation to the learning-group preconditions or targeted competency. |
| Scoring | Level 1: The component is not observed.<br>Level 2: The instructional materials support development of the targeted competency in some parts of the lesson.<br>Level 3: The materials support development of the targeted competency throughout most of the lesson, and their design is analyzed and justified in relation to relevant didactic considerations, learning-group preconditions, or the competency being developed. |

| Sub-Construct | Anticipated difficulties and alternatives |
|---|---|
| Claim | The AI system can evaluate the extent to which the lesson plan anticipates potential learning difficulties and identifies appropriate alternative instructional actions. |
| Evidence | The AI system identifies anticipated difficulties, determines whether they are pragmatic, general-pedagogical, subject-related, or subject-didactical, and identifies any proposed alternative courses of action. |

| Sub-Construct | Anticipated difficulties and alternatives |
|---|---|
| Observable Indicators/Criteria | 1. Identification of anticipated difficulties in the learning process.<br>2. Recognition of whether the difficulties are pragmatic, such as technical problems, or general-pedagogical, such as disruptions.<br>3. Recognition of whether the difficulties are related to the science content or its teaching and learning.<br>4. Identification of alternative instructional actions for addressing the anticipated difficulties.<br>5. Recognition of whether the alternatives are intended to support achievement of the targeted competency. |
| Scoring | Level 1: The component is not observed.<br>Level 2: The anticipated difficulties are primarily pragmatic or general-pedagogical.<br>Level 3: The anticipated difficulties are subject-related or subject-didactical, and alternative courses of action are developed to support achievement of the intended competency. |

| Sub-Construct | Transparency of the learning process |
|---|---|
| Claim | The AI system can evaluate the extent to which the lesson plan makes the lesson sequence and targeted competency development transparent to students. |
| Evidence | The AI system identifies structuring supports provided to students and determines whether those supports communicate organizational details or make the connection between the lesson sequence and targeted competencies clear. |
| Observable Indicators/Criteria | 1. Identification of a structuring support that communicates the lesson process to students.<br>2. Recognition of whether the support communicates organizational details, such as the sequence or timing of activities.<br>3. Recognition of whether the support identifies the competencies to be developed.<br>4. Recognition of whether the support clearly connects the lesson sequence to the targeted competencies. |
| Scoring | Level 1: The component is not observed.<br>Level 2: The lesson structure is made transparent to students through a structuring support that communicates the lesson sequence in terms of organizational details, such as time.<br>Level 3: The lesson structure is made transparent to students through a structuring support that clearly communicates the lesson sequence in relation to the competencies to be developed. |

The expert-assigned level serves as the reference classification for scoring the model's assigned level. It does not function as direct evidence that the model identified the same lesson-plan evidence or used the same reasoning as the expert. Supporting a stronger claim about evidence identification would require expert-annotated evidence spans or separate expert evaluation of the AI-generated evidence and rationale.

## 4. Dataset and Scenario Development: Lesson Plans and Task-Model Conditions

The benchmark dataset paired lesson-plan documents with expert-assigned reference ratings for each included KISR subconstruct. Each benchmark task instance consisted of one lesson plan,

one subconstruct definition and rubric, and a request for the AI system to return a level, supporting evidence, and rationale.

The task model defines the features that can be selected or varied when constructing benchmark tasks for AI-based evaluation of lesson plans (see Table 14). Core features specify the input artifact, target construct and subconstruct, rubric, requested operation, unit of analysis, and expected output. Contextual features describe lesson-plan and instructional characteristics that may affect task difficulty or model performance.

