CSC311: Introduction to Machine Learning

Project Proposal

Submitted by

Purav Gupta, Vennise Ho, Tianyu Luo, Sean Woo

University of Toronto
Fall 2025

1 Data Exploration

The dataset includes 825 survey responses from 275 students, with each student providing one response per AI model (*ChatGPT*, *Claude*, and *Gemini*). Each entry describes how the student used the model and their perception of its strengths and weaknesses. The target variable `label` identifies the model, forming a three-class classification task.

Feature Types and Cleaning

- **Ordinal (1–5):** `academic_use_likelihood`, `suboptimal_frequency`, `reference_expectation`, `verify_frequency` — numeric Likert responses, imputed with median values.
- **Categorical:** `best_task_types`, `suboptimal_task_types` — multi-choice fields cleaned, split, and one-hot encoded.
- **Text:** `tasks_use_model`, `suboptimal_example`, `verify_method` — open-ended responses kept for potential NLP features.

Missing values and Excel artifacts (e.g., "#NAME?") were standardized as `NaN`. Ordinal columns were converted to numeric values (1–5) and estimated using the median. Text fields were left as strings for potential natural language processing. Categorical features were one-hot encoded, and the target `label` was encoded as integers 0–2. No numerical outliers were found due to fixed Likert scales.

Data Splitting and Leakage Prevention

Each student contributed three related responses, so records were grouped by `student_id` before splitting with `GroupShuffleSplit`. The split used 70% for training, 15% for validation, and 15% for testing. All analysis and plots were done on the training data only.

Exploratory Findings

A few consistent trends appeared in the training data:

- Most students rated academic use around 3–4, showing moderate adoption.
- **ChatGPT** users reported higher reference expectations and verification frequency.
- **Claude** was often noted for explaining complex concepts.
- **Gemini** users highlighted strengths in data analysis and coding.

Summary and Implications

The dataset is balanced and diverse in features—combining ordinal, categorical, and text features. These characteristics make it suitable for models that handle mixed data types. Patterns across the data suggest that both task types and the language students use to describe them could help distinguish between responses.

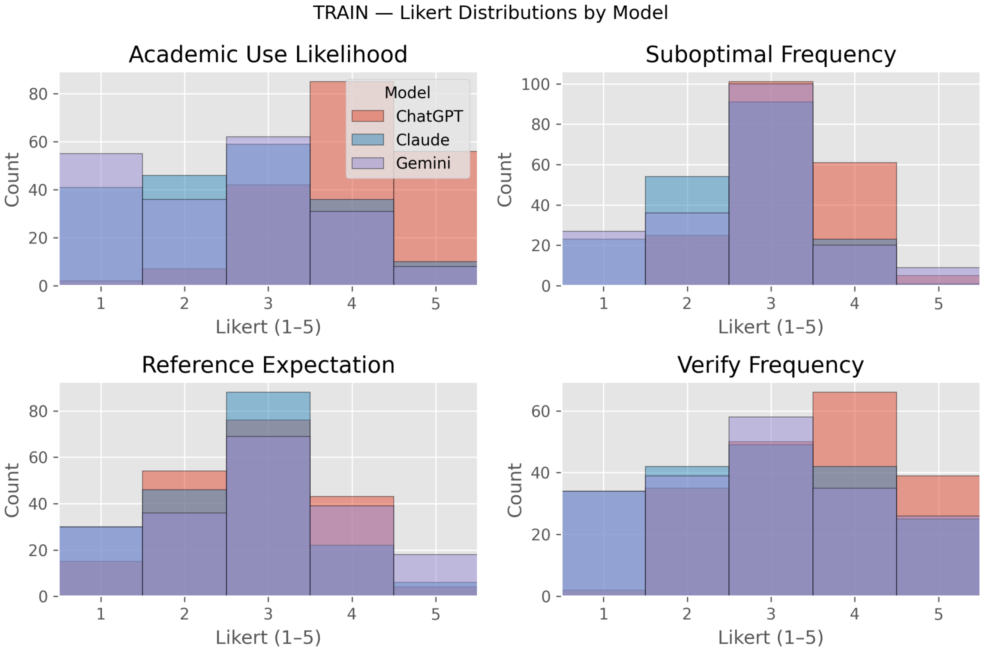

Figure 1: Distribution of Likert-scale responses in the training set, by model label.

2 Methodology

Model Families

Three model families that we will explore are: **Decision Trees**, **Logistic Regression** and **Neural Network**.

- Decision trees are naturally suitable for this dataset because many features are ordinal and categorical. Decision trees excel at capturing intrinsic patterns in the data and handles relationships between features very well as it is a universal function approximator.
- Logistic regression, handling multiclass classification through softmax activation, works well with both categorical and numerical features very well. One major advantage of logistic regression in our dataset is its interpretability where the importance of each feature can be easily interpreted through coefficients.
- Neural Network is also suitable for this dataset because it is also a universal function approximator. Neural Network is highly flexible and is great at capturing complex patterns and relationships in the dataset.

Optimization Technique

For both logistic regression and neural network, we will use Adam Gradient Descent as optimization technique. Adam GD combines the first momentum and second momentum of the gradient to make the optimization fast and robust.
For decision trees, we will use the greedy recursive split based on information gain mea-

sured by entropy and an exhaustive search over features and split points at each node.

Validation Method

Given that the dataset size is small, we will use a 5-fold cross validation method where we split the training and validation dataset into 5 equal sets and for each set of hyperparameters, train on 4 sets and validate on 1 set.

Hyperparameter Tuning

- For decision trees, we need to tune the maximum depth of the tree and min samples split to balance to prevent both underfitting and overfitting. We also need to tune the maximum features to be considered to each split to prevent the tree from taking into account too many features in one split which will significantly reduce interpretability and might lead to underfitting/overfitting.
- For Logistic regression, we need to tune the learning rate and regularization strength (λ) for model convergence and performance.
- For Neural Networks, we need to tune the number of hidden layers and the number of neurons per layer.

Evaluation Metrics

In this dataset, some important metrics beyond accuracy are:

- **Precision** calculated by $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$: It reflects the true proportion truly belonging to one class among all samples predicted as that class . This is valuable to this dataset because with accuracy, precision captures the reliability of using our prediction result from the trained model.
- **Recall** calculated by $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$: It reflects the proportion of samples we correctly identified among all samples that truly belonged to one class. This is valuable to this dataset because recall reflects how many true labels we are missing among all data with the same label giving us a clear direction whether we should keep improving the model or terminate training.
- **Specificity** calculated by $\frac{\text{True Negative}}{\text{True Negative + False Positive}}$: It reflects the proportion of samples we correctly exclude among all samples that don't belong to a class. This is valuable to this dataset because in the context of the dataset, it is important not to mislabel one class for another as it will likely adversely impact user satisfaction with the corresponding AI model (i.e. we don't want correlate a model with a task that it is not good at).