Enhancing Peer Assessment through LLM-Guided Feedback Clustering: A Scalable Approach to Improve Coherence and Alignment

Mohamed Khalil Brik
Department of Computer Science
College of Charleston
Charleston, USA
brikm@g.cofc.edu

***Abstract*—This full research paper introduces a novel approach to improving the quality and coherence of formative feedback in peer assessment, a widely recognized pedagogical method that enables students to evaluate each other's work. Typically conducted in a double-blind format to ensure impartiality and independence, peer assessment often results in assessors identifying different issues when reviewing the same artifact, leading to discrepancies in feedback that can confuse reviewees and undermine the credibility of the process. While the double-blind structure promotes fairness, it prevents assessors from discussing and resolving these differences, and since the effectiveness of formative feedback relies on clarity and coherence, such inconsistencies can hinder students' ability to act on the feedback. Traditional solutions, such as instructor-led meta-reviews to identify and resolve disagreements, are labor-intensive and often delay the delivery of timely feedback. To address these challenges, this study proposes a novel method that systematically identifies all issues highlighted by multiple assessors and provides targeted prompts to help reviewers reconcile discrepancies, thereby encouraging more aligned and validated feedback. This approach reduces confusion for reviewees, enhances the reliability of the assessment process, and alleviates the burden on instructors, enabling more efficient and timely feedback delivery. The findings contribute to ongoing research on peer assessment by offering a scalable solution to improve the clarity, consistency, and impact of formative feedback in educational settings.**

***Index Terms*—Peer Assessment, Formative Feedback, GPT Clustering, Feedback Alignment, Educational Technology, Human-AI Collaboration**

I. Introduction

Formative peer assessment is a widely used pedagogical technique that encourages students to critically engage with each other's work by providing constructive feedback. However, challenges related to the coherence, consistency, and clarity of peer-generated feedback often limit its effectiveness. A recent study by Chen et al. [7] highlights that variations in the quality and focus of peer feedback can hinder students' ability to extract actionable insights. This research builds upon such findings by introducing an AI-assisted approach that uses Large Language Models (LLMs) like GPT to cluster feedback comments semantically, reconcile inconsistencies among peer assessments, and enhance overall feedback quality. The goal is to make formative assessment more scalable, interpretable, and pedagogically impactful.

II. Background

A. Large Language Models

Large Language Models (LLMs) are deep neural networks trained on vast amounts of textual data to perform a variety of natural language processing tasks, including translation, summarization, sentiment analysis, and question answering. These models, such as OpenAI's GPT and Google's PaLM, leverage transformer-based architectures and massive datasets to generate human-like text and reason across complex prompts. Their scalability and few-shot learning capabilities have made them central to recent advances in artificial intelligence [4].

B. BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model introduced by Google that achieves state-of-the-art performance on various NLP benchmarks. Unlike previous models, BERT reads text bidirectionally using attention mechanisms, allowing it to understand context more effectively. It is pre-trained on a large corpus and can be fine-tuned for downstream tasks like sentiment analysis or question answering. The attention mechanism is a key component, enabling BERT to dynamically weigh the importance of words in a sentence [5].

C. Formative Feedback

Formative feedback is a pedagogical approach that emphasizes continuous, low-stakes assessment to help learners identify areas for improvement during the learning process. Unlike summative assessment, which focuses on final outcomes, formative feedback supports learning by offering timely and actionable insights. It plays a critical role in peer assessment systems, where clarity and consistency of feedback are essential to student learning outcomes [6].

III. Proposed Method

A. Data Overview

The dataset comprises qualitative feedback collected from students as part of a peer review process on software assignments. Each student provided comments in response to a structured set of review questions, resulting in a dataset of over 1,000 individual comments. Each data point includes the following attributes:

- **question_id**: Identifier for the review question.
- **txt**: The review question presented to the student.
- **reviewer_id**: The student reviewer providing the feedback.
- **reviewee_id**: The student whose work is being reviewed.
- **comment**: The free-form textual response provided by the reviewer.