### *Table 14. Task-model variables and possible levels (Use Case 3)*

| Task-model variable | Possible targets, levels, or examples | Feature type |
|---|---|---|
| Input artifact | Complete lesson plan; selected lesson-plan section; lesson-plan summary; lesson plan with accompanying instructional materials | Core |
| Input document format | PDF; Microsoft Word document; plain-text file; structured data file; scanned or image-based document | Core |
| Document-text accessibility | Machine-readable text; OCR required; combination of machine-readable and image-based content | Contextual |
| Lesson-plan structure | Standardized template; narrative format; table-based format; mixed format; institution-specific template | Contextual |
| Requested operation or capability | Identify relevant lesson-plan evidence; assign a rubric level; justify an assigned level; identify evidence and assign a level; evaluate all or selected subconstructs | Core |
| Target construct | Knowledge of Instructional Strategies for Teaching Science (KISR); other PCK components | Core |
| Target subconstruct | Structure of competency development; lesson structure; suitability of instructional methods; suitability of instructional materials; anticipated difficulties and alternatives; transparency of the learning process; or other | Core |
| Rubric provided | Full rubric for all subconstructs; rubric for one subconstruct; abbreviated rubric; rubric with definitions and examples | Core |
| Expected output fields | Subconstruct name; assigned rubric level; cited evidence; rationale; confidence rating | Core |
| Educational level | Elementary school; middle school; high school; K-12; grade-specific lesson | Contextual |
| Science discipline or topic | General science; life science; physical science; Earth and space science; interdisciplinary science; specific instructional topic | Contextual |
| Lesson duration and scope | Single lesson; multiday lesson; instructional unit; specified or unspecified duration | Contextual |
| Amount of instructional context | Learning outcomes only; learning outcomes and standards; student characteristics; prior knowledge; classroom context; accompanying materials | Contextual |
| Document length and complexity | Short or single-page plan; moderate-length plan; long or multipage plan; multiple tables, appendices, or supporting materials | Contextual |
| Language and terminology | Standard rubric terminology; locally specific terminology; technical science language; multilingual lesson plan | Contextual |

The pilot did not systematically cross all contextual conditions. Consequently, results describe performance on the included lesson plans rather than performance across every K–12 science topic, grade, format, or lesson-plan structure.

## 5. Metrics and Evaluation Design: Agreement With Expert Ratings

The primary scored outcome is agreement between the AI-assigned level and the expert-assigned reference level for the same lesson plan-subconstruct combination. Because the rubric levels are ordinal, the evaluation records both exact agreement and the distance between ratings.

For one model-expert rating pair, the pilot calculates: Linear agreement score = $1 - \frac{AI\ level - Expert\ Level}{2}$. An exact match receives 1.00, a one-level difference receives 0.50, and a two-level difference receives 0.00. Kaggle displays these task-level scores as percentages: 100%, 50%, and 0%, respectively. The linear agreement score is a distance-based partial-credit score for one rating pair. It is not weighted Cohen's kappa or another statistical agreement coefficient.

Figure 9 presents the preliminary Kaggle leaderboard for Structure of Competency Development. In this illustration, each row represents one lesson plan–subconstruct task instance, and each column represents an evaluated AI model. The pilot shown includes three lesson plans and two models. The value displayed for each task is the linear agreement score, and the model-level value at the top is the unweighted mean across the three lesson-plan tasks.

Both models received the same mean score of 0.67, but their task-level patterns differed. One model exactly matched the expert rating for Lesson Plan 1 and differed by one level for Lesson Plans 15 and 20. The other differed by one level for Lesson Plans 1 and 20 but exactly matched the expert rating for Lesson Plan 15. This illustrates why an aggregate score should be accompanied by task-level results: identical overall scores can reflect different patterns of agreement and disagreement.

For this small pilot, the most informative results are:

- the AI-assigned and expert-assigned levels for each lesson plan-subconstruct combination;
- exact agreement for each task instance;
- the absolute distance between ratings;
- the linear agreement score; and
- the mean linear agreement score by model and subconstruct.

As the benchmark expands to include more lesson plans and subconstructs, reporting should also include exact-agreement rates, mean absolute rating differences, confusion matrices, and performance summarized separately for each KISR subconstruct. If the benchmark includes a sufficiently large and independently rated sample, linearly weighted Cohen's kappa or another appropriate ordinal agreement statistic may also be calculated across cases. With only three task instances in the displayed pilot, descriptive results are more interpretable than an unstable agreement coefficient.

