EdgeCon Proceedings

Scaling Student Motivation Analysis with Large Language Models

Yanqui Li
Office of Digital Learning, New Jersey Institute of Technology
yanqui.li@njit.edu

Abstract

This study proposes a replicable pipeline for transforming students' open-ended motivation inputs into structured data using large language models (LLMs) and expectancy–value theory. Analysis of a synthetic dataset suggests that distinct motivational profiles may correspond with varying levels of engagement and performance. Motivational responses can also be analyzed to identify patterns across course subjects, levels, and outcomes. These insights have practical implications for curriculum design, academic advising, and early intervention. Although motivational features may not significantly enhance predictive accuracy or recall, they contribute to interpretation by offering insights into the internal factors that drive student engagement and disengagement. This work showcases a use case of LLMs to power scalable motivation analytics to inform student success efforts.

Keywords: Student Motivation, Learning Analytics, Large Language Models, Expectancy–Value Theory, Synthetic Data

1. Introduction

In the field of learning analytics, researchers have primarily relied on structured data such as demographics, participation logs, course assessment scores, and cumulative GPAs to understand student behavior and predict academic outcomes (Ifenthaler & Yau, 2020; Papamitsiou & Economides, 2014). However, these behavioral signals often fall short in revealing the underlying “why” behind student behavior patterns (Pardo et al., 2015). Student motivation offers a complementary perspective to behavioral and performance-based factors. Research shows that motivation and participation play a critical role in predicting student performance, with situational interest and intrinsic motivation influencing both engagement and outcomes (de Barba et al., 2016).

These motivational inputs offer insight into students’ internal drivers, such as personal interests, career or academic goals, and perceived pressures, such as program or degree requirements. However, because these responses are often open-ended and unstructured, they are analyzed less frequently due to challenges in standardization and automation. Recent advances in large language models (LLMs) and natural language processing have enhanced the ability to analyze free-text data at scale, enabling integration of unstructured student input into learning analytics workflows (Parker et al., 2024).

This project proposes a scalable pipeline for transforming unstructured motivational input into structured, actionable data. Unlike prior studies that rely on fixed-response surveys, this approach integrates naturally occurring open-ended responses and uses LLMs to automate theory-based coding. Overall, the pipeline includes five steps: (1) data gathering through the learning management system (LMS) using consistent prompts; (2) data cleaning; (3) LLM-based feature extraction and encoding; (4) exploratory data analysis to uncover patterns in motivational profiles across engagement, course subjects, and course levels; (5) predictive modeling to examine how motivational factors contribute to identifying students at risk.

As a pre-IRB prototype built on synthetic data, this work demonstrates both feasibility and value. It serves as a proof-of-concept (POC) to integrate motivation into learning analytics by capturing students’ intentions and alignment beyond observable behavior. This approach has the potential to inform early alerts, academic advising, and curriculum development, contributing to more responsive educational practices.

2. Foundations and Related Work

2.1 Review of Relevant Literature

Student motivation has long been recognized as a key factor influencing learning behavior, academic engagement, and performance. A meta-analysis of 344 samples found that intrinsic motivation is positively associated with academic performance, persistence, and well-being, while identified regulation (i.e., valuing the task) correlates strongly with long-term persistence (Howard et al., 2021). In another study in the context of service-learning, Leung et al. (2022) found that student motivation, when coupled with high-quality learning experiences, significantly predicted cognitive learning outcomes. Their findings emphasized that when students understand and value what they are learning, they are more likely to engage deeply and retain knowledge.

Given the well-established role of motivation in student success, it is increasingly important to identify both students and courses where motivation may be lacking. Students with consistently low motivation across courses may benefit from autonomy-supportive practices, which have been shown to foster student motivation in higher education settings (Okada, 2023). Likewise, courses that receive low intrinsic value ratings may benefit from redesign or revised communication strategies—for example, more precise descriptions in enrollment pages to help students better relate to the content (Kember et al., 2008).