Initially, the dataset existed solely as raw comments, without any clustering or categorization. For instance, for the question *"Is the admin able to login using their credentials?"*, several responses were collected, such as:

```
"Yes, admin can login"
"Yes. The admin can login with the
    credentials specified in the README
     document"
"Login works great."
"Nothing is deployed."
"Reservations by admin not visible..."
```

These responses reflect varied themes—some confirm functionality, others report issues, and a few are irrelevant or unclear.

B. Clustering Methodology

To structure this data, we first grouped the comments by `question_id`, collecting all responses related to the same question. This reflects a real-world use case where multiple reviewers evaluate the same student's work on a specific functionality, and we aim to identify common feedback patterns.

Next, human annotators manually clustered these grouped comments into semantically coherent categories. Each cluster was labeled with a short title that captured its central theme. An example of such a human-generated clustering is shown below:

Following this, we applied GPT-based clustering using a few-shot prompting strategy, where GPT grouped the same set of comments into clusters and assigned thematic titles. Human annotators then evaluated the GPT-generated clusters using a three-point persuasion scale:

- **0** → No alignment: Human prefers their original clusters.
- **1** → Persuaded: Human adopts the GPT clustering instead.
- **2** → Exact match: Human clustering already matched GPT's.

This qualitative evaluation was complemented by quantitative metrics to better assess the alignment between human and GPT clustering. Specifically, we computed the **pairwise F1-score**, where all pairs of comments within the same cluster were compared between GPT and human outputs. This method captures how often both systems agree on grouping the same comments together.

```
{
  "Admin Login Works": [
    "Yes, admin can login",
    "Login works great.",
    "Yes. The admin can login with the
        credentials specified
     in the README document."
  ],
  "Admin Login Issues or Errors": [
    "Nothing is deployed.",
    "Reservations by admin not visible"
  ]
}
```

Fig. 1. GPT Sample Output

Other potential metrics discussed and considered include the **Jaccard Index**, the **Adjusted Rand Index (ARI)**, and **Normalized Mutual Information (NMI)**, each offering complementary views of clustering similarity.

Together, these qualitative and quantitative techniques allow for a robust evaluation of GPT's ability to generate human-aligned clusters of formative feedback, shedding light on its utility in educational peer assessment contexts.

C. Keywords Extraction Methodology

To better understand how language models identify important elements in feedback, we compare keyword extraction results from two approaches: SHAP-based interpretation of GPT sentiment output and KeyBERT, a transformer-based keyword extraction model. This comparison highlights both alignment and divergence in how each model ranks the contribution of tokens within user-generated comments.

IV. Experiment and Results

A. Persuasion Score Distribution

The figure below shows the distribution of persuasion scores across the dataset.

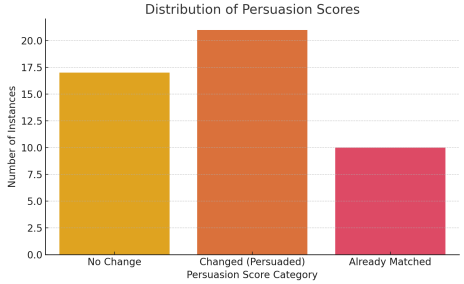

Fig. 3. Distribution of Persuasion Scores

Clustering Prompt for GPT

You are an advanced language model skilled in clustering and labeling textual data. Your task is to analyze a list of user comments, group them into meaningful clusters based on their semantic similarity, and provide a concise yet descriptive title for each cluster. Your output should be a dictionary where the keys are the titles of the clusters, and the values are arrays of strings, each containing the comments that belong to that cluster.

Example Input:

```
"User is able to delete reservations."
"Yes, users can delete reservations."
"YEs flight deletion is implemented"
"Yes"
```

Expected Output:

```
{
    "Reservation Deletion Confirmation":
         [
        "User is able to delete
            reservations.",
        "Yes, users can delete
            reservations."
    ],
    "Flight Deletion Confirmation": [
        "YEs flight deletion is
            implemented"
    ],
    "General Confirmation": [
        "Yes"
    ]
}
```

Fig. 2. Prompt used to guide GPT for clustering and labeling user comments.

