National College of Ireland

LLMs for Structured Educational Data: A Zero and Few-Shot Approach to Student Performance Prediction

MSc Research Project
MSCAI1

Karandeep Singh Mann
Student ID: x23295945

School of Computing
National College of Ireland

Supervisor: Dr. Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Karandeep Singh Mann
Student ID:	x23295945
Programme:	MSCAI1
Year:	2025
Module:	MSc Research Project
Supervisor:	Dr. Abdul Shahid
Submission Due Date:	15/09/2025
Project Title:	LLMs for Structured Educational Data: A Zero and Few-Shot Approach to Student Performance Prediction
Word Count:	XXX
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

LLMs for Structured Educational Data: A Zero and Few-Shot Approach to Student Performance Prediction

Karandeep Singh Mann
x23295945

Abstract

Early and accurate prediction of student performance is crucial for enabling timely academic support to the at-risk students. The past work have used machine learning and deep learning approaches on educational datasets such as OULAD and achieved strong results for prediction of student outcomes. But they relied on feature engineering, required retraining and did not offered any interpretability. This study tries to check whether large language models (LLMs) can be used to overcome these limitations by using zero and few-shot prompting strategies on structured tabular datasets. Evaluation has been done on three open-source LLMs across binary, multiclass and per-class settings. The models reached up to accuracies of 89.6% and 75.9% on binary and multi-class predictions respectively, all without the need for fine-tuning. They even produced short natural language explanations to describe why a particular prediction was made. The findings show that LLMs can not only become an alternative to the traditional prediction models but can also guide educators to make informed decisions through their explanation based insights.

1 Introduction

Predicting student performance is one of the important aspect of learning analytics especially in online education. It allows educators to early identify at-risk students and provide any necessary support to these students to improve their academic results (Torkhani[1] and Rezgui; 2025). Studies on Massive Open Online Courses (MOOCs) have showed course completion rates to be as low as 22%, which implies that dropout rates to be 78% (Jha et al.; 2019). Such figures have led researchers to look into data-driven approaches so as to analyze and support student academically. A huge amount of data is being generated with the rise of Virtual Learning Environments (VLEs). This data can range from their login information, clickstream information and their assessment submissions. But it isn't easy to learn meaningful insights from this data as it is often unstructured and varies a lot between courses.

In response, a number of data-driven approaches have been devised in past few years to predict student outcomes. Early approaches have made the use of conventional ML algorithms like Logistic Regression and Decision trees to predict student outcomes (Torkhani[1] and Rezgui; 2025). They used features like demographic information, assessment records and platform activity for training. More recently, deep learning models such as CNN, LSTM etc have been in use to capture the temporal patterns of student VLE data. Graph-based and ensemble based approaches have also shown high predictive accuracy to find at-risk students (Zhao et al.; 2025; Huang and Zeng; 2024). Overall, all these techniques have shown strong performances to early identify any struggling student.

However, there are some limitations with these approaches. First, most of the models focus mainly on binary predictions which is either pass/fail or at-risk/no-risk (Al-Azazi and Ghurab; 2023) and overlook other categories such as Withdrawn and Distinction. Second, these models are typical black-box systems which do not provide any kind of explainability to educators as to why a particular student is predicted as Fail. Finally, most of the models are specific to courses and generally prone to overfitting (Al-Azazi and Ghurab; 2023). They require retraining before they can be used on new courses that restricts their scalability.

To tackle these challenges, there have been advancements in finding alternative solution which is Large Language Models. There are LLMs such as GPT, LLaMa etc which have shown strong performances across domains that includes education sector when it comes to tasks like reasoning and classification (Severino et al.; 2025; Cohn et al.; 2024). Unlike traditional methods, they take natural language text as a form of input and generates meaningful results and predictions as an output using zero-shot and few-shot prompting. This opens the possibility to pass student profile information and their activity data into textual summaries to LLMs and predict their outcomes as an output from LLMs. Some studies have already started their research in this direction. They have shown how LLMs could be used to achieve high level performance on complex tasks such as the capability to grade student responses or predict student outcomes: Pass/Fail or at-risk/no-risk or dropout/no-dropout (Diyab et al.; 2025). Moreover, LLMs are also able to reach the state-of-the-art results on many benchmark datasets for classification as well as reasoning tasks for explainability (Fang et al.; 2024). However, most of these work in educational sector are limited to binary classification and there is a little work done to explore multi-class classification in datasets such as OULAD dataset.

The main objective of this presented study is to understand whether LLMs can be used to predict student outcomes especially in zero-shot and few-shot settings. Rather than relying on feature engineering, this study checks how LLMs can process the textual summaries of the student data to draw predictions and provide good reasoning to those predictions. The focus of this work is on the classification problem based upon the educational benchmark OULAD dataset where students are classified in one of the four categories: Withdrawn, Fail, Pass and Distinction. Accordingly, the study will be answering two major research questions as follows:

1. *To what extent can LLMs be used to predict student performance using zero-shot and few-shot prompting on the OULAD dataset compared to the traditional ML approaches?*

2. *How effective is the reasoning generated by the LLMs aligns with actual factors affecting student performance and if it can help educators understand those predictions?*

To build upon the above-mentioned objectives, this study tries to offer the following contributions to the field of LLM based student performance prediction.

- It tries to show the potential of LLMs to predict student outcomes by using zero and few-shot prompting without any dependency on feature engineering.

- It will also be among the first to apply this particular approach on the OULAD dataset to do a multi-classification task and go beyond binary predictions.

- It also creates an interpretability layer to explain each of its prediction. This will incorporate transparency and build trust of the end users in these AI systems.

The remainder of this study is structured in a following way: Section 2 includes the discussion of the past work which has been done related to predicting student performance. Section 3 provides details of the proposed methodology which includes data preparation, choice of models etc. The design specifications for the system such as prompt design, summary generation is discussed with Section 4. Section 5 gives the details of the implementation such as the technical pipeline and the response handling. Section 6 includes the experimental results of work done, compare the performance of LLMs to traditional methods and assess the reasoning that LLMs provide. Finally, the conclusion, limitations and future work for this study has been explored in section 7.

2 Related Work

Over the past years, many studies have used machine and deep learning techniques on OULAD and similar datasets to early predict failure or dropouts. Recently, LLMs have opened a new dimensions where prompt based learning can also be used to make predictions and explanations. This literature review discusses on these developments, highlighting about the achievements and limitations of these studies. It also discuss how this research builds upon these studies.

2.1 Traditional ML and DL approaches

Early studies have shown usage of machine and deep learning approaches on the OULAD dataset for predicting student performance. For instance, Torkhani[1] and Rezgui (2025) conducted a study where they tested performance of the standard Machine Learning models such as Logistic Regression, Random Forest, SVM etc alongside Deep Learning models such as CNNs and LSTMs. Among all the models, LSTM gave the best results with an accuracy of 83.41% to identify at-risk students. Moreover, a notable performance was shown by Random Forest model, specially in terms of recall score. Building on, He et al. (2020) developed a hybrid RNN-GRU model. It combined the weekly VLE logs of students with their demographic features. It was found that the model achieved accuracy of more than 80%. It even performed better than standalone LSTM and GRU model to early predict at-risk students by the end of semester. Even though these work proved the importance of temporal patterns analysis, but were limited to binary predictions. Adding to this, Adnan et al. (2021) performed a study using OULAD to predict performance at different stages of course timeline (0% to 100%). Their work used many models such as Random Forest, AdaBoostClassifer etc to perform multi-class (72.3%), binary (92%) as well per-class (65-74%) classifications. Notably, their results have showed improvement in accuracy when there more student activity data is available throughout the course.

Later on, researchers tried to do multi-class classification using the OULAD dataset. They classifed the result into one of the four labels: Distinction, Pass, Fail and Withdrawn. For example, Shou et al. (2024) introduced a time-series model called MTAPSP. They combined assessment, demographic and VLE data of students to form a dataset. 74% accuracy and a F1-score of 73% was achieved by their model for four-class predictions. In addition, their model was capable to flag at-risk/no-risk students with 99% accuracy. Similarly, Al-Azazi and Ghurab (2023) proposed a hybrid ANN-LSTM architecture that classified students in three performance risk bands: high, medium and low using their activity data. Their model achieved 70% accuracy by mid completion of course. Together, these studies show importance of student engagement on learning platforms as a factor to predict their performance.

