"Oh, so that's what went wrong!": Targeted Feedback through Concept and Category Aware Prompting in Educational Settings

Shyam Agarwal[1*] and Ali Moghimi[2]

[1]Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, 95616, CA, USA.
[2]Department of Biological and Agricultural Engineering, University of California, Davis, One Shields Avenue, Davis, 95616, CA, USA.

*Corresponding author(s). E-mail(s): shyagarwal@ucdavis.edu;
Contributing authors: amoghimi@ucdavis.edu;

Abstract

Providing timely and detailed feedback to student assessments is crucial for a holistic educational experience. However, reviewing assessments and providing feedback to assessment takers is a challenging problem because of the limited availability of instructor time, large class sizes, and other resource constraints. This problem is difficult since manual solutions are labor intensive, and automated solutions are difficult to generalize to scenarios where questions are open-ended. Unlike objective questions like multiple-choice questions or fill-in-the-blanks, free-response short-answer questions help better understand a learner's capability as they encourage generative processing and test a learner's understanding of core concepts. Thus, it is important to develop systems that can provide automated feedback on such open-ended questions. This paper proposes a novel and practical approach to providing feedback on short-answer responses to open-ended questions by utilizing our concurrent work to classify student responses as either correct, incorrect, or incomplete using near-domain data and then uses semantic clustering with human assistance to analyze the key concept of student responses. These pieces of information are then fed into a generative pre-trained model to generate targeted and personalized feedback for the students. Our automated grading method outperforms the SOTA machine learning models as well as some large language models in providing summative feedback by a huge margin of over 10-20%. Our feedback generation framework extends on this leverage, does not require the use of any pre-written grading rubrics, and is specifically designed keeping in mind the practical classroom settings. We believe

that this is the first work that undertakes a semantic clustering approach along with utilizing a large language model and grade scores based on near-domain data to provide automated feedback.

Keywords: Automated Feedback Generation, Short Answer Feedback, Prompt Engineering, Concept and Category Aware Prompting

1 Introduction and Related Works

Feedback plays a crucial role in a student's learning journey and overall growth. As defined by Ambrose and others, feedback is "information given to students about their performance that guides future behavior" [1]. As per Boud and Molly, feedback is "the process whereby learners obtain information about their work in order to appreciate the similarities and differences between the appropriate standards for any given work, and the qualities of the work itself, in order to generate improved work". In simple words, feedback is any information that can inform an individual about their relative performance and can guide them towards strategies to improve themselves.

Feedback is said to be formative when it aims to modify the thinking or behavior of the learner with the aim of improving their learning. These feedbacks are usually given as part of short learning cycles where a student is supposed to understand the feedback and either redo the assignment or build on top of it. Constructive feedback can not only have a positive impact on student's comprehension of concepts, but it also encourages the development of critical thinking abilities and fosters metacognitive skills [2]. Timely, specific, non-evaluative, supportive and meaningful feedback is very strongly related to increased learner motivation and improved academic performances [3]. Many researches support these impacts on improving knowledge and acquiring skills [4–9] or on acting as a motivating factor in the learning process [10, 11].

Given that feedback plays such a crucial role in a student's learning process, the ability to provide quality feedback is quite important. A good feedback must provide specific information to improve the gap between what is understood by a student and what is the ground truth [12]. In recent times, the focus has shifted on providing more personalized feedback that is tailored to the specific needs, styles, pace, and knowledge gap of the student. Generic feedback might be helpful in some cases, but it often fails to guide students appropriately because in most cases, it is no better than providing student with a score or the source text from a book. In contrast, personalization increases the applicability and relevance of feedback, making it more impactful for learner growth. For example, feedback targeted towards student's misconception reinforces self-directed learning [13] that can encourage self-regulation and self-confidence. Similarly, when feedback is combined with correctional review, the feedback and instruction are intertwined such that "the process itself takes on the forms of new instruction, rather than informing the student solely about correctness" [14].

The scope of this personalization depends on the type of question. Questions like multiple-choice, one word or fill in the blank provide little to no room to understand student's thought process and learning. In contrast, essay-style questions or constructed-response questions (CRQ) provide more room for both, the learners and the instructors to engage in deeper cognitive processes. They require students to articulate their understandings into their own words which offers valuable insights into their thought processes and misconceptions, especially because there is more room for creativity and critical thinking [15]. They also allow instructors more opportunity to understand how to change their teaching style, and make it more targeted towards their audience. Out of all CRQs, short-response ones are quite interesting. They do not constraint the student thinking with pre-defined answers, but at the same time, owing to their shorter length, they do not provide much opportunity to students to explain in detail, their understanding. At the same time, the possible correct answers are usually limited because of the closed-form nature of questions [16] as opposed to opinion-based long-form essays.

In practical scenarios, achieving the above mentioned quality feedback is quite difficult because it is a laborious and time-consuming task, especially when the instructor to student ratio is small and when the formative feedback cycles need to be shorter in length. This problem of being able to scale quality feedback to a wide audience is often referred as the "scaling feedback" problem. It is challenging because [17, 18]:

1. Student work displays significant diversity in most cases and is usually characterized by a distribution that has a long tail aka the majority of solutions are infrequent.
2. Supervised techniques are not very feasible because the process of labeling student work with detailed feedback is both arduous and costly.
3. Generating wrong feedback is not acceptable because it could mean wrong learning outcomes for the student.
4. The feedback produced must be easily understood, and also explainable to the student and the instructor.

Since the response evaluation is highly subjective, mainly depending on the evaluator's interpretation of student response, the end result can be inconsistent feedback. Moreover, students might also misinterpret the question on their end and provide response to something that was not asked in the question. The problem worsens in short-response questions because of shorter response length.

Different people have approached the problem of feedback generation from different perspectives. Some have used supervised learning methods like regression analysis and have achieved high accuracies [19], but this approach requires availability of huge amounts of data depending on the specific application and hence, cannot be used widely.

Others have used techniques where humans can intervene and design rubrics which can be used for automated grading by a rubric-focused learning algorithm [20–22]. There have been numerous works that rely on these grading rubrics in the form of a model answer, and measure the syntactic and semantic similarities between the two of them [23]. These works differ in the way they define similarity with some simply using word matching [24], some using knowledge and corpus based methods [25], some using

semantic vector embeddings [26], and some using semantic similarity and dependency graph alignments [27]. Some recent works have also tried targeted feedback generation at key point level. Zhao and others extract key points from a response and match them to a model answer to identify the missing key points that are used to generate feedback based on a template [15]. However, depending on the domain, it might not be possible to come up with a model answer or there might be a lot of possible model answers.

Another set of researchers employ the use of probabilistic programs to write generative descriptions of student cognition and synthesize millions of labeled examples of student responses which are then used to infer feedback on real student solutions [18]. However, it is neither easy nor generalizable to model student cognition using probabilistic programs for all domains of questions. Wu and colleagues also propose a human-in-the-loop rubric sampling approach that is zero-shot in nature [17]. However, the results would depend widely on the complexity of the task and their method also suffers from the need to involve instructors to mimic student thinking which is not very feasible.

The personalization aspect has also been studied well, especially in many works by Linn and her students. They investigate how automated guidance can support short answer responses from middle school students. They have compared automated feedback alone and in combination with information about personalized nature of feedback [28]. They have also compared feedback alone and in combination with students providing feedback on sample essay [29] and in combination with a revision process modeling interface [30]. In all these cases, they used C-rater-ML tool from ETS [31] which depends on model/reference answer or grading rubrics and provides mainly templated feedback.

More recently, people have employed large language models (LLMs) to generate these feedbacks. Aggarwal and others generate Engineering Short Answer Feedback (EngSAF) dataset where they generate synthetic feedback using LLMs like Gemini and ChatGPT [32]. Scarlatos and others use GPT-4 to annotate human and LLM-generated feedback along with proposing a reinforcement-learning based framework to optimize the correctness and alignment of the feedback [33]. Meyer and others compared the effectiveness of LLM-generated feedback to no feedback and highlighted that LLM-generated feedback has a positive correlation to students' cognitive and affective-motivational outcomes [34].

