

Final Report – 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. It allows students to critique each other’s work and provide constructive feedback. It is a two side learning exercise. The reviewee learns from their mistakes through the comments of the reviewer. The reviewer learns from other students’ work and develops better communication skills through writing feedback, which can help them later in their careers.

However, challenges related to the coherence, consistency, and clarity of peer-generated feedback often limit its effectiveness in real-life classrooms. Students can provide different, sometimes contradicting, feedback to their peers. As a result,

the student becomes confused about interpreting the reviewers’ comments, which decreases the entire activity’s efficiency.

A recent study by Chen et al. [6] 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) to cluster feedback comments semantically, reconcile inconsistencies among peer assessments, and enhance overall feedback quality. The goal is to make the formative assessment more scalable, interpretable, and pedagogically impactful.

Building on the clustering methodology explored in this study, a promising future direction involves the development of a recommender system designed specifically for peer reviewers. The objective of this system is to support reviewers in generating more consistent, complete, and pedagogically valuable feedback by recommending related feedback comments authored by other reviewers on the same artifact or question.

Once comments are clustered, the recommender system would identify clusters that contain semantically similar responses. When a reviewer submits a comment, the system can analyze it in real-time, determine its most likely cluster assignment, and then retrieve other representative comments within that cluster. These retrieved comments can be presented as suggestions to the reviewer, either for inspiration, comparison, or direct incorporation.

This system can serve multiple purposes:

- **Enhance Feedback Consistency:** By exposing reviewers to peer comments that address the same issue or feature, it reduces the likelihood of conflicting or redundant feedback.
- **Foster Reflective Reviewing:** Reviewers may reconsider or refine their original comment when presented with alternative perspectives, leading to higher-quality, more thoughtful feedback.
- **Encourage Feedback Merging:** The interface could allow reviewers to selectively merge parts of other comments into their own, producing more comprehensive and well-rounded evaluations.

- **Improve Review Equity:** By promoting awareness of different angles of assessment, the recommender system helps mitigate the effects of individual reviewer bias or oversight.

The system would leverage semantic similarity scoring (e.g., cosine similarity of sentence embeddings or attention-weighted transformer outputs) to rank feedback suggestions and could optionally incorporate metadata such as review context, assignment rubric, or reviewer expertise. Over time, interaction data could be used to fine-tune the recommendation strategy using reinforcement learning or collaborative filtering techniques.

Such a recommender system represents a natural extension of LLM-driven clustering and opens the door for richer, AI-supported collaboration in peer assessment workflows.

II. BACKGROUND

A. Large Language Models

Large Language Models (LLMs) are deep neural networks trained on extensive corpora of text to perform a wide range of natural language processing (NLP) tasks, such as translation, summarization, text classification, question answering, and text generation. These models are typically built on the transformer architecture, introduced by Vaswani et al. [7], which uses self-attention mechanisms to model long-range dependencies and contextual relationships between tokens in a sequence.

Among the most well-known LLMs is OpenAI's GPT (Generative Pretrained Transformer) family, which includes GPT-2 [8], GPT-3 [9], and GPT-4 [10]. These models are trained using autoregressive objectives and demonstrate remarkable capabilities in few-shot and zero-shot learning through prompt-based interaction.

Another prominent model is PaLM (Pathways Language Model) by Google [11], which scales up to 540 billion parameters and leverages the Pathways system for efficient training across multiple tasks. PaLM exhibits strong performance on a variety of benchmarks, including reasoning, code generation, and multilingual understanding.

Additional noteworthy LLMs include:

- **BERT (Bidirectional Encoder Representations from Transformers)** [12], a masked language model pre-trained bidirectionally, widely used for classification and QA tasks.
- **T5 (Text-To-Text Transfer Transformer)** [13], which frames all NLP problems as text-to-text tasks and achieves state-of-the-art results across benchmarks.
- **OPT (Open Pretrained Transformer)** by Meta AI [14], an open-source alternative to GPT-3 with comparable performance.
- **LLaMA (Large Language Model Meta AI)** [15], a family of foundation models focused on efficiency and open-access research.

These LLMs represent a paradigm shift in AI, where scale, data diversity, and architectural innovations converge

to produce models capable of performing complex reasoning and language tasks with minimal supervision. Their emergence continues to shape the landscape of AI applications in education, medicine, law, and more.

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 [12].

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 [5].