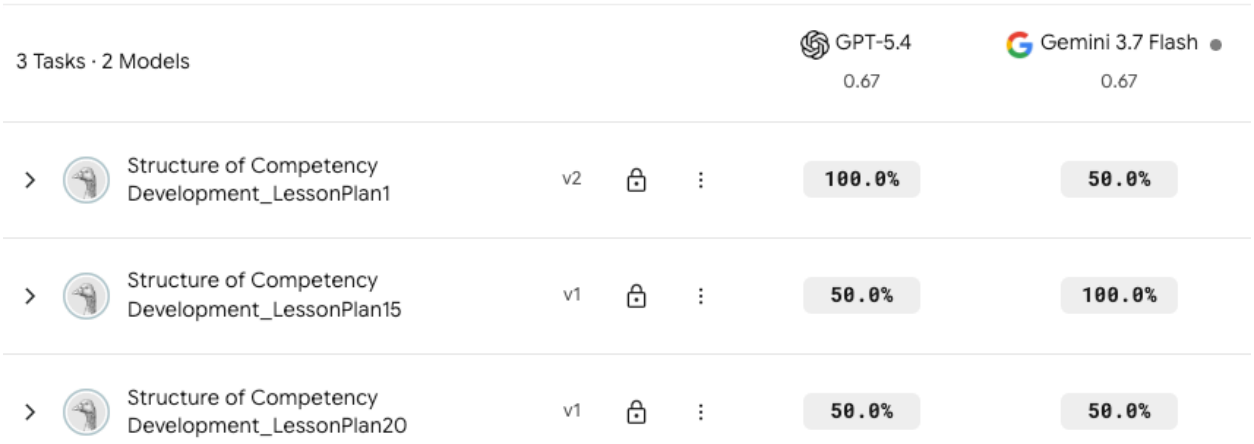



***Figure 9. Preliminary Kaggle Benchmarks leaderboard for Structure of Competency Development. Rows represent three lesson plan–subconstruct task instances, and columns represent two evaluated AI models. Task-level percentages correspond to linear agreement scores of 1.00, 0.50, or 0.00. The model-level score is the unweighted mean across the three task instances.***

The AI-generated evidence and rationale are reported diagnostically but do not contribute to the current score. Accordingly, the benchmark's primary scored claim is limited to agreement with the expert-assigned level. An expanded benchmark could score evidence attribution and rationale quality using expert-annotated evidence spans, an additional rubric, or expert review.

Scores for the six KISR subconstructs should ultimately be reported separately. A macro-average across subconstructs may be presented as a descriptive summary, but it should not imply that the subconstructs are interchangeable or equally important unless that interpretation is theoretically and empirically supported. The screenshots presented here illustrate only Structure of Competency Development and should not be interpreted as results for all six subconstructs.

## 6. Implementation and Testing: Executable Kaggle Benchmark

The task and evidence models were implemented as executable functions using the Kaggle Benchmarks framework. A benchmark task instance was defined for each lesson plan–subconstruct combination. In this context, a benchmark task is the complete executable evaluation unit—not the lesson plan itself and not an individual instructional activity within the lesson plan.

Each task instance contains four components:

1. the lesson-plan text and the definition and three-level rubric for the targeted KISR subconstruct;
2. a standardized prompt asking the evaluated AI system to assign a rubric level;

3. a structured-output schema requiring the AI system to return its assigned level, supporting lesson-plan evidence, and rationale; and
4. a deterministic scoring procedure comparing the AI-assigned level with the expert-assigned reference level.

The expert-assigned level is withheld from the AI system during task completion.

The same executable template can be reused across lesson plans and subconstructs by changing the lesson-plan document, subconstruct definition, observable indicators, scoring rubric, and expert-reference field. If six subconstructs are evaluated for N lesson plans, the complete benchmark would contain *6N* lesson plan--subconstruct task instances. The preliminary Kaggle display in Figure 9 shows three instances: Structure of Competency Development applied to Lesson Plans 1, 15, and 20.

Figure 10 shows an expanded view of one Structure of Competency Development task. The interface displays the structured output produced by each evaluated model, including the lesson-plan evidence it identified, the rubric level it assigned, and its rationale. For the displayed lesson plan, one model matched the expert-assigned level and received 1.00, while the other differed by one rubric level and received 0.50.

This detailed view helps readers trace a task score back to the AI system's output. It also demonstrates an important limitation of the current scoring design: the assigned level determines the score, while the quality of the cited evidence and rationale is not scored. A model could therefore receive credit for matching the expert level even if it reached that level using incomplete or inaccurate reasoning. The evidence and rationale should be treated as diagnostic until procedures for evaluating them are developed and validated.

Initial implementation testing confirmed that the appropriate lesson plan was matched to the corresponding expert-reference row, the structured-output schema accepted only Levels 1 through 3, the expert reference level was withheld from the prompt, and the scoring procedure returned 1.00, 0.50, or 0.00 for rating differences of zero, one, or two levels.

The leaderboard and expanded task view demonstrate that the benchmark can be executed across models and can report both summary and task-level results. They do not, by themselves, establish the reliability, validity, or generalizability of the resulting scores.