Going beyond the attention-based models, Huang and Zeng (2024) made a dual GNN structure. It combined two different graphs: an interaction graph which captured student's online activity and attribute graph which was based on similarities among features. By using the architecture, their model got 83.96% accuracy for binary Pass/Fail prediction and 90.18% accuracy for Pass/Withdraw prediction. Such accuracies were the state-of-the-art under same conditions.

For boosting the performance, many studies have tried using ensemble techniques as well. For instance, Zhao et al. (2025) took temporal features from the dataset with the help of CNNs and fed them into a RF-SVM ensemble model. They were able to attain 98% accuracy score to determine binary outcomes: Pass/Fail. They also did an experimentation where they combined multiple decision trees with logistic regression to predict student outcome. It also achieved 98% accuracy which was approximately 4% more than the baselines. These results indicate that ensemble methods can be used to get more accurate results from the models when applied to a dataset like OULAD.

Even though, these studies showed incredible performance, most of them took this problem as binary classification (at-risk/not or Pass/fail). They do not discuss on the multi label outcomes of the OULAD dataset. Moreover, there was no explainability and transparency to explain the model's behaviour. This is where LLMs comes into the picture.

2.2 Emergence of LLMs

The rise of large language models have completely transformed the field of Natural language processing. It is because, LLMs have the ability to understand and produce text in a natural human-like way. They are models that are pre-trained using diverse text corpora. Recent work have been trying using LLMs in educational domain such as for grading, providing feedback and doing score predictions. For instance, Lee et al. (2024) gathered around 1650 responses from science assessments of middle school. They tried to assign grades to the responses through six different prompting methods on GPT-3.5 and GPT-4 models. It was found that few-shot did better than zero-shot in terms of accuracy score (67% vs 60%). With the chain-of-thought reasoning, the performance of LLMs also went up (+13.4% in zero-shot, +3.7% in few-shot). Apart from accuracy, textual justifications were also provided by the LLMs for its predictions. The research points to the transparency of LLMs and their performance abilities. In a similar manner, Diyab et al. (2025) developed a system known as AI Assess, which was based on ChatGPT. It was designed to evaluate performance of the students in Software Engineering and CS courses. The system could do four things: assign grades automatically, generate feedback, create questionnaires and identify any weakness among students. They applied zero-shot and few-shot prompting via LMQL to evaluate the system on 25 students. The experimentation showed that zero-shot gave 16% accuracy for grading while few-shot technique gave 80% accuracy. The results which were non-matching only had a maximum of 5% difference. The results depict how examples in a prompt design play a major role to improve the performance of LLMs.

In a wider predictive context, Neshaei et al. (2024) used LLMs to check if students will be able to answer questions correctly or not based upon their last performances. Three different datasets were used for this task: Statics, ASSISTment 2009 and ASSISTment 2017. They applied zero-shot prompting technique on GPT-3.5 and tested performance of the LLM

against performance of some fine-tuned LLMs (BERT, GPT-2 and GPT-3) alongside a few old approaches such as BKT, DKT and Logistic Regression. The experimentation found that zero-shot prompting gave an AUC score of 0.5 similar to traditional methods while fine-tuned models were able to achieve an AUC score in the range of 0.68-0.71. Limitations of zero-shot strategy in tasks when no contextual information is available is demonstrated in the study.

LLMs have also shown explainable nature with the usage of prompting techniques. A study done by Cohn et al. (2024) explored if GPT-4 can be used to score and provide explanation to the responses obtained from middle school science assessments. They used chain-of-thought reasoning together with few-shot prompting and active learning to improve accuracy as time goes by. On simple concepts, the method gave an accuracy up to 100% but achieved 81-85% with complex ones across three questions. Notably, the approach did not only assign scores but gave detailed explanations and proper feedback. Hence, LLMs can go beyond predictions by showing transparent nature.

Finally, there have been hybrid models that integrated LLMs to some other learning strategies. Oh et al. (2024) developed a system known as LMgMF (Language Model-guided Matrix Factorization). It was a combination of LLMs with matrix factorization to predict student performance by using both scores as well as course description. They converted the multi-modal information in sentences and sent them as input into LLMs such as GPT-J and LLaMA 2. The prediction models made use of embeddings generated by LLMs for the task of performance prediction. Analysis was done on i-ScreamEdu and ASSISTments 2009 datasets and it showed that LMgMF did better than standard MF (82.2% to 83.0% on one dataset, and 74.26% to 74.76% on other one respectively). It also showed a gain of 6% in case of cold-start scenarios. The study depicted how black box LLMs can be used without knowing its internal parameters for educational tasks.

2.3 LLMs for Tabular Data

While the above mentioned studies have shown promising results from LLMs in educational domain, most of them focus on question-answering tasks and structured responses. There is still very less work that has been done on structured tabular data using LLMs. This gap motivated Hegselmann et al. (2023) to introduce a framework called as TabLLM, which applied LLMs on tabular data. The approach transformed each record into natural language and then the text was sent as an input to LLMs for classification task. The experimentation was done on nine different benchmark datasets. Not only, TabLLM achieved an AUC score more than 0.6 for several datasets in a zero-shot setting but got up to 98% accuracy in few-shot setting (128 shots) for some datasets. The approach performed better than some of the state-of-the-art methods such as XGBoost, LightGBM etc. They also tried nine different serialization strategies and proved that design of the prompt strongly affects the accuracy of the LLM.

Building upon the work done for TabLLM, Fang et al. (2024) did a review of over 90 studies which applied LLMs on tabular data. They found that LLMs were able to achieve accuracies in the range of 66% to 70% for some standard benchmark datasets such as "Blood". The results of LLMs were as good as performance of state-of-the-art methods like XGBoost. Their survey also showed how LLMs worked well with small number of examples in case of few-shot prompting, but performance decreased as number of examples increased. There were some challenges working with LLMs that were pointed in the study such as data had to be precisely

serialized and taking care of token limits. The insights show that LLMs have huge potential and accuracy of LLMs can be significantly improved with better prompts design.

Study	**Dataset**	**Approach / Model**	**Features**	**Results**	**Limitations**
Torkhani & Rezgui (2025)	OULAD	Logistic Regression, RF, SVM, CNN, LSTM	VLE activity logs	LSTM: 83.41% accuracy	Binary classification, No explainability
He et al. (2020)	OULAD	Hybrid RNN-GRU	VLE logs + demographics	more than 80% accuracy	Binary Classification, No explainability
Shou et al. (2024)	OULAD	MTAPSP (Attention Time-series)	VLE + assessments + demographics	74% accuracy, 73% F1 Score	No interpretability
Al-Azazi & Ghurab (2023)	OULAD	Hybrid model of ANN-LSTM	VLE activity data	70% accuracy mid-course	No explanations
Adnan et al. (2021)	OULAD	Random Forest, AdaBoostClassifier, ExtraTreeClassifier, KNN, SVM	VLE, Assessment and Demographic data	Multi-class: 72.3%, Binary Class - 92%, Per-class: 65-74%	Limited interpretability, Lower Accuracy at early stages
Huang & Zeng (2024)	OULAD	Dual Graph Neural Network	Graphs of activity and attributes	83.96% (Pass/Fail); 90.18% (Pass/Withdraw)	Binary classification, complex to scale
Zhao et al. (2025)	OULAD	CNN + RF-SVM Ensemble	Temporal features extracted by CNN	98% accuracy	Binary classification, No Interpretability
Lee et al. (2024)	Science Assessments	GPT-3.5, GPT-4	Zero-/few-shot, CoT prompting	Few-shot 67% vs Zero-shot 60%	No tabular data, Small Domain
Diyab et al. (2025)	CS/SE Courses (25 students)	ChatGPT (AI Assess)	Zero-/few-shot LMQL prompting	Few-shot: 80% vs Zero-shot: 16% accuracy	Small-scale test, Structured responses only
Neshaei et al. (2024)	Statics, ASSISTments	GPT-3.5 Zero-shot vs Fine-tuned LLMs	Zero-shot prompting	Zero-shot AUC: 0.5; Fine-tuned AUC: 0.68–0.71	Poor performance with zero-shot

