"Did my figure do justice to the answer?": Towards Multimodal Short Answer grading with Feedback (MMSAF)

Anonymous ACL submission

Abstract

In education, the traditional Automatic Short Answer Grading (ASAG) with feedback problem has focused primarily on evaluating text-only responses. However, real-world assessments often include diagrams alongside written explanations. To address this limitation, we introduce the Multimodal Short Answer Grading with Feedback (MMSAF) problem, which requires jointly evaluating textual and diagrammatic content while also providing explanatory feedback. Collecting data representative of such multimodal responses is challenging due to both scale and logistical constraints. To mitigate this, we develop an automated data generation framework that leverages LLM hallucinations to mimic common student errors, thereby constructing a dataset of 2,197 instances. We evaluate 4 Multimodal Large Language Models (MLLMs) across 3 STEM subjects, showing that MLLMs achieve 55% accuracy in predicting answer correctness (correct/partially correct/incorrect) and 75% accuracy in assessing image relevance. This also includes a human evaluation with 9 annotators across 5 parameters, including a rubric-based approach. Our rubric-based methodology provides a scalable mechanism for quantifying feedback quality, thereby reducing annotation costs through effective human–LLM collaboration. The rubrics also serve as a way to evaluate the feedback quality semantically rather than using overlap-based approaches.

1 Introduction

In educational assessments, the traditional Automatic Short Answer Grading (ASAG) with feedback problem focuses on developing scalable solutions for automatically grading short textual responses along with feedback. Such feedback is important (Deeva et al., 2021) for a student's growth. Moreover, corrective, motivational and informative feedback can drastically speed up a student's learning process and help the student develop an innate curiosity.

However, assessments often contain open-ended responses that have both a textual part and a diagram. These questions play a crucial role in assessments, as they require students to express their understanding through textual and visual elements, leading to deeper levels of cognitive engagement. While in some scenarios, students are required to draw such diagrams (E.g. ray diagrams in physics), in others, they are required to label certain aspects of it (E.g., a diagram of a body part, such as the heart in biology). In such a diagram, each small component, such as an arrow, holds a semantic value that contributes to evaluating the answer as a whole. Grading such responses at scale while providing individual feedback to students is often difficult, especially in classrooms with a low teacher-to-student ratio (Burrows et al., 2015).

This leads to the question: Can we develop scalable assessment tools that can assist teachers in evaluating multimodal responses while providing quality feedback? To facilitate research towards developing such systems, we propose the Multimodal Short Answer grading with Feedback (MMSAF) problem. The problem is challenging for MLLMs as it tests their diagram understanding, natural language comprehension, and comparative reasoning capabilities.

Our contributions are as follows -

1. Introduction of the MMSAF problem, along with a dataset of 2,197 instances. (Section 3, Section 4.1)
2. MMSAF-DGF, a framework to generate an MMSAF dataset from a given set of standard label questions and reference answer pairs. It simulates common student errors using LLM hallucinations as a tool. (Section 4.1)
3. A rubric-based approach to evaluate the quality of feedback coupled with extensive zero-shot evaluation on 4 existing MLLMs across

3 different STEM subjects. We observe the following: (Section 6)

- Accuracy of up to 62.5% using existing MLLMs on generating the correctness level of a response (correct/partially correct/incorrect).
- Accuracy of up to 80.36% using existing MLLMs on generating the image relevance level (relevant/irrelevant) of a diagram in a response.
- Human evaluations indicate feedback generated by such MLLMs is closely related to how teachers provide feedback to students.

2 Related Work

In recent years, there has been growing interest from both the natural language processing (NLP) and education research communities in the task of Automatic Short Answer Grading (ASAG) with feedback. A notable milestone in this direction came in 2022, when Filighera et al. (2022) introduced the first dataset for ASAG with feedback problem. This bilingual dataset focused on short textual responses to questions across various topics, primarily in computer science. However, the dataset was limited to only about 2,000 responses, and lacked diversity across different engineering disciplines.

To address these shortcomings, Aggarwal et al. (2025) introduced the EngSAF dataset, which contained about 5,800 student answers drawn from 25 different engineering courses spanning multiple subfields. This dataset laid the groundwork for more robust benchmarking and model development in the ASAG with the feedback task.

Following this, research began to shift towards more advanced methods for the ASAG with the feedback problem. While earlier approaches primarily leveraged carefully crafted prompts, Fateen et al. (2024) proposed a retrieval-augmented generation (RAG) based approach to enhance response quality and contextual relevance.

Any feedback generated by the model serves as a way to explain its grading rationale, which can add a layer of explainability to the grading task. Li et al. (2023) introduced the Automated Explainable Student Response Assessment (AERA) framework, which generates scoring rationales using ChatGPT. AERA demonstrated rationales comparable in quality to human explanations and achieved a Quadratic Weighted Kappa (QWK) score of 11% on benchmark datasets. Similarly, Tornqvist et al. (2023) presented ExASAG, an explainable grading framework that integrates SHAP (Shapley Additive exPlanations) with SciBERT to introduce interpretability to the process of grading such assignments.

While these efforts significantly advanced the field of text-based ASAG, they fall short in handling multimodal responses containing only visual models or diagrams. Visual models, as part of student answers, play an important role in demonstrating the student's proficiency level on a particular concept. To address this, Leong et al. (2018) and later Sagherian et al. (2022) proposed automated grading systems for scientific visual models, evaluated on proprietary datasets from the Educational Testing Service (ETS). However, these visual models were created using predefined shapes such as boxes, arrows, and fish, limiting student expressiveness and creativity. Similarly, Lee and Zhai (2023) showed that GPT-4V is capable of evaluating such visual models. However, none of these works deal with the scenario of evaluating responses containing both text and diagrams.

To bridge this gap and handle real-world scenarios where students provide multimodal responses that include textual answers accompanied by diagrams, we introduce the MMSAF problem in the next section.

3 The Multimodal Short Answer Grading with Feedback (MMSAF) Problem

We introduce the Multimodal Short Answer Grading with Feedback (MMSAF) problem, which focuses on evaluating student responses containing both textual and visual content. Given a Question (Q), a Reference Answer (RA), and a Student Answer (SA), the objective is to assign a Level of Correctness (LC) label, an Image Relevance (IR) label, and a Feedback (F) which provides rationale behind the LC and IR labels.

Figure 1 provides an illustrative example. In the given question, the student is asked to *explain* the flow of blood in the human heart *using a diagram*. The question demands a supporting diagram, which adds clarity and depth to the textual response. Note that creating such diagrams using only simple visual models is difficult, making it significantly more challenging than standard visual model scoring tasks. This highlights the motivation

