2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #50889

Can Large Language Models Recognize Neurodiverse Student Challenges and Strengths? A Complementary Study with ChatGPT and Machine Learning

Dr. Zhuwei Qin, San Francisco State University

Dr. Zhuwei Qin is currently an assistant professor in the School of Engineering at San Francisco State University (SFSU). His research focuses on designing efficient AI computing algorithms and human-centered AI systems for real-world deployment, with applications such as health technology, rehabilitation, and AI education. Dr. Qin serves as the director of the Mobile and Intelligent Computing Laboratory (MIC Lab) at SFSU. Dr. Qin's research endeavors are dedicated to addressing the inherent challenges related to efficiency and robustness in the practical application of deep learning within real-world environments. A central focus of MIC Lab is to achieve computational acceleration for deep learning, including deep neural networks and large language models, on low-power hardware without reliance on heavy compute or storage resources. His research addresses the efficiency, reliability, and security challenges in the algorithm design and system optimization, as well as application development to create intelligent mobile edge computing systems.

Yiyi Wang, San Francisco State University

Yiyi Wang is an assistant professor of civil engineering at San Francisco State University. In addition to engineering education, her research also focuses on the nexus between mapping, information technology, and transportation and has published in Accident Analysis & Prevention, Journal of Transportation Geography, and Annuals of Regional Science. She served on the Transportation Research Board (TRB) ABJ80 Statistical Analysis committee and the National Cooperative Highway Research Program (NCHRP) panel. She advises the student chapter of the Society of Women Engineers (SWE) at SFSU.

MARY K REQUA, San Francisco State University

Mary K. Requa received her PhD from the UC Berkeley/San Francisco State Joint Doctoral Program in Special Education. She is an associate professor and program coordinator in the mild/moderate support needs program. Her primary teaching responsibilities include courses related to understanding, assessing, and instructing individuals with mild and moderate support needs including methods and materials to promote successful learning outcomes for typical and atypical learners of all ages. Dr. Requa's primary research interests include pre-service teacher education and inclusive practices. She is currently co-principal investigator on an NSF Improving Undergraduate STEM Education (IUSE) grant. This project aims to develop a prototype of an AI-driven teamwork platform and explore its feasibility in (1) providing neurodiversity-informed content to students and (2) assisting professors triage or mentor students during teamwork and increasing neurodiversity awareness among faculty and students.

Zhenyu Lin, San Francisco State University

Can Large Language Models Recognize Neurodiverse Student Challenges and Strengths? A Complementary Study with ChatGPT and Machine Learning

Abstract

Large language models (LLMs) such as ChatGPT are increasingly used by college students as on-demand learning companions, producing student-authored language that can surface how learners describe difficulties and successes. For neurodiverse students (e.g., autism, ADHD, specific learning differences), recognizing both challenges and strengths is essential for designing inclusive, strength-based educational technologies. This paper examines whether LLM-based methods can classify short qualitative excerpts related to neurodiverse students' experiences as expressing a challenge, a strength, or both.

As a pilot study, we curate a dataset of 284 literature-grounded excerpts describing neurodivergent experiences in academic contexts and label each excerpt through team-based human coding. We then evaluate (1) ChatGPT as a structured-output classifier in an agreement setting against human annotations, and (2) a supervised machine learning approach (contrastive learning) in a generalization setting using an 80/20 train–test split. ChatGPT achieves 0.71 accuracy (F1 = 0.77) with high mean self-reported confidence (0.91), demonstrating the strongest agreement for challenge-oriented excerpts and lower reliability for strength and mixed-category expressions. The contrastive learning model achieves 0.88 test accuracy, with strong separation between Challenge (0.97 class-wise accuracy) and Strength (0.81), but limited performance on the sparsely represented categories.

Overall, the results indicate that supervised representation learning can more reliably distinguish challenge- and strength-oriented expressions in small qualitative datasets, while LLM-based classification offers a useful but limited reference point for interpretive alignment with human judgment. These findings support the cautious use of AI as an assistive analytic tool for surfacing student-expressed challenges and assets, informing more inclusive and strength-based approaches to engineering education without inferring diagnoses or replacing human judgment.

Introduction

The adoption of large language models (LLMs) such as ChatGPT has accelerated rapidly in higher education. Students increasingly use these systems as tutoring companions to request explanations, practice problem solving, and clarify conceptual misunderstandings [1]-[2]. Beyond their instructional utility, interactions with LLMs generate rich, student-authored language that reflects learners' struggles, strategies, and accomplishments. When analyzed with appropriate care and human oversight, such language offers a promising avenue for understanding student experience at scale.