2.2 Theoretical Framework

This study applies the Expectancy–Value Theory (Eccles & Wigfield, 2000) to interpret students' self-reported motivations for enrolling in a course. The theory posits that achievement-related choices are influenced by two key factors: students' expectations of success and the subjective value they assign to the task. In the context of course registration, value can be intrinsic (e.g., personal interest in the subject), utility-based (e.g., relevance to career or significance), imposed (e.g., fulfilling a degree requirement), or cost-related (e.g., anticipated workload or fear of failure).

Importantly, students may hold multiple value beliefs simultaneously—a course may be interesting, helpful, and required at the same time. These value types can coexist and interact to shape students' motivation to enroll and persist. By using this theoretical lens, the study provides a structured framework for interpreting students' diverse motivational inputs in an educational context.

2.3 Identified Gaps in Practice

Despite the widespread use of open-ended surveys, LMS discussion forums, and introductory course reflections that ask students to explain their reasons for enrolling, institutions rarely analyze these responses systematically or at scale. This underutilization is mainly due to the unstructured nature of the data and the lack of scalable, theory-informed frameworks for interpretation. While motivation theory is well established in educational psychology, few studies have operationalized these constructs using free-text responses collected in authentic academic contexts. This reveals a notable gap between theoretical understanding and practical implementation, particularly in applying LLM tools capable of automated, context-sensitive interpretation of student input.

3. This Work's Contribution

This study addresses these gaps by introducing a POC pipeline that integrates Expectancy–Value Theory with LLMs to process and analyze free-text student motivation inputs. Using synthetic data as a stand-in for real-world responses, the project illustrates the feasibility of applying this framework to large-scale, institutionally collected data. It demonstrates several types of analysis that can be conducted, including: (1) correlation among motivation, engagement, and outcomes; (2) motivation profiles across subjects and course levels; and (3) predictive modeling of academic outcomes incorporating motivational dimensions.

By showing how theoretical constructs of motivation can be embedded into a scalable workflow, this work contributes to the intersection of learning analytics and LLM-powered educational data science. The proposed pipeline is designed to be replicable and adaptable for institutions interested in leveraging free-text student feedback to align curricula, support advising, and enhance retention strategies.

4. Project Description and Current Status

To demonstrate feasibility before IRB approval, we currently use synthetic data, which enables development, demonstration, and stakeholder engagement without compromising student privacy. Potential stakeholders include curriculum developers, academic advisors, and departmental leadership, who can use the resulting insights to understand student intent better and improve educational support. Our progress to date includes creating a synthetic motivation dataset, aligning with motivation theory, and developing a scalable, replicable pipeline. We plan to submit an IRB application in Fall 2025, with pilot testing to follow upon approval.

5. Methodology and Approach

5.1 Data Collection Strategy

In the formal phase, data will be collected through the LMS using tools such as Canvas Blueprint to deliver consistent prompts across courses. The dataset will include students' free-text responses explaining their reasons for registering for the course, along with participation metrics and final grades. Only students who provide informed consent will have their data included in the study, ensuring ethical compliance through an opt-in approach.

In the POC phase, synthetic data were generated using LLMs, including Gemini and Claude. These models simulated student responses based on predefined engagement and performance profiles. The dataset spans six courses: five 100-level introductory courses (Math, Physics, Chemistry, English, and Introductory Computer Science) and one 500-level advanced course (Advanced Computer Science). In total, 600 synthetic student records were created, covering combinations such as high engagement/high performance and low engagement/high performance.