Metric	Value
Total Comments	50
No Change (0)	17
Changed (1)	21
Already Matched (2)	10
Percentage No Change	10.06%
Percentage Changed	12.43%
Percentage Already Matched	5.92%

TABLE I
SUMMARY OF PERSUASION SCORE STATISTICS

It is important to note that a persuasion score of 0 does not necessarily imply that the human and GPT clusterings are entirely divergent. Rather, it simply indicates that the human did not adopt the exact same clustering as GPT after exposure. In many cases, the two clusterings may still be highly similar or overlapping in structure, even if not identical. To capture these nuances, more fine-grained similarity metrics—such as the Pairwise F1-Score or Adjusted Rand Index—can be employed to quantify the degree of alignment between clustering schemes beyond strict identity. These measures provide a richer understanding of how closely aligned the clusterings are in practice, even when the persuasion score does not reflect an exact match.

B. Cluster Agreement Metrics

To more accurately assess the similarity between human and GPT clusterings, we employ clustering agreement metrics that go beyond exact match comparisons. These metrics evaluate the structural alignment between clusterings and are especially useful when two sets of clusters are similar but not identical. We focus on three widely recognized metrics:

1) Pairwise F1-Score: The Pairwise F1-Score treats clustering as a pairwise classification task: for every pair of comments, it determines whether they are in the same cluster in both clusterings [1].

- Let TP be the number of comment pairs clustered together in both clusterings.
- Let FP be the number of pairs clustered together in the predicted clustering but not in the reference.
- Let FN be the number of pairs clustered together in the reference clustering but not in the predicted one.

Then:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

2) Adjusted Rand Index (ARI): The Adjusted Rand Index (ARI) measures similarity by counting pairwise agreements between clusterings, while adjusting for the possibility of agreement by chance [2].

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where:

- n is the total number of elements,
- n_{ij} is the number of items in both cluster i of the first clustering and cluster j of the second,
- $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$.

3) Normalized Mutual Information (NMI): Normalized Mutual Information (NMI) measures the amount of shared information between the predicted and reference clusterings [3].

$$\text{NMI}(U, V) = \frac{2 \cdot I(U; V)}{H(U) + H(V)}$$

where:

- $I(U; V)$ is the mutual information between clusterings U and V,
- $H(U)$ and $H(V)$ are the entropies of U and V respectively.

The NMI score ranges from 0 (no mutual information) to 1 (perfect correlation).

Figure 4 summarizes the performance of the clustering alignment between GPT-generated and human-adjusted clusterings using three evaluation metrics: Normalized Mutual Information (NMI), F1-score, and Adjusted Rand Index (ARI).

The results indicate a high degree of similarity between the model and human feedback structures, with an NMI of 0.816 and an F1-score of 0.806. These suggest that the clusters are not only semantically similar but also maintain strong overlap in pairwise grouping. The ARI score of 0.680, while lower, still reflects meaningful structural alignment beyond random chance.

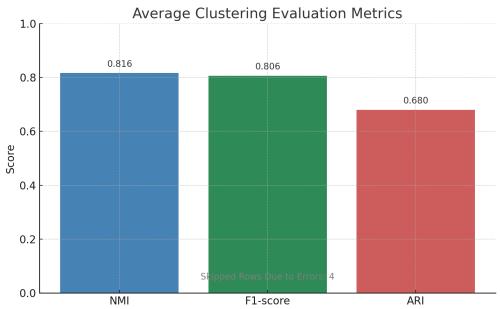

Fig. 4. Average clustering agreement scores between GPT and human clusterings.

C. Keyword Extraction Results

To explore how important terms are identified in student feedback, we analyze keyword extraction using the KeyBERT model. KeyBERT leverages BERT embeddings to rank the most relevant tokens in a sentence based on their contextual similarity to the entire text. Figures 5 and 6 illustrate two different visualizations derived from KeyBERT's scoring of token relevance within a single feedback sentence.