Study	Dataset	Approach / Model	Features	Results	Limitations
Cohn et al. (2024)	Science Assessments	GPT-4 CoT + Few-shot	Active learning, CoT prompting	100% accuracy (simple tasks); 81–85% (complex tasks)	Limited to structured text responses
Oh et al. (2024)	i-ScreamEdu, ASSISTments	LMgMF	LLM embeddings + Matrix Factorization	83.0% (first dataset); 74.76% (second dataset)	Not tabular education data
Hegselmann et al. (2023)	9 tabular datasets	TabLLM framework	Zero- and Few-shot Prompting	AUC >0.6 (zero-shot); 98% (few-shot)	No temporal data, Prompt design sensitive
Fang et al. (2024)	Review of 90+ studies	Survey of LLMs on tabular data	Few-shot vs many-shot prompting	Accuracy: 66–70% (Blood dataset)	Token limits, Serialization needed

Table 1: Summary of the studies related to student performance prediction

2.4 Current Research

The current research also makes use of LLMs to understand temporal patterns of educational data such as in OULAD dataset which hasn't been explored in above-mentioned studies. They also lack domain specific prompting to handle classification on number of different categories. Building on this path, this study makes use of LLMs to perform any prediction or classification on the OULAD dataset. It changes the temporal information and student activity data into a natural language format. It then checks if the information can be used with prompting designs to predict performance for students. Unlike other studies, it tries to do multi-class classification: Withdrawn, Fail, Pass and Distinction.

3 Methodology

The primary goal of this study is to find the effectiveness of LLMs to predict student performance using zero and few shot prompting. The task includes classification into four categories: Pass, Distinction, Fail and Withdrawn based upon behavioral and assessment data of students. Natural language summaries of the student data is derived from the "Open University Learning Analytics Dataset (OULAD)" and these summaries are sent as input to three different LLMs. Figure 1 shows the overall architecture for the research study.

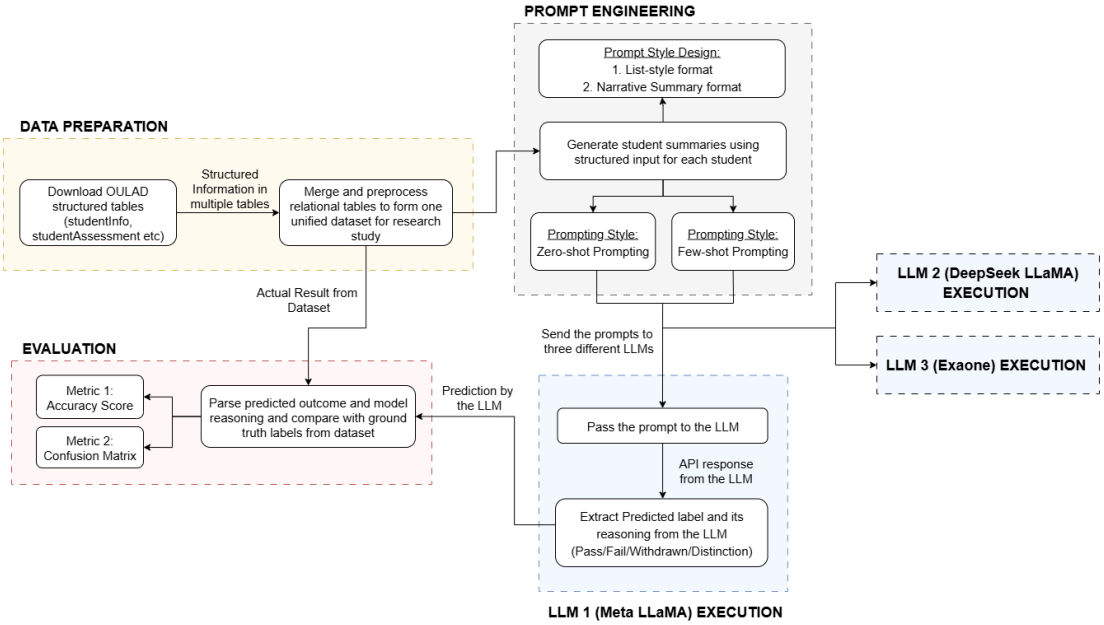


Figure 1: Overall architecture of the research

3.1 Prompting Techniques

The study uses prompt-based strategies to perform 4-class classification of student performance without any prior training. Two popular techniques: zero-shot and few-shot prompting are used to evaluate the capabilities of LLMs.

3.1.1 Zero-Shot Prompting

Within zero-shot prompting, a task description and an input text is provided to the model without giving any labeled examples (Sivarajkumar et al.; 2024; Sivarajkumar and Wang; 2023). The prediction is totally dependent on model's internal knowledge and their instruction-following ability. Prior work (Kojima et al.; 2022) have shown that LLMs have been able to perform complex reasoning tasks without any need of any context. Zero-shot in this study basically checks if LLMs can analyze the student summaries and generate predictions based on its own knowledge.

3.1.2 Few-Shot Prompting

In few-shot prompting, models is given description of the task along with a number of input-output examples without any fine-tuning. This helps the model to fully understand the task structure, expected output format and any relevant input patterns (Min et al.; 2022). This is helpful for prediction of student performance where certain behavioral differences can be found through guided examples. Past research have shown that few-shot prompting can enhance the reasoning capabilities of LLMs which results in higher performances of the model in complex tasks (Wei et al.; 2022).

3.2 Data Source and Preparation

The dataset used in this research study is "Open University Learning Analytics Dataset (OULAD)". It is a large-scale dataset which has been published for educational research by the Open University in UK (Kuzilek et al.; 2017). It contains records of 32593 students

who are enrolled in various modules from 2013 to 2014. The original dataset has multiple relational tables that contains a variety of information: Student Demographics (studentInfo), Course Registrations (studentRegistration, courses), VLE Activity Logs (studentVle, vle) and assessment performance (studentAssessment, assessments). All the tables are linked via primary keys such as id_student. code_module etc, as shown in the ER diagram (Figure 2).

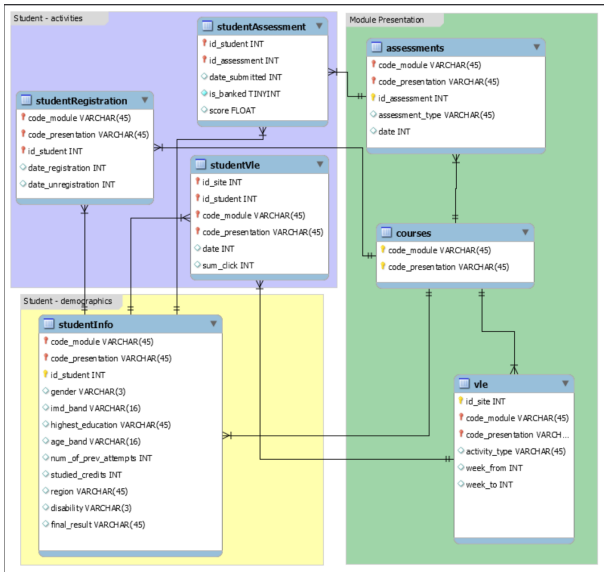


Figure 2: Entity-Relationship diagram of the OULAD dataset

3.2.1 Final Dataset Preparation

The final unified dataset used for this research is shown in Table 2:

Feature(s)	**Source**
`id_student`, `age_band`, `gender`, `disability`, `studied_credits`, `code_module`, `region`, `num_of_prev_attempts`, `highest_education`	Demographic and Academic background details are taken directly from the `studentInfo` table
`date_registration`, `date_unregistration`	Extracted directly from the `studentRegistration` table
`month_1_sum_clicks` to `month_10_sum_clicks`	Student interaction logs are build from `studentVle` table via grouping by month to build ten features. Each feature represents total number of clicks in a given month during the whole course.
`missed_assessments`	Calculated by counting all assessment records in `studentAssessment` where the score is null.
`weighted_avg_score`	Calculated as the weighted average of scores from all the submitted assessments.
`final_result`	It is the target label field from `studentInfo` table and is used as ground truth having four classes: Pass, Fail, Withdrawn and Distinction.

Table 2: Feature Description and their sources

The OULAD dataset is relevant for this research as it contains demographic, assessment and online activity details for over 32000 students. Its rich and structured information makes it easier for conversion of the data into descriptive behavioral summaries. Furthermore, this dataset has been used in educational data mining research for many tasks such as predicting dropout rates, at-risk students, and performance etc (Jin et al.; 2024).

