How effective are VLMs in assisting humans in inferring the quality of mental models from Multimodal short answers?

Pritam Sil[1] **Durgaprasad Karnam**[2] **Vinay Reddy Venumuddala**[3] **Pushpak Bhattacharyya**[1]
[1]Department of Computer Science and Engineering, IIT Bombay, Mumbai, India
[2]Center for Educational Technology, IIT Bombay, Mumbai, India
[3]School of Management, Mahindra University, Hyderabad, India
pritamsil@cse.iitb.ac.in, karnamdpdurga@gmail.com
vinay.venumuddala@mahindrauniversity.edu.in

Abstract

STEM Mental models can play a critical role in assessing students' conceptual understanding of a topic. They not only offer insights into what students know but also into how effectively they can apply, relate to, and integrate concepts across various contexts. Thus, students' responses are critical markers of the quality of their understanding and not entities that should be merely graded. However, inferring these mental models from student answers is challenging as it requires deep reasoning skills. We propose *MMGrader*, an approach that infers the quality of students' mental models from their multimodal responses using concept graphs as an analytical framework. In our evaluation with 9 openly available models, we found that the best-performing models fall short of human-level performance. This is because they only achieved an accuracy of approximately 40%, a prediction error of 1.1 units, and a scoring distribution fairly aligned with human scoring patterns. With improved accuracy, these can be highly effective assistants to teachers in inferring the mental models of their entire classrooms, enabling them to do so efficiently and help improve their pedagogies more effectively by designing targeted help sessions and lectures that strengthen areas where students collectively demonstrate lower proficiency.

1 Introduction

Learning Science, Technology, Engineering, and Mathematics (STEM) requires students to build and manipulate abstract entities, such as mental models[1]. Mental models are "basic units of coherently structured knowledge" (Hestenes, 2009), representing cognitive structures that represent scientific understanding. They are often embodied in what is called physical intuition. However, students do not always construct coherent mental models, which can limit their conceptual understanding. Consequently, assessing the quality of students' mental models can serve as a proxy for evaluating the depth of their learning in STEM topics. Inferring such mental models at scale, particularly from handwritten student responses, presents a significant challenge.

Historically, educational measurement has relied on rubric-based approaches for evaluating student responses (Hillegas, 1912). However, such approaches are often labour-intensive, requiring instructors to create separate rubrics for each question, which limits scalability and generalizability. Recent advances in Artificial Intelligence (AI) offer new possibilities for inferring mental models from open-ended student responses. Natural Language Processing (NLP) techniques enable semantic similarity analysis, automated concept extraction, and discourse-level modelling, allowing evaluation beyond surface-level scoring. While these approaches can generate similarity scores, they do not explicitly reveal the underlying mental model of a student.

To address these limitations, we propose shifting the focus from scoring individual answers to directly inferring students' mental models as a measure of learning quality. In this exploratory study, we investigate whether students' underlying STEM mental models can be inferred from multimodal handwritten responses, which often contain both text and diagrams. This task requires deep reasoning capabilities, whether performed by humans or AI. Inferring mental models in formal STEM domains is challenging, as it involves higher-order cognitive processes that typically require specialised expertise even among human annotators.

[1] We use the terms "mental models" and "STEM mental models" synonymously throughout this work.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1130–1140
March 24-29, 2026 ©2026 Association for Computational Linguistics

Our contributions are as follows -

1. *MMGrader:* An approach for inferring students' STEM mental models from multimodal responses. (Section 4)

2. MMGrader decomposes topics into fundamental units, called concept links, and evaluates student responses using a provided mapping between questions and concept links, along with a predefined scoring scale. (Section 4)

3. Extensive Evaluation of how close are openly available vision–language models (VLMs) to approximating human annotators in reasoning over students' multimodal responses to infer mental models. (Section 5)

4. Evaluations reveal that the best performing model achieves an accuracy of $\tilde{4}0\%$ along a prediction error of 1.1 units, and a scoring distribution fairly aligned with human scoring patterns. (Section 5.2)

2 Related Work

Kenneth Craik is credited with introducing the term "mental models". In education, mental models serve as an effective tool for understanding a student's proficiency in a topic. However, researchers have quantified these models in various ways, leading to diverse approaches for assessing student understanding.

Based on this concept, Rus et al. (2009) inferred whether a student possesses a high, intermediate, or low mental model from student-generated paragraphs written during prior knowledge activation, a self-regulatory process. The answers came from the topic of the circulatory system in biology. Their approach was evaluated using traditional machine learning techniques such as Naive Bayes (NB), Bayes Nets (BNets), Support Vector Machines (SVM), Logistic Regression (LR), and two variants of decision trees. While effective, this line of work focused on only labels, and thus, a more detailed structure of mental models is still necessary for complex tasks such as grading.

To capture more detailed representations, Maharjan and Rus (2019) used concept maps as a tool for modelling students' mental models. They proposed a novel method called DT-OpenIE to extract tuples from student answers. This was achieved by converting sentences into short clauses and then extracting tuples, which were compared to ground-truth tuples provided by human annotators to grade student answers. Although this approach provided more structured insights, its dataset contained only short textual answers (1–2 sentences), which does not reflect the broader variety of student responses.