In this work, we try to utilize large language models (LLMs) along with the traditional non-supervised technique of semantic clustering in an attempt to produce targeted feedback on a dataset of short-response questions about the biological concept of central dogma. We build on top of the approach of grading with near-domain data that Agarwal and Moghimi present using the same dataset[35]. Our key contributions are as follows:

1. We present a novel framework to produce automated feedback for short-response questions.
2. We compare the performance of the results from this technique with standard baseline prompts.

3. We present a new technique of prompting, Concept and Category Aware Prompting (C2AP).

2 Methods

In this section, we will explain the dataset that we are using, the terminology involved as well as the different steps involved in our Targeted Feedback through Concept and Category Aware Prompting (TF-C2AP) framework. Overall, we identify the category (correct, incomplete, or incorrect), the concept that each response relates to, and feed that into a large language model like GPT-4 using Concept and Category Aware Prompting (C2AP).

Table 1 Distribution of the Collected Data for Replication, Transcription, and Translation

Category	Label	Training	Validation	Testing
Replication	Correct	785	219	554
	Incomplete	240	64	169
	Incorrect	405	146	279
Transcription	Correct	759	217	554
	Incomplete	220	79	146
	Incorrect	451	133	302
Translation	Correct	940	272	665
	Incomplete	310	96	203
	Incorrect	180	61	134

2.1 Dataset

Information flow is considered as a fundamental concept in biology [36, 37], but research suggests that it is usually tough for students to grasp [38]. Some potential reasons that have been identified are: difficulty in understanding the storage and exchange of information [39, 40], challenges with understanding how DNA is encoded into protein through RNA intermediate like misinterpreting that the amino acids used in the translation process are created during translation [41], and confusion between DNA, gene, and protein [40, 42, 43]. Smith and others find out further issues by studying students' interpretation on the impact on transcription and translation [44, 45]. Several researchers have also tried to understand why this is such a difficult topic to understand, and many owe this to lack of usage of insightful visual diagrams in classes [46] or to multiple complex steps involved in each step. [38]

The dataset used for this work comes from GCA which is a concept inventory about information flow [45] involving nine learning goals through 25 multiple choice questions. Two of these questions evaluate the understanding about comparing different mutation types and describing the effects on on genes, corresponding mRNAs and proteins. Prevost and others modified this into a constructed short-answer question [38] that was later used by Agarwal and Moghimi [35]. The dataset used consisted of

three open-ended prompts asking students to explain how a mutation that resulted in a premature stop codon affected replication, transcription, and translation. A total of 1043 responses were collected from undergraduate biology majors from two large public research universities. These students were taught the relevant central dogma concepts in their introductory cell and molecular biology-focused courses. The questions were admininstered as a part of their online homework assignment, and students were rewarded one point for completion so they were asked to give their best shot. Later, more responses were collected, bringing the total to 2861. The collected data represented multiple institutions, class populations and time periods. This was then filtered to remove some nonsensical responses (like "IDK") and the filtered data was marked as either correct, incomplete, or incorrect by human experts who had an intergrader agreement of 80%. In cases of disagreement, the majority vote was taken into consideration. The data distribution is shown in 1.

This is the question: *The following DNA sequence occurs near the middle of the coding region of a gene. DNA 5' A A T G A A T G G* G A G C C T G A A G G A 3' There is a G to A base change at the position marked with an asterisk. Consequently, a codon normally encoding an amino acid becomes a stop codon. How will this alteration influence DNA replication? How will this alteration influence DNA transcription? How will this alteration influence DNA translation?*

Table 2 Examples of student responses and rubric coding, with each response shown given by a different student.

Question	<sample_correct>	<sample_incomplete>	<sample_incorrect>
Replication	It will not have any effect on the replication process.	This would be an example of a nonsense mutation.	The DNA will stop replicating when it reaches the stop codon.
Transcription	It will not have any effect on transcription.	This will cause a mutation in the transcription process.	In the process of transcribing DNA into RNA, the newly added stop codon will inhibit the rest of the chain from being transcribed into RNA.
Translation	This will influence translation because the stop codon will cause the amino acid sequence to end before it should. This will create a different polypeptide or protein that will either not function or function differently than it should have.	The code will be translated with a different base and will be read differently. This will result in a different protein being built.	This will have no influence on translation.

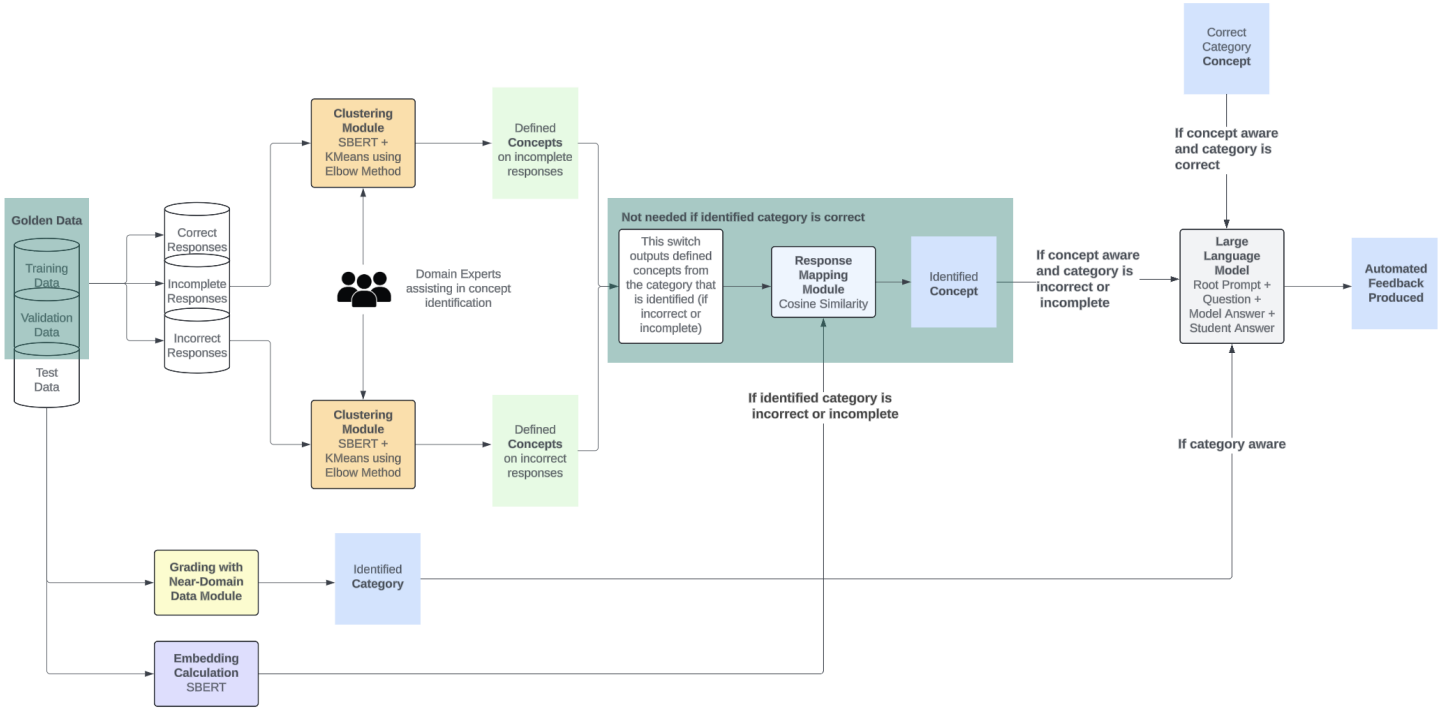

Fig. 1 Approach for Targeted Feedback Through Concept and Category-Aware Prompting and its Variantions

2.2 Terminology and Approach

Our dataset consists of responses from students to three different questions. Let's denote each student as $S_1, S_2, \ldots, S_n$. Q_1 refers to the effect on replication, Q_2 refers to the effect on transcription, and Q_3 refers to the effect on translation. For a student S_k, the responses are referred as $S_{k|1}$, $S_{k|2}$, $S_{k|3}$. We reserve responses from 50% students for training (T_1', T_2', T_3' for Q_1, Q_2, and Q_3 respectively), responses from 15% students for validation (V_1, V_2, V_3 for Q_1, Q_2, and Q_3 respectively), and the responses from the remaining 35% students for testing (T_1, T_2, T_3 for Q_1, Q_2, and Q_3 respectively).