3.3 Choice of models

Three different LLMs have been selected within this study for performing the task of student prediction using zero-shot and few-shot prompting. This study makes use of LLMs that not only are open-source but also accessible through APIs. Furthermore, all experiments with LLMs have been conducted on personal device with internet access and no fine-tuning is performed for this research.

3.3.1 Meta LLaMA 3

Meta's LLaMa 3.3 is an open-source and a state-of-the-art LLM that is made of 70 billion parameters. It is selected for this research because of its performance on benchmark evaluations:

- **Instruction following:** The instruct variant of LLaMA 3 is fine-tuned specifically for tasks based upon instructions. According to the technical report by Meta (Dubey et al.; 2024), it has outperformed other models of similar size on instruction-based benchmarks such as IFEval. Thus, it becomes useful for the OULAD dataset in this study that needs prompt-based classification.
- **Advanced Inference:** In a medical question-answering study (Severino et al.; 2025), LLaMA 3 got 77.5% success rate and it was the highest among all other tested open source models. The model not just only overtook the average score of human performance but also competed with top models such as GPT-4o and Claude. As predicting student performance requires understanding activity and assessment patterns, LLaMA 3 will be helpful because of its strong inference-based reasoning tasks.

3.3.2 DeepSeek R1 Distilled LLaMA

DeepSeek R1 Distilled is a reasoning-based LLM which has been derived from LLaMA 70B. Of all the chosen models, this LLM is specifically chosen because of its reasoning capabilities to justify student performance predictions.

- **Chain-of-thought Reasoning:** DeepSeek R1 naturally generates step-by-step explanations while making its predictions (Moëll et al.; 2025). This helps to fully understand the "why" behind a particular classification made. Based upon the explained output, teachers can take appropriate actions making this LLM ideal for teacher-facing tools.
- **GPT-4 Level performance:** A peer-reviewed study (Chan et al.; 2025) showed that DeepSeek R1 achieved near GPT-4 level state-of-the-art performance in an inference based medical diagnosis task. This shows the model's strong inference capabilities and since the OULAD dataset contains behavioral patterns which need interpretative reasoning, this model will be relevant for this research.

3.3.3 Exaone 3.5

Exaone 3.5 is a large language model which is developed by LG AI Research. It is made of 32 billion parameters and is selected for this study because of its strengths as follows:

- **Instruction tuned:** Exaone has been fine-tuned for instruction based tasks. It ranked among top models such as LLaMA-2 and Falcon etc when evaluated on seven instruction-based benchmarks like BBH, MT-Bench etc (An et al.; 2024). This makes it suitable for prompt based classification within this research.
- **Strong Reasoning:** On academia based reasoning tasks such as GSM8K, CSAT etc, Exaone outperformed models such as Google's Gemma 27B An et al. (2024). This makes it useful to understand the patterns in the OULAD dataset and reason its predictions.
- **Long-context support:** Exaone has a token limit of 32k and has achieved better performance in long context tasks such as LongBench, NIAH etc if compared to the models within its size class. (An et al.; 2024). This shall be useful for this research as OULAD dataset contains lengthy clickstream logs and assessment records.

3.4 Evaluation Strategy

Several measures are present to check the performance of LLMs. A single metric, Accuracy Score has been primarily used within this study to analyze the performance of models. Accuracy Score can be used to find the total correct predictions that is made by the model out of total number of records present in the dataset. This will be useful as it provides clear and an interpretable measure to compare different models and prompting strategies.

4 Design Specification

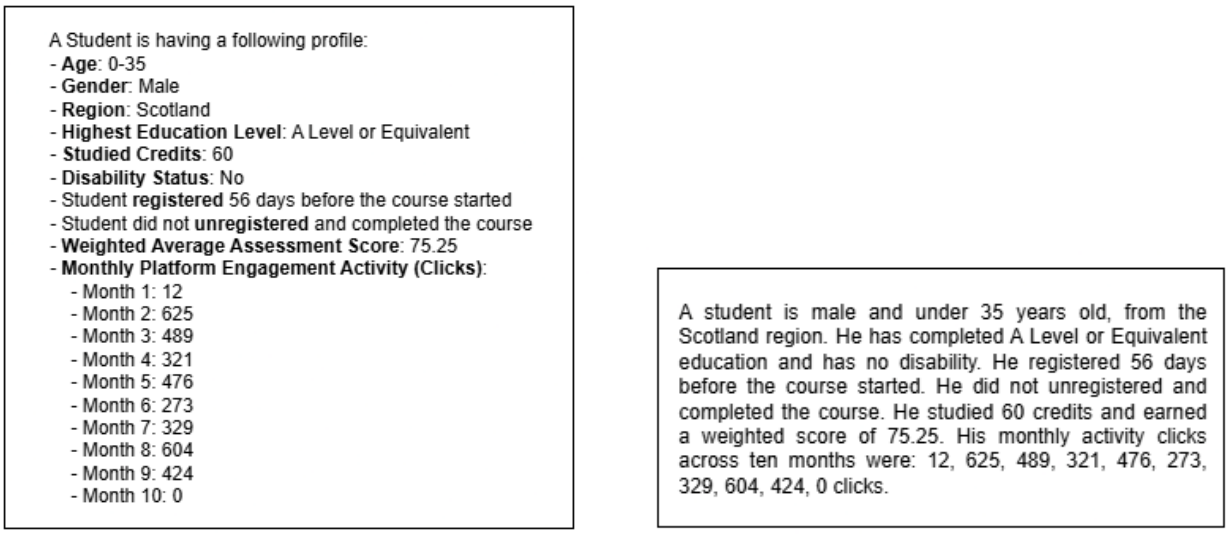
The focus of this section is to discuss how the system is designed such that LLMs are able to predict the student outcomes using structured tabular data. This is done by transforming tabular data into description natural language summary which further is processed using prompting techniques. This section outlines the design logic behind data transformation, prompt creation, APIs and prediction handling that altogether lays the base of this research.

4.1 Student Summary Generation

The structured student records from OULAD dataset are converted into text format to be sent as inputs to LLMs. There are two different summary formats which have been used within this study: List-Style format and Narrative Summary format.

4.1.1 List-Style format

The list-style format presents the information in a structured and bullet-point style with labelling done for each field. This allows LLMs to easily interpret and understand the true meaning of each and every value. The goal of this style is to preserve tabular structure of the student information but still allow the language based interaction with LLMs. Figure 3 (a) shows an example of transformation of one student record into list-style format.



(a) List-style prompt design

(b) Narrative-style prompt design

Figure 3: Different Prompts Design

4.1.2 Narrative Style format

The narrative-style format presents the information of students in a free-flowing paragraph. Unlike the list-style format, this format do not use field-value labels but rather embeds information within coherent sentences. The resultant summary resembles human written description which will align closely with the type of data LLMs are usually trained on. The goal is to see whether LLMs can understand behavioral patterns and relationships among attributes when the input is given in a natural language which is used by humans. Figure 3 (b) shows an example of transformation of one student record into narrative-style format.

4.2 Prompting Design and Prediction Logic

There are two prompting techniques: zero and few-shot prompting that have been used within this study. A specific logic has been used for both strategies so as to accurately classify across four classes: Pass, Withdrawn, Distinction and Fail.

4.2.1 Zero-shot Prompting

Within zero-shot prompting, the LLM is given task description along with the summary of a student (list-style or narrative-style) without any examples. The prompt asks the LLM to use its own knowledge and classify the student into of the four classes. For this study, a definition for each class is included within the prompt given to LLM. This is done to have clear picture of what each category meant actually. The model is also asked to give the response with only a single label prediction.

MAIN PROMPT

You are a classification assistant. Given this student profile:

<student summary (list-style/narrative-style)>

Definitions:
- **Pass**: The student completed the course and demonstrated an acceptable level of performance and engagement
- **Fail**: The student completed the course but did not meet academic requirements. This may include low performance, gaps in engagement, or factors such as prior education, late registration, or other challenges.
- **Withdrawn**: The student disengaged early and did not complete the course. This is indicated by a withdrawal date or a pattern of minimal participation, possibly influenced by personal circumstances.

What will be the student's final result? Respond with only one: Pass, Fail or Withdrawn.