5.2 Implementation Workflow

The implementation consists of five primary stages: (1) Data Gathering: Either synthetic or real LMS data are collected, including motivational responses, participation metrics, and academic outcomes. (2) Data Cleaning: All data is cleaned and de-identified to ensure privacy. (3) LLM-based Feature Extraction and Encoding: Free-text responses are processed using LLMs to classify motivational responses according to the Expectancy–Value Theory framework. The models return structured JSON outputs, which are then encoded into analyzable variables. Value components—including intrinsic value, utility value, imposed value, and cost—are represented as binary features (1 = mentioned, 0 = not mentioned). Expectancy, reflecting students' self-perceived ability to succeed, is encoded as a categorical variable: 1 for high confidence, 0 for low confidence, and null if no mention is made. This approach preserves subtle differences in student self-efficacy while avoiding misclassification when the construct is not explicitly stated. To improve categorization accuracy, human raters reviewed sample outputs, and prompt wording was adjusted to address observed inconsistencies. This iterative process helped refine the LLM's performance and will continue during the real-data phase to ensure theoretical alignment. (4) Exploratory Data Analysis: Motivation patterns are examined across subjects and course levels to identify trends, variations, and potential gaps in alignment between course design and student intent. (5) Predictive Modeling: Machine learning models (e.g., Random Forest Classifier) are implemented to assess whether early motivational indicators can help predict course failure.

The POC phase used a synthetic dataset to validate the pipeline, confirming the robustness of each processing stage and providing valuable feedback for refining the workflow before applying it to real student data. Table 1 provides examples of motivational inputs and their corresponding feature labels, illustrating how open-ended responses are structured for analysis.

Table 1

Examples of Motivation Statements and Expectancy–Value Annotations

Motivation Statement	Expectancy	Utility Value	Intrinsic Value	Imposed Value	Cost
I've always enjoyed math, and Calculus I is a natural progression. I'm looking forward to the challenge and gaining a deeper insight into how mathematical principles govern the world around us.	null	0	1	0	0
I'm taking this course to strengthen my analytical skills, which are important for my business major. I'm finding it tougher than I expected, but I'm dedicated to improving my understanding and performance.	1	1	0	0	1

Note. Expectancy represents self-perceived ability to succeed: 1 = high confidence, 0 = low confidence, null = not mentioned.

6. Preliminary Findings

6.1 Motivation, Engagement, and Outcome Patterns

Analysis of the synthetic dataset reveals several motivation–engagement–outcome trends consistent with expectancy–value theory. As shown in Figure 1A, intrinsic value and utility value were positively correlated with behavioral indicators such as participation level ($r = .39$ and $r = .28$, respectively) and activity time ($r = .26$ and $r = .27$, respectively), suggesting that students who find courses interesting or career-relevant tend to engage more actively.

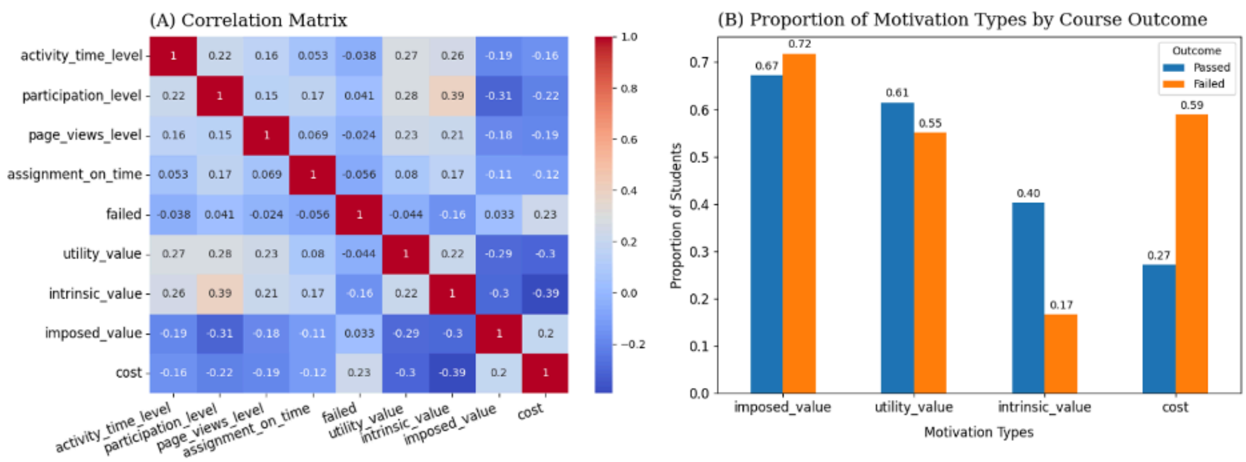
Conversely, imposed value was negatively correlated with intrinsic value ($r = -.30$) and utility value ($r = -.29$), suggesting a potential motivational conflict when courses are taken for external rather than personal reasons.