For neurodiverse students, including those with autism spectrum disorder (ASD), attention-deficit/hyperactivity disorder (ADHD), and specific learning differences, these linguistic signals are particularly salient. Executive function challenges such as planning, working memory, time management, emotional regulation, and sensory processing are frequently cited as barriers to academic success. At the same time, stigma surrounding disability disclosure often limits access to formal accommodations, particularly in post-secondary settings where supports diminish sharply compared to K–12 education [3]. Even when accommodations are

available, many neurodiverse students report that existing supports do not adequately address social and sensory needs, contributing to isolation, reduced self-esteem, and elevated attrition.

Despite these barriers, neurodiverse students contribute important strengths to engineering and STEM learning environments, including creativity, persistence, attention to detail, and analytical or system-level thinking. However, educational technologies and institutional support structures frequently emphasize deficits, overlooking these assets or recognizing challenges only after students reach points of crisis. Capturing both challenges and strengths is therefore essential for developing inclusive, strength-based educational supports that align with neurodiversity scholarship and equity-centered engineering education [4].

The purpose of this study is to examine whether LLM-based methods and conventional supervised machine learning approaches can recognize neurodiverse student challenges and strengths as expressed in qualitative descriptions of experience. LLM-based classification offers insight into the interpretive capabilities of contemporary foundation models trained on large-scale, internet-derived text, while supervised machine learning provides a complementary perspective on how curated data and learned representations influence model outputs. Rather than diagnosing neurodivergence or predicting disability status, this study focuses on classifying expressions of experience as reflecting a Challenge, Strength, or Both.

This study asks the following research questions:

- RQ1: Can an LLM (e.g., ChatGPT) identify neurodiverse student challenges and strengths from qualitative text in agreement with human coding?
- RQ2: Can contrastive learning produce embeddings that distinguish challenge-oriented and strength-oriented student expressions?

To this end, we adopt a hybrid methodology that integrates human-coded qualitative excerpts, ChatGPT structured-output classification evaluated as an agreement task, and contrastive learning–based classification evaluated on a held-out test set. This design enables comparison between interpretive alignment with human judgment and generalization to unseen data.

By addressing these questions, this work contributes empirical evidence on the capabilities and limitations of AI-assisted analysis for small, qualitative datasets in engineering education. The findings inform the design of inclusive, strength-based AI support systems that surface patterns in student experience while remaining grounded in human judgment and ethical use.

Background and Related Work

AI Tutors and Affective Dimensions of Learning

Research on AI tutors has traditionally emphasized content delivery, personalization, and adaptive feedback. More recent work, however, has begun to explore how AI systems engage with meta-cognitive and affective dimensions of learning, such as reflection, motivation, and emotional regulation [5]. In adjacent domains, including mental health and counseling, LLM-driven conversational agents have been shown to engage with users' self-disclosed challenges in ways that are perceived as sensitive or empathetic [6]. While these systems are not substitutes for professional care and raise important ethical concerns, such findings provide

evidence that LLMs can recognize and respond to nuanced expressions of difficulty and strength in natural language. When carefully constrained and interpreted, this capability may have relevance for educational contexts that seek to support students' learning experiences without assuming diagnostic or therapeutic roles.

LLMs in Qualitative Research

LLMs are increasingly used to assist qualitative research tasks such as coding, summarization, and theme identification. Prior studies report moderate to high alignment between LLM-generated codes and human annotations when prompts are well structured and category definitions are explicit [7]. At the same time, concerns persist regarding bias, consistency, and the risk of over-automation, particularly when LLM outputs are treated as authoritative rather than assistive. As a result, hybrid human–AI approaches are frequently recommended, in which LLMs support but do not replace human interpretation. Moreover, agreement with human coding alone is insufficient; evaluating robustness and generalization remains an important methodological challenge, especially for small, qualitative datasets.

Contrastive Learning for Text Representation

Contrastive learning has emerged as an effective supervised machine learning approach for learning text embeddings, particularly in limited and imbalanced data settings [8]. By encouraging semantically similar samples to cluster while separating dissimilar ones, contrastive learning produces representations suited for downstream classification and clustering tasks. This approach is especially relevant for nuanced qualitative categories, such as expressions of student challenges and strengths, where semantic boundaries are subtle, context-dependent, and difficult to capture using traditional feature-based methods. As such, contrastive learning provides a complementary methodological lens for evaluating whether challenge- and strength-oriented student expressions can be meaningfully separated in a learned representation space.

Methods

Data Collection and Annotation

This study draws on qualitative evidence from prior research to construct a dataset that captures how neurodiverse students' challenges and strengths are described in academic contexts. Rather than collecting new student data, we grounded the dataset in peer-reviewed literature to ensure that the excerpts reflect well-documented, ethically vetted, and theoretically informed accounts of neurodiverse experiences.