This is done for the category identification using grading with near-domain data, and it is in alignment with what Agarwal and Moghimi [35] do, using the exact same partitions from their work. For the concept identification and other parts, we do not need a separate validation and training dataset, so we combine them into a golden dataset which consists of the remaining 50% responses that were not used for testing (hereby, referred as G_1, G_2, G_3 for Q_1, Q_2, and Q_3 respectively). We further create three subsets of G_i: $G_{i|correct}$, $G_{i|incomplete}$, and $G_{i|incorrect}$ which consists of the responses in G_i that were marked correct, incomplete and incorrect respectively in the golden dataset (i = 0, 1, 2 for Q_1, Q_2, and Q_3 respectively). In more mathematical terms, we get that without loss of generality, given that we have S_1, S_2, …, S_n students in some random order and let $p = 0.5n$ and $q = 0.15n$, then for o = 1, 2, 3 for some Q_o, we have

$$T_o' = \{S_{1|o}, S_{2|o}, \ldots, S_{p|o}\} \tag{1}$$
$$V_o = \{S_{p+1|o}, S_{p+2|o}, \ldots, S_{p+q|o}\} \tag{2}$$
$$T_o = \{S_{p+q+1|o}, S_{p+q+2|o}, \ldots, S_{n|o}\} \tag{3}$$

which gives us,

$$T_1' = \{S_{1|1}, S_{2|1}, \ldots, S_{p|1}\} \tag{4}$$
$$T_2' = \{S_{1|2}, S_{2|2}, \ldots, S_{p|2}\} \tag{5}$$
$$T_3' = \{S_{1|3}, S_{2|3}, \ldots, S_{p|3}\} \tag{6}$$
$$V_1 = \{S_{p+1|1}, S_{p+2|1}, \ldots, S_{p+q|1}\} \tag{7}$$
$$V_2 = \{S_{p+1|2}, S_{p+2|2}, \ldots, S_{p+q|2}\} \tag{8}$$
$$V_3 = \{S_{p+1|3}, S_{p+2|3}, \ldots, S_{p+q|3}\} \tag{9}$$
$$T_1 = \{S_{p+q+1|1}, S_{p+q+2|1}, \ldots, S_{n|1}\} \tag{10}$$
$$T_2 = \{S_{p+q+1|2}, S_{p+q+2|2}, \ldots, S_{n|2}\} \tag{11}$$
$$T_3 = \{S_{p+q+1|3}, S_{p+q+2|3}, \ldots, S_{n|3}\} \tag{12}$$
$$\tag{13}$$

For o as above, we also have

$$G_o = T_o' \cup V_o \tag{14}$$

which gives us,

$$G_1 = \{S_{1|1}, S_{p+q+2|1}, \ldots, S_{p+q|1}\} \tag{15}$$
$$G_2 = \{S_{1|2}, S_{p+q+2|2}, \ldots, S_{p+q|2}\} \tag{16}$$
$$G_3 = \{S_{1|3}, S_{p+q+2|3}, \ldots, S_{p+q|3}\} \tag{17}$$

Please note that,

$$G_{o|correct} = \{x \in G_o \mid \text{x's category is labeled correct}\} \tag{18}$$
$$G_{o|incomplete} = \{x \in G_o \mid \text{x's category is labeled incomplete}\} \tag{19}$$
$$G_{o|incorrect} = \{x \in G_o \mid \text{x's category is labeled incorrect}\} \tag{20}$$

It is important to mention that technically, we do not need to reserve all responses (Q_1, Q_2, Q_3) from a student in either T_i, T_i', V_i or G_i (i as above) for rest of the framework. This is to say that each of the questions could have been considered separately (for instance, for a student S_m, $S_{m|1}$ could have been in T_1, $S_{m|2}$ could have been in V_2, and $S_{m|3}$ could have been in T_3', which means G_1 would have $S_{m|1}$, G_2 would have $S_{m|2}$, and G_3 would not have $S_{m|3}$) since the framework is separate for each of the three processes (replication, transcription, and translation). However, given that we use the grading with near-domain data approach in this framework which utilizes data from similar questions to grade a particular question, the authors decided to keep the sets separately to ensure that a student's response to one question is not used to calculate their response to another question as practical academic settings would abide by it. Thus, we decide to keep all responses from a student in either of the sets for the rest of the framework too, and instead of T_p and V_p, we use G_p (for p = 1, 2, 3 for Q_1, Q_2, and Q_3 respectively). In a more generalized setting where grades are calculated without depending on near-domain data, the rest of the framework can still be used with any response being in either of T_p or G_p (p as above).

Overall, there are 5 steps in this process.

1. We first use semantic clustering to group the incorrect and incomplete training and validation data separately. This is to say that we cluster $G_1|incorrect$, $G_1|incomplete$, $G_2|incorrect$, $G_2|incomplete$, $G_3|incorrect$, and $G_3|incomplete$ separately.
2. We ask human experts to look at some responses in each cluster to define common themes/concepts.
3. Then, using Agarwal and Moghimi's grading with near-domain data approach [35], we produce category-labels (correct, incomplete, or incorrect) for test responses.
4. Based on the category, we use cosine similarity to map the test responses to the human-expert defined concept from clustering.
5. The extracted concept and category are then fed into GPT-4o along with other pieces of information to create the finalized feedback.

Also, we find the need for clustering correct responses as unnecessary because to generate feedback, the reference concept from the model answer can be used instead. The approach is also graphically represented in figure 1.

2.3 Semantic Clustering

In this section, we share some common challenges in clustering short-response texts as well as how we cluster the responses as a part of the framework.

2.3.1 Challenges with clustering short-response texts

Clustering is the process of grouping together entities into multiple groups (one group is trivial) such that entities in one group are similar to each other in some aspect, and farther from entities in the other groups in that aspect. In our case, the entities are the student responses (vectorized into embeddings as explained later) and this aspect is defined as the semantic similarity between the responses because we want responses with similar themes/misunderstandings/gap to be clustered together.

Owing to the short length of short-answer text responses, clustering them is a more challenging task, especially because traditional similarity measures (like cosine similarity) rely heavily on co-occurrence of terms between the responses. This means that even if responses are topically related, but they do not have similar words, they are more likely not going to be clustered together. This problem is more relevant in scenarios where vocabulary size is huge and text length is short, like in tweets [47]. We realize that this is not an issue with our dataset, because owing to the narrow focus of the question topic (Central dogma), the vocabulary size is limited and thus, many keywords are repeatedly used by students as shown in SOME FIGURE. Many other works have also shown promising results with clustering short-texts using similarity measures like cosine similarity [48–50]

Some other challenges with clustering short texts include lack of context [51] since it is hard to interpret the meaning of certain words/phrases without proper context. In that case, responses might have similar words but different meanings. We realize that this is not an issue because the responses given by learners share the same context as the question asked to them. Also, although students come from different backgrounds, class populations, universities and time periods, since they all were exposed to similar class content, would have solved similar problems, and would have read similar material, they actually share a lot more context than we would expect, which is also highlighted in another work on clustering student responses [52]. Thus, we do not think this to be a major problem with our data.

2.3.2 Clustering the G_i datasets

We first convert the train text responses into numerical vector embeddings using Sentence-BERT (SBERT) model [53]. We choose SBERT over other conventional embeddings (like Word2Vec [54], GloVe [55], BERT [56] etc.) because the latter focus on generating embeddings based on individual words. They produce token-level embeddings which are then aggregated using sum, average or [CLS] token embedding. In contrast, SBERT generates fixed-size sentence embeddings by fine-tuning BERT on sentence-pair tasks. Thus, it is able to capture the rich semantic relationships at a sentence-level which is quite significant for short text responses. In particular, we used the *paraphrase-MiniLM-L6-v2* model from Hugging Face's Sentence Transformer library because it is a standard model that is pre-trained for sentence similarity tasks.