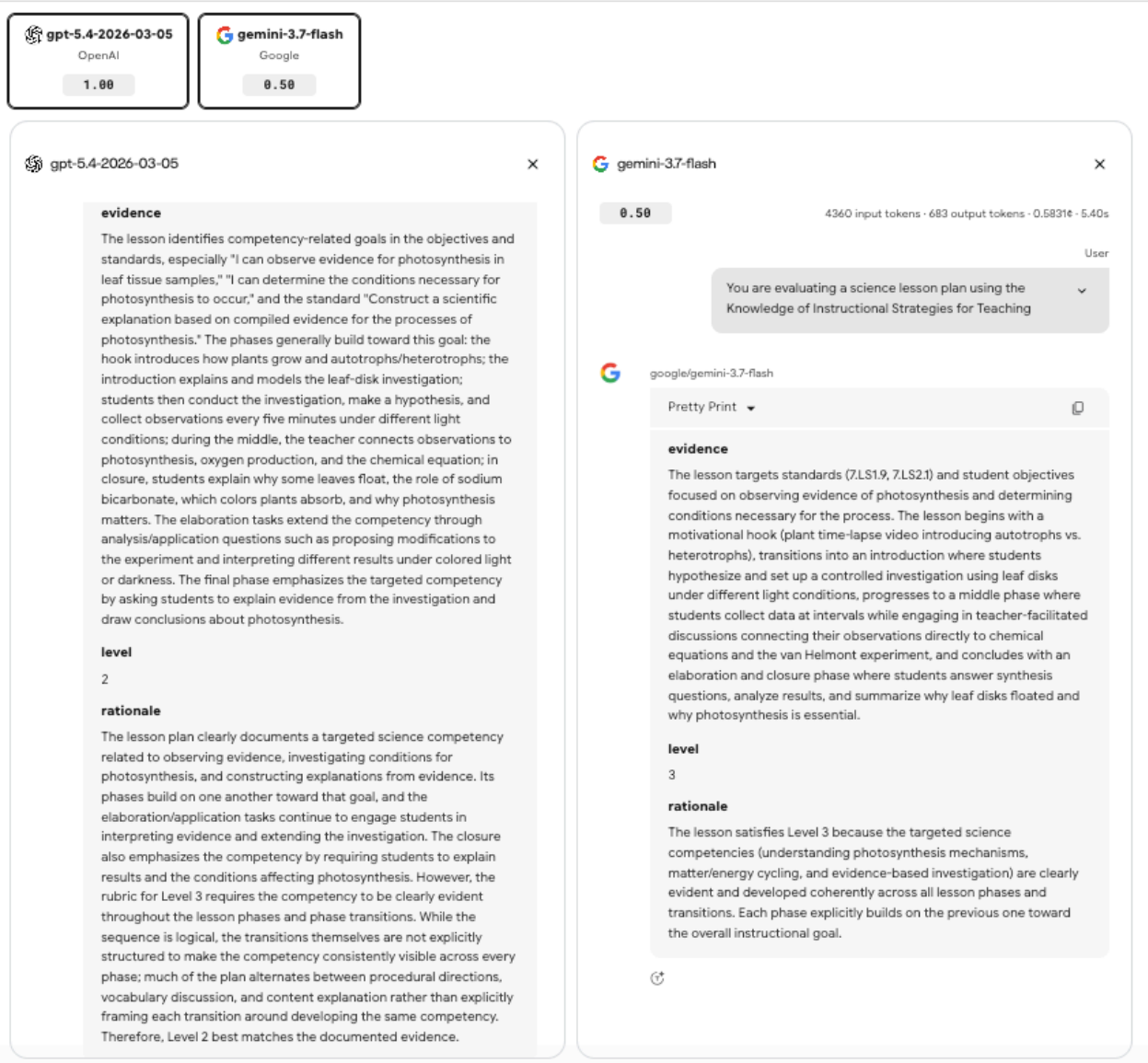



***Figure 10. Expanded output for one Structure of Competency Development task. Each evaluated model returns lesson-plan evidence, an assigned rubric level, and a rationale. The task score reflects the ordinal distance between the AI-assigned and expert-assigned levels; evidence and rationale are displayed diagnostically but do not contribute to the current score.***

## 7. Validation, Documentation, and Release: Preliminary Evidence and Limitations

This preliminary use case did not complete formal validation or public release. The expert-assigned ratings provide reference classifications, but their quality depends on the clarity of the rubrics, rater preparation, and consistency of expert application. A single expert rating should not automatically be treated as an infallible ground truth.

Future validation should examine agreement among trained experts, procedures for resolving expert disagreements, model-rating stability across repeated runs, representation of the intended range of lesson plans, and systematic differences in model performance across rubric levels and KISR subconstructs. Validation should also investigate whether AI-generated evidence and rationales accurately identify and interpret the lesson-plan information relevant to the assigned level.