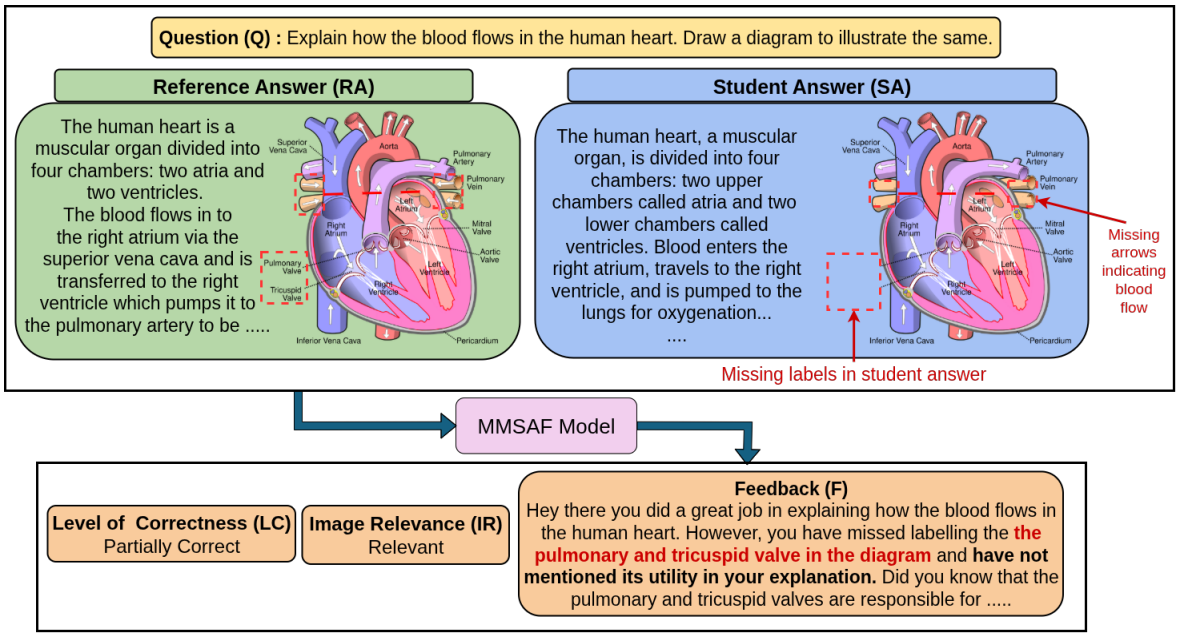


Figure 1: Illustration of the MMSAF problem with an example. (Heart diagram taken from edurev.in)

for introducing this problem.

3.1 Problem Formulation

We decompose the MMSAF problem into two sub-problems: a classification task for LC and IR labels, and a feedback generation task.

- **Classification Task:** Given the SA, the MLLM M must evaluate it with respect to the given Q and RA and determine its correctness level that is correct (C) or partially correct (PC) or incorrect (I):

$$(Q, SA, RA) \xrightarrow{M} \{C, PC, I\}$$

Similarly, given the SA, the model has to determine whether the image in the student response is relevant (R) or irrelevant (IRel) to the given Q and RA:

$$(Q, SA, RA) \xrightarrow{M} \{\text{R, IRel}\}$$

- **Feedback Generation Task:** Here given the SA, the MLLM (M) has to evaluate it with respect to the Q and SA and generate the feedback. The feedback should point out the errors present in SA, ways to mitigate them and additional information which might assist the student. This acts as a small explanation as to why the LC and IR levels were assigned by the model to the SA and increases trust of scoring systems.

Thus, it can be formulated as :

$$(Q, SA, RA) \xrightarrow{M} \text{Feedback}$$

So, **why is this challenging for MLLMs?** The MMSAF problems require the model to understand the semantics present in both the text and diagram. This involves the model to perform the following -

- **Natural language understanding:** The model has to analyse the semantics of the student response, including topic-specific details. This means the model must have prior knowledge of educational data to perform this task effectively (Peris et al. (2022)).

- **Diagram understanding:** The model has to interpret and capture finer details present in the diagram, such as arrows, labels and even topic-specific objects, such as lenses in the case of physics. It depends on their ability to make use of their background knowledge as shortcuts to identify and reason about the relational information (Giledereli et al. (2024)).

- **Comparative reasoning:** Now that it has analysed the text and diagram present in both the SA and RA, it has to now compare between them. Thus, the model's comparative

reasoning (Yu et al., 2023) abilities play a crucial role in generating corrective, motivational, and informative feedback.

In addition to these technical challenges, collecting data representative of such responses is often a significant challenge. This involves collaborations with different educational institutions, which may take a longer time to evolve, a lack of individualised feedback provided to students, and addressing privacy concerns related to sharing such data. To reduce the delay in curating such datasets, we propose an MMSAF-DGF in the next section.

4 MMSAF dataset generation framework (MMSAF-DGF)

MMSAF-DGF provides a solution to generate the MMSAF dataset by leveraging standard questions and answer pairs, as well as LLM hallucinations. This reduces the time and monetary effort required in curating such datasets.

4.1 MMSAF Dataset

We formally define the **MMSAF dataset** where each data point is composed of 7 elements, namely: Q, RA, SA, LC, IR, F and Rubrics for error detection in feedback (FR). As per the problem, given a (Q, RA, SA) tuple, the task is to generate feedback that evaluates both the correctness and image relevance of the student's answer. Moreover, the rubric-based feedback evaluation strategy represented by the FR value serves the following purpose:

- A way to semantically evaluate feedback quality rather than relying on existing metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), which rely on word overlap.
- An approach towards reducing annotation cost by employing both humans and the LLM-as-a-judge paradigm to evaluate feedback quality.

Another aspect pertaining to FR is that it is specific to a particular datapoint and not shared, since the errors introduced in the responses are all question-specific. Note that in each datapoint, the Q and RA pairs are standardised, indicating that they have been carefully curated by subject matter experts (SMEs). The challenge is to generate the SA, LC, IR, F and FR components. Before we explain how the same is done, we explain how LLM-based hallucinations can be used to mimic student errors.

4.2 Leveraging LLM hallucinations to simulate student errors

In this subsection, we elaborate the psychology behind common errors students make in examinations. Marsh and Eliseev (2019) primarily categorises them as follows:

- *Errors made with confidence:* This happens when students answer questions confidently but are unaware of the answers they provide. This means that the student is either confidently placing a correct fact as the response or making up answers to attempt it. E.g. "Haemoglobin consists of calcium and globin." Here, given the context, calcium is an incorrect response, and the student is unaware of it.
- *Misunderstanding:* Such errors are often made by students when they are unable to understand the question properly or even the related concept. This indicates that the provided answer again contains some partially correct fact. E.g. "Bats shout at objects, hence mammals can too perform echolocation" Here, the student misunderstands the correct fact that bats have special organs to enable echolocation.
- *Conceptual Change:* These kinds of errors are made when the student has understood a particular concept but is unable to apply it in a different context. This means that the student is confused and ends up using a fact or concept that is unrelated to the problem. E.g. "Diffusion happens because the oxygen and carbon dioxide molecules move willingly." Here, diffusion occurs due to a difference in concentrations, but the student has written a wrong fact.

Additional examples and explanations, as provided by Marsh and Eliseev (2019), can be found in Appendix M. We argue that LLM hallucinations, particularly factual fabrication and inconsistency (Huang et al., 2025), can be used as a tool to simulate such responses. As per Huang et al. (2025), factual inconsistency involves using a correct fact in a wrong context, and factual fabrication involves making up incorrect facts. This phenomenon can be easily used to mimic student responses, as shown in the earlier examples. Using this notion, we explain how MMSAF-DGF generates an MMSAF dataset.

4.3 Overview of MMSAF-DGF

Figure 2 illustrates the overall workflow of MMSAF-DGF. The framework consists of the following steps:

Step 1: It takes in a set of standard questions and reference answer pairs. Note that the reference answer contains both a text and a figure or diagram.

Step 2: The framework introduces errors to the correct answers to generate the partially correct and incorrect student responses. We take a divide-and-conquer approach by introducing errors in the text and visual part separately and then combining them using a correctness matrix to generate a data point.

For the textual part, hallucinations are introduced to generate the partially correct and incorrect responses, while the correct answers are rephrased to generate the correct student response. Gemini (Team, 2024) was employed to generate hallucinations due to its free availability. To assess its effectiveness in simulating student errors, we evaluated 50 instances each for Partially Correct (PC) and Incorrect (I) responses. Our results show that Gemini successfully produced hallucinated responses in 86% of the Incorrect cases and 80% of the Partially Correct cases.

For the provided diagram, we treat it as a correct response diagram. Partially correct images are generated by removing labels and objects, replacing closely related objects or swapping labels to generate a , and randomly assigning another image from the subject pool to generate the incorrect image.

Step 3: The C, PC and I textual and visual components are combined together to generate the final student response using the provided correctness matrix (Table 3 in Appendix D). When introducing hallucinations, we record the incorrect fact introduced and formulate it as a question for the rubrics and as a text for the sample feedback.