Extending this direction, Agarwal et al. (2022) introduced the use of Abstract Meaning Representation (AMR) graphs. In these graphs, nodes represent concepts or predicates, while edges capture relations such as subject/object. These elements were extracted from the student's answer to construct the AMR graph. The same is performed on the reference answer, converted to embeddings and compared to generate the final score. This approach, again, can be seen as inferring mental models from student answers. Despite this detailed methodology, the answers in this study were also limited to textual responses.

Similarly, Sahu and Bhowmick (2025) proposed constructing answer graphs that resemble AMR graphs, thereby representing students' mental models in a structured way to identify gaps in student responses. Once again, however, the answers considered were short (1–2 sentences) and textual.

Apart from short textual answers, Fan et al. (2023) extracted concept graphs from code submitted as answers to CS-1 programming assignments. These graphs were compared with those derived from reference solutions to automatically grade programming answers. While this demonstrates the applicability of concept graph–based approaches to other domains, it still does not address multimodal or longer responses.

Prior work has employed a range of representations for mental models, including simple labels (high/intermediate/low), concept maps, concept graphs, and AMR graphs. However, these approaches share several limitations as they primarily focus on short textual answers. Moreover, they do not account for symbols and notations that are common in domains such as physics, and they lack generalizability across disciplines. To address these challenges, we propose *MMGrader*, a novel approach for inferring mental models from multimodal student answers using concept graphs as an analytical framework. We begin by explaining how mental models can be represented in educational contexts in the next section.

3 How to represent mental models in Education ?

As mentioned by Hestenes (2009), mental models are "units of coherently structured knowledge" formed in the human mind when interacting with real-world objects. They can be "directly compared with physical things and processes" and are "embodied in physical intuition." This raises a fundamental question: how can we represent them? In the context of education, one effective representation is through concept graphs.

3.1 Concept Hierarchy

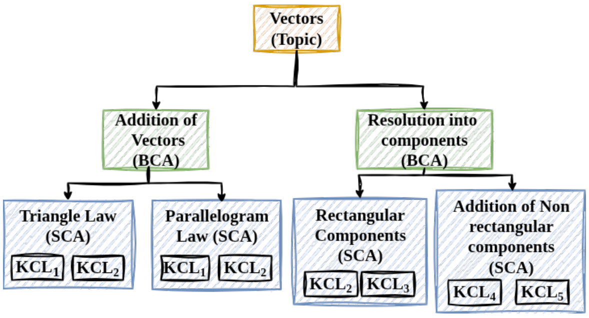

Vectors (Topic)
Addition of Vectors (BCA)
Resolution into components (BCA)
Triangle Law (SCA)
KCL1
KCL2
Parallelogram Law (SCA)
KCL1
KCL2
Rectangular Components (SCA)
KCL2
KCL3
Addition of Non rectangular components (SCA)
KCL4
KCL5

Figure 1: Concept hierarchy for the topic of vectors

In STEM education, each topic can be organised according to a concept hierarchy. A topic can be divided into broader concept areas (BCAs) or subtopics, which are further refined into subconcept areas (SCAs). SCAs, in turn, contain fundamental units called Key Concept Links (KCLs). When an educator teaches a topic or subtopic, these KCLs serve as the learning outcomes for that topic. A coherent mental model acquired by a student revolves around these KCLs.

Figure 1 illustrates a concept hierarchy for the topic of vectors. For example, when an educator explains the triangle law of vector addition, the learner strengthens their understanding of the direction (KCL_1) and magnitude (KCL_2) of vectors. Once incorporated into their mental model, these KCLs enable students to reason about how an object will move when multiple forces act upon it.

3.2 Concept Graph

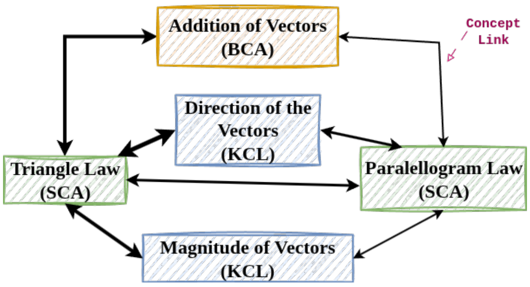

Addition of Vectors (BCA)
Concept Link
Direction of the Vectors (KCL)
Triangle Law (SCA)
Paralellogram Law (SCA)
Magnitude of Vectors (KCL)

Figure 2: A concept graph for addition of vectors

Building on the idea of concept hierarchies, we adopt the methodology of Karnam et al. (2021) to define a concept graph $G(V, E)$, where the vertices V are conceptual units and the edges E are concept links. The set of vertices includes components from the concept hierarchy: $V = BCA \cup SCA \cup KCL$. Each edge represents a relationship between two nodes. For instance, the link between the KCL "direction of vectors" and the "triangle law" reflects that understanding the triangle law strengthens a student's grasp of vector direction.