The research team conducted a comprehensive literature review spanning both discipline-specific education research (e.g., engineering education and STEM higher education) and neurodiversity-focused special education literature. The objectives of this review were twofold: (1) to identify challenges and strengths associated with neurodiverse student experiences reported in evidence-based research, and (2) to extract representative textual excerpts that illustrate how these challenges and strengths are articulated in the literature.

Challenges identified during the review included executive function–related difficulties such as planning, working memory, time management, emotional regulation, and sensory overload. Strengths included persistence, attention to detail, creativity, systematic thinking, and deep

engagement with areas of interest. From the reviewed sources, the team extracted representative quotes that conveyed these experiences in narrative form. Each quote was treated as an individual excerpt, representing either a direct quotation or a synthesized description grounded in peer-reviewed literature rather than direct student testimony.

All extracted excerpts were manually reviewed and annotated by the research team. Each excerpt was assigned one of four labels: Challenge, Strength, or Both. The *Challenge* label was used when an excerpt primarily described barriers or difficulties experienced by neurodiverse students. The *Strength* label was applied when an excerpt emphasized assets or positive capabilities. The *Both* label was used when an excerpt meaningfully conveyed the coexistence of challenges and strengths within the same narrative.

Figure 1 summarizes the distribution of high-level labels in the curated dataset (N = 284). As shown in the figure, excerpts labeled as Challenge comprise the majority of samples (65.1%, n = 185), followed by Strength excerpts (31.3%, n = 89), with a smaller proportion labeled as Both (3.5%, n = 10). This imbalance reflects the tendency of the existing literature to emphasize barriers faced by neurodiverse students more frequently than their strengths.

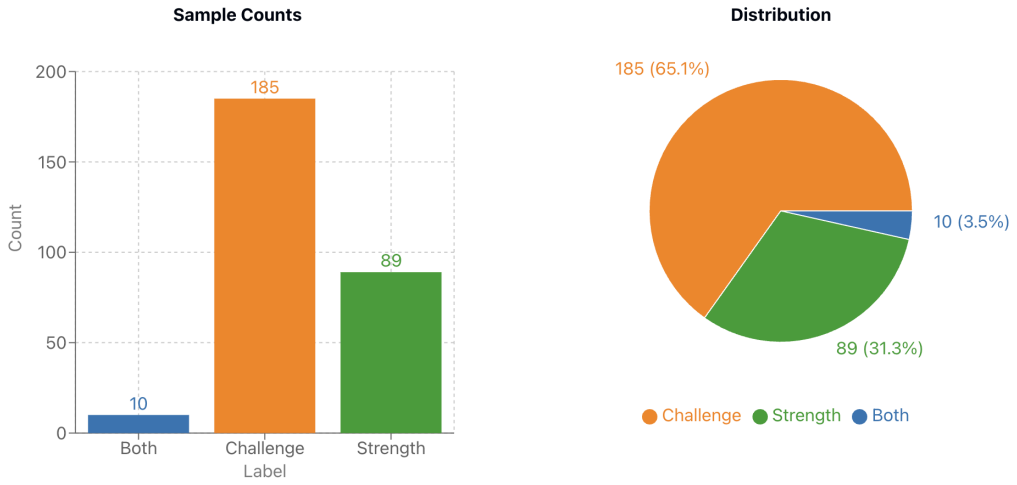

Figure 1. Distribution of labeled excerpts in the curated dataset (N = 284). The bar chart (left) shows sample counts for each label, and the pie chart (right) shows their proportional distribution. Challenge excerpts constitute the majority of the dataset (65.1%, n = 185), followed by Strength excerpts (31.3%, n = 89), with a smaller proportion labeled as Both (3.5%, n = 10).

To provide additional context on the specific experiences captured in the dataset, Figure 2 presents the frequency distribution of fine-grained challenge- and strength-related themes extracted from the literature. The figure reveals a long-tailed distribution, with a small number of frequently cited themes and many low-frequency themes. This pattern highlights the heterogeneity of neurodiverse student experiences reported in prior research and motivates the use of higher-level labels (Challenge vs. Strength) in the classification framework, as many individual subthemes are sparsely represented.