Figure 5 shows a bar chart where words like `not`, `working`, and `functionality` appear as negatively associated, while `Create`, `properly`, and `feedback` are highlighted as strong contributors. This view emphasizes the directional importance of individual words based on their semantic and positional role.

Figure 6 complements this with a sentence-level visualization that contextualizes each word's influence through color-coded arrows, helping illustrate the flow of semantic weight across the sentence.

While KeyBERT provides strong domain-relevant keywords through contextual similarity, GPT-based SHAP explanations offer a more sentiment-aware breakdown of tokens. GPT's interpretability methods allow not only for relevance identification but also for understanding whether a word contributes positively or negatively to the model's output. This distinction is valuable in clustering applications, especially when prioritizing emotional tone versus semantic content. The GPT output for the same sentence is shown in Figure **??**

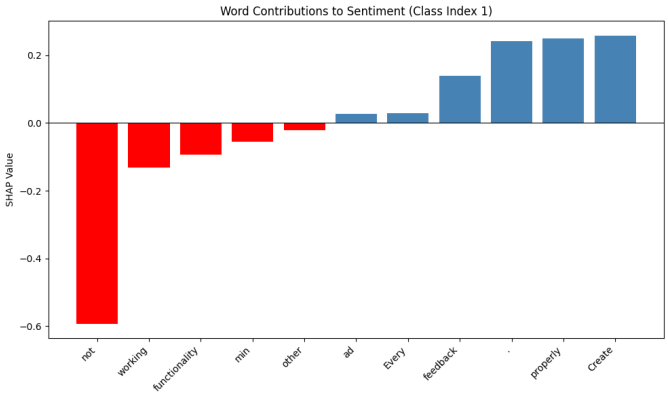

Fig. 5. KeyBERT-style bar plot highlighting the token contributions within a feedback comment.

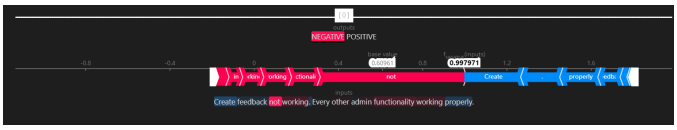

Fig. 6. Token-level contribution visualization from KeyBERT based on contextual embeddings.

REFERENCES

[1] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, 2000.

[2] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[3] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[4] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[6] V. J. Shute, "Focus on Formative Feedback," *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008.

[7] G. Chen, R. F. Kizilcec, J. J. Williams, K. McKeown, and M. Naaman, "PeerStudio: Supporting Peer Feedback in Online Learning at Scale," in *Proceedings of the Seventh ACM Conference on Learning @ Scale*, 2020, pp. 51–60.

```
Prompt:

Extract the key terms from the following
    sentence. For each key term, provide a
    brief explanation of why it was selected
    and how it contributes to the meaning of
    the sentence.

Sentence: "[Insert your sentence here]"

Format your output as follows:

Key Terms and Explanations:
1. [Term 1] - [Explanation of its importance]
2. [Term 2] - [Explanation of its importance]
...
```

Fig. 7. Prompt used to guide GPT-based key term extraction and explanation.

```
Key Terms and Explanations:
1. Create - Indicates the intended action or
    feature being referenced. It sets the
    focus of the first clause, identifying the
     capability that is expected to be
    functional.
2. feedback - Refers to the specific component
     or feature that is the subject of the
    issue. It gives context to what "Create"
    is meant to act upon.
3. not - A strong negative indicator
    highlighting that the "Create feedback"
    feature is not functioning, central to
    understanding the issue.
4. working - Used twice first negatively ("not
     working") and then positively ("working
    properly").  I t s  essential to
    distinguish between what is broken and
    what is functional.
5. admin functionality - A broader category
    that situates "Create feedback" within a
    set of administrative tools, reinforcing
    the context of what is functioning
    correctly.
6. properly - Emphasizes the contrast between
    the malfunctioning "Create feedback" and
    the rest of the system, supporting the
    notion that the issue is isolated.
```

Fig. 8. Key term extraction and explanation based on a GPT-style analysis prompt.