

Learning Progression-Guided Scientific Model Assessment Using Foundation and Supervised Vision Models

Leonora Kaldaras¹ and Yasasvi Nalla²

¹University of North Carolina at Chapel Hill, ²University of Houston

Abstract

Artificial intelligence can scale assessment of student-generated scientific models, but valid educational use requires algorithms to identify features that meaningfully represent the intended learning construct. This study compares foundation and supervised computer vision approaches for evaluating learning progression (LP)-aligned evidence in approximately 1,600 high-school students' electroscopes models. Grounding DINO combined with the Segment Anything Model (SAM) was used for zero-shot detection without task-specific labeled training data, whereas a CustomCharge convolutional neural network (CNN) was trained on human-scored models. Both approaches were evaluated against expert scoring across 13 LP-aligned analytic categories. Grounding DINO+SAM achieved higher Cohen's κ in 12 of 13 categories, with strong performance across both charge- and force-related evidence. Categories involving less frequent or more complex cross-scenario representations remained comparatively challenging. Findings demonstrate complementary strengths of foundation and supervised approaches and illustrate how LPs can provide a theoretically grounded framework for developing and validating AI assessment of scientific models.

1 Introduction

Contemporary science education emphasizes students' ability to use scientific knowledge and practices to explain and predict phenomena rather than simply recall disciplinary knowledge [1]. Scientific modeling is central to this goal because students must represent relevant components of a system, relationships among those components, and causal mechanisms that account for observed phenomena. Evaluating student-generated models can therefore provide important evidence of students' knowledge-in-use.

Scientific models are challenging to evaluate at scale. Students may express scientifically meaningful ideas through diverse visual representations, making manual scoring time-intensive and automated scoring difficult. Computer vision offers a potential mechanism for identifying relevant components of students' drawings, but predictive performance alone is insufficient for establishing validity. An algorithm may reproduce human classifications while relying on visual characteristics that are unrelated to the intended educational construct [2,3].

Learning progressions (LPs) provide a mechanism for addressing this challenge. LPs describe increasingly sophisticated forms of student understanding and can guide specification of the evidence expected in student responses. In LP-guided AI assessment, the LP defines what constitutes meaningful evidence of proficiency, an LP-aligned rubric operationalizes that evidence, and AI is used to identify the corresponding features in student work. This creates an explicit connection between learning theory, assessment design, and AI evaluation.

The present study examines this approach using student-generated models of electrical interactions. We compare two substantially different computer vision approaches: Grounding DINO combined with the Segment Anything Model (SAM), which supports zero-shot detection without task-specific labeled training data, and a CustomCharge convolutional neural network (CNN), trained through supervised learning on human-scored student models. Rather than asking which architecture is universally superior, we examine how the approaches perform across different forms of LP-aligned evidence and what those differences imply for valid and scalable scientific-model assessment.

The study addresses three research questions: (1) To what extent do foundation and supervised computer vision models reproduce expert human scoring of LP-aligned components in student-generated scientific models? (2) How does performance differ across types of LP-aligned

visual evidence? (3) What are the relative strengths and limitations of foundation and supervised approaches for scalable LP-guided scientific model assessment?

2 Learning Progression-Guided AI Assessment

The study is grounded in a previously validated, NGSS-aligned LP for knowledge-in-use of electrical interactions [2]. The progression integrates disciplinary ideas related to electric forces and charge transfer with scientific modeling and causal reasoning and describes increasingly sophisticated ways students coordinate these ideas to explain electrostatic phenomena.

To operationalize the LP for assessment, an Evidence-Centered Design approach was used to specify claims about what students should understand and the observable evidence needed to support those claims. These evidence statements were translated into analytic rubric categories. The categories therefore were not selected simply because they were visually detectable; they represented evidence derived from the underlying construct and used by trained human scorers.

This distinction is important for AI-supported assessment. The AI models in this study did not directly predict LP levels. Instead, they identified 13 LP-aligned analytic categories representing construct-relevant evidence in student models. Combinations of these categories can subsequently be converted to LP levels using predetermined, deterministic scoring rules. Thus, AI performs evidence detection, while developmental interpretation remains governed by the LP. This separation makes it possible to examine where automated evaluation agrees with human judgment and where particular forms of evidence remain difficult to detect.

3 Methods

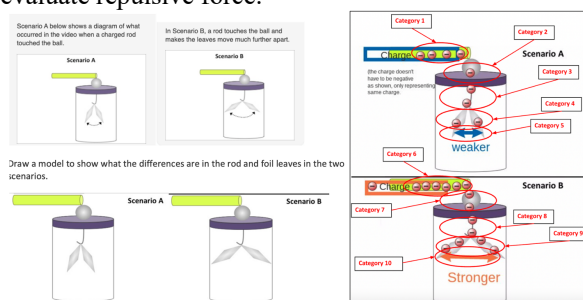
3.1 Data and Assessment Task

The dataset consisted of approximately 1,600 scientific models produced by ninth-grade students participating in a high-school Physical Science curriculum focused on electrical interactions. Students completed a validated LP-aligned task in which they modeled an electroscope phenomenon under two conditions.