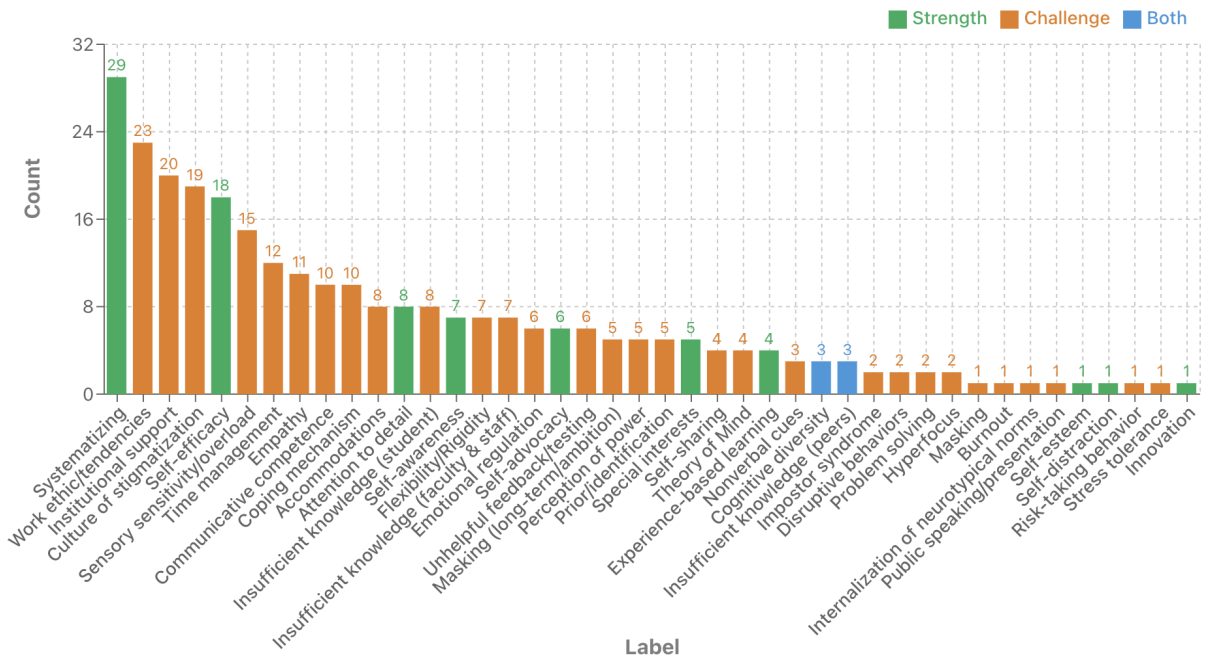

Figure 2. Frequency distribution of fine-grained challenge- and strength-related labels extracted from the literature. Bars indicate the number of excerpts associated with each label, colored by category (Challenge, Strength, or Both). The distribution highlights substantial variability in how the curated dataset served two complementary purposes.

As shown in Figure 2, first, it provided qualitative evidence of how neurodiverse challenges and strengths are framed in prior research. Second, the excerpts were used to construct structured inputs for model evaluation. For example, as shown in Table 1, an excerpt such as "I often lose track of instructions in fast-paced discussions" was paired with a structured annotation such as *Challenge → Working memory*. All 284 excerpts were used for ChatGPT-assisted classification, and a subset was used to train and evaluate the contrastive learning model described in the following sections.

Table 1. Data example of neurodiverse students' challenges and strengths

Excerpt	Label	Fine-grained Label
"I often lose track of instructions in fast-paced discussions."	Challenge	Working Memory
"I'm meticulous and often find errors others miss."	Strength	Attention to detail

ChatGPT Structured-Output Classification

To examine whether a large language model could identify neurodiverse student challenges and strengths in alignment with human coding, we employed ChatGPT using a structured-output prompting strategy. The model was positioned as a classification assistant and provided with explicit category definitions to reduce ambiguity and encourage consistent reasoning.

Specifically, ChatGPT was prompted with the following instructions:

You are a classification expert analyzing neurodivergent experiences in academic settings.

Classify the given text into one of the following categories:

- Challenge: Text describes difficulties, obstacles, or negative experiences

- Strength: Text describes positive attributes, capabilities, or beneficial aspects

- Both: Text contains both challenges and strengths

- Neither: Text doesn't clearly fit into Challenge or Strength categories

Provide your classification with a confidence score (0-1) and a brief explanation of your reasoning.

Be precise and objective in your classification.

For each of the 284 excerpts, ChatGPT returned a predicted label, a numerical confidence score, and a brief natural-language rationale. All excerpts were classified using the same prompt to ensure consistency across samples. Because ChatGPT was applied to the full dataset without a train–test split, its performance was evaluated as an agreement task relative to the human-coded ground truth rather than as a measure of predictive generalization.

Contrastive Learning Classification

While ChatGPT-assisted classification provides an efficient means of assessing alignment between large language model outputs and human-coded labels, it does not evaluate generalization beyond the observed dataset. To address this limitation and to examine whether challenge- and strength-oriented student expressions can be distinguished in a learned representation space, we employ a contrastive learning–based supervised classification approach.