DOI: [to be assigned]

Figure 1B illustrates outcome disparities across motivation types. Students who passed exhibited higher rates of intrinsic value (40% vs. 17%) and utility value (61% vs. 55%) compared to those who failed. In contrast, failed students were more likely to report cost-related concerns (59% vs. 27%) and imposed value (72% vs. 67%), indicating that external pressure and perceived difficulty may be risk signals for poor performance.

Figure 1

Motivation, Engagement, and Outcome Patterns in the Synthetic Dataset.



Note. (A) Correlation matrix displaying relationships among student motivation features, behavioral engagement, and course outcome. (B) Proportion of students exhibiting each motivation type, separated by course outcome. Expectancy was excluded from the analysis here due to a high proportion of null values, which could bias the results.

6.2 Variation in Student Motivation Across Subjects and Course Levels

Figure 2 reveals meaningful variation in motivational profiles across academic subjects and course levels. Figure 2A shows that English courses exhibit the highest imposed value (.76) and the lowest intrinsic value (.18), suggesting that students often enroll in English courses for external reasons rather than personal interest. This pattern may warrant closer attention to how such courses are framed and communicated, especially in general education contexts.

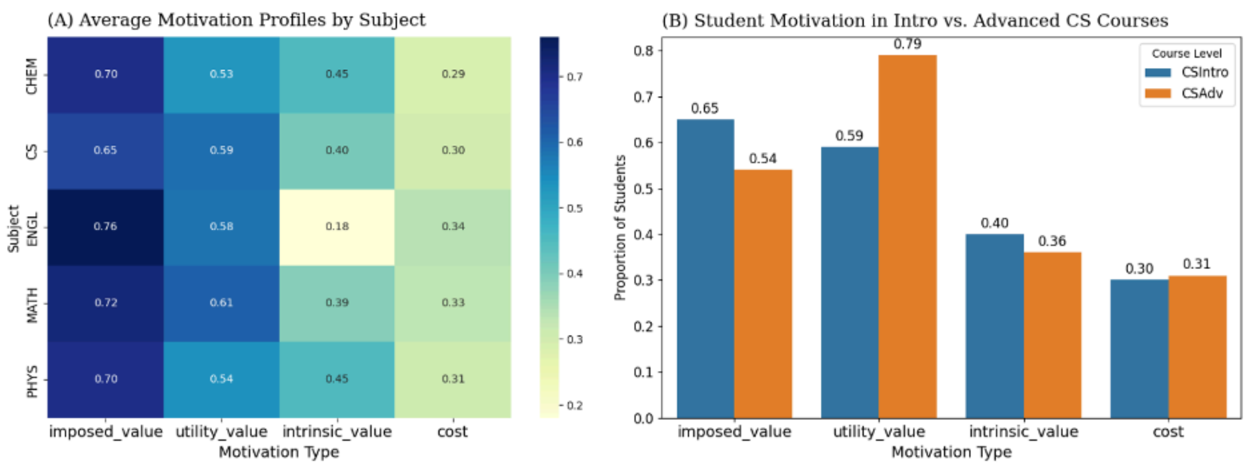
In contrast, STEM subjects such as Chemistry and Physics display more balanced profiles, with relatively higher intrinsic value (.45) and moderate levels of utility and imposed value. Mathematics emerges as a particularly career-relevant subject, exhibiting the highest utility value (.61), indicating strong alignment with students' academic or professional goals.