This model is inspired from Microsoft's MiniLM work [57]. It is trained using a masked language modeling objective on large text corpra, followed by a siamese network setup with cosine similarity as the loss function to fine-tune it on paraphrase identification task (identifying whether two texts convey the same meaning even if they have different words). It produces dense 384-dimensional vector representations.

Then, we use KMeans clustering algorithm to partition the embeddings into a predefined number of clusters (k). KMeans works by iteratively assigning data points to the nearest cluster centroid followed by recalculating the centroids based on each cluster at the given step. In particular, we used Lloyd's algorithm which uses euclidean distance to find the nearest cluster and calculates the centroids for each cluster using mean. We used Scikit-learn tool to perform this clustering. The assignment and update steps are shown below.

Assignment Step:

$$c_i = \arg \min_{j \in \{1,2,\ldots,k\}} \|\mathbf{x}_i - \mu_j\|^2$$

Update Step:

$$\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

The value of k was determined using the Elbow method which is used to find the ideal number of clusters such that the compactness of clusters as well as the separation between different clusters is balaced. It involves plotting the Within-Cluster Sum of Squares (WCSS) for different values of k and selecting the point where the WCSS curve flattens (like an "elbow"). The WCSS curves and the chosen number of clusters are present in figure 2.

We cluster the following sets separately: $G_1|incorrect$, $G_1|incomplete$, $G_2|incorrect$, $G_2|incomplete$, $G_3|incorrect$, and $G_3|incomplete$. There are two reasons for this. Firstly, clustering different questions separately ensures the generality and makes sure that responses from a particular question do not impact the clustering of other responses. Secondly, our understanding is that since incomplete responses were hard for human experts to identify as either correct or incorrect, there presence during the clustering process could bias the clustering of responses that were clearly identified as incorrect.

We do not cluster the correct responses because it is an unnecessary step. When the result is correct, we can always use the reference concept (from model answer) to produce the feedback instead, utilizing the power of the grading technique. This is especially true because the grading depends on content instead of the writing quality, and in most cases, the possible correct answers are limited by the closed-form of the question [16] as shared earlier, although there can be several reasons for it being incorrect.

2.4 Human Defined Concepts

The clustered responses were then studied by a human expert manually who identified overarching themes in the responses. The identified themes were ran through two

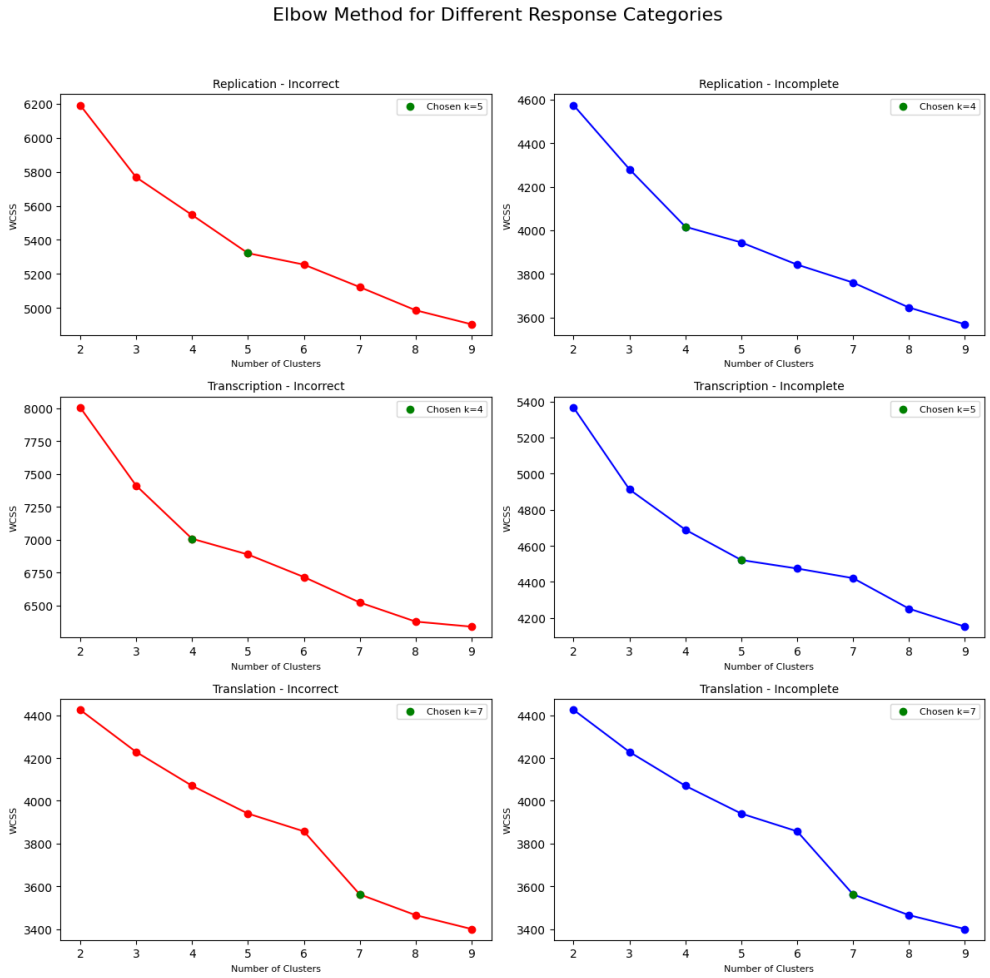

Fig. 2 The Within-Cluster Sum of Squares (WCSS) is plotted against the number of clusters to visualize the diminishing gains in clustering quality as k increases. Green markers indicate the selected k values for each category: 5 and 4 for replication, 4 and 5 for transcription, and 7 and 7 for translation.

more subject experts later. All the experts were undergraduates majors in biology and related fields (namely, Genetics, Biotechnology, and Biomedical Engineering) and were in their senior year at University of California, Davis. The identified concepts and their detailed description can be found in tables 3 for incorrect replication, 4 for incorrect transcription, 5 and 6 for incorrect translation, 7 for incomplete replication, 8 for incomplete transcription and 9 and 10 for incomplete translation.

Cluster	Explanation
0	The answer is incorrect because a premature stop codon doesn't halt or generate a partial strand of DNA. It remains unaffected by the stop codon and replicates the complementary strand irrespective of the mutation. However, if the mutation is related to a single or double-stranded break in the DNA, then it might generate an incomplete replicate.
1	The answers are incorrect because the DNA strand is unaffected by the stop codon. The process of replication won't pause due to the mutation because the stop codon doesn't have a function in replication that affects the DNA. Thus, no pause or shortening of DNA strands will occur. The stop codon will be replicated as usual as in other replications.
2	The answer is incorrect because the process of DNA replication won't stop and hence won't be affected due to stop codon introduction, it will still carry out the replication process normally. Thus, the process of replication will still occur irrespective of the premature stop codon.
3	The answers are incorrect because DNA replication is just focused on producing a complementary strand irrespective of the presence of the STOP codon. The role of the stop codon comes into play during the process of translation, where the process of translation will stop upon encountering a premature stop codon, and generate a shorter stretch of amino acids. Thus, the process of translation terminates early due to the stop codon and doesn't impact DNA replication.
4	The answers are incorrect because the addition of a premature stop codon does not affect the DNA strand or the RNA transcript. A premature stop codon only affects an amino acid during translation. A shorter amino acid will be produced due to the premature stop codon.