Contrastive learning is particularly well-suited to this study for three reasons. First, the dataset is relatively small and imbalanced, with a sparse representation of certain categories. In such low-resource settings, contrastive learning is effective because it leverages pairwise similarity relationships rather than relying solely on large quantities of labeled examples. Second, the semantic boundary between Challenge and Strength expressions is often subtle and context-dependent. By learning embeddings that pull semantically similar excerpts closer together while pushing dissimilar excerpts apart, contrastive learning aligns naturally with the qualitative and nuanced nature of the data. Third, contrastive learning supports evaluation within a supervised learning paradigm using a held-out test set, enabling assessment of generalization performance on unseen samples. Together, these considerations motivate contrastive learning as a complementary method that strengthens the methodological rigor of the study by evaluating generalizability in addition to interpretive alignment.

As illustrated in Figure 3, labeled text excerpts are first encoded using a pre-trained sentence-transformer model and fine-tuned through contrastive learning, where semantically similar excerpts are pulled closer together, and dissimilar excerpts are pushed apart in the embedding space. A linear classification head is then trained on top of the learned embeddings using labeled training samples, and model performance is evaluated on a held-out test set. This two-stage process enables learning discriminative representations from limited, imbalanced qualitative data and supports evaluation of generalization beyond agreement with human coding.

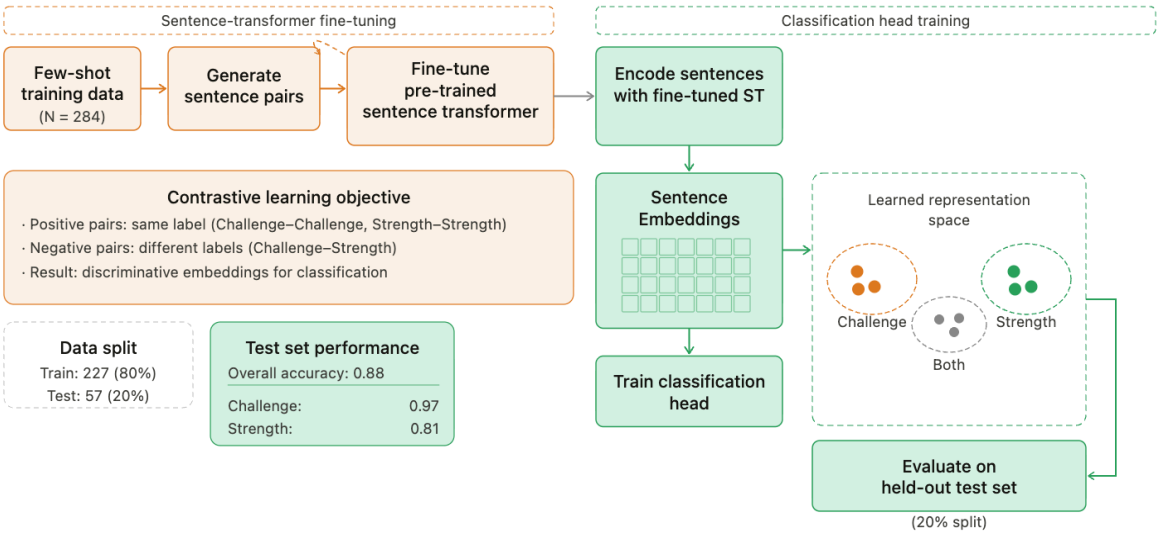

Figure 3. Overview of the contrastive learning–based classification pipeline. The pipeline consists of two main stages: (1) Sentence-transformer fine-tuning using contrastive learning to learn discriminative embeddings, where positive pairs are pulled together, and negative pairs are pushed apart in the embedding space, and (2) classification head training on the learned embeddings using an 80/20 train-test split. The model achieves 0.88 overall test accuracy with strong class-wise performance for Challenge (0.97) and Strength (0.81) categories.

SetFit was selected because prior work has demonstrated that it can achieve strong classification performance with minimal labeled data [8]. According to prior studies, SetFit attains high accuracy with limited supervision; for example, with only eight labeled examples per class on the Customer Reviews sentiment dataset, SetFit performs comparably to fine-tuning RoBERTa Large on the full training set of approximately 3,000 examples. This data efficiency makes SetFit particularly well-suited for the present study, where labeled qualitative excerpts are limited and class imbalance is present.

For this study, text samples were encoded using a pre-trained sentence-transformer model and fine-tuned using a contrastive loss objective. During training, pairs of excerpts sharing the same high-level label (e.g., Challenge–Challenge or Strength–Strength) were treated as positive pairs, while excerpts from different categories were treated as negative pairs. After contrastive fine-tuning, a linear classification head was trained on top of the learned embeddings to predict high-level labels. The dataset was split into 227 excerpts (80%) for training and 57 excerpts (20%) for held-out testing.