Students first observed that a charged rod caused the electroscope leaves to move apart. In Scenario

A the leaves moved apart slightly; in Scenario B they moved farther apart. Students were expected to represent charge across the rod and electroscope components, greater charge in Scenario B than Scenario A, and a correspondingly greater repulsive force between the leaves. The task focuses on macroscopic representations of charge transfer and qualitative Coulombic interactions rather than atomic-level mechanisms or energy; therefore, LP Level 2 is the highest level elicited by this item. Figure 1 shows the two electroscope scenarios, a complete Level 2 model, and the rubric-relevant charge and force representations.

Figure 1: Electroscope modeling task and example of a complete scientific model representing the highest LP level measured by the item. Categories 1–4–9 evaluate charge representations; Categories 5 and 10 evaluate repulsive force.



3.2 LP-Aligned Analytic Rubric and Human Scoring

As illustrated in Figure 1, the modeling rubric contains 13 analytic categories. Categories 1-4 identify point charges on the rod, sphere, hook, and leaves in Scenario A. Categories 6-9 evaluate whether greater amounts of charge are represented on the corresponding components in Scenario B. Categories 5 and 10 assess repulsive force between the leaves, including whether greater force is represented in Scenario B. Categories 11-13 capture common incomplete or inaccurate representations, including problematic representations of both charge types, similar amounts of charge across scenarios, and models indicating that the rod or electroscope is uncharged in Scenario A.

Three trained human raters independently evaluated student models using the same LP-aligned analytic rubric applied by the AI systems. Each category was coded according to the presence or absence of the specified evidence. Krippendorff's α exceeded .80 for every analytic category, indicating strong consistency among

trained human scorers. Human classifications served as the reference criterion for evaluating AI performance.

3.3 Computer Vision Approaches

Grounding dino + SAM. Grounding DINO [7] and SAM [4] were combined to identify and localize LP-aligned visual evidence in student models. Grounding DINO detected charge symbols and electroscope components, while SAM supported segmentation and spatial analysis. Detected features were assigned to Scenario A or B and to regions such as the rod, sphere, hook, and leaves, then combined to evaluate the 13 LP-aligned analytic categories. The foundation-model pipeline was applied without task-specific supervised training.

CustomCharge CNN. A CustomCharge CNN was developed in PyTorch [6] using supervised convolutional neural network principles [5]. Unlike the zero-shot foundation-model approach, the CNN was trained on human-labeled examples of student-generated electroscope models to classify the same LP-aligned analytic categories. Performance was evaluated on held-out data. The approaches therefore represent contrasting mechanisms for linking computer vision to theory-guided assessment: zero-shot detection of explicitly specified evidence versus supervised learning of task-specific representations from human-scored examples.

3.4 Analysis

AI classifications were compared with expert human scores separately for each of the 13 analytic categories. Cohen’s κ evaluated agreement beyond chance, F1 scores characterized classification performance while accounting for precision and recall, and precision-recall area under the curve (PR-AUC) assessed discrimination. Category-level analyses were used because identifying LP-aligned evidence is the AI-dependent component of the assessment; subsequent LP placement follows deterministic rules rather than additional AI inference.

We also inspected representative visual outputs to examine scoring-process validity [3]. Because the LP and rubric specify a priori which visual components constitute construct-relevant evidence, localized AI outputs can be compared with the features identified by trained human raters. This analysis does not claim that AI and humans

use identical internal reasoning; rather, it examines whether AI classifications are associated with the same substantive model components that define the scoring criteria.

4 Results

Both approaches identified multiple forms of LP-aligned evidence with substantial agreement with expert scoring, but performance varied by category (Table 1). Grounding DINO+SAM achieved higher Cohen’s κ than the CustomCharge CNN in 12 of 13 categories. Its strongest performance occurred across multiple LP-aligned categories, including Categories 1 ($\kappa=.868$), 2 (.853), 4 (.850), 5 (.865), and 9 (.866). The CNN also showed substantial agreement for several categories, although κ was generally lower.

As shown in Table 1, Grounding DINO+SAM also showed particularly strong agreement for force representations. Grounding DINO+SAM showed particularly strong agreement for force-related Categories 5 ($\kappa=.865$) and 10 ($\kappa=.822$), outperforming the CNN on both categories ($\kappa=.394$ and .432, respectively). The CNN achieved slightly higher agreement only for Category 13 ($\kappa=.781$ vs. .748). Both approaches showed lower agreement for Categories 11 and 12, with Grounding DINO+SAM achieving $\kappa=.599$ and .536, respectively. These categories involve less frequent and more complex patterns of evidence across model components and scenarios.