DOI: [to be assigned]

Within computer science, the synthetic dataset suggests notable distinctions between introductory and advanced courses. As shown in Figure 2B, advanced courses exhibit a higher average utility value (.79 vs. .59), indicating that career relevance becomes more emphasized in later stages. Cost levels, however, remain comparable across levels, suggesting that increased course complexity does not necessarily translate into a heightened sense of burden.

Figure 2

Variation in Student Motivation Across Subjects and Course Levels.



Note. (A) Average proportion of students exhibiting each motivation type across 100-level introductory courses. (B) Proportion of students showing each motivation type, separated by course level (introductory vs. advanced).

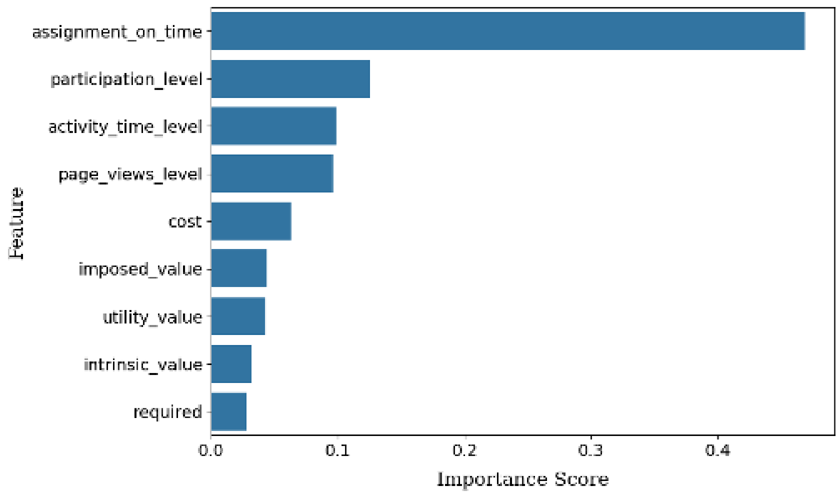
6.3 Predictive Modeling with Motivational and Behavioral Features

To test the predictive performance of this pipeline, we trained a Random Forest classifier on a synthetic dataset with an overall failure rate of 13%. The model achieved an accuracy of approximately 83%, which is expected given that participation and motivation are generally weaker predictors than performance-based measures such as current scores or GPA. As shown in Figure 3, behavioral indicators, especially assignment submission timeliness, were the most influential features. The model's recall for identifying students who failed was low (19%), likely due to class imbalance and the difficulty of detecting at-risk students.

These results suggest that while motivational data may not improve predictive performance, it adds interpretive value. Understanding that a student failed not only because of low participation but also due to external pressure or anticipated difficulty can inform more targeted and empathetic interventions.

Figure 3

Feature Importance Scores from a Random Forest Classifier Predicting Course Failure.



7. Preliminary Findings and Stakeholder Implications

The POC analysis suggests that integrating motivational inputs with behavioral data can improve the interpretability of learning analytics. While motivational traits may not significantly enhance predictive accuracy or recall, they offer practical context, highlighting why students engage, struggle, or disengage. These insights can inform more targeted advising, curriculum adjustments, and instructional strategies by helping identify misalignments between student expectations and course structures.

Synthetic data provides an ethical way to prototype new approaches; however, this approach has limitations. For example, it cannot fully capture the linguistic ambiguity, noise, and variability found in authentic student responses. Additionally, LLMs may introduce subtle biases during categorization, particularly when motivations are less clearly expressed. These limitations highlight the need for rigorous validation and human oversight during real-world deployment.