3.3 Concept Graphs as tools to represent mental models

Mental models are not directly accessible but are indicated by markers. Identifying such markers present in student responses is a challenging task. We present concept graphs as analytical frameworks to effectively perform this task.

This is because concept graphs capture strong associations between concepts within a topic, aligning closely with the definition of mental models as "units of coherently structured knowledge." Each edge in the graph can be assigned a strength score, reflecting how well a student has learned a particular concept. A higher score indicates more coherent and high-quality mental models and a greater understanding, which in turn suggests that the student can successfully apply that concept across other contexts.

Moreover, the concept graphs used for a particular topic need not be unique. An instructor can easily customise them according to the learning outcomes of a course. They can be seen as data structures to infer the completeness, correctness and strength of the underlying mental models. Building on this idea, we propose a new approach called MMGrader to infer students' mental models from multimodal responses.

4 MMGrader: An approach to infer students' mental models

Concept graphs can be effective analytical frameworks to infer students' mental models from their responses, as demonstrated by Karnam et al. (2021). Using these analytical frameworks on students' multimodal responses requires deep expertise and reasoning capacities, even among humans. By multimodal, we mean that the answers may involve text, diagrams, or a combination of the two. MMGrader addresses this by employing an analytical framework that involves breaking down the topic into concept links, establishing a relevancy mapping between questions and the concept links, and providing a rubric for evaluation. The approach is designed in such a way that instructors do not need to prepare any additional input beyond what is already present in standard STEM education.

The details of MMGrader[2] (Figure 3) are as follows:-

Input: MMGrader takes in the following as the input:-

- *Student assessment:* A student assessment containing handwritten responses containing text and diagrams.
- *Concept graph:* The concept graph associated with the topic. It serves as an analytical framework to analyse student responses.
- *Question–concept link mapping:* A mapping between a question to its relevant concept links.

Each topic includes a predefined concept hierarchy that is typically available and standardised in STEM curricula. During assessment design, the instructor maps each question to its relevant concept links. These concept links can be directly linked to the learning outcomes of a topic. Hence, the approach is flexible and can customized according to the instructor's requirements.

Step 1 (Evaluation): Given the question–concept link mapping, the proposed approach evaluates student answers to estimate a strength score for each concept link associated with a question. This is a challenging task as it requires diagram understanding capabilities along with deep reasoning skills. While humans can perform this task easily, we investigate whether openly available VLMs can replicate this ability.

Step 2 (Generation): Using the VLM-generated strength scores, the proposed approach constructs a concept graph representing the strength of the student's mental model. To generate the strength scores, we adopt a bottom-up approach along the concept hierarchy, starting from KCL–SCA pairs. The resultant strength score is computed as the average of all scores shown by the student across relevant questions. For example, say CL_2 (Figure 3) is being measured in question 1 and question 2, then its strength is computed as the average of the scores obtained in both. Similarly, we compute the strength scores of other CLs as average of the scores obtained on their mapped questions. The intuition is that once a student has grasped a fundamental concept, that is, a concept link, they should apply it consistently across all their answers. Similarly, we compute the other strength scores. This concept graph, built with the strengths of each link for a student, stands as a representation of the student's mental model.

The performance of MMGrader depends on the diagram interpretation and reasoning capabilities of the underlying VLMs. Therefore, evaluating MMGrader is equivalent to evaluating the performance of VLMs, details of which are provided in the next section.

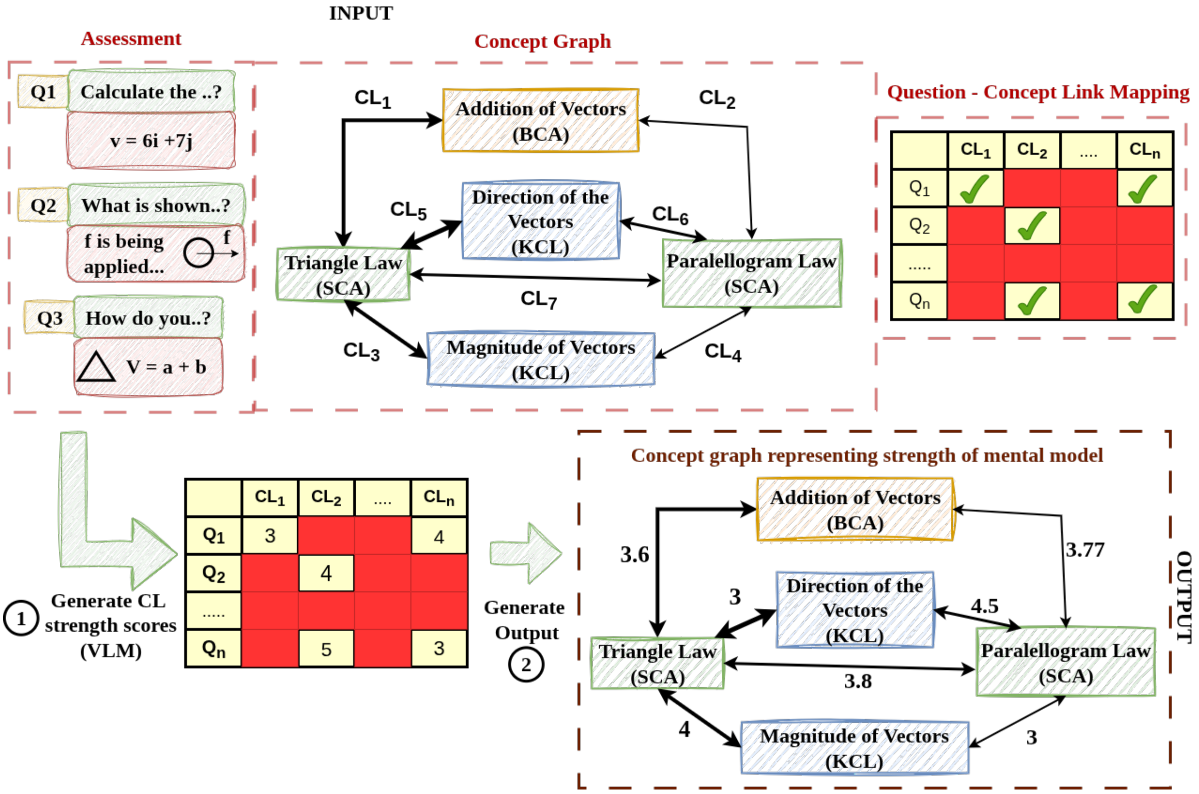
5 How close are VLMs to Humans ?