Table 3 Cluster Explanations for Incorrect Replication Responses

Cluster	Explanation
0	The answer is incorrect because, despite a mutation that adds a stop codon, the process of transcription of DNA to RNA will still occur without any premature stop in the process. The stop codon doesn't affect the process of transcription. It just stops prematurely only while translating and produces a shorter amino-acid sequence.
1	The case that a stop codon prevents proper transcription of DNA to RNA is wrong because, like replication, the process of transcription is also unaffected by the presence of a stop codon. An RNA will be transcribed from DNA without any pauses or stops and isn't affected by its presence. It only affects the level of translation because the codons are read by the tRNA from the mRNA at this point.
2	The answer that shorter mRNA strand is incorrect because the stop codon, or any other codon mutation, doesn't affect the mRNA sequence that is produced as a result of transcription. Transcription is simply converting the DNA sequence into an RNA format that can be processed into an amino acid. Thus, it's unaffected by the stop codon, and no reduction in the transcript length occurs.
3	The answer is incorrect because the stop codon doesn't have an effect on either of the processes of replication or transcription and hence won't affect the DNA or the RNA strand. Both the strands will be replicated and transcribed completely irrespective of the stop codon. Thus, the DNA will be fully replicated and will produce a fully transcribed mRNA irrespective of the stop codon.

Table 4 Cluster Explanations for Incorrect Transcription Responses

2.5 Category identification via Grading with Near-Domain Data

In Grading with near-domain data [35], Agarwal and Moghimi show that since in most practical academic situations, the availability of large amount of manually-labeled data is less, one can grade short-response questions by fine-tuning the model on data coming

Cluster	Explanation
0	The answer is wrong here because a premature stop codon serves the purpose of introducing an early stop while translating. Thus, a full amino acid sequence is not translated, and a premature stop codon causes a shorter peptide chain to be produced. Furthermore, the mutation is not a missense but a nonsense mutation. Had it been a missense mutation, a change in the amino acid would still have produced the same length of peptide and translated for a different amino acid. But a nonsense mutation is the one that introduces a premature stop. So, it's a nonsense mutation.
1	These answers are incorrect because overall the function of a stop codon is to bring a stop or end the process of translation when it's encountered by the ribosomes while translating. A premature stop ends the process of translation earlier than a normal translation process for that particular sequence of codons, and it's not supposed to affect the DNA or the RNA sequence during the process of replication or transcription respectively. Also, each stop codon has the same function, just like other amino acids that have multiple sets of codons coding for a particular amino acid, a stop codon codes for a stop and occurs in 3 sets of codons.
2	The process of translation will occur but it will just come to a stop earlier than usual for a particular sequence of codons or genes for which we are encoding a protein because of the introduction of a premature stop codon. Thus, it's incorrect to say that the process of translation won't occur. *Note: any mutation associated with the start codon in translation may prevent translation (depending on the mutation), but a premature stop doesn't prevent translation.
3	The answers stated by the students are incorrect because there is no frameshift mutation occurring here. The only mutation occurring is a nonsense mutation, in which a functional amino acid is converted to a stop codon. No shifting of codons is seen due to any indels (or any other mutation) that may contribute to a frameshift mutation. Furthermore, there is no missense mutation happening as well that changes the amino acid being coded for to another amino acid. The claim that a shorter mRNA transcript will be produced is false, the RNA sequence will just have a variation in the arrangement of the codons, where post the premature stop codon, there might be sequences that continue to code for the amino acid in a non-mutated situation. Thus, it's at the level of translation that the ribosome detects a stop codon and introduces a stop in translation, producing a smaller peptide chain, and leaving behind the rest of the transcript unread.

Table 5 Cluster Explanations for Incorrect Translation Responses - Part 1/2

from similar domain. For Q_2 and Q_3, they used very small number of responses from that particular question and still achieved similar or better accuracies by utilizing data from all previous questions. They also evaluate the performance of base BERT and large language models (fine-tuned and non-fine-tuned) on this task and share some other insights as well, but we mainly use the categories identified from their base model on Q_1, and previous question pre-trained model for Q_2 and Q_3 to identify the category for our responses in test datasets (T_1, T_2, and T_3).This is done to simulate the academic settings where neither much data would be available to train from scratch, nor much computational resources to fine-tune large language models. Anyway, the performances are all very similar across all three settings with data, time, and computational trade offs. For Q_1 though, there is no previous question pre-trained model so we have to use the fine-tuned base BERT model.

2.6 Concept identification via Response mapping

Cosine similarity is a popular metric used to measure similarity between two vectors in multidimensional space. It does so by calculating the angles between the vectors

Cluster	Explanation
4	The answers by the students are incorrect because the size of the RNA remains unaffected by the introduction of a premature stop codon. It will still copy the complementary sequences from the DNA until it receives a signal to stop transcription (which is different from the stop codon). It's also wrong that an individual will lack RNA post transcription due to the stop codon because it's independent of the stop codon sequence for translation. Thus, the individual won't lack an mRNA for a premature stop codon mutation. *Note: until the mutation of a stop codon is highly deleterious and during the process of translation, there might be an occurrence of a nonsense-mediated decay and the RNA is destroyed, but this occurs while translation where the ribosome encounters a premature stop codon and if it's highly deleterious the mRNA will be destroyed later. But overall RNA is produced post-transcription. However, yes, an RNA copy can be said to be incorrect because the transcript produced for the translation of amino acid is not normal and has a premature stop that affects the peptide synthesis but the RNA copy is still incorrect.
5	The answers are incorrect because the protein, if translated, will have an effect due to the introduction of a premature stop codon and the ribosome thus might miss out on reading important sequences essential for the functioning of the organism. Thus, there will be an effect on the protein, but it cannot be said that there is no effect of the mutation on the protein. Furthermore, the mRNA is fully transcribed from the DNA irrespective of the stop codon, which only functions during translation. Thus, it's wrong to say that it'll be short. Thus, the mRNA will retain its size and it will be processed due to transcription. *Note: The effect may vary depending on the location, and the importance of the codons that failed to be translated. Thus, it can be the case that there was little effect on the protein.
6	The answers stated by the students are incorrect because the process of translation will be affected by a premature stop codon. Translation will fail to produce a whole length of peptide by stopping earlier due to the PSC (premature stop codon). Therefore, translation is impacted.

Table 6 Cluster Explanations for Incorrect Translation Responses - Part 2/2

Cluster	Explanation
0	The answers provided by the student are incomplete because, although there will be a variation in the DNA sequence due to the mutation that eventually codes for a stop codon, there is no effect of that mutation on the process of replication. The process of replication will continue as is.
1	These answers stated by the students are incomplete because there is no reference to the effect of premature stop codon on the process of DNA replication. The process of replication is supposed to have no effect on DNA synthesis and will only affect the process of translation, where it produces a shorter protein or a non-functional gene product.
2	The answers here are incomplete because these answers are more focused on the fact of the effect of the introduction of a stop codon on the DNA sequence rather than on the process of replication. Thus, the answers are missing the fact that the process of replication will still occur normally, irrespective of a nonsense mutation that affects the sequence at the translational or protein level.
3	The claim that a mutation occurs and the DNA will be error-prone is correct but the students fail to suggest the effects of these changes on the process of replication and hence it's incomplete. The process of replication despite the mutation will still occur and produce another complementary strand of DNA.

Table 7 Cluster Explanations for Incomplete Replication Responses

to capture the orientation instead of the magnitude. For a test response, once we know the category, we calculate the vector embeddings using the same procedure as explained in 2.3.2 and find the closest cluster by calculating the cosine similarity with

Cluster	Explanation
0	The answer stated by the student that the process of transcription will be incorrect is correct but it doesn't suggest the effect of the mutation on transcription. Therefore it's incomplete. Thus, the process of transcription will remain unaffected and transcribe or copy the incorrect DNA sequence.
1	These answers are incomplete because, although the transcript has a variation introduced in the sequence due to the premature stop codon mutation, the answers don't state that the process of transcription will still occur producing an mRNA. Furthermore, some answers also suggest the process of early stoppage of the process of translation due to the stop codon, which is correct but fails to mention the effects of the mutation on the process of transcription, and hence the answers are incomplete.
2	These answers stated by the student are incomplete because the answers don't focus on the effect of mutation on the process of transcription. The answers provided by the students are rather focused on the variations produced in the DNA or the transcripts as a result of the mutation. Furthermore, students also focus a lot on the effects of the mutation on the end product of translation, the peptide chain which will be shorter and will be unable to function properly. Thus, all these claims made by the students are correct (majorly) but they are missing the effect of the mutation on the process of transcription.
3	The answer suggesting a different mRNA sequence being produced is true, however, the answers are incomplete because they lack the explanation that the process of transcription is unaffected by the mutation of the introduction of a premature stop codon, and hence they can transcribe the DNA sequence like the non-mutated sequence.
4	The answers stated by the students are incomplete because they tend to show the variation in the mRNA sequence due to the mutation and hence the introduction of the stop codon. However, they fail to explain the effects of the mutation on the process, which is the process of transcription remains unaffected and encodes for an mRNA which is processed for the translation of the amino acid.