8. Anticipated Outcomes and Significance

This project is expected to produce two primary deliverables: (1) a replicable pipeline for collecting, structuring, and analyzing open-ended student motivation responses using the Expectancy–Value Theory and LLMs, and (2) a set of stakeholder-specific insights that illustrate how motivational patterns vary across student populations, course types, course levels, and performance outcomes.

The broader impact of this work lies in expanding learning analytics to incorporate unstructured, student-generated motivational input—an often overlooked but valuable source of insight. By systematically integrating these motivational inputs into existing analytics frameworks, institutions can gain a deeper, more contextualized understanding to support proactive, personalized interventions.

9. Next Steps

Upon IRB approval, the full data pipeline will be applied to a real dataset to validate and refine model outputs, with special attention to generalizability and robustness across subjects and course levels. Evaluation will occur along two dimensions: (1) technical performance (e.g., feature extraction consistency, classification recall) using motivational data; and (2) stakeholder utility, measured through structured feedback from instructors, advisors, and institutional research staff on the usefulness and interpretability of the outputs.

Looking ahead, the project has several potential extensions. These include incorporating temporal analysis to track shifts in motivation over time and adapting the pipeline for other open-ended data (e.g., post-semester reflections).

10. Discussion Points

As institutions consider using LLMs to analyze unstructured student input, questions arise around appropriateness, ethics, and institutional readiness. Even with offline or self-hosted LLMs, concerns related to privacy, transparency, and responsible use remain. Given that many institutions adopt a cautious stance toward LLMs, we seek input on how others have navigated internal policies, stakeholder concerns, and approval processes for LLM-based analysis of open-ended student inputs.

We also welcome feedback on refining the classification of motivational constructs. Are there alternative or hybrid theories that could enhance the operationalization of student motivation across disciplines? Finally, we invite lessons from practitioners who have incorporated free-text student input into learning analytics. What technical or organizational barriers were faced, and what strategies proved effective in addressing them?

11. Conclusion

This POC demonstrates how open-ended student motivation statements can be transformed into structured data using LLMs and the Expectancy–Value Theory. Synthetic data experiments revealed how motivational patterns vary across subjects, course levels, and performance outcomes, providing actionable insights for targeted interventions and curricular refinement. While motivational traits alone may not be the strongest predictors of course outcomes, they offer meaningful context to complement behavioral indicators. This project presents a replicable model for integrating motivational insights into advising, instruction, and program design. As institutions increasingly prioritize human-centered analytics, this work-in-progress underscores the value of incorporating student voice at scale.

Future phases of the project will focus on validating the pipeline in real-world contexts, with attention to model robustness, stakeholder relevance, and responsible implementation. The vision is to develop a scalable, theory-informed framework for transforming unstructured student feedback into actionable insights that support institutional decision-making.

References

de Barba, P. G., Kennedy, G. E., & Ainley, M. D. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning, 32*(3), 218–231. https://doi.org/10.1111/jcal.12130

Eccles, J. S., & Wigfield, A. (2000). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Howard, J. L., Bureau, J., Guay, F., Chong, J. X., & Ryan, R. M. (2021). Student motivation and associated outcomes: A meta-analysis from self-determination theory. *Perspectives on Psychological Science, 16*(6), 1300–1323. https://doi.org/10.1177/1745691620966789

Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development, 68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Kember, D., Ho, A., & Hong, C. (2008). The importance of establishing relevance in motivating student learning. *Active Learning in Higher Education, 9*(3), 249–263. https://doi.org/10.1177/1469787408095849

Lo, K. W., Ngai, G., Chan, S. C., & Kwan, K. P. (2022). How students' motivation and learning experience affect their service-learning outcomes: A structural equation modeling analysis. *Frontiers in Psychology, 13*, 825902. https://doi.org/10.3389/fpsyg.2022.825902

Okada, R. (2023). Effects of perceived autonomy support on academic achievement and motivation among higher education students: A meta-analysis. *Japanese Psychological Research, 65*(3), 230–242. https://doi.org/10.1111