Using this framework, we generate a dataset of *2,197 instances* using *181 high school-level questions across physics, chemistry, and biology*. The dataset split statistics are in Appendix K. We first justify the choices of the 4 MLLMs in Section 5 and use the generated dataset to benchmark them in Section 6.

5 LLMs in Consideration

The MMSAF problem involves evaluating multiple images and text as a whole. As mentioned in Section 3.1, this is a complex task that relies on the MLLM's diagram understanding, natural language understanding, and comparative reasoning capabilities. This indicates that only models that support multiple images as inputs, are well-versed in complex tasks, and have educational data in their prior knowledge can be used. Given this motivation, we select 4 such models: ChatGPT, Gemini, Pixtral, and Molmo. Open-source models (Molmo and Pixtral) were accessed via the Huggingface library,[1], while APIs were used for ChatGPT[2] and Gemini.[3]

ChatGPT: ChatGPT (GPT-4o mini) from OpenAI is a well-established multilingual and multimodal model. It has demonstrated strong performance in grading and educational tasks, including diagram-based scoring (Lee and Zhai, 2023), making it a suitable candidate for MMSAF.

Gemini: Gemini, developed by Google, is at par with ChatGPT on multimodal reasoning benchmarks like MMMU. It has outperformed ChatGPT in various tasks.[4]. We use the freely available *gemini-1.5-flash* for our experiments. Further experiments with *gemini-2.5-flash* have been added in Appendix O.

Pixtral: Pixtral 12B (Agrawal et al., 2024), by Mistral.ai, is an open-source multimodal model that outperforms models like LLaVA, Claude-3, and Gemini-1.5 Flash on benchmarks like MMMU.[5] Built on Mistral Nemo 12B and a custom vision encoder (Pixtral-ViT), it combines efficiency with strong performance, making it a robust MLLM. Addtionally, it is trained to interpret diagrams while providing detailed and structured interpretations.

Molmo: Compared to other MLLMs, Molmo (Deitke et al., 2024) by AllenAI has been specifically trained on academic datasets, which makes it a possible candidate for solving the MMSAF problem. According to their website,[6] Molmo has outperformed existing leading models, such as GPT -4, Gemini 1.5, and Claude-3, on 11 different academic benchmarks and in human evaluations. In particular, we select the *Molmo-7B-D-0924* model variant, which is suitable for our dataset's academic and multimodal nature.

[1] https://huggingface.co/

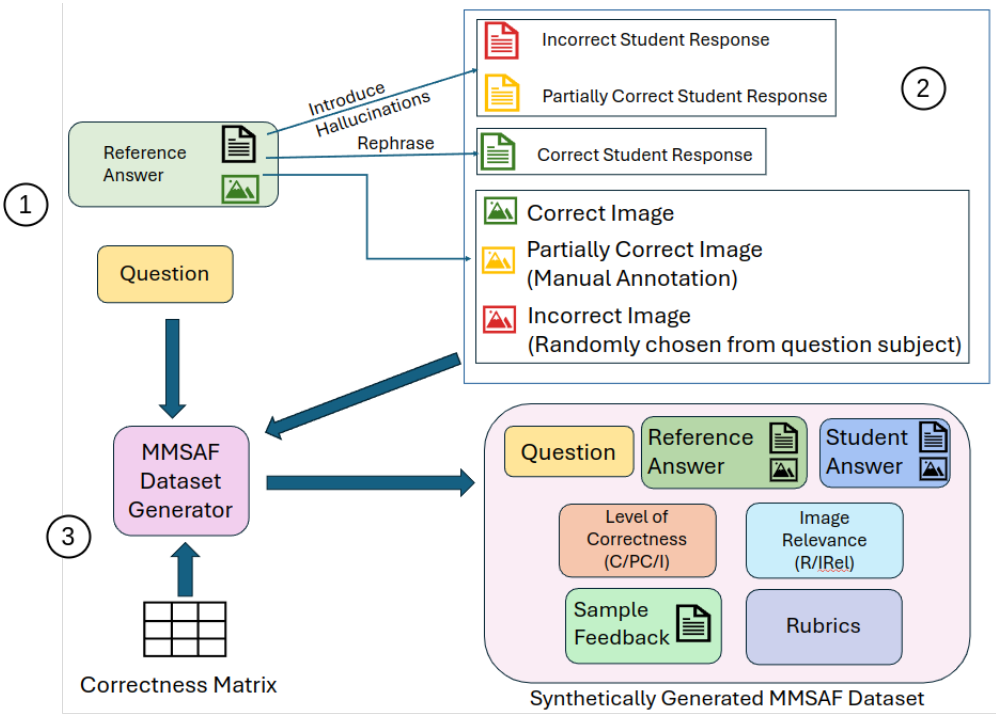
[2] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[3] https://ai.google.dev/gemini-api/docs/api-key

[4] https://blog.google/technology/ai/google-gemini-ai/#performance (Last Accessed: 06-10-2025)

[5] https://mistral.ai/news/pixtral-12b/ (Last Accessed: 05-12-2024)

[6] https://molmo.allenai.org/blog (Last Accessed: 06-10-2025)



C - Correct PC - Partially Correct I - Incorrect R - Relevant IRel - Irrelevant

Figure 2: Overview of the MMSAF dataset generation framework (MMSAF-DGF)

The next section details how these models were evaluated on our dataset.

6 Evaluation of LLM Generated Feedback

The goal of this evaluation is to benchmark the zero-shot performance of existing MLLMs on this dataset and grade their capabilities. To do so, 221 data points (130 from biology, 56 from chemistry and 35 from physics) were sampled randomly and fed to different LLMs using the prompt provided in Appendix E. The so generated level of correctness, image relevance labels and feedback was collected and analyzed in the later sections.

Note that the experimental setup can be found in Appendix L. The corresponding LC , IR and feedback values generated were collected and then analysed in the following subsections.

6.1 Analysis of Correctness and Relevance levels

To evaluate the performance of the models in predicting Level of Correctness and Image Relevance labels, we report macro-averaged accuracy values across each subject, namely physics, chemistry, and biology.

As shown in Table 1, Gemini outperforms all other models across the evaluated metrics, suggesting higher reliability in predicting LC labels and fewer false positives. This is due to equal strength in the 3 capabilities (as in Section 3.1) necessary to solve this task. When it comes to ChatGPT, it had a tendency to label most answers as "Partially Correct" class, which impacted its overall performance. Molmo, in contrast, exhibited a strong bias towards labelling answers as "Incorrect" indicating that it is not so adept at this complex task. Pixtral, although more lenient, frequently confused "Incorrect" and "Partially Correct" responses, resulting in reduced performance but close to that of Gemini. However, this behaviour indicates potential for performance improvement through fine-tuning.

Table 1 also presents the results for IR prediction. ChatGPT achieves the highest performance across all metrics, indicating strong reasoning and diagram understanding capabilities. However, Molmo frequently predicted most images as relevant, resulting in poor performance, indicating scope for improvement in its performance. Pixtral also suffered from false positives, while Gemini often misclassified relevant images as irrelevant.

Further breakdowns and confusion matrices can be found in Appendix G and Appendix H. Next, we analyze whether they feedback generated by them are close to human judgement or not.

Model	Level of Correctness			Image Relevance		
	Physics	Chemistry	Biology	Physics	Chemistry	Biology
ChatGPT	48.57	58.93	47.29	**77.14**	**80.36**	**72.09**
Molmo	40.00	39.29	43.85	34.29	26.79	29.23
Pixtral	51.43	58.93	49.23	74.29	60.71	66.15
Gemini	**57.14**	**62.50**	**50.77**	**77.14**	51.79	55.38

Table 1: Accuracy of models on generating LC and IR levels

6.2 Evaluation Task for Experts

The LLM-generated feedback, comprising 200 to 300 words, was recorded and presented to 9 Subject Matter Experts (SMEs), comprising 3 experts each for the domains of physics, chemistry, and biology. Relevant details about SMEs are mentioned in Appendix F.