MMGrader leverages concept graphs as analytical frameworks to represent a student's STEM mental models. As a result, the task reduces to a simple scoring task. The scoring involves assigning a strength score to a concept link given the question, answer and scoring guidelines. To evaluate whether VLMs are comparable to human annotators on this task, we first construct a dataset as described in the next section.

5.1 Dataset Construction

The dataset was collected from an assessment consisting of *10 questions*, jointly prepared by educators and education researchers. The assessment was attempted by *6 students*, each of whom had varying scholastic levels, and the answer sheets were collected and annotated by 6 human annotators. All human annotators have at least 2-4 years of experience in education practice or research, and were working in a premier center of education research.

[2] Our code is available at : https://github.com/psil123/MMGrader


INPUT
Assessment
Q1
Calculate the ..?
v = 6i +7j
Q2
What is shown..?
f is being applied...
f
Q3
How do you..?
V = a + b
Concept Graph
CL1
CL2
Addition of Vectors (BCA)
CL5
Direction of the Vectors (KCL)
CL6
Triangle Law (SCA)
Paralellogram Law (SCA)
CL7
CL3
Magnitude of Vectors (KCL)
CL4
Question - Concept Link Mapping
1
Generate CL strength scores (VLM)
Generate Output
2
Concept graph representing strength of mental model
Addition of Vectors (BCA)
3.6
3.77
3
Direction of the Vectors (KCL)
4.5
Triangle Law (SCA)
Paralellogram Law (SCA)
3.8
4
Magnitude of Vectors (KCL)
3
OUTPUT

	CL_1	CL_2	....	CL_n
Q_1	✓	🟥	🟥	✓
Q_2	🟥	✓	🟥	🟥
.....	🟥	🟥	🟥	🟥
Q_n	🟥	✓	🟥	✓

	CL_1	CL_2	....	CL_n
Q_1	3	🟥	🟥	4
Q_2	🟥	4	🟥	🟥
.....	🟥	🟥	🟥	🟥
Q_n	🟥	5	🟥	3

Figure 3: Overview of MMGrader

Each question was manually mapped to a subset of *12 concept links* on the topic of addition of vectors and resolution of vectors from the 11th-grade curriculum. To establish ground-truth scores, each annotator was responsible for rating 2 answer sheets. After completing their individual annotations, all annotators met to resolve discrepancies in the assigned scores through offline discussion over the provided rubrics to reach a consensus. As a result, each *question–concept link–student answer* triplet was assigned a ground-truth strength score. The final dataset contained *895 data points*, where each data point corresponds to such a triplet with its associated score. This controlled annotation process provides more reliable ground-truth labels and ensures that the gathered data is of high quality.

All questions and answers in the dataset were handwritten. Each question included a diagram, while answers could be text-only, diagrammatic, or a combination of both. To extract the handwritten text from questions and answers, we used Google's Gemini model (`gemini-2.5-flash`) for optical character recognition (OCR). Gemini has been shown to outperform existing models on several OCR benchmarks[3]. Figure 4 shows a single data point from the dataset.

We plan to expand this dataset with additional samples collected from real-world examinations across various subjects, thereby increasing its scale and complexity and enabling a broader evaluation of multimodal student assessments.