Table 2. Performance Comparison of ChatGPT and Contrastive Learning

Overall Result			Class-wise Accuracy		
	ChatGPT	Contrastive Learning		ChatGPT	Contrastive Learning
Accuracy	0.7077	**0.8772**	Challenge	0.7838	**0.9706**
Precision	0.8568	N/A	Strength	0.5955	**0.8095**
Average Confidence	0.9104	N/A	Both	0.3	N/A

Table 2 compares the performance of ChatGPT-assisted classification and contrastive learning–based classification. ChatGPT achieved an overall accuracy of 0.71 with a high average confidence score (0.91), indicating substantial agreement with human-coded annotations. However, ChatGPT's performance varied across categories, with higher accuracy for Challenge excerpts (0.78) and lower accuracy for Strength (0.60), reflecting the dataset's imbalance and the nuanced nature of strength-oriented expressions.

In contrast, the contrastive learning model achieved higher overall accuracy (0.88) on a held-out test set and demonstrated strong class-wise performance for the two dominant categories. Accuracy for Challenge excerpts reached 0.97, while Strength accuracy improved to 0.81, suggesting that contrastive learning effectively separates these categories in the learned embedding space. Performance for the Both and Neither categories was poor for both methods, reflecting their sparse representation in the dataset and underscoring limitations associated with class imbalance.

Taken together, these results illustrate the complementary strengths of the two approaches: ChatGPT provides high-confidence, interpretable agreement with human judgments, while contrastive learning offers improved generalization and class separation for the most prevalent categories.

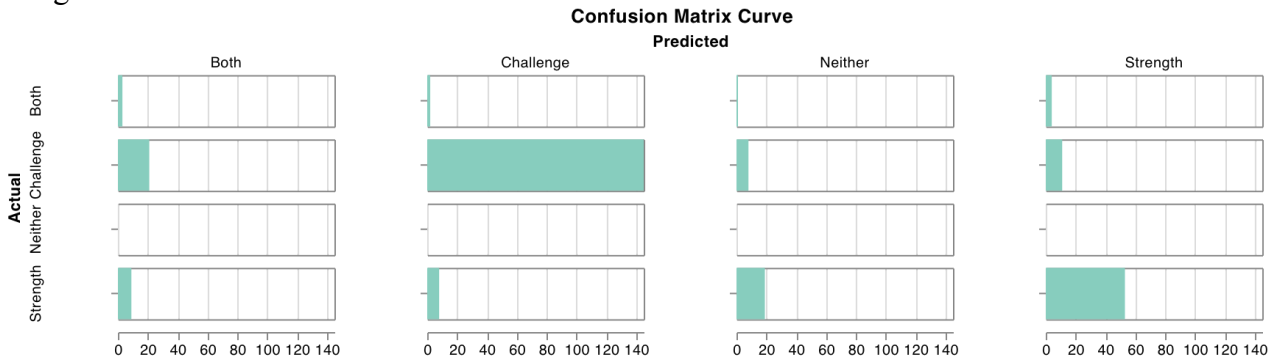

Figure 4. Confusion Matrix for ChatGPT Classification. Rows represent human-coded ground truth labels, and columns represent ChatGPT-predicted labels. The visualization highlights strong alignment for the dominant Challenge category, moderate separation for Strength, and substantial confusion for sparsely represented categories.

Figure 4 provides a detailed view of ChatGPT's classification behavior relative to human-coded labels. The confusion matrix shows that excerpts labeled as Challenge by human coders were most frequently and consistently classified correctly by ChatGPT, with the majority of Challenge instances mapped to the Challenge category. This pattern aligns with the higher class-wise accuracy observed for Challenge excerpts and reflects the explicit and frequently documented nature of challenge-oriented language in the dataset.

For Strength excerpts, ChatGPT demonstrated moderate classification accuracy. While many Strength instances were correctly identified, a noticeable proportion were misclassified as Challenge, suggesting that strength-oriented expressions are often more implicit or embedded within narratives of difficulty. This confusion mirrors the lower class-wise accuracy for Strength relative to Challenge.

Performance on the Both and Neither categories was limited. Excerpts labeled as Both were frequently misclassified as either Challenge or Strength, indicating difficulty in recognizing co-occurring challenges and strengths within a single excerpt. Neither excerpts were rarely predicted correctly, reflecting both their sparse representation and the lack of strong linguistic cues distinguishing them from the other categories.

Overall, the confusion matrix reinforces that ChatGPT is effective at identifying dominant challenge-related language but less reliable for minority and ambiguous categories. These patterns support the interpretation of ChatGPT results as agreement with human judgment rather than robust generalization and motivate the complementary use of contrastive learning to improve separability for the dominant categories.

