DEPARTMENT OF COMPUTER SCIENCE
INDIAN INSTITUTE OF TECHNOLOGY MADRAS
CHENNAI – 600036

Advanced Feature Extraction and Optimization Techniques for Student Response Prediction and Question Quality Ranking

A Thesis

Submitted by

GAURAV KANWAT

For the award of the degree

Of

MASTERS OF TECHNOLOGY

May 2025

THESIS CERTIFICATE

This is to undertake that the Thesis titled **ADVANCED FEATURE EXTRACTION AND OPTIMIZATION TECHNIQUES FOR STUDENT RESPONSE PREDICTION AND QUESTION QUALITY RANKING**, submitted by me to the Indian Institute of Technology Madras, for the award of **Masters of Technology**, is a bona fide record of the research work done by me under the supervision of **Prof. Chandrashekar Lakshminarayanan**. The contents of this Thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

In order to effectively convey the idea presented in this Thesis, the following work of other authors or sources was reprinted in the Thesis with their permission:

Chennai 600036 **Gaurav Kanwat**

Date: May 2025

Prof. Chandrashekar Lakshminarayanan
Research advisor
Professor
Department of Computer Science
IIT Madras

ABSTRACT

In the realm of educational data mining, assessing the quality of educational content plays a pivotal role in enhancing personalized learning and improving student outcomes. This project addresses Task 3 of the NeurIPS 2020 Education Challenge, which focuses on ranking the quality of questions based on their ability to elicit informative and high-quality responses from students. The primary objective was to develop a model that can predict question quality by leveraging large-scale metadata, student interaction data, and associated images of questions.

Our approach began with comprehensive feature extraction from the available datasets. We engineered diverse features capturing question difficulty, subject hierarchies, answer entropy measures, student engagement statistics, and text clarity scores. Special emphasis was placed on designing metadata-driven features, group-level entropy metrics, and incorporating image-based clarity analysis to enrich the feature space.

Multiple machine learning models were experimented with, including XGBoost, CatBoost, and LightGBM classifiers, optimized through rigorous hyperparameter tuning. We evaluated model performance using pairwise accuracy and custom leaderboard metrics tailored to the competition's ranking evaluation. Model explainability was enhanced through feature importance analyses, enabling deeper insights into the attributes most influential in determining question quality.

Our final model achieved strong generalization performance and demonstrated the effectiveness of combining student response dynamics, question metadata features, and image-derived clarity scores for predicting question quality. The outcomes of this work not only contribute to the challenge goals but also offer scalable strategies for improving content curation in educational platforms.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND PROBLEM STATEMENT

1.1 PROBLEM STATEMENT

In digital education platforms, students frequently interact with a wide variety of questions, each varying in complexity, clarity, structure, and pedagogical intent. However, not all questions contribute equally to a student's learning experience. Some questions may be ambiguous, poorly framed, or visually unclear, which can hinder comprehension and reduce the quality of learning outcomes.

The NeurIPS 2020 Education Challenge [7] Task 3 focuses on ranking educational questions based on their quality, using student response patterns as a proxy for question clarity and pedagogical effectiveness. The core objective is to build a model that can automatically compare any two questions and determine which one is of higher quality. This judgment must align with expert annotations, which serve as the ground truth.

The problem is particularly challenging due to the multi-modal nature of the data: each question is associated with both structured metadata and an image containing the actual question content. Furthermore, quality is inherently subjective and influenced by factors such as clarity of expression, image legibility, subject hierarchy, and student engagement patterns. The task requires not only traditional machine learning modeling, but also deep feature engineering and image processing to assess question quality accurately.

This problem has practical significance in the deployment of large-scale intelligent tutoring systems, where automated filtering and ranking of questions can directly impact student learning experiences and curriculum design.

1.2 OBJECTIVE OF THE WORK

The primary objective of this work is to develop a robust and interpretable machine learning pipeline capable of automatically ranking educational questions by quality, in alignment with expert-provided judgments. Given a pair of questions, the model should accurately predict which of the two is of higher quality.

To achieve this, our work is centered around the following goals:

- **Feature-Rich Modeling:** Design and extract meaningful features [3] from diverse sources, including question metadata, student response patterns, and image content, to capture the multifaceted notion of question quality.