Table 8 Cluster Explanations for Incomplete Transcription Responses

the centroid (mean of embeddings of all responses in a cluster). The human-identified concept for that particular cluster is then utilized.

2.7 Feedback Generation using C2AP and variations

We utilize the generative capabilities of the state-of-the-art large language models to produce feedback for a given response. The prompt consists of a root prompt (inspired from work by Aggarwal and others [32]) which is as follows:

You are an automatic short-answer feedback generator. Given the question, the student answer, the sample answer across all three categories, the evaluation of student response (correct, incomplete, or incorrect), and the concept that the student got wrong most likely, provide constructive feedback in about 3-4 lines. Ensure feedback guides the learner without invoking negative emotions and does not reference the provided reference answer.

Then, we provide the question, the student answer, the sample answer, the evaluation, as well as the human-identified concept in the rest of the prompt, where <substring>represents the answer provided by the student and <sample_correct>, <sample_incomplete>, and<sample_incorrect>refer to the sample correct, sample incomplete, and sample incorrect responses for each <question_type>(replication, transcription, translation). <response_category>is the identified categoy (correct, incomplete, or incorrect) and <response_concept>is the human-identified concept that

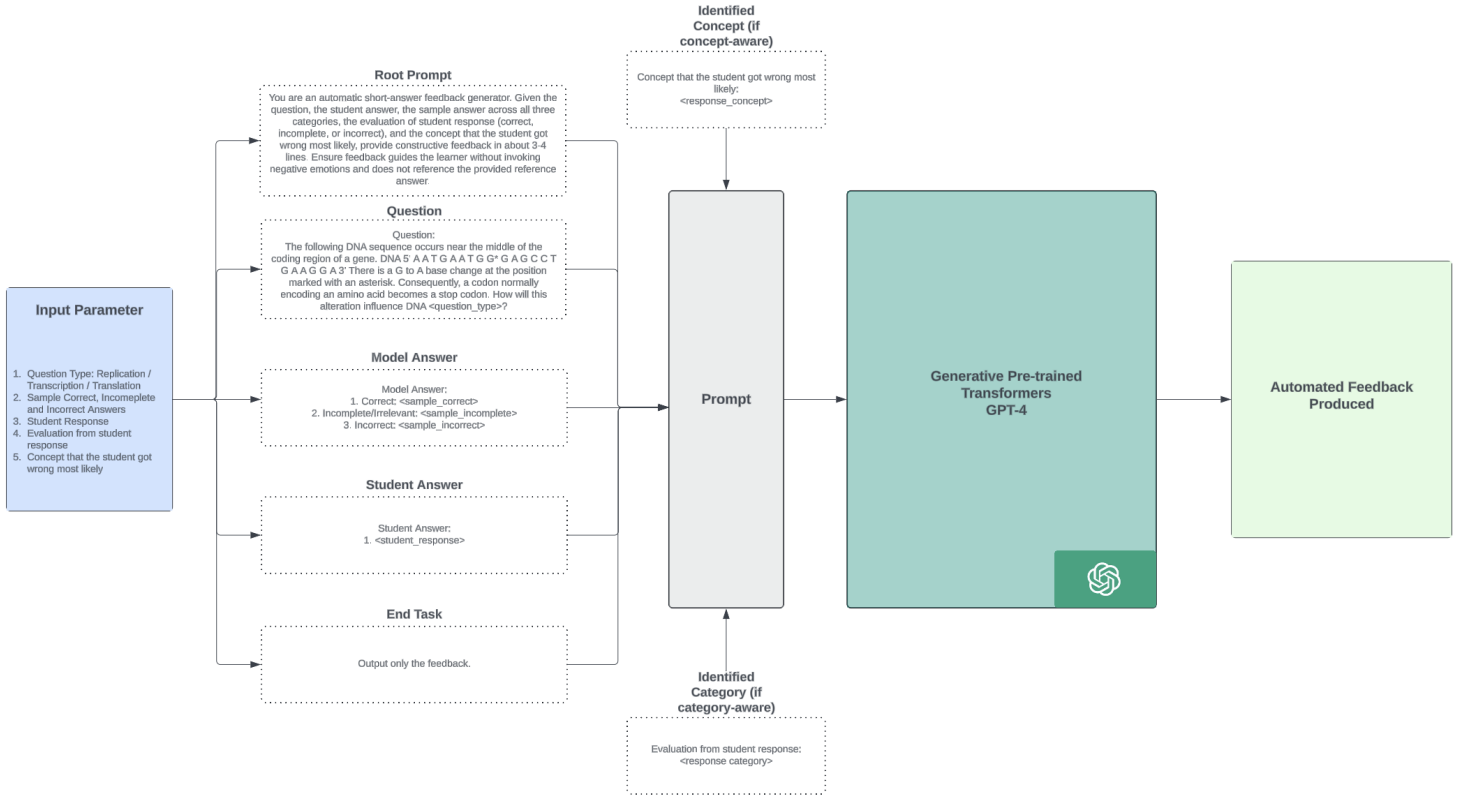

Fig. 3 GPT 4 Model Prompting

Cluster	Explanation
0	The answer is incomplete because it suggests the final result of the translation process. Initially, the ribosomes will be able to translate the mRNA to a peptide chain until it reaches a stop codon. And, in this case, the process of translation ends early since the ribosome encounters a premature stop codon. Thus, the polypeptide chain expressed is incorrect due to the mutation which is a result of the change in the sequence of nucleotides.
1	The answer that the stop codon prevents the process of translation from occurring is true but incomplete, because the stop codon prevents the process of translation from occurring beyond the point of a stop codon, and here since it's a premature stop codon, translation fails to occur beyond that. Furthermore, the answers concerning a mutation changing codon to different amino acid are incomplete because, the mutation that led to the substitution of a G to A (hence forming a stop codon), led to a change in the production of amino acid, changing from Trp to a stop codon. Thus, the answer is incorrect in the sense that the mutation does bring about a change in the coding of an amino acid, which is eventually a stop codon (premature), that stops the process of translation earlier.
2	The answer that translation won't occur, or it will be unable to translate is incomplete and is true only after the introduction of a stop codon, in this case, a premature stop codon. Thus, the process of translation won't occur after that point of the stop codon. *Note: Translation will occur but if the protein produced is deleterious, it might lead to degradation of the transcript and the peptide chain.
3	The answers are incomplete because they tend to focus on the replication process and the changes in the DNA that will produce a different final product, however, the missing part is that the process of translation will occur irrespective of a premature or a normal stop codon. Thus, a different or mutated genetic material will be translated due to the premature stop codon.

Table 9 Cluster Explanations for Incomplete Translation Responses - Part 1/2

Cluster	Explanation
4	The answers are incomplete because the answers focus on the RNA sequence (mutated or incorrect) that in consequence produces a shorter amino acid upon translation. The answers lack the fact that a fully transcribed RNA(irrespective of the location of the stop codon, but here it's a premature stop codon), which is incorrect due to the presence of a premature stop codon sequence, will lead to earlier stoppage of translation since the ribosome will encounter a UGA stop earlier than a normal sequence. Thus, it's incomplete in the fact that the answers don't state that these incomplete or incorrect RNA will carry a premature stop codon that will in turn affect the early stoppage of translation, producing an incomplete peptide sequence.
5	The answers are incomplete because the students are seen to draw a connection between the RNA or the DNA with the final product, the protein. This isn't incorrect but the answers (most of them) fail to mention the effect of the stop codon that will influence the process of translation, stopping it early, and producing a reduced peptide chain or protein. This may or may not be non-functional depending upon the amino acid sequences present in the chain.
6	The answer is incomplete because the students fail to state that there won't be any influence in the process of translation and that the process of translation will occur normally until it encounters a premature stop codon. Upon encountering a premature stop codon, the process of translation stops earlier.