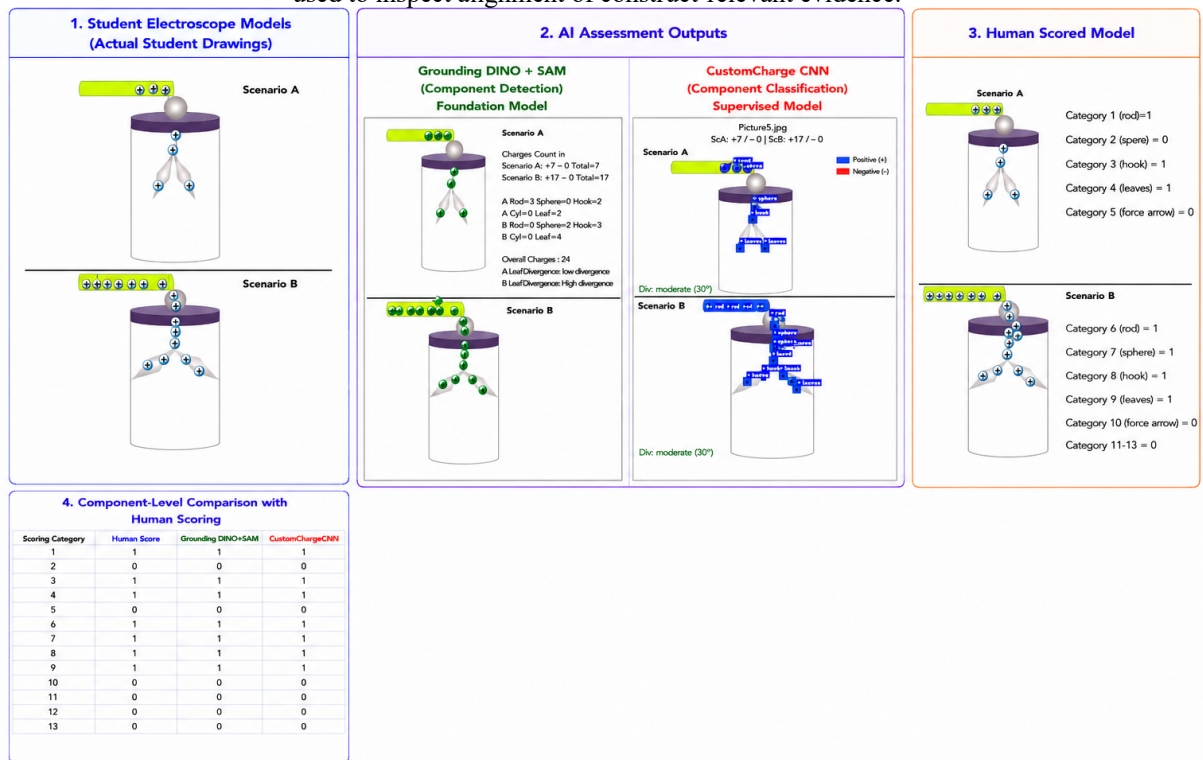
Categories 11 and 12 were also highly imbalanced, with substantially fewer positive than negative examples. This imbalance may contribute to the lower agreement observed for these categories and makes accuracy alone less informative. Therefore, Cohen’s κ and F1 were emphasized to better characterize performance beyond overall classification accuracy.

F1 and PR-AUC values provided complementary information about category-level performance. Strong charge-related categories generally showed high F1 and PR-AUC values, while Categories 11 and 12 remained more challenging. Representative output inspection further showed that, in high-agreement cases, detected features corresponded to the charge and other rubric-relevant components identified by trained human raters (Figure 2). Disagreement was concentrated primarily in less frequent and more complex model patterns.

Table 1: Human-AI performance across the 13 LP-aligned analytic categories.

Cat.	DINO κ	DINO F1	DINO PR-AUC	CNN κ	CNN F1	CNN PR-AUC
1	0.868	0.909	0.853	0.759	0.847	0.933
2	0.853	0.875	0.785	0.706	0.759	0.898
3	0.736	0.773	0.915	0.671	0.706	0.869
4	0.850	0.877	0.865	0.720	0.774	0.856
5	0.865	0.892	0.268	0.394	0.577	0.483
6	0.834	0.870	0.704	0.653	0.760	0.858
7	0.798	0.815	0.813	0.656	0.714	0.811
8	0.783	0.800	0.812	0.679	0.710	0.874
9	0.866	0.889	0.931	0.679	0.731	0.853
10	0.822	0.857	0.770	0.432	0.592	0.480
11	0.599	0.651	0.560	0.132	0.226	0.153
12	0.536	0.581	0.470	0.130	0.211	0.150
13	0.748	0.937	0.900	0.781	0.940	0.963

Figure 2: Representative comparison of student electroscope models, AI assessment outputs, and human scoring used to inspect alignment of construct-relevant evidence.