The SMEs were instructed to score each feedback on a scale of 1 to 5 based on the following parameters -

1. *Fluency and Grammatical Correctness (FGE):* This metric denotes the fluency and grammatical correctness of the LLM-generated feedback. The idea is to check if the LLM-generated sentences are grammatically correct or not. A score of 1 denotes that the FGE level of the feedback is extremely poor, while a score of 5 indicates that it is excellent.
2. *Emotional Impact (EI):* This metric is to check whether the LLM-generated feedback will have a positive impact on the student or not, that is, whether the feedback is more encouraging and assistive for the student or not. A score of 1 denotes a negative impact, while a score of 5 denotes a positive impact.
3. *Level of Feedback Correctness (LFC):* This metric is to determine whether the feedback has properly captured all the errors present in the student's answer. A score of 1 denotes that no error has been captured in the feedback, while a score of 5 denotes that all the errors have been captured in the feedback.
4. *Error Mitigation in Feedback (EM):* This metric evaluates whether the feedback has properly addressed each and every error present in the student's answer and suggested ways to correct them. A score of 1 denotes no such error mitigation has been done, while a score of 5 denotes that all the ways necessary to correct all errors are present.
5. *Rubrics for error detection (FR):* While traditional NLP metrics like ROUGE-2 (Ganesan, 2018), SCAReBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020) rely on n-gram overlap, they often miss the semantic accuracy of feedback, particularly in identifying and addressing student errors. To address this, we propose a rubric-based evaluation that checks whether the feedback captures all relevant errors. Annotators assess each rubric as a True/False question based on the LLM-generated feedback.

6.3 Analysis of Expert Evaluation

Once the three SMEs had completed their respective evaluation tasks, all the scores for each metric were collected and averaged out to present the final data. Any disagreements were resolved through majority voting among the three raters. Table 2 summarises the average ratings assigned by annotators to the feedback generated by different LLMs over various criteria mentioned in Section 6.2 for each and every subject.

Physics: Physics questions involve interpreting abstract diagrams containing objects such as arrows, circles and rectangles. It also involves interpreting certain domain-specific objects such as lenses and mirrors. Apart from diagrams, such questions involve calculations and physics concepts based on reasoning. ChatGPT outperformed others in all areas, with SMEs highlighting its strength in identifying calculation errors and providing step-by-step explanations. This is reflective of its strong reasoning and math skills, also validated by benchmarks like MATH-500 and MMMU.[7] The scores from Table 2 indicate that it is most closely aligned

[7] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Model	Physics					Chemistry					Biology				
	FGE	EI	LFC	EM	FR	FGE	EI	LFC	EM	FR	FGE	EI	LFC	EM	FR
ChatGPT	**5.00**	**4.59**	**4.53**	**4.56**	**0.78**	**4.98**	**4.98**	**4.42**	**4.51**	0.63	5.00	3.07	3.24	3.15	0.50
Molmo	4.75	2.66	2.69	2.25	0.53	4.70	3.25	2.93	2.98	0.48	4.92	3.11	3.04	2.69	0.54
Gemini	5.00	4.56	4.12	4.09	0.68	4.95	4.67	4.37	4.42	**0.65**	5.00	3.59	**3.43**	3.07	0.53
Pixtral	5.00	4.15	4.24	4.24	0.55	4.95	4.71	4.10	4.05	0.49	**5.00**	**3.87**	3.40	**3.48**	**0.58**

Table 2: Average expert evaluation scores of different metrics on each subject

to how humand provide feedback in Physics. However, in contrast, Molmo often hallucinated, mislabeled correct answers, and struggled with reasoning, though it handled concrete diagrams reasonably well. Pixtral delivered structured feedback with clearly structured explanations, aligning with its training data where it had to interpret graphs and diagrams and provide structured analysis.

Chemistry: Chemistry questions involve domain-specific chemical formulas as part of their questions and diagrams. They also include interpreting graphs. Some diagrams involve interpreting domain-specific objects such as a beaker, a scientific fork and others. When compared to other models, ChatGPT performs the best on all parameters, which indicates that it also provides human-like feedback in chemistry. Additionally, SMEs note that ChatGPT provides detailed explanations. The reason is the same as for physics questions, as these questions test the reasoning skills of such MLLMs and how well they can interpret abstract diagrams. However, they also point out that Molmo's tone was too direct for students and did not provide proper explanations for its assigned labels, particularly in terms of correctness and image relevance. This is because Molmo is not particularly trained to interpret such domain-specific notations and symbols involved in chemistry.

Biology: Biology questions are more factual, involving real-life diagrams such as a heart. Interpreting such diagrams is much more complex as it requires interpreting domain-specific objects coupled with minute details such as arrows and labels which add semantic value to the diagram. MLLMs trained on such data will perform better in identifying key concepts. Pixtral excelled, providing structured, concept-based feedback and maintaining a polite tone, leading to higher emotional impact scores. This is also attributed to its training data, which consists of chart interpretation, allowing it to capture minute semantic details present in the diagrams. Although Molmo demonstrated a decent understanding of the diagram, it failed to connect errors to relevant concepts, highlighting a gap in its reasoning. ChatGPT and Gemini performed well, but not as effectively as Pixtral which is attributed to its inability to interpret minute details present in such diagrams.

To summarise, ChatGPT is most aligned towards human judgement in generating feedback for physics and chemistry due to its strong reasoning skills and diagram interpretation skills. However, Pixtral is more effective in biology, which is attributed to its ability to capture minute elements present in such biological diagrams. It is also due to the fact that it provides detailed, empathetic feedback and structured analysis. Again, SMEs noted Molmo's feedback often included Chinese alphabets, was overly direct, and sometimes mislabeled correct answers, suggesting a need for fine-tuning. More details are included in Appendices I and J and ablation studies in Appendix N.

7 Conclusion

Grading of multimodal responses and providing individualised feedback at scale is challenging. TO aid designing scalable solutions to this problem, this work introduces the MMSAF problem, coupled with a data generation framework (MMSAF-DGF) that can generate such datasets using standard Q and RA pairs. The framework leverages LLM hallucinations as tools to simulate erroneous student responses. Using this framework, we generate a dataset of 2,197 instances, built on questions from physics, chemistry, and biology textbooks used in high school.

We also establish a baseline using 4 MLLMs across all 3 subjects. Evaluations indicate that certain MLLMs in each subject are more aligned towards human judgment owing to their superior diagram understanding, natural language understanding and comparative reasoning capabilities.

Limitations

This work follows a 3-point scale for grading the answers as correct, partially correct, or incorrect. This is done to limit the label space and verify whether MLLMs are performing well on this on a larger scale, which is the case in real-life assessments. However, such systems can still be used to provide feedback for small-scale assessments. The current work relies on Gemini to introduce hallucinations. However, other MLLMs are yet to be explored in this context.

While the dataset is currently restricted to physics, chemistry, and biology questions at the high school level, adding data points from other subjects and university-level courses can increase its complexity and richness. Another limitation of the automated pipeline pertaining to partially correct images is that it needs to be manually annotated, which can lead to a scalability issue. This problem can be automated using simple text manipulation operations performed via OpenCV[8] or by generating diagrams using solutions such as DiagrammerGPT (Zala et al., 2024) once their code is released. Apart from this, MMSAF-DGF is a concept-heavy domain and remains to be extended into other generic domains, such as reasoning.

Ethical Considerations

The questions and reference answers have been extracted from the National Council of Educational Research and Training (NCERT) textbooks for 10th standard science, 11th standard, and 12th standard physics, chemistry and biology. NCERT is an autonomous organisation established by the Government of India that provides guidance and recommendations to both central and state governments on policies and initiatives aimed at improving the quality of school education. The textbooks have been downloaded from `https://www.ncrtsolutions.in/`. We adhere to NCERT guidelines, which state that *"NCERT books can also be downloaded free of cost from our website for non-commercial purposes."*.