Table 10 Cluster Explanations for Incomplete Translation Responses - Part 2/2

the response was previously mapped to. The rest of the prompt is as follows:
Question:
The following DNA sequence occurs near the middle of the coding region of a gene.
DNA 5' A A T G A A T G G G A G C C T G A A G G A 3'*
There is a G to A base change at the position marked with an asterisk. Consequently,

a codon normally encoding an amino acid becomes a stop codon.
How will this alteration influence DNA <question_type>?
Model Answer:
1. Correct: <sample_correct>
2. Incomplete/Irrelevant: <sample_incomplete>
3. Incorrect: <sample_incorrect>.
Student Answer:
1. <substring>
Evaluation from student response:
<response_category>
Concept that the student got wrong most likely:
<response_concept>
Output only the feedback.

We also experiment different settings, with and without the category, with and without the concept and their possible combinations to analyze if this has any impact on the results. In that case, we change the root prompt to only suggest the information that is actually present. The GPT model prompting cycle has been visually shown in 3. The sample responses are shown in 2.

2.8 Evaluation of Generated Feedback

To evaluate how credible and reliable the produced feedback is, we randomly sample 30% of the responses in each question type (replication, transcription, and translation) and distribute them equally across the three categories to create a representative sample of our feedback results. We ask three human experts to evaluate the produced feedback on a scale of 1-5 across the following metrics, as done by Aggarwal and others while creating EngSAF dataset [32]:

1. Grammatical Correctness: This aspect tests whether the feedback is grammatically correct and fluent in English. This ensures that the model generated responses sound like human responses to the learners. A higher score represents more fluent response.
2. Feedback Accuracy: This aspect tests the overall quality of the provided feedback across relevancy to the student's response, content produced and justification of the explanation provided. A higher score represents higher quality.
3. Emotional Impact: This aspect tests the impact of the feedback on student's emotional state. It ensures that the feedback produced does not trigger any negative emotions that might distress them or discourage them (like when the word "fail", "fool", or "dumb" is used). A higher score represents more emotionally aware feedback that encourages students to learn from their mistakes.

Researchers have used metrics like Rouge-2 [58], BLEU [59], METEOR [60], and BERTSCORE [61] as metrics to evaluate feedback in the past. Rouge-2 measures the overlap of bigrams, BLEU evaluates precision of n-grams (usually, upto 4), METEOR assesses the alignment using synonymy, stemming, and word order, and BERTScore uses BERT embeddings and calculates cosine similarity between generated feedback

and a reference feedback. Unfortunately, due to lack of human feedback responses, none of these metrics could be calculated.

Table 11 Metrics for feedback evaluation across all experimental settings

	Human evaluation		
Experimental Setting	GC	FA	EI
Human Baseline	3.6	4.3	4.4
SP	3.5	2.2	4.5
CAP	3.7	3.2	4.4
C2AP	3.5	4.7	4.6

3 Experiments, Results and Discussion

We use the following experimental settings so we can compare the feedbacks produced and understand what are some important factors in generating quality feedback.

- Standard Prompting (SP) - We neither include the identified concept nor the identified category in our prompt to the large language model.
- Category Aware Prompting (CAP) - We only include the identified category, but not the concept in our prompt to the large language model.
- Concept and Category Aware Prompting (C2AP) - We include both, the identified concept as well as the identified category in our prompt to the large-language model.

It is important to note that we also compare the performance with human baseline, which refers to providing the concepts identified by human graders to the students as the feedback. In the baseline experiment, we got a score of 3.6, 4.3, and 4.4 for grammatical correctness, feedback accuracy, and emotional impact respectively. With the SP setting, we got a score of 3.5, 2.2, and 4.5 for grammatical correctness, feedback accuracy, and emotional impact respectively. With the CAP setting, we got a score of 3.7, 3.2, and 4.4 for grammatical correctness, feedback accuracy, and emotional impact respectively. Finally, with the C2AP setting, we got a score of 3.5, 4.7, and 4.6 for grammatical correctness, feedback accuracy, and emotional impact respectively. The metrics are also shown in table 11.

This suggests that overall, grammatical correctness and emotional impact were similar across all the feedback categories. This suggests that the GPT-4o model was able to keep up with the instructions and ensure that feedback is encouraging for students. However, the main difference across all the four settings came in the feedback accuracy. The standard prompting performs the worst with a score of 2.2, which is reasonable because based on the grading results from Agarwal and Moghimi's work [35] on the same dataset, the non-fine tuned models performed far worse on category identification task itself. In such a case, it is reasonable for us to assume that the feedback calculated won't be accurate either. However, given the category of the response, the model seems to perform better with a score of 3.2 as expected. The human baseline from the concepts is still better with a score of 4.3. However, providing the category

and the concept improved the results considerably, even better than the human baseline with a score of 4.7/5. Our understanding is that this is reasonable because the human baseline has the exact same feedback for all the responses in a cluster, despite clusters having some outlying responses. It seems from the results that the large language model was able to understand the overarching idea and refine the feedback for even these outlying responses to provide a higher accuracy, which was the intent behind the approach.

4 Conclusion

In this paper, we presented a novel framework for automated feedback generation that is based on Grading with Near Domain data to identify categories, semantic clustering to identify concepts (with human assistance), and prompting large language models to generate the feedback. Based on our results, we conclude that providing both categories and concepts are useful for the generation of good quality feedback using LLMs.

Acknowledgements. We owe special thanks to our human-experts who spent their valuable time reading through the different responses, and evaluating the generated feedback. Special thanks to Maanvi Kaushal Shah, a Genetics and Genomics Senior at University of California, Davis for helping analyze the student responses and identifying overarching themes and concepts. This work wouldn't have been possible without their support and kind assistance.

References

[1] Nyquist, J.G., Jubran, R.: How learning works: Seven research-based principles for smart teaching. The Journal of Chiropractic Education **26**(2), 192–193 (2012) https://doi.org/10.7899/JCE-12-022

[2] Hattie, J., Timperley, H.: The power of feedback. Review of Educational Research **77**(1), 81–112 (2007)

[3] Shute, V.J.: Focus on formative feedback. Review of Educational Research **78**(1), 153–189 (2008) https://doi.org/10.3102/0034654307313795

[4] Azevedo, R., Bernard, R.M.: A meta-analysis of the effects of feedback in computer-based instruction. Journal of Educational Computing Research **13**(2), 111–127 (1995)

[5] Bangert-Drowns, R.L., Kulik, C.-L.C., Kulik, J.A., Morgan, M.T.: The instructional effect of feedback in test-like events. Review of Educational Research **61**(2), 213–238 (1991)

[6] Corbett, A.T., Anderson, J.R.: Feedback timing and student control in the lisp intelligent tutoring system. In: Bierman, D., Brueker, J., Sandberg, J. (eds.) Proceedings of the Fourth International Conference on Artificial Intelligence and Education, pp. 64–72. IOS, Springfield, VA (1989)

[7] Epstein, M.L., Lazarus, A.D., Calvano, T.B., Matthews, K.A., Hendel, R.A., Epstein, B.B., *et al.*: Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. The Psychological Record **52**, 187–201 (2002)

[8] Moreno, R.: Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. Instructional Science **32**, 99–113 (2004)

[9] Pridemore, D.R., Klein, J.D.: Control of practice and level of feedback in computer-based instruction. Contemporary Educational Psychology **20**, 444–450 (1995)

[10] Lepper, M.R., Chabay, R.W.: Intrinsic motivation and instruction: Conflicting views on the role of motivational processes in computer-based education. Educational Psychologist **20**(4), 217–230 (1985)