5.2 Experimental Evaluation

The effectiveness of MMGrader depends on how well VLMs can assign strength scores on various concept links. To determine this, we evaluate a total of 9 openly available models that support multiple images along with complex prompts. Out of these models, Granite (Granite Vision Team, 2025) is a small 2B model while Molmo (Deitke et al., 2024), Qwen (Wang et al., 2024) and LLaVa (Wang et al., 2024) are 7B models. Apart from this, we have LLamaVision (Liu et al., 2023), which is an 11B model and Pixtral (Agrawal et al., 2024) and Gemma (DeepMind, 2025b), which are 12B models. We have also considered Gemini (DeepMind, 2025a), as it is closed-source but openly available. More details have been added in Appendix B. The main objective of this set of experiments is to determine whether existing VLMs are capable of eval-

[3] `https://benchmarking.nanonets.com/` (Last accessed: 06/10/2025)

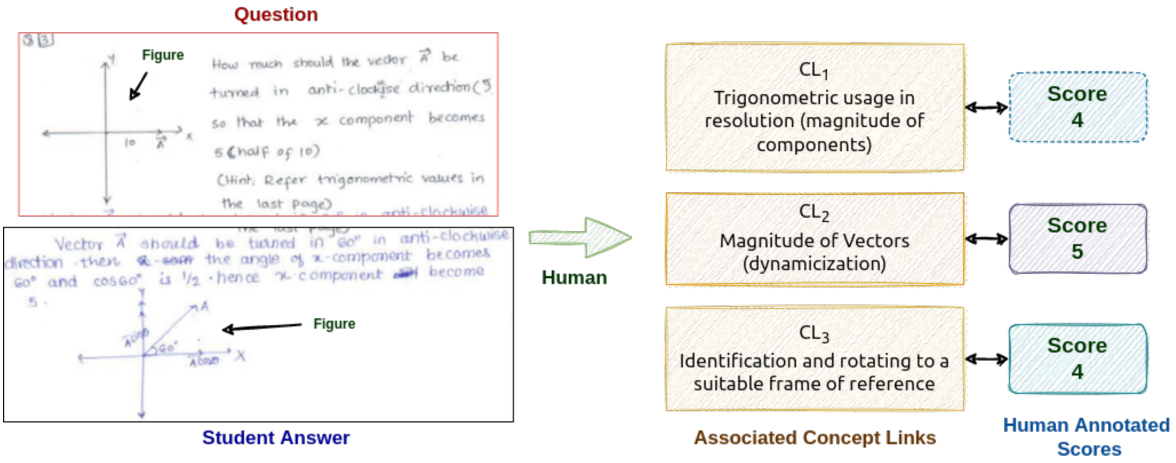

Question
3 B
Figure
How much should the vector A⃗ be turned in anti-clockwise direction (5 so that the x component becomes 5 (half of 10)
(Hint: Refer trigonometric values in the last page)
Vector A⃗ should be turned in 60° in anti-clockwise direction then the angle of x-component becomes 60° and cos60° is 1/2 · hence x-component become 5.
Figure
Student Answer
Human
CL₁ Trigonometric usage in resolution (magnitude of components)
CL₂ Magnitude of Vectors (dynamicization)
CL₃ Identification and rotating to a suitable frame of reference
Associated Concept Links
Score 4
Score 5
Score 4
Human Annotated Scores

Figure 4: Sample from our dataset

Model	Base Scoring Scenario			Generic Scoring Scenario			Detailed Scoring Scenario		
	Accuracy	**RMSE**	**EMD**	**Accuracy**	**RMSE**	**EMD**	**Accuracy**	**RMSE**	**EMD**
Molmo	29.66	1.42	0.42	**39.98**	**1.1**	**0.58**	**32.5**	**1.26**	**0.17**
Pixtral	**29.82**	**1.5**	**0.45**	29.82	1.5	0.22	28.7	1.45	0.49
InternLM	20.4	1.78	0.74	26.93	1.66	0.58	22.78	1.68	0.77
Gemma	22.54	1.66	0.8	24.44	1.6	0.61	18.62	1.7	0.93
LLama-Vision	19.77	1.98	0.93	24.27	1.7	0.5	22.71	1.71	0.53
Gemini	23.01	1.6	0.72	22.89	1.6	0.69	19.95	1.66	0.75
Qwen	18.15	1.83	3.89	18.15	1.83	0.91	26.33	1.51	0.64
LLava	13.17	2.0	1.23	13.76	1.95	1.21	13.4	1.99	1.21
Granite	8.9	2.25	1.18	10.79	2.24	1.3	9.37	2.2	1.13

Table 1: Performance of openly available VLMs on our dataset

uating handwritten multimodal answers and inferring the underlying mental models.

Three different experimental settings were considered, for which the detailed prompts can be found in Appendix A. The experimental settings were as follows:

- **Base Scoring Scenario:** The VLM is instructed to generate an integer strength score between 1 and 5, without specifying what each score represents. The prompt required the models to rely entirely on prior knowledge from pretraining when evaluating answers.