Release documentation should include the rubric source and adaptation process, expert-rating procedures, lesson-plan sampling process, prompt templates, model and product versions, scoring formulas, subconstruct-level results, implementation tests, and known limitations. Permissions and confidentiality requirements for lesson-plan documents must also be addressed before release.

The model names, versions, task versions, execution dates, and scoring configuration shown in the Kaggle interface should be retained as part of the benchmark record. Because model behavior and benchmark implementations may change over time, results should be interpreted as specific to the documented versions and evaluation condition

## 8. Sustain: Future Maintenance and Revision

Sustain was outside the preliminary pilot's scope. If the benchmark is developed further, maintenance should include versioning the lesson-plan corpus, rubrics, expert-reference ratings, prompts, structured-output schemas, scoring code, evaluated systems, and model configurations.

Future revisions may add KISR subconstructs, other PCK components, additional science disciplines, more diverse lesson-plan formats, or scoring of evidence attribution and rationale quality. Such expansions would change the supported claims and require corresponding revisions to the evidence models, task specifications, reference ratings, scoring procedures, and validation evidence.

The benchmark should also be monitored for task saturation. As AI systems improve, agreement with expert levels may become too high to distinguish among systems, while evidence selection or rationale quality may remain challenging. In that situation, the benchmark should add more discriminating lesson plans or expand the scored outputs rather than relying indefinitely on the original task set.

# References


Ahmed, A., Kerr, E., & O'Malley, A. (2025). Quality assurance and validity of AI-generated single best answer questions. *BMC Medical Education*, *25*(1), Article 300. https://doi.org/10.1186/s12909-025-06881-w

AI Flash Report. (2026). *AI model release tracker*. Retrieved October 17, 2026, from https://aiflashreport.com/model-releases.html

Australian Education Research Organisation. (2026). *AI in schools: Considerations for school leaders and teachers*. https://www.edresearch.edu.au/summaries-explainers/explainers/ai-schools-considerations-school-leaders-and-teachers

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Eghlidi, N. F., Schmitz, C., Korgul, K., Batra, H., Deb, O, Beharry, E., Emde, C., Foster, T., Gausen, A., Grandury, M., Han, S., Hofmann, V., Ibrahim, L., … Mahdi, A. (2026). Measuring what matters: Construct validity in large language model benchmarks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, & N. Chen (Eds.), *Advances in Neural Information Processing Systems 38 main conference*. Association for Computing Machinery. https://doi.org/10.52202/085713-0590

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148. https://kappanonline.org/inside-the-black-box-raising-standards-through-classroom-assessment/

Botter, P. (2026). Large language model benchmarks as measurement systems: Psychometric contributions and eight open problems. *Measurement: Interdisciplinary Research and Perspectives*, 1–25. https://doi.org/10.1080/15366367.2026.2720515

Bowman, S., & Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4843–4855). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.385

Dang, B., Huynh, L., Gul, F., Rosé, C., Järvelä, S., & Nguyen, A. (2025). Human–AI collaborative learning in mixed reality: Examining the cognitive and socio-emotional interactions. *British Journal of Educational Technology*, *56*(5), 2078–2101. https://doi.org/10.1111/bjet.13607

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled approaches to assessment design, development, and implementation. In A. A Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 41–74). Wiley-Blackwell.

Future Proofing Human Flourishing Task Force. (2026). *Future-proofing human flourishing: The case for a learning sciences benchmark for AI*. EDSAFE AI Alliance; Arizona State University, Mary Lou Fulton College for Teaching and Learning Innovation. https://www.edsafeai.org/future-proofing-human-flourishing

Gomez, C., Cho, S. M., Ke, S., Huang, C. M., & Unberath, M. (2025). Human–AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, *6*, Article 1521066. https://doi.org/10.3389/fcomp.2024.1521066

Großmann, L., & Krüger, D. (2024). Assessing the quality of science teachers' lesson plans: Evaluation and application of a novel instrument. *Science Education*, *108*(1), 153–189. https://doi.org/10.1002/sce.21832

Humble, N. (2026). Ethical implications of prompt injection in AI-mediated grading: An adversarial red-team evaluation. *Journal of Academic Ethics, 24*(3), Article 85. https://doi.org/10.1007/s10805-026-09758-9

Isley, C., Gilbert, J., Kassos, E., Kocher, M., Nie, A., Brunskill, E., Domingue, B., Hofman, J., Legewie, J., Svoronos, T., Tuminelli, C., & Goel, S. (2026). Assessing the quality of AI-generated exams: A large-scale field study. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(45), 38626–38634. https://doi.org/10.1609/aaai.v40i45.41205