[11] Narciss, S., Huth, K.: How to design informative tutoring feedback for multimedia learning. In: Niegemann, H.M., Leutner, D., Brunken, R. (eds.) Instructional Design for Multimedia Learning, pp. 181–195. Waxmann, Munster, New York (2004)

[12] Sadler, D.R.: Formative assessment and the design of instructional systems. Instructional Science **18**, 119–144 (1989) https://doi.org/10.1007/BF00117714

[13] Nicol, D., Macfarlane-Dick, D.: Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education **31**, 199–218 (2006) https://doi.org/10.1080/03075070600572090

[14] Kulhavy, R.W.: Feedback in written instruction. Review of Educational Research **47**(2), 211–232 (1977) https://doi.org/10.3102/00346543047002211

[15] Zhao, J., Larson, K., Xu, W., Gattani, N., Thille, C.: Targeted feedback generation for constructed-response questions. (2021). https://api.semanticscholar.org/CorpusID:233295226

[16] Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education **25**(1), 60–117 (2015) https://doi.org/10.1007/s40593-014-0026-8

[17] Wu, M., Mossé, M., Goodman, N.D., Piech, C.: Zero shot learning for code education: Rubric sampling with deep learning inference. ArXiv **abs/1809.01357** (2018)

[18] Malik, A., Wu, M., Vasavada, V., Song, J., Coots, M., Mitchell, J., Goodman, N., Piech, C.: Generative Grading: Near Human-level Accuracy for Automated Feedback on Richly Structured Problems. https://arxiv.org/abs/1905.09916v5

[19] Sultan, M.A.-M., Salazar, C., Sumner, T.: Fast and easy short answer grading with high accuracy. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1070–1075 (2016)

[20] Basu, S., Jacobs, C., Vanderwende, L.: Powergrading: a clustering approach to amplify human effort for short answer grading. Transactions of the Association for Computational Linguistics **1**, 391–402 (2013)

[21] Süzen, N., Gorban, A.N., Levesley, J., Mirkes, E.M.: Automatic short answer grading and feedback using text mining methods. Procedia Computer Science **169**, 726–743 (2020)

[22] Lan, A.S., Vats, D., Waters, A.E., Baraniuk, R.G.: Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In: Proceedings of the Second (2015) ACM Conference on Learning@Scale, pp. 167–176 (2015)

[23] Leacock, C., Chodorow, M.: C-rater: Automated scoring of short-answer questions. Computers and the Humanities **37**(4), 389–405 (2003)

[24] Selvi, P., Bnerjee, A.K.: Automatic Short -Answer Grading System (ASAGS) (2010). https://arxiv.org/abs/1011.1742

[25] Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 567–575 (2009)

[26] Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhgatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: Use both. In: International Conference on Artificial Intelligence in Education, pp. 503–517 (2018). Springer

[27] Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 752–762 (2011)

[28] Tansomboon, C., Gerard, L.F., Vitale, J.M., Linn, M.C.: Designing automated guidance to promote productive revision of science explanations. International Journal of Artificial Intelligence in Education **17**, 729–757 (2017) https://doi.org/10.1007/s40593-017-0145-0

[29] Gerard, L., Linn, M.C., Madhok, J.: Examining the impacts of annotation and automated guidance on essay revision and science learning. In: Looi, C.-K., Polman, J.L., Cress, U., Reimann, P. (eds.) Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) (2016)

[30] Gerard, L., Linn, M.C.: Computer-based guidance to support students' revision of their science explanations. Computers & Education **176**, 104351 (2022) https://doi.org/10.1016/j.compedu.2021.104351

[31] Heilman, M., Madnani, N.: ETS: Domain adaptation and stacking for short answer scoring. In: Manandhar, S., Yuret, D. (eds.) Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pp. 275–279. Association for Computational Linguistics, Atlanta, GA (2013). https://aclanthology.org/S13-2046

[32] Aggarwal, D., Bhattacharyya, P., Raman, B.: "i understand why i got this grade": Automatic short answer grading with feedback. arXiv preprint **arXiv:2407.12818** (2024)

[33] Scarlatos, A., Smith, D., Woodhead, S., Lan, A.: Improving the validity of automatically generated feedback via reinforcement learning. arXiv preprint **arXiv:2403.01304** (2024). License: CC BY 4.0

[34] Meyer, J., Jansen, T., Schiller, R., Liebenow, L.W., Steinbach, M., Horbach, A., Fleckenstein, J.: Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence **6**, 100199 (2024) https://doi.org/10.1016/j.caeai.2023.100199

[35] Agarwal, S., Moghimi, A.: Grading with near-domain data: Towards human-level accuracy for automated feedback on short response questions. arXiv preprint (2024). Preprint

[36] American Association for the Advancement of Science: Vision and Change in Undergraduate Biology Education: A Call to Action. American Association for the Advancement of Science, Washington, DC (2011)

[37] NGSS Lead States: Next Generation Science Standards: For States, by States. National Academies Press, Washington, DC (2013)

[38] Prevost, L.B., Smith, M.K., Knight, J.K.: Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. CBE Life Sciences Education **15**(4), 65 (2016) https://doi.org/10.1187/cbe.15-12-0267

[39] Lewis, J., Leach, J., Wood-Robinson, C.: All in the genes? young people's understanding of the nature of genes. J Biol Educ **34**, 74–79 (2000)

[40] Mills Shaw, K.R., Van Horne, K., Zhang, H., Boughman, J.: Essay contest reveals misconceptions of high school students in genetics content. Genetics **178**, 1157–1168 (2008)

[41] Fisher, K.M.: A misconception in biology: amino acids and translation. J Res Sci Teach **22**, 53–62 (1985)

[42] Wood-Robinson, C.: Young people's understanding of the nature of genetic information in the cells of an organism. J Biol Educ **35**, 29–36 (2000)

[43] Marbach-Ad, G.: Attempting to break the code in student comprehension of genetic concepts. J Biol Educ **35**, 183–189 (2001)

[44] Smith, M.K., Knight, J.K.: Using the genetics concept assessment to document persistent conceptual difficulties in undergraduate genetics courses. Genetics **191**, 21–32 (2012)

[45] Smith, M.K., Wood, W.B., Knight, J.K.: The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ **7**, 422–430 (2008)

[46] Wright, L.K., Fisk, J.N., Newman, D.L.: Dna → rna: what do students think the arrow means? CBE Life Sci Educ **13**, 338–348 (2014)

[47] Pinto, D., Benedí, J.-M., Rosso, P.: Clustering narrow-domain short texts by using the kullback-leibler distance. In: Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, pp. 611–622. Springer, Berlin, Heidelberg (2007)

[48] Kang, J.H., Lerman, K., Plangprasopchok, A.: Analyzing microblogs with affinity propagation. In: Proceedings of the First Workshop on Social Media Analytics, pp. 67–70. ACM, New York, NY, USA (2010)

[49] Ni, X., Quan, X., Lu, Z., Wenyin, L., Hua, B.: Short text clustering by finding core terms. Knowledge and Information Systems **27**(3), 345–365 (2011)

[50] Yang, H., Mysore, A., Wallace, S.: A semi-supervised topic-driven approach for clustering textual answers to survey questions. In: Advanced Data Mining and Applications. Lecture Notes in Computer Science (LNCS), vol. 5678, pp. 374–385. Springer, ??? (2009)

[51] Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. In: Proceedings of the 29th European Conference on IR Research, pp. 16–27. Springer, Berlin, Heidelberg (2007)

[52] Hämäläinen, W., Joy, M., Berger, F., Huttunen, S.: Clustering students' open-ended questionnaire answers. arXiv preprint arXiv:1809.07306 (2018). Accessed: 2024-11-29

[53] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (2019)

[54] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013). https://arxiv.org/abs/1301.3781

[55] Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1162 . https://aclanthology.org/D14-1162

[56] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)

[57] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers (2020). https://arxiv.org/abs/2002.10957

[58] Post, M.: A call for clarity in reporting bleu scores. arXiv preprint arXiv:1804.08771 (2018)

[59] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (2002). https://api.semanticscholar.org/CorpusID:11080756

[60] Banerjee, S., Lavie, A.: 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, pp. 65–72 (2005)

[61] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)