- **Generic Scoring Scenario:** The VLM is instructed to generate an integer strength score between 1 and 5, with generic specifications provided for each score. This setting offers a more guided scale, but the model must still interpret the concept link, question, and answer based on its prior knowledge. The scoring rubric added to the prompt was:
  - 1: No indication of ability to handle the link
  - 2: Very limited familiarity with the concept
  - 3: Inconsistent procedure, reproducing textbook understanding without modification
  - 4: Partial conceptual understanding, applying with some deviations from standard textbook usage
  - 5: Strong conceptual understanding

- **Detailed Scoring Scenario:** The VLM is instructed to generate an integer strength score between 1 and 5 with a concept-link–specific rubric. For example, for the concept link *magnitude of vectors*, the scoring scale used was:
  - 1: No indication of understanding regard-

ing the manipulation of magnitudes

  - 2: Very limited understanding of magnitudes
  - 3: No ability to dynamically simulate vectors
  - 4: Partial conceptual understanding of vector manipulation
  - 5: Strong conceptual understanding of vector manipulation

  This was the most guided setting, where each score definition was explicitly tied to the concept link, and the model's task was to map the student's answer to the appropriate score. This is the exact scenario which were provided to humans during the scoring task.

- **Chain-of-Thought (CoT) Scoring Scenario:** CoT is a prompting technique introduced by (Wei et al., 2022), which is effective in encouraging models to perform detailed reasoning before responding to a task. In our CoT prompt, we provide a generic scoring scale along with detailed reasoning steps. This set of experiments was designed to evaluate whether model performance improves when the models are explicitly instructed on how to think before generating the strength score.

We evaluated model performance using three metrics: exact match accuracy, root mean squared error (RMSE), and earth mover's distance (EMD) Rubner et al. (1998). Exact match accuracy measures whether the model's score exactly matches the human-annotated score, with higher values indicating closer alignment with human judgment. RMSE quantifies the deviation of model scores from human scores, while EMD evaluates whether the distribution of scores produced by the model resembles that of humans. A lower RMSE or EMD indicates closer alignment in scoring patterns.

Across all experiments, the best performance was achieved by **Molmo**, with an accuracy of 39.98%, RMSE of 1.1, and EMD of 0.58. This suggests that Molmo's predictions were typically within one unit of the human-assigned score and that its overall scoring distribution was fairly close to that of human annotators. We attribute this to Molmo's training on educational data, which has also enabled it to outperform other models on educational benchmarks.

Pixtral consistently achieved a low EMD score across all experimental scenarios, indicating that its scoring distribution closely resembled human scoring. However, with an RMSE of approximately 1.5, its predictions tended to deviate further from the ground truth on an individual level, resulting in lower accuracy. This suggests that Pixtral could easily mimic human scoring patterns if fine-tuned on this data. In comparison, **InternLM** also performed reasonably well, likely due to its pretraining on scientific notations and tokens. However, its performance lagged behind Molmo because of overthinking and other reasoning-related issues, as discussed in Section 5.3. **Gemma** and **LLamaVision** followed a similar trend.

For **Qwen**, we observed consistent improvements in performance as more fine-grained details were added to the prompt. Although Qwen is not specifically trained on educational data, it is able to pick up the additional details provided in the prompt and use them to evaluate the answers.

Gemini showed relatively strong performance, particularly in interpreting handwritten diagrams. However, its reasoning abilities sometimes introduced unnecessary confusion and overthinking into scoring decisions. This was evident in its generated rationales, as discussed in Section 5.3.

LLaVa, trained primarily on general-purpose data rather than educational datasets or hand-drawn diagrams, was limited in its performance on this task.

Finally, **Granite** performed the worst overall. As a smaller model trained on relatively simple tasks and without any prior hand-drawn images in its training data, its predictions were the least aligned with human annotations. This also explains why it frequently repeats the same sentence in its outputs.

	Accuracy	**RMSE**	**EMD**
Molmo	**34.16**	**1.27**	**0.63**
Qwen	29.66	1.44	0.6
InternLM	25.27	1.62	0.53
Pixtral	24.32	1.65	0.48
Gemini	23.13	1.61	0.58
Gemma	20.4	1.66	0.81
LLamaVision	20.47	1.97	0.78
LLava	10.56	2.09	1.19
Granite	0.95	3.93	3.63

Table 2: Performances of VLMs on CoT-based scenario

We also performed a similar set of experiments using a chain-of-thought (CoT) prompting strategy (Table 2). The results followed the same trend, further validating our earlier observation that **Molmo** is the most effective candidate among all VLMs. One interesting observation is that **Qwen** exhibited improved performance when provided with more fine-grained details. This suggests that, Qwen is able to leverage the step-by-step reasoning cues to perform better on the task.

To summarise, openly available VLMs still fall short of human-level performance at this task. With improved accuracy, these can be highly effective assistants to teachers in inferring the mental models of their entire classrooms, enabling them to do so efficiently and help improve their pedagogies more effectively by designing targeted help sessions and lectures that strengthen areas where students collectively demonstrate lower proficiency.

5.3 Qualitative Analysis of VLM Outputs

While the quantitative results provide an overall picture of the current state of existing VLMs, we also investigate why certain VLMs perform poorly.

When examining the predicted values for InternLM, Qwen, and Pixtral, we observed that they produced incorrect strength scores. We conjecture that this behaviour arises from overthinking and self-contradiction. An illustrative example from InternLM is shown in the box below.

Output generated by InternLM

...