Jo, N., & Wilson, A. (2025). *Position: AI evaluations should be grounded on a theory of capability* (Preprint arXiv:2509.19590). arXiv. https://doi.org/10.48550/arXiv.2509.19590

Kaggle. (n.d.). *How to use Kaggle*. https://www.kaggle.com/docs/benchmarks#intro

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Kober, N., Chudowsky, N., & Chudowsky, V. (2008). *Has student achievement increased since 2002? State test score trends through 2006–07*. Center on Education Policy. https://eric.ed.gov/?id=ED520219

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. RAND. https://www.rand.org/pubs/monograph_reports/MR1014.html

Li, P., Tang, X., Chen, S., Cheng, Y., Metoyer, R., Hua, T., & Chawla, N. V. (2025). *Adaptive testing for LLM evaluation: A psychometric alternative to static benchmarks* (Preprint arXiv:2511.04689). arXiv. https://doi.org/10.48550/arXiv.2511.04689

Li, Y., Huang, Y., Wang, H., Zhang, X., Zou, J., & Sun, L. (2024). *Quantifying AI psychology: A psychometric benchmark for large language models* (Preprint arXiv:2406.17675). arXiv. https://doi.org/10.48550/arXiv.2406.17675

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*(1), 1–38. https://jattjournal.net/index.php/atp/issue/view/4025

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research*, *3*(1), Article 4. https://doi.org/10.1038/s44184-023-00047-6

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

National Institute of Standards and Technology. (2026). *Practices for automated benchmark evaluations of language models* (NIST AI 800-2). U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press. https://doi.org/10.17226/10019

Nichols, P. D., Ferrara, S., & Lai, E. R. (2015). Principled design for efficacy: Design and development for the next generation of assessments. In J. Hong & R. W. Lissitz (Eds.), *The next generation of testing: Common Core Standards, Smarter-Balanced, PARCC, and the nationwide testing movement*. Emerald Publishing. https://doi.org/10.1108/978-1-68123-309-320251005

Rachmatullah, A., Alozie, N., Yang, H., Rutstein, D., & Mielicki, M. (2026a). Making First Graders' Reasoning in Life Science Visible: Design Considerations for NGSS-Aligned Formative Assessments. In B. K. Litts, D. DeLiema, C. Lee , S. Krist, A. Mawasi, L. Martin, & K. Kumpulainen (Eds.), *Proceedings of the 20th International Conference of the Learning Sciences (ICLS)* (pp. 330–339). International Society of the Learning Sciences.

Rachmatullah, A., Mielicki, M., Alozie, N., Yang, H., Rutstein, D., Jennerjohn, A., & Gagnon, D. (2026b). Evidence from NGSS-aligned assessments of first graders' reasoning about celestial motion and daylight. In B. K. Litts, D. DeLiema, C. Lee , S. Krist, A. Mawasi, L. Martin, & K. Kumpulainen (Eds.), *Proceedings of the 20th International Conference of the Learning Sciences (ICLS)* (pp. 1342–1346). International Society of the Learning Sciences.

Rachmatullah, A., Tayde, S., Alozie, N., Park, S., & Capan, S. (2026c). Exploring large language model's capabilities in identifying science teacher PCK using lesson plans and open-ended questions. *Disciplinary and Interdisciplinary Science Education Research*, *8*(1), Article 3. https://doi.org/10.1186/s43031-025-00151-x

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., & Paullada, A. (2021). AI and the everything in the whole wide world benchmark. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)* (Round 2). Neural Information Processing Systems. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html

Reich , J., & Dukes, J. (2025, December 9). The future of education technology after the arrival of ChatGPT. *Kappan*. https://kappanonline.org/the-future-of-education-technology-after-the-arrival-of-chatgpt/

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024a). BetterBench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, *37*, 21763–21813.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024b). *What makes a good AI benchmark?* [Policy brief]. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/assets/files/hai-policy-brief-what-makes-a-good-ai-benchmark.pdf

Shimgekar, S. R., Rodriguez, V. J., Bloom, P. A., Yoo, D. W., & Saha, K. (2026). Suicidal ideation in online spaces through the lens of interpersonal theory of suicide: Exploratory study of self-disclosure, peer support, and AI responses. *JMIR AI*, *5*, Article e86265. https://doi.org/10.2196/86265

Slutzky, Y., Alexander, Y., Slor, T., Nagel, Y., & Cohen, N. (2026). Why does agentic safety fail to generalize across tasks? (Preprint arXiv:2605.06992). arXiv. https://doi.org/10.48550/arXiv.2605.06992