5 Discussion and Conclusion

The findings demonstrate that both foundation and supervised computer vision approaches can identify theoretically specified evidence in student-generated scientific models, but their strengths differ across forms of visual evidence (Table 1). Grounding DINO+SAM achieved higher agreement with expert scoring than the CustomCharge CNN in 12 of 13 categories, with strong performance across both charge- and force-related evidence. The CNN achieved slightly higher agreement only for Category 13. Categories 11 and 12 remained comparatively challenging and were characterized by highly imbalanced class distributions, with relatively few positive examples. This imbalance, together with the more complex patterns of evidence represented by these categories, may have contributed to their lower agreement.

These differences highlight the distinct characteristics of foundation and task-specific supervised approaches. Foundation models provide a potentially efficient starting point when construct-relevant evidence is visually distinguishable and labeled educational data are scarce. Supervised models require human-scored examples but can learn task-specific characteristics of student representations that are difficult to specify through zero-shot prompts. The force categories further demonstrate the effectiveness of the Grounding DINO+SAM approach. Despite variability in hand-drawn force representations, Grounding DINO+SAM achieved substantially higher agreement than the CNN for Categories 5 and 10. These findings suggest that foundation-model-based detection can effectively identify rubric-relevant force representations in student-generated models.

A central contribution of this study is the separation of AI-based evidence detection from LP-based developmental interpretation. The AI models do not learn an opaque mapping from a drawing directly to an LP level. Instead, the LP and ECD process determine which evidence is educationally meaningful, and AI is evaluated on its ability to recognize that evidence. Once analytic categories are assigned, LP placement follows predetermined scoring rules. This structure preserves the role of learning theory in determining what counts as more

sophisticated scientific modeling while using AI to reduce the burden of identifying evidence at scale. This separation also supports a stronger validity argument than predictive accuracy alone. Human–AI agreement indicates whether an automated score corresponds with expert classification, but educational validity additionally requires evidence that the system is responding to features relevant to the intended construct. Here, the analytic categories were defined before model training or application, and visual outputs could be inspected relative to those categories, as illustrated in Figure 2. High agreement accompanied by localization of the expected charge or force representations provides more informative evidence than an end-to-end proficiency prediction whose basis cannot be examined.

The category-level results further make weaknesses transparent. Rather than treating a single aggregate metric as evidence that a system is or is not valid, the framework identifies which substantive parts of the construct are well supported and which remain problematic. In the present study, charge and force representations were detected relatively reliably, while Categories 11 and 12 remained challenging due to class imbalance and less frequent, more complex response patterns. This information can guide targeted refinement of the AI component without changing the theoretical definition of the construct or the rules used to interpret evidence developmentally.

The complementary pattern also suggests a direction for future system design. A hybrid architecture could use foundation models for categories that can be detected reliably without labeled data and introduce supervised models selectively for features requiring task-specific specialization. Such an architecture may reduce annotation demands while retaining strong performance on difficult evidence types. The present study did not test a hybrid model, so this remains a design implication rather than an empirical finding.

Several limitations should be noted. The study focused on a single electroscope modeling task within one scientific domain. Performance may differ for other modeling tasks, age groups, or representational conventions. Several analytical categories, particularly categories 11 and 12, occurred relatively infrequently, limiting the number of positive examples available for both

supervised learning and stable evaluation. In addition, some rubric categories require relational interpretation across multiple visual components and across Scenarios A and B rather than detection of an isolated object. Future work should test the approach across additional LP-aligned tasks and investigate methods for representing heterogeneous and uncommon forms of student thinking without narrowing the range of scientifically acceptable representations. Overall, this study demonstrates a theory-guided approach in which learning progressions define what evidence matters and AI supports scalable identification of that evidence. The contribution is therefore not simply a comparison of two computer vision models. It is a framework for connecting developmental learning theory, human assessment criteria, and interpretable AI evaluation of complex student-generated scientific models.

References

- [1] National Research Council. 2012. A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press, Washington, DC.
- [2] Kaldaras, L., Akaeze, H., & Krajcik, J. (2021). Developing and validating Next Generation Science Standards-aligned learning progression to track three-dimensional learning of electrical interactions in high school physical science. *Journal of Research in Science Teaching*, 58(4), 589-618.
- [3] Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024, August). Developing valid assessments in the era of generative artificial intelligence. In *Frontiers in Education* (Vol. 9, p. 1399377). Frontiers Media SA.
- [4] Alexander Kirillov et al. 2023. Segment Anything. In *Proceedings of ICCV 2023*.
- [5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521:436–444.
- [6] Adam Paszke et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32.
- [7] Shilong Liu et al. 2024. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of ECCV 2024*.