- **Image-Based Clarity Assessment:** Develop a mechanism to quantify the visual clarity of question images using text segmentation and OCR-based [6] techniques, contributing a novel modality to the quality prediction task.

- **Behavioral Signal Integration:** Leverage answer distributions, entropy, and subject hierarchies to understand how students interact with questions and infer latent quality indicators.

- **Pairwise Classification Framework:** Frame the ranking task as a pairwise classification problem and train models using tree-based ensemble methods to learn relative quality judgments.

- **Evaluation and Interpretability:** Rigorously evaluate model performance using the pairwise accuracy metric and analyze key features influencing predictions, ensuring both effectiveness and transparency.

Through this comprehensive approach, the work aims not only to achieve high predictive accuracy but also to uncover insights into what makes a question pedagogically valuable.

1.3 RELEVANCE TO EDUCATIONAL DATA MINING

Automatically assessing and ranking question quality sits at the intersection of educational data mining (EDM) and learning analytics, which aim to use data to improve learning outcomes.

In today's digital learning systems, vast numbers of questions are generated across subjects and levels, yet scalable tools to evaluate their clarity or pedagogical value remain limited. Manual review is impractical, and expert-only assessments face bias and scalability issues. Student interaction data, meanwhile, provides a rich but underused signal that EDM can leverage. This work contributes to EDM in several key ways:

- **Data-Driven Quality Assessment:** By modeling student response patterns (e.g., correctness rates, entropy, subject-level performance), we enable objective and scalable evaluation of question quality, rooted in real-world student behavior.

- **Multi-Modal Feature Fusion:** Combining structured metadata with image clarity scores introduces a novel multimodal dimension to question quality analysis, aligning with current EDM efforts to utilize diverse data sources such as clickstreams, visual content, and learner-generated data.

- **Insight into Instructional Design:** The interpretability of our model allows educators and curriculum designers to understand which question attributes — such as clarity, difficulty balance, or hierarchical structure — impact learning most effectively.

- **Towards Intelligent Tutoring Systems:** Our approach paves the way for automated filtering, improvement, and recommendation of questions in intelligent tutoring systems, where adaptive content delivery is critical.

In essence, this work strengthens the feedback loop between students and content by allowing question-level analytics to guide educational design, which is a core goal of educational data mining.

CHAPTER 2

DATASET OVERVIEW

2.1 DATASET DESCRIPTION

The dataset used in this work is provided as part of the Challenge [7] and is sourced from a large-scale digital learning platform. It is designed to support the task of predicting question quality by leveraging both structured metadata and visual content.

The dataset is composed of the following main components:

- **Question Metadata (`question_metadata_task_3_4.csv`):** Contains attributes such as `QuestionId`, `SubjectId`, `Level`, and `QuestionType`, providing essential categorical information for each question.
- **Answer Data (`train_task_3_4.csv`):** Includes student responses to questions, with each row containing a `StudentId`, `QuestionId`, and `AnswerValue` (0 or 1). This forms the basis for modeling student behavior and correctness distribution.
- **Student Metadata (`student_metadata_task_3_4.csv`):** Provides student-specific attributes such as `StudentId`, `Gender`, `PremiumPupil`, and `DateOfBirth`, useful for behavioral profiling and demographic analysis.
- **Subject Hierarchy (`subject_metadata.csv`):** Defines a hierarchical structure among subjects, allowing us to derive features like subject depth and ancestry for each question.
- **Question Images (`/data/images/*.jpg`):** Each question is associated with an image file named using its `QuestionId` (e.g., 10.jpg for `QuestionId` = 10). These images visually present the actual content of the questions, including diagrams, text, and equations.

- **Expert Judgments (`quality_response_remapped_public.csv` and `quality_response_remapped_private.csv`):** Pairs of questions labeled by educational experts with a binary outcome indicating which of the two is considered higher quality. These form the ground truth for training and evaluation.

Key Characteristics

- The dataset comprises thousands of questions, millions of student responses, and hundreds of question images.
- The data is multi-modal and partially noisy, especially in student answers and image quality, demanding careful preprocessing.
- The pairwise structure of expert labels transforms the prediction problem into a ranking or comparative classification task rather than a direct regression or multiclass classification task.

This dataset provides a unique opportunity to integrate student interaction data with content-based features and image processing to tackle the real-world challenge of question quality assessment.