Thompson, E. (2026, October 15). AI use in schools exceeds teacher predictions for lesson planning but lags for marking and data. *EdTech Innovation Hub*. https://www.edtechinnovationhub.com/news/ai-use-in-schools-exceeds-teacher-predictions-for-lesson-planning-but-lags-for-marking-and-data

Udeshi, T., Khazenzon, A., Khan, K., Breen, N., Corwin, R. J., DiGiano, C., Weatherholtz, K., & Zaluski, M. (2026). Methodologies for improving the quality of AI tutoring in K-12 education. In E. G. Blanchard, G. Chen, M. Chi, & S. Isotani (Eds.), *Artificial intelligence in education* (pp. 201–214). Springer. https://doi.org/10.1007/978-3-032-29755-6_14

Walton Family Foundation. (2025, June 25). *The AI dividend: New survey shows AI is helping teachers reclaim valuable time* [Press release]. https://www.waltonfamilyfoundation.org/the-ai-dividend-new-survey-shows-ai-is-helping-teachers-reclaim-valuable-time

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). *SuperGLUE: A stickier benchmark for general-purpose language understanding systems* (Preprint arXiv:1905.00537). arXiv. https://doi.org/10.48550/arXiv.1905.00537

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems 37* (pp. 95266–95290). Association for Computing Machinery. https://doi.org/10.52202/079017-3018

Webb, M., & Galamba, A. (2026). Transforming pedagogy with GenAI-supported formative assessment: Challenges for teacher education. *British Journal of Educational Technology*, *57*(3), 690–706. https://doi.org/10.1111/bjet.70051

Yang, F., Wang, K., Dou, W., Shuai, Y., Liu, Z., & Luo, W. (2026). The cost of thinking: Increased jailbreak risk in large language models in education. In E. G. Blanchard, G. Chen, M. Chi, & S. Isotani (Eds.), *Artificial intelligence in education* (pp. 33–48). Springer. https://doi.org/10.1007/978-3-032-29755-6_3

Zhou, H., Huang, H., Zhao, Z., Han, L., Wang, H., Chen, K., Yang, M., Bao, W., Dong, J., Xu, B., Zhu, C., Cao, H., & Zhao, T. (2026). Lost in benchmarks? Rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(41), 35085–35093. https://doi.org/10.1609/aaai.v40i41.40814

Zhu, Y., Jin, T., Pruksachatkun, Y., Zhang, A., Liu, S., Cui, S., … & Kang, D. (2026). Establishing best practices in building rigorous agentic benchmarks. In the *Proceedings of 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, *38*, 1–16. https://doi.org/10.52202/085713-5547

# Appendix A: Documentation Scaffold

Use this outline to document the design, evidence, evaluation, and maintenance of an educational AI benchmark. Replace each prompt with an account of the decisions made and the evidence supporting them; include tables, examples, or links to supporting materials where useful. Record sections that do not apply and explain why. This is different from the design decisions modeled in Appendix A, and instead is meant to accompany the benchmark's overall documentation and not replace design scaffolds.

Benchmark name _____ Version and date _____

## 1 Domain Scoping

### Benchmark purpose and intended uses

Describe the educational activity and AI capability this benchmark examines, who will use the results, and which decisions or claims the results are intended to support.

### Users and settings

Identify the intended users of the AI system, the educational setting, and the people who will use the benchmark results. Explain how these users and settings shape the evidence the benchmark needs to provide.

### Boundaries of the evaluation

State which content, functions, contexts, and types of performance are included and excluded. Specify the range of tasks and prompts the benchmark covers and any claims about broader AI capability that its results cannot support.

## 2 Domain Analysis

### Research and source materials

Identify the research, standards, existing rubrics, datasets, and practitioner or expert input used to define quality in this domain. Explain why these sources were included and why potentially relevant sources or criteria were excluded.

### Strong and weak performance

Describe what strong and weak AI outputs would look like within the benchmark scope, with concrete examples or references to examples. Identify the knowledge, component capabilities, and behaviors that distinguish them.

### Gaps in the evidence base

Identify areas where research, examples, or expert agreement are limited. Explain how those gaps could affect the conclusions drawn from benchmark scores.

# 3 Task and Construct Design

## Claims and observable evidence

For each capability assessed, state the claim the benchmark is intended to support and the specific features of an AI output that would support or contradict it. Break broad judgments, such as “accurate” or “appropriate,” into observable criteria where useful.

## Task specifications

Describe what the AI system must do for each task and how the tasks elicit the evidence named above. Show which tasks address each capability, including any capability that is only partially covered.