References

Dishank Aggarwal, Pritam Sil, Bhaskaran Raman, and Pushpak Bhattacharyya. 2025. "i understand why i got this grade": Automatic short answer grading (asag) with feedback. In *Artificial Intelligence in Education*, pages 304–318, Cham. Springer Nature Switzerland.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Galina Deeva, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt. 2021. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162:104094.

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, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *Preprint*, arXiv:2409.17146.

Menna Fateen, Bo Wang, and Tsunenori Mine. 2024. Beyond scores: A modular rag-based system for automatic short answer scoring with feedback. *Preprint*, arXiv:2409.20042.

[8] https://opencv.org/

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. 2022. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8577–8591, Dublin, Ireland. Association for Computational Linguistics.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *Preprint*, arXiv:1803.01937.

Buse Giledereli, Yifan Hou, Yilei Tu, and Mrinmaya Sachan. 2024. Do vision-language models really understand visual language? *ArXiv*, abs/2410.00193.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Gyeong-Geon Lee and Xiaoming Zhai. 2023. Nerif: Gpt-4v for automatic scoring of drawn models. *Preprint*, arXiv:2311.12990.

Chee Wee Leong, Lei Liu, Rutuja Ubale, and Lei Chen. 2018. Toward large-scale automated scoring of scientific visual models. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, L@S '18, New York, NY, USA. Association for Computing Machinery.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6007–6026, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Elizabeth J. Marsh and Emmaline Drew Eliseev. 2019. *Correcting Student Errors and Misconceptions*, page 437–459. Cambridge Handbooks in Psychology. Cambridge University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Charith Peris, Lizhen Tan, Thomas Gueudre, Turan Gojayev, Pan Wei, and Gokmen Oz. 2022. Knowledge distillation transfer sets and their impact on downstream NLU tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 128–137, Abu Dhabi, UAE. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ari Sagherian, Suhasini Lingaiah, Mohamed Abouelenien, Chee Wee Leong, Lei Liu, Mengxuan Zhao, Blake Lafuente, Shu-Kang Chen, and Yi Qi. 2022. Learning progression-based automated scoring of visual models. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 213–222, New York, NY, USA. Association for Computing Machinery.

Ashita Saxena. 2024. Hallucination detection in machine generated text. Master's thesis, Indian Institute of Technology Bombay.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman, and Alexandra Farazouli. 2023. ExASAG: Explainable framework for automatic short answer grading. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 361–371, Toronto, Canada. Association for Computational Linguistics.

Mengxia Yu, Zhihan Zhang, Wenhao Yu, and Meng Jiang. 2023. Pre-training language models for comparative reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12421–12433, Singapore. Association for Computational Linguistics.

Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2024. Diagrammergpt: Generating open-domain, open-platform diagrams via llm planning. *Preprint*, arXiv:2310.12128.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Prompt used for synthetically generating correct responses

As mentioned in Section **??**, the prompt used to generate the correct responses is as follows. Note we place the question within the `<QUESTION>` and the `</QUESTION>` tags. Similarly, we place the reference answer within the `<ANSWER>` `</ANSWER>` tags.

Prompt 2: Prompt for generating correct responses

Prompt: You are a student who is attempting an examination.

Task: Given a question and its original answer, rewrite the original answer as a paragraph in a different tone so that it is correct and captures all the necessary facts in the original answer.

<QUESTION></QUESTION>

<ANSWER></ANSWER>

Output Requirements:

<HANSWER>The correct

answer</HANSWER>

B Prompt used for synthetically generating partially correct responses

As mentioned in Section **??**, the prompt used to generate the partially correct responses by introducing factual fabrication is as follows. Note we place the question within the `<QUESTION>` and the `</QUESTION>` tags. Similarly, we place the reference answer within the `<ANSWER>` `</ANSWER>` tags. It is similar to the one used by Ashita et al in (Saxena, 2024).

Prompt 3: Prompt for generating partially correct responses (Factual Fabrication)

Factual fabrications in text can be created by introducing contextually relevant information or facts that are not verifiable by established real-world evidence. This technique involves adding elements that fit logically within the context but are factually incorrect or unverifiable.

Task: Given a question and its corresponding original factual answer, you are to create a factual fabrication in the answer by introducing a new, contextually relevant information or fact that cannot be verified against real-world knowledge. You should rewrite the answer (the hallucinated answer) by adding the fabricated information in the original answer as well as the information/-fact you introduced in the original answer to generate the hallucinated answer.

<QUESTION></QUESTION>

<ANSWER></ANSWER>

Output Requirements:

<HANSWER>The hallucinated answer</HANSWER>

<RDETAILS>The additional information/-fact introduced to generate the hallucinated answer in a single sentence .<RDETAILS>

The prompt used to generate the partially correct responses by introducing factual inconsistency is as follows. Note we place the question within the `<QUESTION>` and the `</QUESTION>` tags. Similarly, we place the reference answer within the `<ANSWER>` `</ANSWER>` tags. It is similar to the one used by Saxena (2024).

Prompt 4: Prompt for generating partially correct responses (Factual Inconsistency)

Prompt: In the field of dependency parsing, modifiers are defined as words or phrases that provide additional information about other elements in a sentence. One technique to generate deliberate factual inconsistencies in text, termed the "Termite Strategy" targets these modifiers. This strategy involves replacing modifiers with alternative words or phrases that are factually inconsistent yet still maintain the overall coherence of the sentence.

Task:

Given a question and its original answer, apply the Termite Strategy to introduce a factual inconsistency in the original answer. Replace a modifier in the original answer with an alternative that contradicts the factual information in the answer, but still retains sentence coherence. You must rewrite the "complete" answer with the modifications (the "hallucinated answer") and also provide the replacement details.

<QUESTION></QUESTION>

<ANSWER></ANSWER>

Output Requirements:

<HANSWER>The hallucinated answer</HANSWER>

<RDETAILS>Describe the original modifier and the replacement word or phrase used to create the inconsistency in a single sentence<RDETAILS>

C Prompt used for synthetically generating incorrect responses

As mentioned in Section **??**, the prompt used to generate the incorrect responses by introducing factual fabrication is as follows. Note we place the question within the `<QUESTION>` and the `</QUESTION>` tags. Similarly, we place the reference answer within the `<ANSWER>` `</ANSWER>` tags. It is similar to the one used by used by Saxena (2024).

Prompt 5: Prompt for generating incorrect responses (Factual Fabrication)

Prompt: Factual Fabrication refers to instances where the LLM's output contains facts that are unverifiable against established real-world knowledge.

Task:

Given a question, the task is to generate an incorrect answer using the techniques of factual fabrication.

<QUESTION></QUESTION>

You should follow the following output format:

Output Requirements:

<HANSWER>The hallucinated answer</HANSWER>

<RDETAILS>Mention why the answer is incorrect in a single sentence<RDETAILS>

The prompt used to generate the incorrect responses by introducing factual inconsistency is as follows. Note we place the question within the `<QUESTION>` and the `</QUESTION>` tags. Similarly, we place the reference answer within the `<ANSWER>` `</ANSWER>` tags. It is similar to the one used by used by Saxena (2024).

Prompt 6: Prompt for generating incorrect responses (Factual Inconsistency)

Prompt: Factual Inconsistency refers to situations where the LLM's output contains facts that can be grounded in real-world information, but present contradictions.

Task:

Given a question , the task is to generate an incorrect answer using the techniques of factual inconsistency .

<QUESTION></QUESTION>

You should follow the following output format:

Output Requirements:

<HANSWER>The hallucinated answer</HANSWER>

<RDETAILS>Mention why the answer is incorrect in a single sentence<RDETAILS>

D Correctness Matrix

The correctness mentioned in Section **??** has been explained here. Given the correct/partially correct/incorrect textual answer and images, the final set of student answers is generated using the correctness matrix as below.