2.2 KEY STATISTICS AND OBSERVATIONS

To gain an initial understanding of the dataset and guide feature engineering, we conducted an exploratory analysis across all major data components. The goal was to identify patterns, imbalances, and noteworthy trends that could influence the design of our modeling pipeline.

1. **Student-Question Interaction (`train_task_3_4.csv`)**
    - **Total Responses:** ~19.8 million
    - **Unique Students:** ~118,971

- **Unique Questions:** ~27,613
- **Answer Distribution:**
  - Correct (`AnswerValue` = 1): ~52%
  - Incorrect (`AnswerValue` = 0): ~48%
- **Observation:** The answer distribution is fairly balanced, but certain questions have extreme correctness ratios (either too easy or too hard), which may indicate quality issues or misaligned difficulty.

2. **Question Metadata (question_metadata_task_3_4.csv)**
   - **Total Questions:** 27,613
   - **Distinct Subjects:** ~200+
   - **Levels:** Levels range from 0 (broad/general subjects) to 5 (specific topics).
   - **Observation:** The subject hierarchy is deep and uneven — some topics have many leaf nodes while others are shallow. This impacts subject-level aggregation and generalizability of question features.
3. **Student Metadata (student_metadata_task_3_4.csv)**
   - **Total Students:** 118,971
   - **Gender Breakdown:** Slight male dominance (~55–60%)
   - **Premium Pupil Proportion:** ~40%
   - **Age Distribution:** Skewed towards younger students (ages 10–17)
   - **Observation:** A non-uniform distribution across demographic variables requires care in modeling to avoid bias and overfitting to dominant subgroups.

4. **Subject Metadata (subject_metadata.csv)**

    - **Total Subjects:** ~260
    - **Max Hierarchical Depth:** 5 levels
    - **Observation:** Certain question clusters appear only at deeper levels (more specific subjects), and these are often underrepresented, potentially influencing model learning.

5. **Image Data (images/)**

    - **Total Images:** ~27,613 (1 per question)
    - **Format:** .jpg, grayscale or color, with varied resolution and layout
    - **Text Coverage:** Varies significantly — some images are clear and well-formatted, others are noisy, dense, or cluttered.
    - **Observation:** OCR accuracy depends heavily on image quality. Questions with poor legibility may correlate with lower clarity or expert-assigned quality.

6. **Quality Response Pairs (quality_response_remapped_*.csv)**

    - **Total Pairs (Public + Private):** ~40,000
    - **Label Distribution:** Balanced (~50% each class)
    - **Observation:** Expert judgments show diversity — some comparisons are between very similar questions, while others are between distinctly different ones (e.g., clean vs. noisy formatting).

General Observations

- The dataset is high-volume and high-variance, demanding robust preprocessing and careful validation.
- Long-tail distributions exist in both question frequency and subject coverage.
- The presence of visual noise and behavioral outliers requires specific attention in feature engineering.
- Initial analysis reinforces the need to incorporate both content features (image/text clarity, structure) and contextual signals (student behavior, subject difficulty) for effective quality ranking.

2.3 CHALLENGES IN THE DATASET

The dataset [7] offers a rich, multi-modal foundation for modeling question quality, but it also presents several non-trivial challenges that must be addressed to ensure effective learning and generalization. These challenges arise from data sparsity, heterogeneity, noise, and the inherent subjectivity of the task.

1. **Multi-Modal Complexity**
    - **Issue:** The dataset combines structured metadata, free-form images, and behavior logs, requiring separate pipelines for tabular and visual data.
    - **Impact:** Feature extraction becomes significantly more complex, particularly for images where no textual metadata is available.
2. **Image Noise and Variability**
    - **Issue:** Question images vary widely in quality, resolution, and layout. Some contain well-formatted text and diagrams, while others are blurry, cluttered, or contain handwritten or dense content.

- **Impact:** OCR-based [6] clarity estimation is error-prone, leading to potential misrepresentation of visual clarity and degraded model input.

3. **Subject Hierarchy Imbalance**

   - **Issue:** The subject tree is deep and uneven, with some subjects represented by thousands of questions and others by only a few.

   - **Impact:** Modeling subject-level effects or learning generalized patterns across topics becomes difficult, especially for rare or overly specific subjects.

4. **Student Behavior Bias**

   - **Issue:** Some students answer thousands of questions while others answer only a few. Behavior also varies across demographics like `age`, `gender`, and `account type` (premium vs. non-premium).