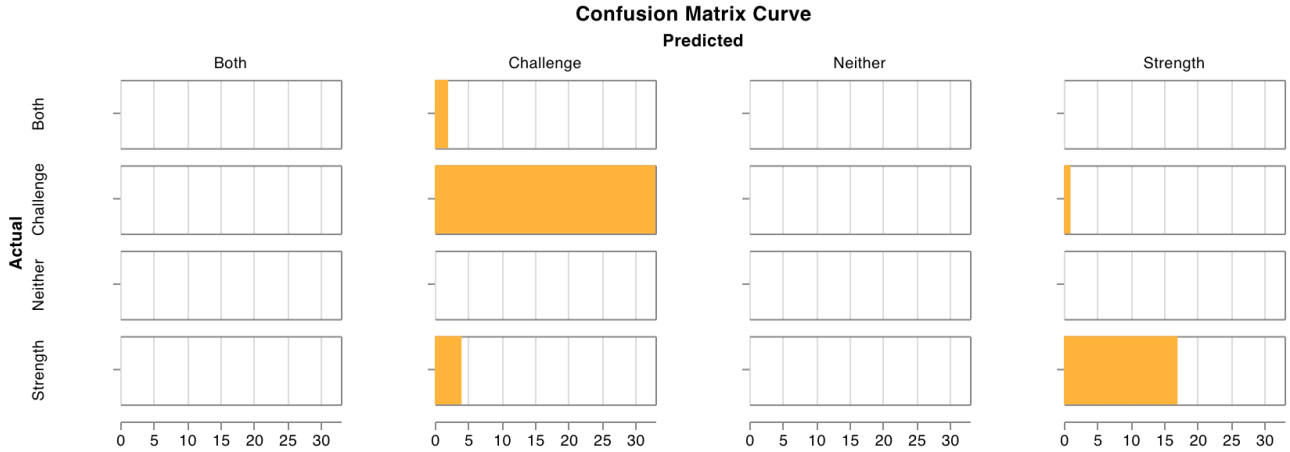

Figure 5. Confusion matrix illustrating the classification performance of the contrastive learning model on the test set. Rows represent human-coded ground truth labels, and columns represent model-predicted labels. The visualization highlights strong separation between the dominant Challenge and Strength categories, with limited correct predictions for sparsely represented categories.

Figure 5 shows the confusion matrix for the contrastive learning–based classifier evaluated on the held-out test set. The model demonstrates strong discrimination between the two dominant categories, with most Challenge excerpts correctly classified as Challenge and most Strength excerpts correctly classified as Strength. This pattern is consistent with the high class-wise accuracies reported for Challenge (0.97) and Strength (0.81).

Misclassifications primarily occur between Challenge and Strength, reflecting residual semantic overlap between difficulty- and asset-oriented language in qualitative narratives. In contrast to ChatGPT, the contrastive learning model rarely predicts the Both or Neither categories, resulting in zero accuracy for these classes. This outcome reflects the extremely limited representation of these categories in the training data and the model's bias toward learning separable representations for the most prevalent labels.

Overall, the confusion matrix indicates that contrastive learning effectively learns a discriminative embedding space for challenge- and strength-oriented expressions and generalizes well to unseen data for these categories. At the same time, the results highlight the limitations of representation learning in low-resource settings for minority and ambiguous classes.

Discussion

This study examined whether large language models (LLMs) and contrastive learning–based supervised machine learning can recognize neurodiverse student challenges and strengths from qualitative text. By integrating human-coded excerpts, ChatGPT structured-output classification, and contrastive learning, the study provides a comparative perspective on interpretive alignment and generalization in the analysis of nuanced student experience data.

The results indicate that ChatGPT demonstrates moderate agreement with human-coded labels rather than strong classification performance. With an overall accuracy of 0.71, ChatGPT correctly classified many challenge-oriented excerpts but showed substantial variability across categories. Performance was notably weaker for strength-oriented and mixed-category excerpts, suggesting that strength-related language is less salient or less consistently interpreted by the model. These findings indicate that, even with structured prompting and explicit definitions, ChatGPT's classifications should be interpreted cautiously and should not be treated as reliable or generalizable predictions. Instead, ChatGPT's outputs are best understood as an exploratory reference point for how a widely used foundation model interprets qualitative descriptions of student experience.

In contrast, the contrastive learning model demonstrated stronger empirical performance on a held-out test set, achieving higher overall accuracy and substantially improved class-wise separation for the dominant Challenge and Strength categories. These results suggest that supervised representation learning, when paired with human-labeled data, can more reliably distinguish between challenge- and strength-oriented expressions in small qualitative datasets. However, performance remained poor for the sparsely represented Both category, underscoring persistent challenges associated with class imbalance and semantic overlap in low-resource settings.