Student Response Text	**Student Response Image**	**Overall Correctness Label**
C	C	C
C	PC/I	PC
PC	C/PC	PC
PC	I	I
I	C	PC
I	PC/I	I

C - Correct **PC** - Partially Correct **I** - Incorrect

Table 3: Matrix for determining the Level of Correctness

E Prompt to Generate Feedback

As mentioned in Section 6.2, the prompt used to generate the feedback given the question, reference answer, and the student answer is as follows. Note we place the question within the `<QUESTION>` and the `</QUESTION>` tags. Similarly, we place the reference answer within the `<ANSWER>` `</ANSWER>`

tags and the student answer within the <STUDENT> </STUDENT> tags. Additionally, we place the reference answer image and the student answer image one after the other, just after the prompt and pass it to the corresponding LLM API.

Prompt 7: Feedback Prompt

Act as a teacher and grade the student answer given the question, reference answer and the student answer.

Task : You have to generate the level of correctness, the image relevance and the feedback. The feedback should point out any errors in the text as well as the image. It should also provide the reason for the level of correctness and image relevance. It can contain some additional information and facts to complement the student's understanding as well. It should be a conversation between you as a teacher and a student. It should be of 500 words.

Input format -

<QUESTION>The question</QUESTION>

<ANSWER>The reference answer. Note that the first image corresponds to the image in the reference answer</ANSWER>

<STUDENT>The student answer. Note that the second image corresponds to the image in the reference answer</STUDENT>

You have to strictly follow this output format-

<CORRECTNESS>Predict whether the answer is Correct, Partially Correct or Incorrect. Note you should evaluate both the text and image of the student answer as a whole. </CORRECTNESS>

<RELEVANCE>Predict whether the second image is relevant or irrelevant to the question</RELEVANCE>

<REASON>The feedback</REASON>

Here is the input :

<QUESTION></QUESTION>

<ANSWER></ANSWER>

<STUDENT></STUDENT>

F Subject Matter Expert Details

The expert evaluation of LLMs involved the help of six Subject Matter Experts (SMEs) (as per Section 6.2), where two SMEs were present each for biology, physics, and chemistry.

Biology: SME 1 has been a high school educator with teaching experience for over 25 years. SME 2 is a research scholar with more than 2 years of experience. SME 3 is a researcher with more than 5 years of teaching experience.

Physics: SME 1 is currently a professor at a reputed institute with over 13 years of teaching experience. SME 2 is a research scholar with more than 2 years of experience. SME 3 is a research scholar with over 2 years of teaching experience.

Chemistry: Both our SMEs are research scholars with more than 3 and 4 years of teaching experience, respectively. SME 3 is an educator with over 20 years of teaching experience.

The LLM-generated feedback had an average of 1758 words and a maximum of 3538 words. Each reviewer had to go through 880 instances of LLM-generated feedback and rate them based on the evaluation parameters as mentioned in Section 6.2. As pointed out by SMEs, this was a very tiring and time-consuming task as they had to meticulously go through a single feedback multiple times for each parameter. Hence, each of them has been compensated appropriately as per the norms for their evaluations and comments.

G Additional Information for Level of Correctness Labels

This appendix analyses the confusion matrix obtained for all the LLMs over the generated level of correctness labels. The major observations have already been mentioned in Section 6.1.

Figure 3 displays the confusion matrix obtained for Gemini over the Level of Correctness labels. While it has correctly predicted the "Correct" labels, it has often labelled "Partially Correct" answers as "Incorrect" and vice versa. However, the generated trend indicates that the model was being lenient in labelling most "Incorrect" answers as "Partially Correct". However, for some answers, it failed to

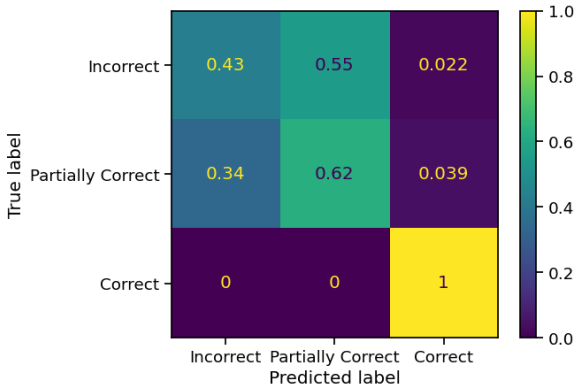


Figure 3: Confusion Matrix for Gemini after True Class Normalization

evaluate the answer properly and marked "Partially Correct" answers as "Incorrect".

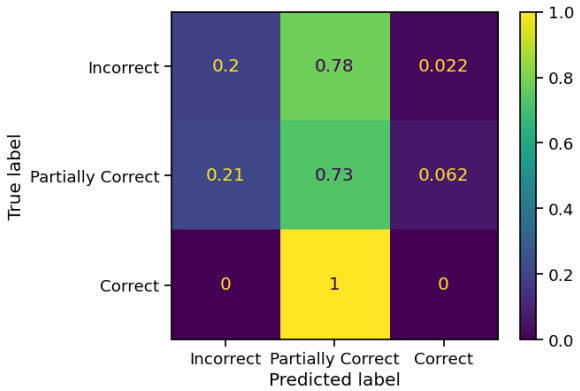


Figure 4: Confusion Matrix for ChatGPT after True Class Normalization

Figure 4 displays the confusion matrix obtained for ChatGPT over the Level of Correctness labels. It can be seen that ChatGPT was lenient in grading the student's answer as it labelled most "Incorrect" answers as "Partially Correct" while it made a mistake of labelling "Correct" answers as "Partially Correct" as well.

Figure 5 displays the confusion matrix obtained for Pixtral over the Level of Correctness labels. Similar to ChatGPT, Pixtral was lenient and labelled "Incorrect" answers as "Partially Correct" which also led to the reduced metrics.

Figure 5 displays the confusion matrix obtained for Molmo over the Level of Correctness labels. As it can be seen, it labelled almost all answers as "Incorrect" indicating that it was unable to show considerable performance while performing complex comparative reasoning over multiple modalities.

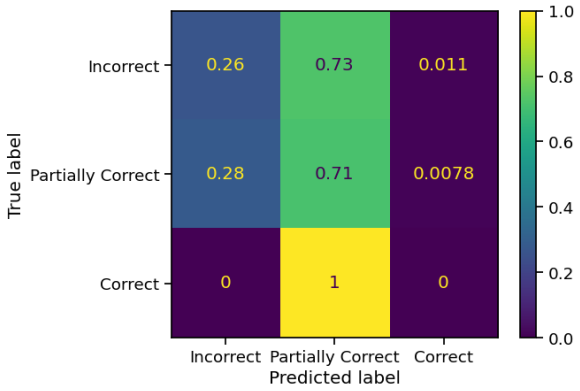


Figure 5: Confusion Matrix for Pixtral after True Class Normalization

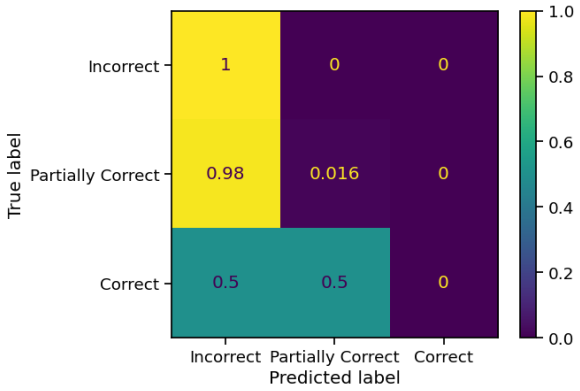


Figure 6: Confusion Matrix for Molmo after True Class Normalization

H Additional Information for Image Relevance Labels

This appendix analyses the confusion matrix obtained for all the LLMs over the generated image relevance labels. The major observations have already been mentioned in Section 6.1.