   - **Impact:** Aggregated behavioral features may be biased toward high-frequency users or certain student subgroups.

5. **Label Subjectivity in Pairwise Judgments**

   - **Issue:** Expert labels for question quality are inherently subjective. The criteria used by annotators may vary or be inconsistently applied.

   - **Impact:** Training a model to match expert judgment requires careful generalization to avoid overfitting to potentially noisy or non-uniform labels.

6. **Data Sparsity in Rare Combinations**

   - **Issue:** Many combinations of `SubjectId`, `Level`, and `QuestionType` appear infrequently in the dataset.

   - **Impact:** Sparse feature representations reduce the reliability of frequency-based statistics and hinder the generalizability of certain

categorical encodings.

7. **Evaluation Design Constraint**

    - **Issue:** The problem is framed as a pairwise ranking task rather than direct quality scoring.

    - **Impact:** Traditional classification/regression metrics are not applicable. Care must be taken to engineer comparative features and interpret model predictions in relative terms.

2.4 MAPPING BETWEEN IMAGE, QUESTIONID, AND METADATA

A crucial aspect of working with this dataset is ensuring that all components — especially visual and structured data — are accurately linked via a shared key. In this challenge, QuestionId serves as the central identifier that binds together question images, metadata, student responses, and expert-labeled comparisons.

1. **Linking Images to Questions**

    - **Image File Format:** All question images are stored as .jpg files within the /images/ directory.

    - **Naming Convention:** Each image is named using its `QuestionId`. For example, the file 27.jpg corresponds to `QuestionId = 27`.

    - **Mapping Strategy:** During preprocessing, a new column `QuestionImage` was created by appending .jpg to each `QuestionId`. This column was then used to join image-based features (e.g., clarity scores) with the main question-level dataset.

2. **Mapping Metadata to Questions**

    - Question Metadata (`question_metadata_task_3_4.csv`) is directly

indexed by `QuestionId`, allowing straightforward joins with:

  - Image-derived features (via filename)
  - Student responses (via shared `QuestionId`)
  - Subject information (via `SubjectId` hierarchy)

- Answer Data (`train_task_3_4.csv`) is joined using `QuestionId` to enrich student-question interactions with additional question-level features.

3. **Mapping Subject Hierarchies**

   - Each `QuestionId` is associated with a `SubjectId`, which in turn can be linked to:
     - Its parent subject (`ParentId`)
     - Its level in the subject tree (`Level`)
   - A recursive mapping process was implemented to extract the full ancestry of each `SubjectId`, allowing us to build features such as:
     - `SubjectDepth`
     - `SubjectPath`
     - `SubjectId_with_parents` (multi-level ancestry list)

4. **Mapping Expert Judgments to Feature Pairs**

   - In the expert-labeled datasets (`quality_response_remapped_*.csv`), each row contains a pair: `QuestionId_1` and `QuestionId_2`, along with a binary label.
   - To generate features for these pairs:
     - Individual features are extracted for both questions.

  - Pairwise features (e.g., difference in clarity score, subject depth, entropy) are then computed.

- The resulting dataset forms the basis for the pairwise classification model.

5. **Handling Missing or Invalid Links**

   - **Image Mismatches:** Some `QuestionIds` may not have corresponding image files (e.g., due to naming errors or missing data). These are excluded or imputed based on available structured features.

   - **Metadata Gaps:** In rare cases, a question may exist in the answer log but be absent from the metadata file. Such inconsistencies were filtered out during the merging phase to maintain data integrity.

By establishing a consistent and lossless mapping across all sources, we ensured that each question — whether considered individually or as part of a comparison — could be enriched with all available structural, behavioral, and visual information. This unified representation was critical for training a reliable and explainable quality ranking model.

CHAPTER 3

DATA PREPROCESSING

3.1 HANDLING MISSING AND NOISY DATA

Educational datasets collected at scale are often prone to incompleteness and inconsistencies due to the diversity of content sources and user interactions. Prior to feature engineering and modeling, we implemented targeted data cleaning strategies to ensure a reliable and consistent foundation. The key issues and resolutions are summarized below:

1. **Missing Question Images**

   - **Issue:** Some QuestionId entries in the metadata or training files had no corresponding image file in the /images/ directory.

   - **Action Taken:**

     – Verified presence of `<QuestionId>`.jpg for all required records.

