“A real person will always be better:” Student Perceptions of GPT-produced Feedback on a CS1 Non-Coding Assignment

Matthew Nadar, John Aromando

University of Delaware

Introduction & Purpose

Using Large Language Models (LLMs) in feedback generation has mainly focused on coding problems, but many assessments involve non-coding tasks. Additionally, formative feedback is crucial in education, enhancing learning and supporting instructors. While effective, LLM-generated feedback faces challenges like inconsistencies and incorrect suggestions.

To address this, we developed a pipeline to generate feedback for various problem types, including matching, labeling, fill-in-the-blank, and short response questions.

Our study expands the use of LLMs by:

- Developing a system for immediate feedback on non-programming tasks.
- Analyzing student perceptions of AI-generated feedback.
- Offering design recommendations for AI feedback systems.

Subjects, Methods & Analysis

Survey: five core questions—each followed by an optional open-ended prompt, demographic questions, and a general feedback prompt at the end.

Subjects: 223 students from a Introductory CS (Python) course for both majors and non-majors.

Procedure: Students completed an assignment and received GPT-generated feedback on one of the five questions, randomly chosen. They were then asked to review the feedback and complete the survey.

Data: 72 responses usable for analysis. Additionally, we collected 334 comments from the optional open-ended questions.

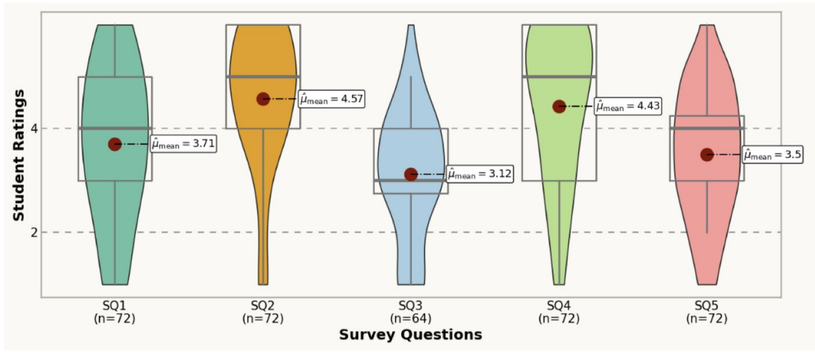
Sentiment Analysis: We used the SiEBERT model for sentiment analysis on 269 comments and 67 feedback items.

Thematic Analysis: Two authors familiarized with comments individually codified comments, then met to resolve conflicts and refine candidate themes.

Results

SQ2 and SQ4 received the highest ratings with consistent evaluations. Ratings for SQ1, SQ3, and SQ5 were more mixed compared to SQ2 and SQ4.

SQ1: How useful is the given feedback for the associated question?
SQ2: If you did this assignment again, how well would you do?
SQ3: Thinking about the experience of receiving feedback on other assignments, how does this compare?
SQ4: How easy to understand is the feedback?
SQ5: Based on your prior experiences, how do you feel about getting GPT-generated feedback?



Mann-Whitney U tests between survey question ratings and demographic variables revealed significant differences between men and women for SQ1 and SQ3. Women rated SQ1 and SQ3 more positively than men, with mean values of 4.41 vs. 2.94 for SQ1, and 3.70 vs. 2.52 for SQ3.

Sentiment analysis showed positive results for most feedback items. Overall, survey questions had a 58% positive sentiment rate, and feedback items had 88% positivity.

Section	Positive(%)	Negative(%)
Feedback	59 (88%)	8 (12%)
SQ1	34 (63%)	20 (37%)
SQ2	31 (54%)	27 (46%)
SQ3	29 (58%)	21 (42%)
SQ4	39 (70%)	17 (30%)
SQ5	22 (43%)	29 (57%)

Thematic analysis resulted in 22 themes identified across the 334 total comments. Per individual survey question, Comprehension had the most associated comments with 52 in SQ4. Similarly, as the theme is highly relevant to each of the 5 survey questions, the most common theme across multiple survey questions was Clarity with 110 comments. Parenting and Pattern are 2 out of 8 feedback -specific themes connected with the most feedback items.

Thematic Labels	Section	Definitions
Accessibility	SQ4	Students indicate signs of jargon usage or simple language.
Accuracy	SQ1	Students indicate correctness of the feedback.
Balance	Feedback	Feedback is an exemplar of balance in tone/sentiment.
Clarity	SQ1, SQ2, SQ3, SQ4, SQ5	Students indicate the feedback is clear/unclear, straightforward/complex, and/or easy/difficult to understand.
Comprehension	SQ4	Students indicate the feedback is understandable or not.
Concise	SQ1, SQ4	Students state feedback length.
Consistency	SQ5	Students indicate the feedback is consistent or inconsistent.
Detail	SQ1, SQ4	Students indicate the feedback is detailed or lacks value.
Error Prone	SQ5	Students comment on generative AI making mistakes.
Grading	SQ3, SQ5	Students mention grading.
Helpfulness	SQ1, SQ3, SQ4	Students indicate the usefulness and/or value of the feedback.
Human Preference	SQ5	Students prefer human feedback.
Improvement	SQ2	Students improve from feedback.
Indifferent	SQ5	Students express neutral feelings/reactions to the feedback.
Optimistic	SQ3, SQ5	Students indicate generative AI feedback appears promising.
Parenting	Feedback	Feedback adds reminders and/or shallow mentions of importance.
Patterns	Feedback	Feedback follows usual structure.
Personalization	SQ3, SQ5	Students comment on personalizing human and/or AI feedback.
Phrasing	Feedback	Feedback includes odd diction.
Picky	Feedback	Feedback stresses small details.
Repetition	SQ2	Students state they could improve due to repetition not feedback.
Satisfaction	SQ1	Students feedback satisfaction.
Sentiment/Tone	Feedback, SQ1, SQ4	Feedback adds complements and encouragement. Students mention the feedback's authenticity.
Stances	SQ3	Students state an opinion about the feedback.
Struggle/Self-efficacy	SQ2	Students address their confidence and/or ability level.
Suggestions	Feedback, SQ1	Feedback and/or Students acknowledge suggestions.
Transparency	Feedback, SQ1	Feedback adds answers and/or explanations. Students obtain explanations and/or answers.

Legend: 0–19 comments | 20–39 comments | 40–69 comments

Conclusions

Student perceptions and trust in generative AI feedback generation is mixed.

Student feedback can be attributed to existing narratives, assumptions, fears, and preconceived notions as shown through the thematic labels of Error Prone, Grading, Human Preference, Optimism, Stances.

The theme "Grading" was prominent despite not being part of the experiment, with concerns over human vs. AI grading leniency and empathy.

Women found the feedback more favorable when compared to men.

A consistent pattern in generative AI feedback was discovered in our thematic analysis.

Generative AI feedback often included reminders, focusing on solutions rather than explanations.

Generative AI feedback frequently contained encouragement and compliments, ending with recommendations or suggestions.

Directions for Future Research

Exploit the characteristics of generated feedback items to achieve better feedback quality.

Employ formative feedback fully and see how students interact with the assignment again and what they produce.

Delve deeper into LLM-specific parameters, such as top_p and temperature, to determine if there are nuanced effects that we may have overlooked.

Acknowledgements

We thank and acknowledge the work of Dr. Austin Cory Bart, Dr. Matthew Mauriello, Dr. Teomara Rutherford, Sammy Alashoush, Olivia Karney, Giovanna Scozzaro, and Mantra Yang, completed.