FOLLOW-UP PROMPT

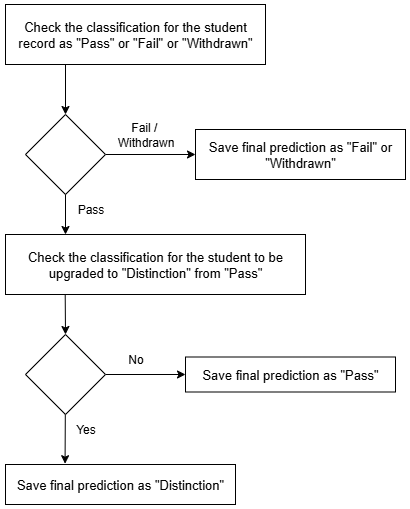
The student profile was previously classified as Pass. Based on their full profile:

<student summary (list-style/narrative-style)>

Would you upgrade this to Distinction based on exceptional performance and engagement?

Respond only with only one label: "yes", "no".

(a) Main and follow-up prompt for zero-shot prompting



(b) Sequential prompting structure for all classes

Figure 4: Zero-shot Prompting Design and flow

However, there is some complexity with "Distinction" class of dataset. It was observed during the initial trials that model got confused between "Pass" and "Distinction" because of their shared characteristics. Hence, a two step approach, also known as chain-of-thought prompting approach, is used to address the concern. First, the LLM is asked to classify across three classes: Pass, Fail and Withdrawn. If model predicts the student record as "Pass", a binary follow-up prompt is given to the LLM to check whether the student should be classified as "Distinction" or not. This helps the LLMs to accurately classify for high-performing students. This step helps the model to differentiate between "Pass" and "Distinction" based upon many indicators such as high scores and good amount of engagement. A sample of zero-shot prompting message is shown in Figure 4.

4.2.2 Few-shot Prompting

If compared to zero-shot prompting, few-shot is more of a guided approach. In this design, an individual prompt message is created for each class which contains set of four input and output examples. Two examples will belong to one class (ex: Fail) and other two will not belong to that class (ex: Pass/Withdrawn). The student summary (list-style or narrative style) is appended after the examples and the LLM is asked to respond by carefully studying the provided examples.

A sequential binary query structure is followed for this prompting technique. The LLM is first prompted with examples and asked if student is "Withdrawn" or "No Withdrawn". If response from the LLM is negative, the model is asked for "Fail" in a similar manner. If response is still negative, the student record is checked for "Distinction" or "No Distinction". The student is finally marked as "Pass" by default, if no match is found.

You are a classification assistant.

Example 1: A student is female and between 35 and 55 years old, from the North Western Region region.
She has completed A Level or Equivalent education and has a disability.
She registered 92 days before the course started. She unregistered 12 days after the course started.
She studied 60 credits and earned a weighted score of 0.0.
Her monthly activity clicks across ten months were: 102, 179, 0, 0, 0, 0, 0, 0, 0, 0 clicks.
Outcome: Withdrawn

Example 2: ...
Outcome: Withdrawn

Example 3: ...
Outcome: No Withdrawn

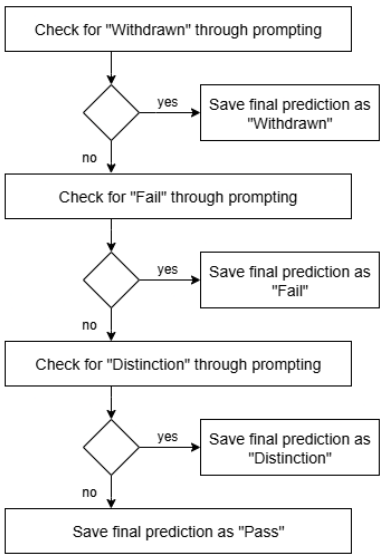
Example 4: ...
Outcome: No Withdrawn

Now evaluate the following student and predict their likely outcome.

<student summary (list-style/narrative-style)>

Respond with only one: "Withdrawn" or "No Withdrawn"

(a) Few-shot prompting example for Withdrawn class



(b) Sequential prompting structure for all classes

Figure 5: Few-shot prompting design and flow

This particular approach is chosen in the aim to achieve high accuracy. During the initial trials, all the classes were combined in a single prompt with a few examples. But, it was found that LLM struggled to make the right prediction because the number of examples for any class was not enough. A sample prompt message only for few-shot prompting for only the "Withdrawn" scenario is shown in Figure 5.

4.3 LLM Interaction Design

LLMs used within this study are online and open-source LLMs and there is need for APIs to interact with LLMs. Hence, all interactions with LLMs within this research is done through the Together.ai API platform. It is selected because of its support for many free public LLM models and high context limits for them. All LLM queries for each student record has been handled in a sequential manner so as to avoid any potential rate limits. There are some basic validation checks implemented in the system to ensure that the responses from LLMs matches the expected format of output.

5 Implementation

This section discusses on how the system has been actually implemented. It covers the tools and technologies that have been used, the execution of data processing pipeline and how the responses were handled and stored.

5.1 Tools and Technologies Used

5.1.1 Programming Language

Python 3.12 programming language is used within this project. This is because of its flexibility and strong support for data processing.

5.1.2 Libraries

Various libraries of python has been used in this project for handling the data and doing analysis on it. Dataset loading and feature engineering has been done using `Pandas` library. The `Numpy` library helped in numerical operations. `Seaborn` and `Matplotlib` library are used for creating visualizations for the results. `Sklearn` library is used to calculate the accuracy score and to build confusion matrix. API calls to communicate with LLMs have been done using the `Requests` library.

5.1.3 Development Environment

Jupyter Notebook has been used for implementation for all the code including debugging and testing within this project.

5.1.4 API Platform

Together.ai is selected for this research to interact with the LLMs. It is selected because of the free access of APIs it provides to access the LLMs and the simplicity of its interface to generate API key.

5.2 Data Storage and Management

All outputs which have been generated that includes both prediction and reasoning are captured and stored safely. These results lays the base to calculate how much effective each LLM is.

In parallel, both list-style and narrative-style behavioral summary for each student record is also stored. This is done in order to ensure transparency as well as to trace back for future reference.

Each output obtained from LLMs is saved within a CSV file. Each file represents a combination of one LLM and respective prompting strategy used. This storage mechanism easily allows to evaluate LLMs using accuracy score and building confusion matrix.

5.3 Output Validation

After receiving response from LLMs, the output is validated using some checks to ensure smooth and a consistent execution. For instance, after each prompt, the system confirmed whether the response from LLM is well-formed, followed the expected structure and returned a clear prediction label and reasoning to it.

Any incomplete or malformed response were skipped to avoid occurrence of errors at run time. This ensured that processing of other student records is not interrupted because of one faulty response from the LLM. Moreover, all errors which occurred were also logged in real-time. This supported faster debugging process and maintain transparency in the system. Overall, the validation checks implemented makes the system fault-tolerant and robust enough to handle different kind of output from LLMs.

6 Evaluation

6.1 Predictive Accuracy of LLMs across prompting techniques

This section of the study contains evaluation of LLMs to predict performance of students. It uses three LLMs for this purpose: Exaone, Meta LLaMa and DeepSeek LLaMa and checks against both zero and few shot prompting. The comparison is done on three tasks, i.e. binary classification, multi-class classification and single class classification. The performance of LLMs is compared to performance of the ML approaches which have been done in the past.

6.1.1 Multi-Class Classification

Multi-class classification means that the student records will be classified in more than 2 categories. In this scenario, it is 4-class classification: Pass, Fail, Withdrawn and Distinction. The accuracies achieved by LLMs in case of multi-classification task is depicted in Table 3. In addition, a comparison with the traditional approaches is shown through a bar graph in Figure 6.