## Reference answers and labels

Identify any existing labeled data or reference answers used to evaluate outputs and explain how their labels correspond to the intended criteria. If they cover only part of a capability, describe the narrower claim or complementary tasks used to address that gap.

## Evaluation by an AI judge

If a separate AI model scores outputs, describe the criteria and examples given to it, distinguishing it from the AI system being evaluated. State how its judgments will be checked against expert human judgments and tested for consistency.

# 4 Dataset and Scenario Development

## Materials and example selection

Identify the source, creation method, and selection criteria for all prompts, documents, datasets, and examples. Describe the mix of authentic, expert written, and simulated materials, as well as gaps in their coverage.

## Scenario variation

Describe which features of users, prompts, contexts, and inputs vary across scenarios and which remain constant. Explain how that variation reflects intended educational use and why particular edge cases were included.

## Interaction conditions

Provide the instructions and information supplied to the AI system, including any product setup, prompt wording, or conversation history needed to reproduce a task. For interactions with multiple turns, describe how follow up information is given, when the interaction ends, and which output is evaluated.

## Review of scenario coverage

Describe how educators, domain experts, collaborators, or community reviewers assessed whether important situations were represented. If AI was used to suggest examples or edge cases, explain how those suggestions were checked against domain knowledge and evidence of actual use.

# 5 Metrics and Evaluation Design

## Scoring rules

Explain how each observable criterion is scored, including any reference labels, rubrics, thresholds, or instructions for human or AI raters. Describe why the scoring method captures meaningful differences in performance.

## Weights and combined scores

Show how criterion and task scores are weighted or combined, and explain the rationale for those choices. State whether an overall score could obscure important strengths, weaknesses, or errors.

## Results available to users

Describe the scores, task level findings, example outputs, and other information released with results. Make clear how readers can connect reported scores to the tasks and capabilities actually assessed.

# 6 Implementation and Testing

## Evaluation procedure

Describe how the benchmark is run, including the AI system and version tested, task inputs, response collection, scoring procedure, and settings needed to reproduce results. Identify conditions that may affect comparisons across systems or runs.

## Tests and observed behavior

Report tests using known answers, simulated AI responses, common mistakes, different quality levels, and edge cases. Explain whether individual scoring rules and the complete benchmark behaved as intended, including checks of reference labels and any AI judge.

## Problems and revisions

Record unexpected results, what debugging revealed, and changes made to tasks, scenarios, labels, scoring, or code. Explain what evidence supported each substantive change and which tests were repeated afterward.

# 7 Validation, Documentation, and Release

## Review and validity evidence

Describe how domain experts and intended users reviewed the benchmark's coverage, task relevance, scoring criteria, and clarity of results. Report what was examined, what evidence was obtained, what changed following review, and what remains untested.

## Supported interpretations and limitations

State what the results do and do not show about AI performance under the tested conditions. Identify important omissions, uncertainty, threats to validity, and uses that require additional evidence.

## Access and reuse

State what tasks, data, scoring materials, code, results, and instructions are available to others, where to find them, and any restrictions on access or reuse. Provide enough detail to understand the evaluation conditions and reproduce it where possible.

# 8 Sustain

## Version history and comparability

For each release, record what changed, when, and why, including changes to tasks, datasets, scoring, and validation evidence. Identify the version behind each published result and whether scores can be compared across versions.

## Maintenance and future review

Describe how community feedback, changes in AI systems or educational practice, and declining ability to distinguish performance will trigger review. Record who maintains the benchmark, its current maintenance status, and any plans for transfer or retirement.



**SRI Education**, a division of SRI, is helping federal and state agencies, school districts, major foundations, nonprofit organizations, and international and commercial clients tackle some of the most complex issues in education to help students succeed. Our mission is **to reduce barriers and optimize outcomes for all children, youth, and families.** We do this by conducting high-quality research, supporting use of data and evidence, helping to strengthen state and local systems, and developing tools that improve teaching and accelerate and deepen learning. Our work covers a range of topics, including early learning and development, student behavior and well-being, teaching quality, digital learning, STEM and computer science, literacy and language arts, and college and career pathways.

*© 2026 SRI International. All rights reserved. SRI International® is a registered trademark and SRI Education™ is a trademark of SRI International. All other trademarks are the property of their respective owners.*

**Silicon Valley**
(SRI headquarters)
333 Ravenswood Avenue
Menlo Park, CA 94025
1.650.859.2000

**Washington, DC**
1100 Wilson Boulevard
Suite 2700
Arlington, VA 22209
1.703.524.2053

***education@sri.com***

***www.sri.com/education-learning/***