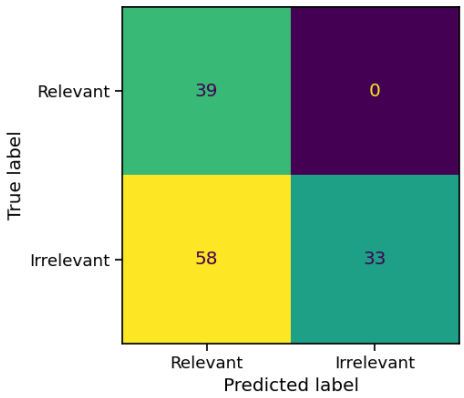


Figure 7: Confusion Matrix for Gemini

Figure 7 displays the confusion matrix obtained for Gemini over the Image Relevance labels. Gemini often predicted "Irrelevant" images as "Relevant", which led to reduced metrics.

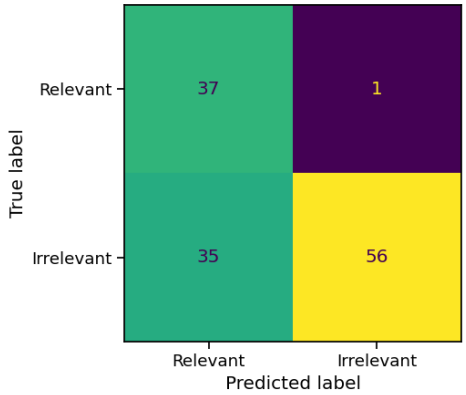


Figure 8: Confusion Matrix for ChatGPT

Figure 8 displays the confusion matrix obtained for ChatGPT over the Image Relevance labels. ChatGPT had the highest accuracy among all the LLMs. However, it still failed to classify a small part of "Relevant" images and "Irrelevant". This is also consistent with the fact that ChatGPT has superior multimodal reasoning capabilities compared to Gemini, Pixtral and Molmo, and this fact has also been verified by different metrics on standard benchmarks such as MMMU.

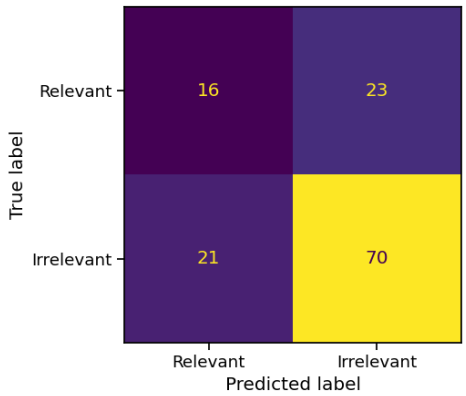


Figure 9: Confusion Matrix for Pixtral

Figure 9 displays the confusion matrix obtained for Pixtral over the Image Relevance labels. When it comes to Pixtral, it has often incorrectly classified "Relevant" images as "Irrelevant" and vice versa.

Figure 10 displays the confusion matrix obtained for Molmo over the Image Relevance labels. Molmo has classified all images as "Relevant", barring one. This indicates that there is a scope to improve Molmo's multimodal reasoning capabilities.

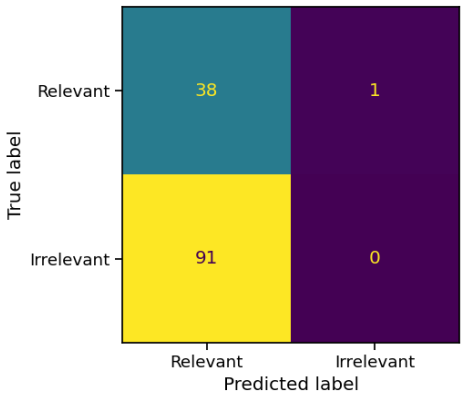


Figure 10: Confusion Matrix for Molmo

I Analysis of Rubrics for error detection

This is an addition to Section 6.3. The synthetically generated student answers contain both text and image parts, which might contain errors. Thus, the generated rubrics can either be used to evaluate errors present in the textual part or in the image part. We first analysed how well the LLMs performed in detecting the textual errors, followed by their performance in the case of image errors. Note that this data was again annotated by the SMEs as stated in section 6.2. The error detection accuracy (Ac_{ED}) for textual or image errors is computed as

$$Ac_{ED} = \frac{\text{Number of rubrics detected}}{\text{Total number of rubrics}}$$

Note that number of rubrics detected is equivalent to number of True values obtained for rubrics divided by total number of rubrics. Also, we use majority voting to resolve any disagreements.

Figure 11 contains the Ac_{ED} pertaining to the text errors of each LLM on each subject and on all subjects as a whole, while Figure 12 contains the same but for image errors of each LLM on each subject and on all subjects as a whole. From both Figure 11 and Figure 12, it can be concluded that ChatGPT performed the best across all subjects with scores of 0.73 and 0.9, respectively. However, if we analyse across individual subjects, ChatGPT was able to beat all except for Gemini in biology for text error-based rubrics and was comparable to other models in biology for image-based rubrics.

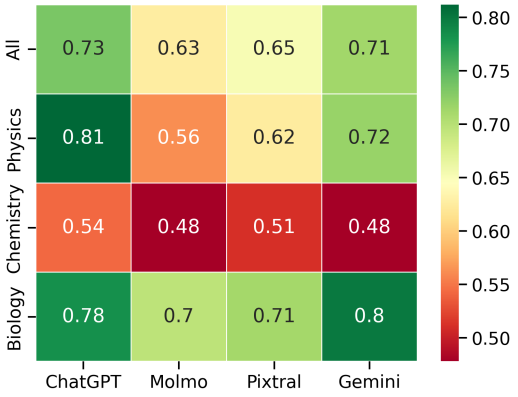


Figure 11: Ac_{ED} values for Text based Rubrics

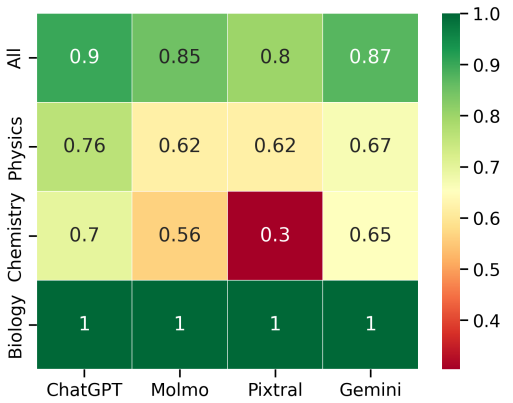


Figure 12: Ac_{ED} values for Image based Rubrics

J Extended Results

This is an addition to Section 6.3. In this appendix, we calculate the difference in performance that arises due to one prompt working better for a given model than others. For the level of correctness, Gemini displayed a 6% improvement over ChatGPT, a 23.64% improvement over Molmo and a 5.43% improvement over Pixtral. For Image relevance labels, ChatGPT displayed a 20.74% improvement over Gemini, 56.14% over Molmo, and 12.2% over Pixtral.

Earlier in Section 6.3, we had shown that ChatGPT performs the best in physics and chemistry, while Pixtral performs the best in biology. Table 4 denotes the percentage improvement of the best model over the other models in each subject.

Physics: It can be seen that ChatGPT significantly beats other models in all parameters except for Gemini in FGE, EI, and Pixtral in FGE. Since LC and EM play a major role in the generated feedback, we can conclude that ChatGPT is the optimal solution for generating feedback for physics.

Chemistry: When it comes to chemistry, ChatGPT again significantly beats other models in all parameter,s which is indicative of the fact that it is the best solution for generating feedback in Chemistry.

Biology: Here, Pixtral significantly beats other models in all parameters except for Gemini and ChatGPT in FGE. However, since LC and EM play a major role in the generated feedback, we can still conclude that Pixtral is the better solution for generating feedback in the case of biology.