Table 3: Overall Multi-Class Accuracy of LLMs (Compared across Prompting Strategies)

Model	**Zero-Shot (List)**	**Zero-Shot (Text)**	**Few-Shot (List)**	**Few-Shot (Text)**
Exaone	70.6%	56.6%	**71.8%**	70.4%
Meta LLaMA 3.3	**73.4%**	69.6%	72.5%	73.3%
DeepSeek LLaMA	73.4%	71.5%	75.6%	**75.9%**

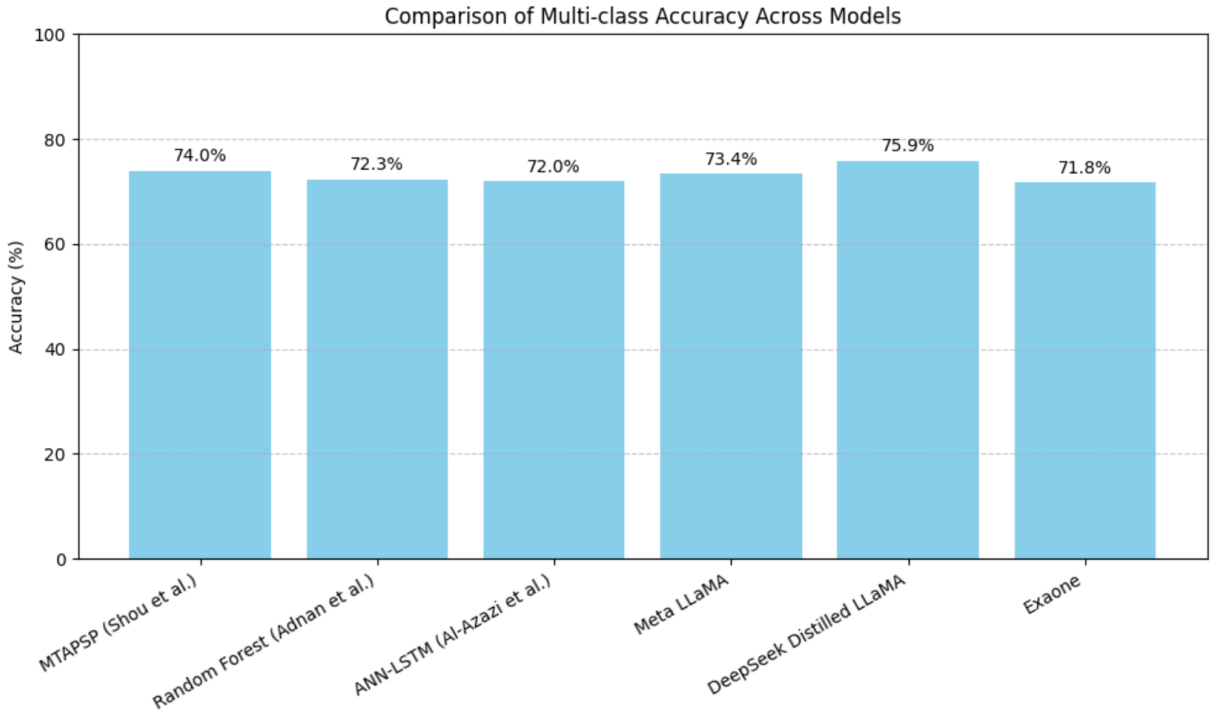


Figure 6: Comparison of accuracies obtained from traditional approaches v/s LLMs

As indicated by the results, best performance among all the LLMs is given by the DeepSeek Distilled LLaMA model with an accuracy score of 75.9% in a few-shot setting. This

accuracy is higher than most traditional approaches that have got accuracies for multi-class classification in the range of 72%-74%. Notably, these results achieved by the LLM is without any fine-tuning or model re-training. It highlights the fact that LLMs can be considered as an alternative to traditional models for multi-class prediction of student performance. The results also show that there is improvement in accuracy for all LLMs when a few examples are provided to them. Just a small amount of guidance helps the LLMs to perform better.

6.1.2 Binary Classification

The binary classification involves predicting if a student is going to pass or fail by the end of course. In the OULAD dataset, the records marked as "Pass" is formed by combining Pass and Distinction records and "Fail" marked records are formed by combining Fail and Withdrawn records. Table 4 gives out the accuracies of LLMs in detail in such situations. In addition, a comparison with the traditional approaches is shown through a line graph in Figure 7.

Table 4: Binary Classification Accuracy of LLMs (Pass vs Fail)

Model	**Zero-Shot (List)**	**Zero-Shot (Text)**	**Few-Shot (List)**	**Few-Shot (Text)**
Exaone (Fail)	**93.2%**	47.2%	86.3%	86.6%
Exaone (Pass)	66.4%	**89.4%**	89.0%	88.9%
Meta LLaMA (Fail)	89.3%	76.8%	**92.9%**	86.8%
Meta LLaMA (Pass)	80.0%	88.8%	84.1%	**91.4%**
DeepSeek (Fail)	**94.4%**	92.2%	87.5%	88.0%
DeepSeek (Pass)	73.0%	72.3%	89.0%	**89.6%**

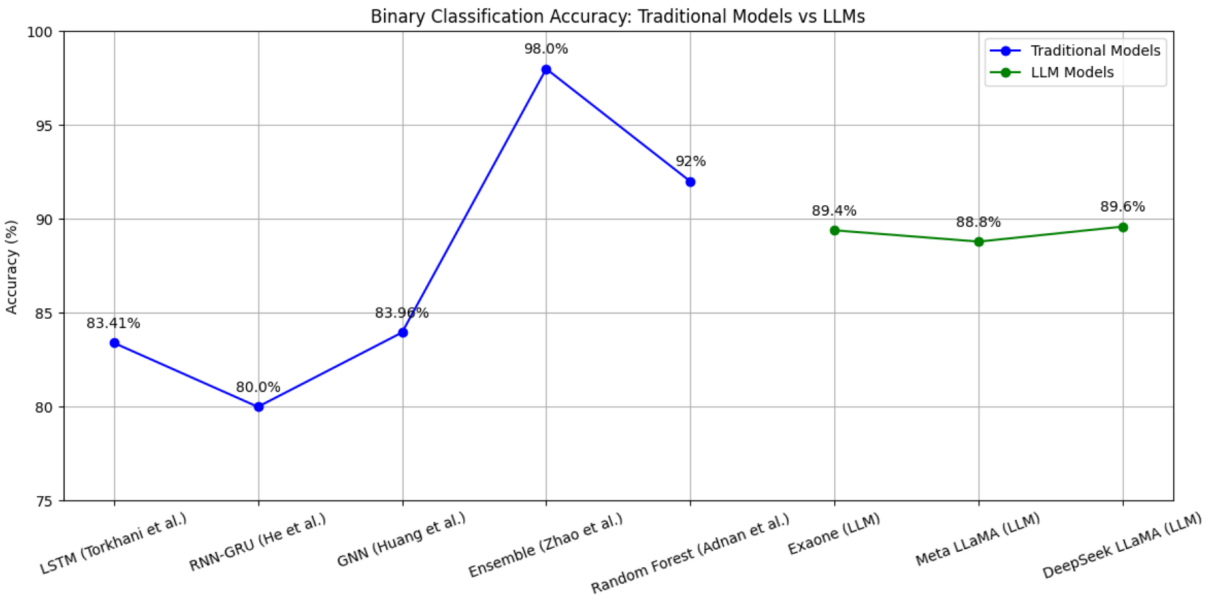


Figure 7: Comparison of accuracies obtained from traditional approaches v/s LLMs

Among all LLMs, DeepSeek Distilled LLaMA performed the best in case of binary settings. It received accuracy of 94.4% and 89.6% on the Fail class and Pass class respectively.

The performance shown by LLMs is quite close if compared to the traditional approaches where models achieved up to 98% accuracy. This is important that LLMs achieved this result without any prior training and yet they were close to the performance of state-of-the-art models. Moreover, it can be seen that narrative-style prompts are more effective for predicting students who passed. It could be that storytelling format helped in revealing subtle positive patterns. On the other side, list-style prompts performed better in predicting students who have failed. It could be because behavioral shortcomings were more explicit in list-style format.

6.1.3 Per-Class Accuracy

Per-class accuracy depicts individual accuracy score for each class label that is achieved by LLMs. There aren't a lot of studies which particularly tells category wise performance on the OULAD dataset. A comparison with the traditional approaches and the LLMs itself is shown through a bar graph in Figure 8.