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:

```
{
  "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

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.

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

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.

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.

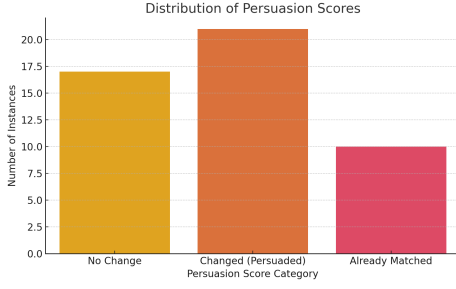


Fig. 3. Distribution of Persuasion Scores

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.

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.

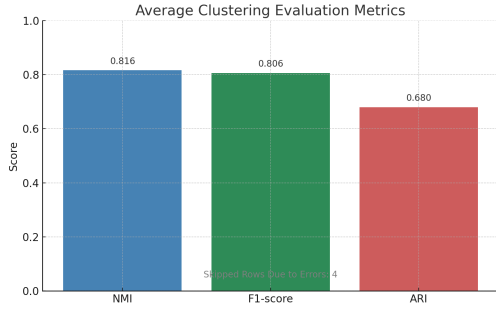


Fig. 4. Average clustering agreement scores between GPT and human clusterings.

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.

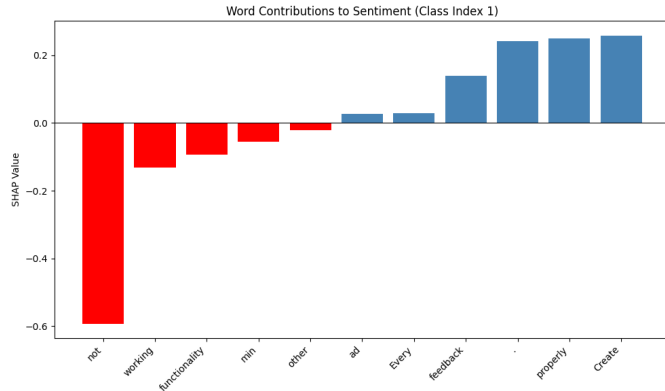


Fig. 5. KeyBERT-style bar plot highlighting the token contributions within a feedback comment.

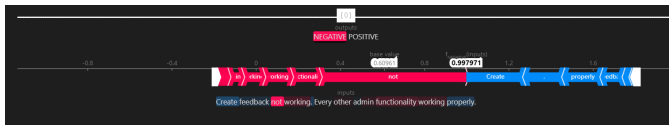


Fig. 6. Token-level contribution visualization from KeyBERT based on contextual embeddings.

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 priori-

tizing emotional tone versus semantic content. The GPT output for the same sentence is shown in Figure ??

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". It'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.

V. CONCLUSION

Our paper presents a novel method for improving the coherence and effectiveness of formative feedback in peer assessment. We prove that LLMs can assist in structuring feedback, aligning assessments, and reducing ambiguity for reviewees. The combination of qualitative persuasion scores and quantitative clustering agreement metrics—such as Pairwise F1-Score, Adjusted Rand Index, and Normalized Mutual

Information—confirms the potential of this human-AI collaboration framework to produce more consistent and actionable feedback to enhance peer assessment.

Furthermore, our analysis of keyword extraction methods, using both KeyBERT and GPT explanations, provides insight into how models interpret feedback content. While educational institutions increasingly adopt peer review at scale, our approach offers a practical solution to improve peer-generated feedback’s clarity, timeliness, and instructional value while reducing the manual burden on instructors. Future work could explore fine-tuning models on domain-specific feedback to specific courses, integrating real-time reviewer assistance, and extending the methodology across disciplines and languages.

SOFTWARE AVAILABILITY

The code and experimental materials used in this study are available at:

https://github.com/khalilbrikTN/Reviews_Research_LLM

ACKNOWLEDGMENT

The author would like to express sincere gratitude to Dr. Parvez Rashid for his invaluable support throughout the development of this work. The insightful meetings held during the semester, along with his constructive feedback and thoughtful criticism, played a crucial role in shaping the direction and quality of this research. His guidance was instrumental in refining the methodology and deepening the analysis presented in this paper.

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] 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.
- [5] V. J. Shute, “Focus on Formative Feedback,” *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008.
- [6] 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.
- [7] A. Vaswani et al., “Attention is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [8] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI Technical Report, 2019.
- [9] T. B. Brown et al., “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.
- [10] OpenAI, “GPT-4 Technical Report,” 2023. [Online]. Available: <https://openai.com/research/gpt-4>
- [11] A. Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways,” arXiv preprint arXiv:2204.02311, 2022.
- [12] 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.
- [13] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.
- [14] S. Zhang et al., “OPT: Open Pre-trained Transformer Language Models,” arXiv preprint arXiv:2205.01068, 2022.
- [15] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” arXiv preprint arXiv:2302.13971, 2023.