First, I need to understand what the vectors $\vec{P_1}$ and $\vec{P_2}$ represent. The notation here seems a bit confusing. Typically, when vectors are written like $\vec{P_1} = 6\cos(50°)$ and $\vec{P_2} = 6\sin(50°)$, they might be components of a vector in a coordinate system.

...

Wait, maybe $\vec{P_1}$ and $\vec{P_2}$ are unit vectors in the x and y directions multiplied by 6.

...

But the original statement just says $\vec{P_1} + \vec{P_2} = \vec{R}$. If $\vec{P_1}$ and $\vec{P_2}$ are components of $\vec{R}$, then yes, their sum would equal $\vec{R}$. However, the notation here is unclear.

...

Wait, perhaps the question is using a different notation where $\vec{P_1}$ and $\vec{P_2}$ are components of a vector in polar coordinates.

...

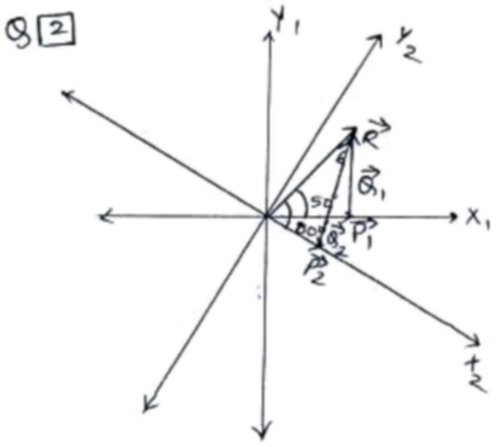


Figure 5: Image provided as part of one of the questions in the dataset.

The predicted sentences indicate that InternLM is not incorporating the given diagram in the question into its reasoning process, even though it provides a clear hint regarding what $\vec{P_1}$ and $\vec{P_2}$ actually represent (Figure 5).

For LLaVa, the model only predicted either 1 or 5 as the strength score, which corresponds to examples of the output format provided in the prompt. This indicates that LLaVa is not interpreting the prompt correctly. Similarly, smaller VLMs such as Granite also failed to interpret the prompt, suggesting that it is mainly trained for simpler tasks such as caption generation. As a result, it generated random strength scores. In one instance where it failed to generate any value, it instead produced the following: `<start of description> The image consists of a geometric diagram with a central point labelled as "O". From this point, three vectors are drawn, labelled as "B", "C", and "Y"`. This again demonstrates its inability to interpret hand drawn diagrams (Figure 5), highlighting the need to finetune it on handwritten data paired with simple instructions.

In the case of closed models, Gemini was able to generate strength scores with an accuracy of 23% for all scenarios. However, instead of following the prescribed output format, it often returned numbers in the form `\boxed{3}`, which is consistent with conventions used in standard benchmarks. While valid in that context, this behaviour is undesirable when the model is explicitly instructed to follow a different format.

6 Conclusion

Students' conceptual understanding in STEM depends on the coherence, accuracy and completeness of underlying STEM mental models. Inferring their mental models is hence a more effective way to design effective pedagogies. Students' responses are not just things to be graded, but as critical markers of the quality of their understanding. We present a method (MMGrader) that addresses the problem of inferring mental models using concept graphs as frameworks to analyse the multimodal responses.

Our evaluations indicated *Molmo* as the best-performing model across all parameters, even though there is a 1.1-unit difference between Molmo and expert scores. These results suggest a promising direction for utilising VLM-based approaches in designing scalable and interpretable auto-grading systems that can effectively quantify student understanding. This opens avenues for areas such as personalised education and Intelligent Teaching Assistants.

Limitations

The VLMs still require training on a real-life dataset to enhance their accuracy and bring them up to par with human annotators.

Ethical Considerations

The dataset was collected with prior permission from the students. The identities of both the students and annotators had been kept anonymous during the whole process.

References

Rajat Agarwal, Varun Khurana, Karish Grover, Mukesh Mohania, and Vikram Goyal. 2022. Multi-relational graph transformer for automatic short answer grading. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2001–2012, Seattle, United States. Association for Computational Linguistics.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Théophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint*.

Google DeepMind. 2025a. Gemini 2.5 flash: Price-performance balanced multimodal thinking model. *arXiv preprint arXiv:2507.06261*. Also refers to Gemini-2.5 model family including Flash, Flash-Lite, Pro, model card available online.

Google DeepMind. 2025b. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*. Variant: Gemma-3-12B-IT.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, and Ranjay Krishna. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*. Model: Molmo-7B-D-0924.

Zhiyu Fan, Shin Hwei Tan, and Abhik Roychoudhury. 2023. Concept-based automated grading of cs-1 programming assignments. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2023, page 199–210, New York, NY, USA. Association for Computing Machinery.

IBM Research Granite Vision Team. 2025. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *arXiv preprint arXiv:2502.09927*. Model: granite-vision-3.3-2B.

David Hestenes. 2009. Modeling theory for math and science education. In *Modeling Students' Mathematical Modeling Competencies: ICTMA 13*, pages 13–41. Springer.

Milo B. Hillegas. 1912. A scale for the measurement of quality in english composition by young people. *Teachers College Record*, 13(4):1–1.