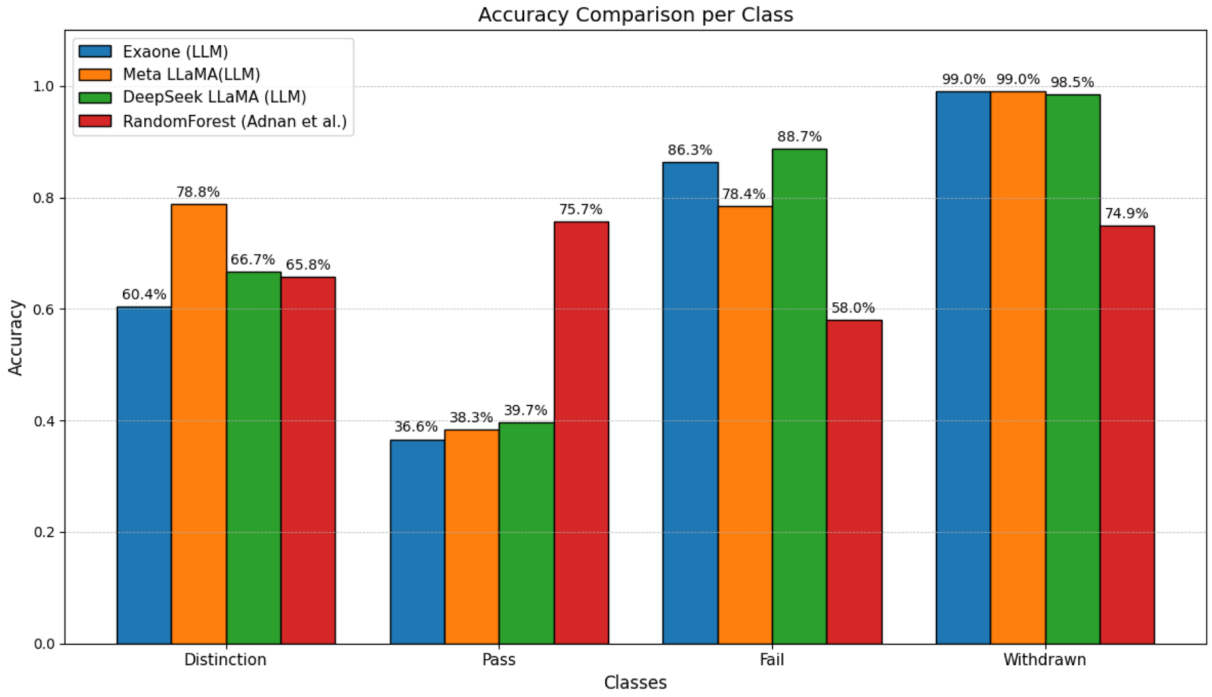


Figure 8: Comparison of accuracies obtained from traditional approaches v/s LLMs

The class-wise analysis indicates that all the three LLMs gave excellent performance (99%) for classification of students in the Withdrawn class. This points out the ability of LLMs to identify dropout signals. However, they poorly performed to classify students in Pass category (36-40%). This could be because of overlapping of behavioral patterns in Pass and Distinction. For Distinction, Meta LLaMa did the best with highest accuracy score of 78.8%. In comparison, Random Forest model from past work has been found to perform better than LLMs with accuracy score of 75.7%. But Fail and Withdrawn students were classified more accurately by LLMs than Random Forest.

6.1.4 Justification to RQ1

The results from the evaluation totally supports RQ1 and shows the actual potential of LLMs through the following points:

- In comparison with traditional approaches, LLMs have attained a competitively high accuracy particularly with DeepSeek Distilled LLaMA version. In the best setting which was narrative style few-shot prompting, LLMs gave almost state-of-the-art results.
- There doesn't require any kind of feature engineering and can work with inputs given in human-readable format that is being used in daily life.
- LLMs offers a low-maintenance and a scalable solution when only a few labeled data is available or retraining entire system is not possible.

6.2 Evaluation of quality of reasoning in LLM Predictions

The reasoning which has been generated by LLMs is analyzed in order to answer the RQ2. Four records have been taken for analysis: one for each class (Pass, Fail, Withdrawn, Distinction). These explanations were checked if they have referenced the most predictive features as shown in Figure 9. These features have been derived by finding feature importances using Random Forest model.

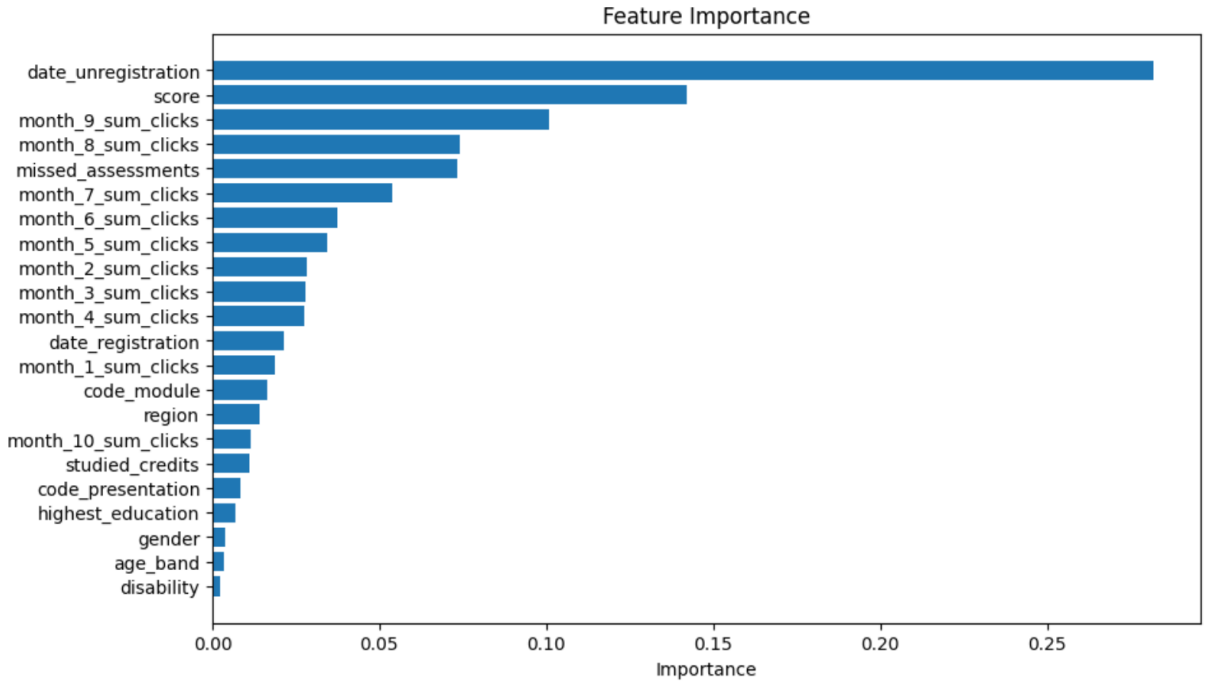


Figure 9: Key factors influencing student performance

Prediction	**Reasoning by LLM**
Pass	*The student is male, under 35, with A Level education and no disability. He* ***registered*** *56 days early and completed the course, so Withdrawn does not apply. His weighted score is* ***75.25****, indicating strong academic performance.* ***Monthly engagement*** *was consistent overall (clicks ranging from 273 to 625), with only a drop to zero in the final month, likely due to course completion. Given the solid academic results and no significant engagement gaps, Pass is the most appropriate label*

Prediction	Reasoning by LLM
Fail	*The student is male, under 35, with A Level education and no disability. He **registered** 45 days before the course began and completed it without withdrawing, so Withdrawn is not applicable. He earned a weighted score of **12.96**, which suggests decent academic performance. However, his **VLE engagement** dropped to zero after the third month (clicks: 40, 299, 134, then all 0s). This sustained disengagement over six months is significant. Although the score appears acceptable, the major gap in engagement suggests the student did not actively participate throughout. Therefore, Fail is the most appropriate label.*
Withdrawn	*The student is female, aged 35–55, with lower-than-A-level education and no disability. She registered well in advance but **unregistered** 130 days after the course began, indicating early disengagement. Her **engagement** dropped significantly after the third month (clicks: 13, 44, 50, 2, then all 0s), and she did not complete the course. Despite having a weighted score of **11.99**, the unregistration and lack of activity suggest she withdrew. Therefore, Withdrawn is the most appropriate label.*
Distinction	*The student is female, under 35, with A Level education and no disability. She **registered** 34 days early and completed the course, showing consistent commitment. Her weighted score of **71.85** is above the typical 70% threshold for a distinction, indicating strong academic performance. While her **engagement** (clicks: 16, 152, 81, 33, 10, 36, 35, 20, 38, 0) fluctuated, she maintained activity during critical months. Since her score exceeds the distinction mark and she completed the course without withdrawing, upgrading her classification to Distinction is justified.*

Table 5: Samples of predictions and LLM reasoning for it

Table 5 shows some predictions made by LLMs and their reasoning. In almost all the cases, reasoning provided by LLM has aligned well with the actual behavior of students. For example, the prediction marked as Fail has highlighted on lack of engagement while the Withdrawn case shows high inactivity and early unregistration. The prediction of Distinction class is supported with reasoning of high score and and very consistent engagement activity. Even, in the pass prediction, the reasoning mentions about moderate click activity with acceptable assessment scores. This proves that not only LLMs can identify both strong and weak academic performance but they can distinguish between moderate and strong performances as well.