K Dataset Split Statistics

As mentioned in Section **??**, the data has been split into the train,test and validation sets in the ratio of 3:1:1 and the statistics is as follows -

The mean and standard deviation of word lengths are as follows -

L Experimental Setup

As mentioned in Section 6, the system specifications used for LLM Inference are as follows -

- CPU: Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz with 112 CPUs
- RAM: 503 GB
- GPU: NVIDIA A100-SXM GPU with 80 GB of memory (Only 1 out of 4 used)

Subject	Model	Evaluation Parameters (% Improvement)				
		FGE	EI	LC	EM	Rubrics
Physics	Gemini	0.00	0.66	9.95	11.49	14.71
	Molmo	5.26	72.56	68.40	102.67	47.17
	Pixtral	0.00	10.60	6.84	7.55	41.82
Chemistry	Gemini	0.61	6.64	1.14	2.04	-3.08
	Molmo	5.96	53.23	50.85	51.34	31.25
	Pixtral	0.61	5.73	7.80	11.36	28.57
Biology	ChatGPT	0.00	26.06	4.94	10.48	16.00
	Gemini	0.00	7.80	-0.87	13.36	9.43
	Molmo	1.63	24.44	11.84	29.37	7.41

Table 4: Percentage improvement of best model over other models in each subject

	Train	Validation	Test
Correct	102	34	35
Partially Correct	692	232	232
Incorrect	522	174	174
Total	**1316**	**440**	**441**

Table 5: Dataset split statistics

	Mean	Standard Deviation
Question	118.61	84.20
Reference Answer	924.30	688.72
Student Answer	888.13	569.12

Table 6: Word statistics of dataset

The relevant models used from Huggingface, OpenAI, and Gemini APIs have already been mentioned in Section 5. All inference operations use default hyperparameters.

M Common errors made by students during exams

This is an elaboration on Section **??**. To facilitate better development of such synthetic datasets, we list down the common mistakes made by students during examinations. These examples have also been suggested by SMEs and categorised as per the categories listed by Marsh and Eliseev (2019). Thus broadly, such errors are categorised as follows -

- *Errors made with confidence:* These errors are commonly made by students where the student confidently provides an answer that is derived from a correct fact, but is totally incorrect given the context or a completely made-up fact. This again contains scenarios such as -
  - *Incorrect Translations:* These errors are made by students when they generate a wrong summarization of a fact. One example from Marsh et al's work includes that while the correct fact is "George Washington has dentures made of bone and other non-wood materials.", the student mistakenly mentions "George Washington has wooden teeth". This is a wrong summarisation of the earlier-mentioned fact.
  - *Factual Errors:* These errors are made by students when they use a correct fact but in the wrong context. An example provided by our SME is as follows: given the fill-in-the-blank type question "Haemoglobin consists of . . . and globin." the student incorrectly mentions calcium instead of haem. While calcium is a known fact and is present in the human body, it is not present in haemoglobin.
  - *High confidence misconceptions or Erroneous guess:* These errors are made by students when they try to fill in the gaps in their knowledge with generic concepts. For example, a student mentioned the Sahara consists of sand, while, in reality, it is a rocky desert. Another example mentioned by our SME is that when a student

is asked to label the parts of a leaf, they mistakenly label the petiole as the stem.

  - *Calculations errors:* Such errors are made by students in the case of numerical questions where they can guess the correct answer but are unable to produce it due to failure to recall certain formulas or another calculation mistake. These errors are common in questions where the students have to prove a theorem. Since the theorem is a well-known fact, students often make up the reasoning process to account for it.

- *Misunderstanding:* These errors are commonly made by students when they fail to understand a question or concept and answer the question incorrectly. Some such scenarios include cases -

  - *Misunderstanding a concept:* Such errors are seen in students' answers when they have a wrong understanding of a certain concept. One such example is when students are asked how echolocation works, they mention that bats shout at objects to judge their distance, and other mammals can do the same as well. However, the correct fact is that bats are able to echolocate because of specialised organs that might be absent in other mammals. As indicated by this example and also pointed out by our SME, the answer is wrong due to wrong reasoning, which is evident by the fact that the student's concept is not clear.
  - *Memory problem:* The student makes such mistakes when they are unable to recall facts relevant to a question due to a wrong conceptual clarity. One such example pointed out by our SME is that a student fails to provide the scientific name of Mango as they do not have concept clarity.
  - *Misunderstanding the question:* Such errors happen when the student fails to interpret a question properly. These errors are commonly seen in passage-based questions where the student is unable to comprehend the passage properly or the question relevant to it, and provides wrong facts from the passage as their answer.

- *Conceptual Change:* These errors happen when a student understands a concept but fails to use it in a different context or determine how it interacts with other concepts. One such example from Marsh et al.'s work is that a student answers that diffusion happens because the oxygen and carbon dioxide molecules move willingly. In reality, the correct fact is that diffusion occurs due to different concentration levels of these molecules and their existing Brownian motion. Another example pointed out by our SME is when the student is asked how cross-pollination occurs, and the student fails to answer because they are unable to recall how pollination takes place.

N Ablation Studies

This is an extension to Section 6.3. To study the variance in the feedback generated after running the same prompt 10 times on the same LLM and input, we do an empirical evaluation and show two such outputs below. The common observation is that the core content of the feedback, which detects inaccuracies or suggests ways to correct them in the student's answer, remains consistent. However, the presentation becomes different.

Feedback 1

... The question asked where to find lactic acid bacteria and some of their uses. Instead, you wrote a detailed description of the human nervous system, including the brain, eye, and ear. The image provided is a cross-section of the cochlea, part of the inner ear and also irrelevant to the topic ... Lactic acid bacteria are found in various fermented foods. The most common example is curd (yogurt). They are also found in other fermented foods like sauerkraut, kimchi, pickles, and some cheeses. Additionally, they naturally exist in our gastrointestinal tracts... Fermentation provides preservation, adds flavor, and creates beneficial nutritional changes...Lactic acid bacteria can improve gut health by promoting the growth of beneficial bacteria in the intestines. They aid in digestion, and some strains produce vitamins like vitamin K and B vitamins....The image you included is a

diagram of the cochlea, a part of the inner ear. This is entirely unrelated to the question about lactic acid bacteria.

Feedback 2

... The question asked about lactic acid bacteria and where they are found, including their applications. Instead, your response is a detailed description of the human nervous system, including the brain, eye, and ear. ... The image you provided is a diagram of the cochlea, a part of the inner ear. A picture of curd or a diagram illustrating fermentation would have been appropriate.... Lactic acid bacteria are found in various fermented foods, including yogurt (curd), kefir, sauerkraut, kimchi, pickles, and many other products. They are also found in some parts of the human gut ... Lactic acid bacteria are crucial in food preservation (fermentation), increasing the nutritional value of food (e.g., vitamin B12 in yogurt), and improving gut health (probiotics). They can also have industrial applications, such as producing lactic acid for various uses.

In the two feedbacks below, the blue regions comment on the question and response provided. Note that while the content is the same, it only varies in the natural language used. Similar to the regions marked using violet, the feedback corrects information pertaining to the question, serving the error mitigation task. However, as seen earlier, they vary in the natural language used.

O Gemini-2.5 Results

Subject	FGE	EI	CL	EM	Rubric
Physics	5.0	4.5	5.0	5.0	1.0
Chemistry	5.0	5.0	5.0	5.0	1.0
Biology	5.0	4.67	3.0	3.33	1.0

Table 7: Human evaluation for Gemini 2.5

Since most the earlier experiments were done using Gemini-1.5, we reevaluate on `gemini-2.5-flash` which is its recent version. The same methodology has been followed over 22 instances. Based on Table 7 and Table 6.3 we conclude that the numbers have significantly improved and has potential to surpass ChatGPT in Physics and Chemistry. However, the pattern remains the same for biology where Pixtral still dominates over Gemini due to similar reasons as mentioned in Section 6.2.