Durgaprasad Karnam, Harshit Agrawal, Pranay Parte, Saurabh Ranjan, Priyanka Borar, Prasanna Prakash Kurup, Amose Jebin Joel, Pattamadai Sankaran Srinivasan, Uddhav Suryawanshi, Aniket Sule, and Sanjay Chandrasekharan. 2021. Touchy feely vectors: A compensatory design approach to support model-based reasoning in developing country classrooms. *Journal of Computer Assisted Learning*, 37(2):446–474.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*. Oral presentation.

Nabin Maharjan and Vasile Rus. 2019. A concept map based assessment of free student answers in tutorial dialogues. In *Artificial Intelligence in Education*, volume 11625 of *Lecture Notes in Computer Science*, pages 244–257. Springer.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV)*, pages 59–66. IEEE / Narosa Publishing House.

Vasile Rus, Mihai Lintean, and Roger Azevedo. 2009. Automatic detection of student mental models during prior knowledge activation in metatutor. In *Proceedings of the 2nd International Conference on Educational Data Mining (EDM 2009)*, pages 161–170, Cordoba, Spain.

Archana Sahu and Plaban Kumar Bhowmick. 2025. Directed graph-alignment approach for identification of gaps in short answers. *arXiv preprint arXiv:2504.04473*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*. Model: Qwen2-VL-7B-Instruct.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837. Curran Associates, Inc.

A Prompts Used

Below are the prompts used for various scenarios. Note that for the Detailed Scoring Scenario, we replace the scoring scale with the concept link-specific details as shown in Section 5.2.

Base Scoring Prompt

You are an expert evaluator of student responses.
The provided questions and student answers belong to the topic of addition and resolution of vectors from 11th standard.
The first image provided belongs to the question.
A second image if present belongs to the student answer.
A strength score is a number between 1 and 5 (both inclusive) which is used to represent how well a concept link has been expressed in the student answer.
Task : Your task is to generate the strength score of the concept link by analyzing the question and student answer pair.
Output Format (strict):
<Score>an integer between 1 and 5</Score>
Examples:
<Score>1</Score>
<Score>5</Score>

Generic Scoring Scenario

You are an expert evaluator of student responses.
The provided questions and student answers belong to the topic of addition and resolution of vectors from 11th standard.
The first image provided belongs to the question.
A second image if present belongs to the student answer.
A strength score is a number between 1 and 5 (both inclusive) which is used to represent how well a concept link has been expressed in the student answer.
Task : Your task is to generate the strength score of the concept link by analyzing the question and student answer pair.
Strength Score Scale (1–5):
1 : (No indication of ability to handle the link)
2 : (Very little familiarity with the skill)
3 : (Inconsistent Procedure) Trying to impose the text book understanding without any modification.
4 : (Inconsistent Concept/ Procedure- applying with some changes from a regular textbook usage)
5 : (Strong Conceptual)
Output Format (strict):
<Score>an integer between 1 and 5</Score>
Examples:
<Score>1</Score>
<Score>5</Score>

Chain-of-Though based Generic Scoring Prompt

You are an expert evaluator of student responses.
The provided questions and student answers belong to the topic of addition and resolution of vectors from 11th standard.

Model Name	Variant	Size
Granite	`ibm-granite/granite-vision-3.3-2b`	2B
Molmo	`allenai/Molmo-7B-D-0924`	7B
Qwen	`Qwen/Qwen2-VL-7B-Instruct`	7B
LLaVa	`llava-hf/llava-v1.6-mistral-7b-hf`	7B
Interlm	`internlm/Intern-S1-mini`	8.54B
LLamaVision	`meta-llama/Llama-3.2-11B-Vision-Instruct`	11B
Pixtral	`mistralai/Pixtral-12B-2409`	12B
Gemma	`google/gemma-3-12b-it`	12B
Gemini	`gemini-2.5-flash`	-

Table 3: Configuration of VLMs used

The first image provided belongs to the question.
A second image if present belongs to the student answer.
A strength score is a number between 1 and 5 (both inclusive) which is used to represent how well a concept link has been expressed in the student answer.
Task Instructions:
1. Review the question carefully.
2. Evaluate the student's answer.
3. Consider the scoring guide, which maps scores to descriptions.
4. Choose the score (1–5) that best reflects how effectively the student's answer demonstrates the concept link.
5. Return only the selected score in the specified output format.
Scoring Guide (1–5):
Each line can be read as "score : general description – concept-link specific description"
1 : (No indication of ability to handle the link)
2 : (Very little familiarity with the skill)
3 : (Inconsistent Procedure) Trying to impose the text book understanding without any modification.
4 : (Inconsistent Concept/ Procedure- applying with some changes from a regular textbook usage)
5 : (Strong Conceptual)
Output Format (strict):
<Score>an integer between 1 and 5</Score>
Examples:
<Score>1</Score>
<Score>5</Score>

B VLMs in Consideration

Table 3 contains the exact model configurations that were used. All the models except Gemini have been accessed via HuggingFace. For Gemini, we have used their API.