Overall, the analysis shows that the reasoning which has been generated by LLMs is not only factually correct but also aligns with the features which matters the most to predict student outcomes. The language which have been used in explanations by LLMs is natural and can be easily interpreted. This becomes highly practical for teachers and educators to not just learn about the prediction but also to why that prediction was made. Therefore, this section answers RQ2 by showing that LLMs can generate meaningful as well as trustworthy reasoning that can help to make informed decision-making in a real-world educational scenario.

6.3 Discussion

This section discusses a few observations and choices that were made during the experimentation so as to ensure the feasibility of the analysis. In the OULAD dataset, there were

student records with more than a score of 70 and strong engagement but have been marked as "Pass". On the contrary, there were records with low scores and yet they are marked as "Distinction". This must be because of natural complexity of the educational datasets but it reduced the accuracy of LLMs. Furthermore, to maintain the consistency of the data, the assessments which had no student records were removed to avoid any kind of wrong calculations. Additionally, there was a high resource cost for sending prompts to LLMs and due to lack of GPU, dataset size was reduced to 5000 instead of using 32000 records. These records included 1250 student records from each class: Distinction, Pass, Withdrawn and Fail. This allowed the experimentation feasible without compromising the goals of this research

7 Conclusion and Future Work

The objective of this study is to check whether LLMs can properly predict student outcomes using zero-shot and few-shot prompting methods. It also tries to find if the reasoning made by LLMs aligns with factors that actually affects student performance. The study makes use of the popular OULAD dataset and three models: Meta LLaMA, Exaone and DeepSeek Distilled LLaMA. It evaluates the predictions made by these models across binary, multi-class and per-class setting and without any feature engineering and training.

The best performing model is DeepSeek LLaMA that achieved 75.9% accuracy in multi-class setting and 89.6% for binary classification. It even showed high results in labels: Withdrawn (98.8%) and Distinction (80%) on a per-class level. The results were very close to the performance given by traditional Machine Learning approaches. Reasoning outputs were found to align with key predictive features which were found by calculating feature importance through Random Forest approach. All these findings proves that LLMs can not only offer an alternative approach to traditional approaches for predictions but can also explain their predictions in a human understandable language.

A major limitation that was observed was that LLMs occasionally misclassified students in edge cases such as for Pass/Fail and Pass/Distinction. This means that LLMs can get confused when students have a very similar behaviour but doesn't belongs to same grade category. Furthermore, the validation of reasoning has been done using feature importance using Random Forest. But this approach don't use real-world factors such as effort made or some personal situations of students. These factors aren't present in the dataset that a human teacher will consider while making any judgement. Lastly, the study makes use of only models with 70 billion parameters and a few prompting strategies. Results can definitely change with models with large number of parameters and more prompting techniques which haven't been explored.

Nonetheless, future work for this project could involve fine-tuning small open-source LLMs using the structured data from the OULAD dataset or use some LLMs with more than 70 billion parameters which can help these LLMs to learn educational patterns much better. Another direction could also include to use LLMs for mid-course performance prediction. It means that predictions can be made at multiple intervals such as monthly or quarterly instead of just doing at the end of the course. This will allow much early interventions if required. Lastly, LLMs could be combined together with traditional models to form a hybrid structure where predictions will be made by ML models joined by reasoning abilities of LLMs. In this way, accuracy can be achieved along with interpretability with the use of only a single framework.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M. and Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models, *Ieee Access* **9**: 7519–7539.

Al-Azazi, F. A. and Ghurab, M. (2023). Ann-lstm: A deep learning model for early student performance prediction in mooc, *heliyon* **9**(4).

An, S., Bae, K., Choi, E., Choi, K., Jungkyu Choi, S., Hong, S., Hwang, J., Jeon, H., Jeongwon Jo, G., Jo, H. et al. (2024). Exaone 3.5: Series of large language models for real-world use cases, *arXiv* .

Chan, L., Xu, X. and Lv, K. (2025). Deepseek-r1 and gpt-4 are comparable in a complex diagnostic challenge: a historical control study, *International Journal of Surgery* **111**(6).

Cohn, C., Hutchins, N., Le, T. and Biswas, G. (2024). A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38.

Diyab, A., Frost, R. M., Fedoruk, B. D. and Diyab, A. (2025). Engineered prompts in chatgpt for educational assessment in software engineering and computer science, *Education Sciences* **15**(2).

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. et al. (2024). The llama 3 herd of models, *arXiv* .

Fang, X., Xu, W., Tan, F. A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S. and Faloutsos, C. (2024). Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey, *arXiv preprint arXiv:2402.17944* .

He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G. and Jiang, B. (2020). Online at-risk student identification using rnn-gru joint neural networks, *Information* **11**(10).

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X. and Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models, *International conference on artificial intelligence and statistics*, PMLR.

Huang, Q. and Zeng, Y. (2024). Improving academic performance predictions with dual graph neural networks, *Complex & Intelligent Systems* **10**(3).

Jha, N. I., Ghergulescu, I. and Moldovan, A.-N. (2019). Oulad mooc dropout and result prediction using ensemble, deep learning and regression techniques., *CSEDU (2)*.

Jin, L., Wang, Y., Song, H. and So, H.-J. (2024). Predictive modelling with the open university learning analytics dataset (oulad): A systematic literature review, *International Conference on Artificial Intelligence in Education*, Springer.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. and Iwasawa, Y. (2022). Large language models are zero-shot reasoners, *Advances in neural information processing systems* **35**.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017). Open university learning analytics dataset, *Scientific data* **4**(1).

Lee, G.-G., Latif, E., Wu, X., Liu, N. and Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring, *Computers and Education: Artificial Intelligence* **6**.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work?, *arXiv preprint arXiv:2202.12837* .

Moëll, B., Sand Aronsson, F. and Akbar, S. (2025). Medical reasoning in llms: an in-depth analysis of deepseek r1, *Frontiers in Artificial Intelligence* **8**.

Neshaei, S. P., Davis, R. L., Hazimeh, A., Lazarevski, B., Dillenbourg, P. and Käser, T. (2024). Towards modeling learner performance with large language models, *arXiv preprint arXiv:2403.14661* .

Oh, C., Park, M., Lim, S. and Song, K. (2024). Language model-guided student performance prediction with multimodal auxiliary information, *Expert Systems with Applications* **250**.

Severino, J. V. B., de Paula, P. A. B., Berger, M. N., Loures, F. S., Todeschini, S. A., Roeder, E. A., Veiga, M. H., Guedes, M. and Marques, G. L. (2025). Benchmarking open-source large language models on portuguese revalida multiple-choice questions, *BMJ Health & Care Informatics* **32**(1).

Shou, Z., Xie, M., Mo, J. and Zhang, H. (2024). Predicting student performance in online learning: a multidimensional time-series data analysis approach, *Applied Sciences* **14**(6).

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S. and Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study, *JMIR Medical Informatics* **12**.

Sivarajkumar, S. and Wang, Y. (2023). Healthprompt: a zero-shot learning paradigm for clinical natural language processing, *AMIA Annual Symposium Proceedings*, Vol. 2022.

Torkhani[1], W. and Rezgui, K. (2025). Oulad mooc student performance prediction using machine and deep learning, *Proceedings of International Conference on Decision Aid and Artificial Intelligence (ICODAI 2024)*, Vol. 12, Springer Nature.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* **35**.

Zhao, S., Zhou, D., Wang, H., Chen, D. and Yu, L. (2025). Enhancing student academic success prediction through ensemble learning and image-based behavioral data transformation, *Applied Sciences* **15**(3).