Taken together, these findings highlight important differences between LLM-based classification and supervised machine learning approaches. ChatGPT provides insight into how a general-purpose foundation model interprets student-authored language based on large-scale pretraining, but its moderate accuracy and uneven category performance limit its suitability for automated classification. In contrast, contrastive learning offers a more robust mechanism for generalization when sufficient labeled data and careful evaluation are available. Neither approach, however, adequately addresses mixed or ambiguous expressions without additional data or modeling strategies.

From an engineering education perspective, these results suggest that AI-enabled systems should not rely on LLMs alone to identify student challenges or strengths. Automated interpretations, particularly of strength-oriented or mixed expressions, risk misclassification and oversimplification. Instead, AI tools may be more appropriately positioned as assistive analytic supports, used to surface potential patterns for further human interpretation rather than to drive autonomous decisions or interventions. Strength-based recognition, in particular, requires careful design to avoid reinforcing deficit-oriented framings of neurodiverse students.

Methodologically, this study underscores the importance of evaluating both interpretive alignment and generalization when applying AI methods to qualitative educational data. Agreement with human coding, as observed with ChatGPT, does not imply predictive reliability,

while higher generalization performance, as observed with contrastive learning, remains constrained by dataset size and label distribution. A hybrid approach that combines human expertise with supervised learning and cautious use of LLMs offers a more defensible pathway for analyzing small, nuanced datasets common in engineering education research.

Limitations and Future Work

This study has several limitations that should be considered when interpreting the results. First, the dataset was derived from excerpts in the existing literature rather than directly from student-generated text. While this approach ensures ethical rigor and theoretical grounding, it may not fully capture the linguistic variability present in authentic student interactions. Future work should validate these findings using student-authored data collected from real educational settings. Second, the dataset is relatively small and imbalanced, with sparse representation of the Both and Neither categories. As a result, both ChatGPT and the contrastive learning model performed poorly on these categories. Future research should explore data augmentation strategies, targeted sampling, or alternative modeling approaches to better capture mixed and ambiguous expressions. Third, ChatGPT classification was evaluated as an agreement task without a held-out test set. While this design was intentional and appropriate for interpretive comparison, it does not support claims about predictive generalization. Future studies could explore prompt variation, cross-model comparisons, or longitudinal stability of LLM classifications. Finally, this study focused on high-level challenge and strength categories. While appropriate for the current dataset, future work could investigate finer-grained subcategories and examine how such distinctions affect both interpretability and downstream educational interventions.

Acknowledgment

This work was supported by the National Science Foundation under Grant No. IUSE-2417557. The views expressed are those of the authors and do not necessarily reflect those of the funding agency.

References

[1] C. K. Lo, “What is the impact of ChatGPT on education? A rapid review of the literature,” Education Sciences, vol. 13, no. 4, p. 410, 2023.
[2] J. Jabbour, K. Kleinbard, O. Miller, R. Haussman, and V. J. Reddi, “SocratiQ: A generative AI-powered learning companion for personalized education and broader accessibility,” in Proc. Workshop on Computer Architecture Education, 2025, pp. 1–10.
[3] National Center for Learning Disabilities, “Understand the issues,” [Online]. Available: https://www.ncld.org/join-the-movement/understand-the-issues/. Accessed: Jan. 12, 2024.
[4] F. Happé and U. Frith, “The weak coherence account: Detail-focused cognitive style in autism spectrum disorders,” Journal of Autism and Developmental Disorders, vol. 36, no. 1, pp. 5–25, 2006.
[5] H. Li, J. Yu, X. Cong, Y. Dang, D. Zhang-Li, Y. Zhan, H. Liu, and Z. Liu, “Exploring LLM-based student simulation for metacognitive cultivation,” arXiv preprint arXiv:2502.11678, 2025.

[6] Z. Iftikhar, A. Xiao, S. Ransom, J. Huang, and H. Suresh, “How LLM counselors violate ethical standards in mental health practice: A practitioner-informed framework,” in Proc. AAAI/ACM Conf. on AI, Ethics, and Society, vol. 8, no. 2, 2025, pp. 1311–1323.
[7] R. H. Tai, L. R. Bentley, X. Xia, J. M. Sitt, S. C. Fankhauser, A. M. Chicas-Mosier, and B. G. Monteith, “An examination of the use of large language models to aid analysis of textual data,” International Journal of Qualitative Methods, vol. 23, p. 16094069241231168, 2024.
[8] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, “Efficient few-shot learning without prompts,” arXiv preprint arXiv:2209.11055, 2022.