     – Excluded such entries from image-based feature pipelines.

     – Retained them in structured feature modeling where applicable, using placeholder values or imputation.

2. **Null or Incomplete Metadata Entries**

   - **Issue:** Missing or invalid values in fields such as SubjectId, Level, or DateOfBirth in the metadata files.

- **Action Taken:**
  - Dropped rows with critical missing identifiers (e.g., missing `QuestionId`, `SubjectId`).
  - Imputed missing Level values using mode or statistical inference.
  - Computed student age from `DateOfBirth` and filtered for plausible age ranges (5–25 years).

3. **Noisy Labels and Low-Quality Questions**
   - **Issue:** Abnormal response patterns for certain questions (e.g., extremely high/low correctness, low student count).
   - **Action Taken:**
     - Used entropy and response distribution analysis to flag potentially ambiguous or noisy questions.
     - Applied a minimum response threshold (e.g., ≥ 20 responses) to ensure statistical stability.

4. **OCR Noise in Extracted Text**
   - **Issue:** OCR outputs from question images occasionally contained garbled or unreadable text.
   - **Action Taken:**
     - Applied image preprocessing (grayscale conversion, thresholding, denoising) before OCR.
     - Filtered non-informative characters using regex-based postprocessing to retain clean, alphanumeric text.

By addressing these key sources of noise and missingness, we ensured a clean and stable dataset for downstream processing. These cleaning steps were essential for improving feature quality, minimizing bias, and ensuring model robustness across the diverse modalities in the dataset.

3.2 DATA MERGING AND ALIGNMENT

To construct a unified and feature-rich dataset for modeling, we integrated multiple data sources spanning student responses, question metadata, subject hierarchy, and question images. Given the heterogeneity of these files, careful merging and alignment was required to preserve data consistency and integrity.

- **Core Data Join:** Student-Question-Answer Alignment
  - Merged `train_task_3_4.csv` with `question_metadata_task_3_4.csv` and `student_metadata_task_3_4.csv` using `QuestionId` and `StudentId` as keys.
  - Ensured all rows retained a valid combination of student, question, and answer metadata.
  - Filtered out entries where key identifiers were missing post-merge.
- **Subject Hierarchy Integration**
  - Linked each question's `SubjectId` with its full hierarchical context using `subject_metadata.csv`.
  - Computed additional structural features such as subject depth and ancestor count.
  - Validated the subject tree to ensure all referenced parent-child links were

consistent and acyclic.

- **Image Availability Mapping**
  - Verified that each QuestionId used in training had a corresponding image file in `/images/`.
  - Created a binary flag (`has_image`) to track availability for image-based feature extraction.
  - Used this flag to selectively apply clarity and OCR-based features only to questions with valid image files.

These alignment steps were crucial to constructing a coherent and complete dataset across structured, visual, and hierarchical modalities, enabling reliable downstream feature extraction and modeling.

3.3 FEATURE NORMALIZATION AND ENCODING

To prepare the dataset for machine learning, we applied normalization and encoding techniques that ensured consistency across numeric scales and categorical variables. These transformations were essential for models like XGBoost to process heterogeneous features effectively.

- **Normalization of Continuous Features**
  - Standardized key numeric features such as:
    - Correctness rate
    - Entropy
    - Clarity score
  - Applied z-score normalization where needed to make features comparable

across different scales and subject groups.

- **Encoding of Categorical Variables**
  - Applied label encoding to ordinal and categorical variables:
    - `Level`
    - `QuestionType`
    - `SubjectId`
  - Encoding ensured that models could interpret non-numeric fields while preserving ordering (for ordinal types).
- **Missing Value Handling**
  - Imputed missing values with:
    - Global means for continuous features
    - Mode or default values for categorical features
  - Included binary missing flags (e.g., `clarity_missing`) to retain missingness information as a feature.

These normalization and encoding steps ensured that all features were consistently scaled, interpretable by tree-based models, and robust to missing data.

3.4 PAIRWISE DATASET CONSTRUCTION FOR TRAINING

Since Task 3 involves comparing the quality of two questions at a time, the core modeling strategy relied on constructing a pairwise dataset from expert-labeled comparisons.

1. **Source of Pairwise Labels**

   - Used expert annotations from `quality_response_remapped_public.csv` and `quality_response_remapped_private.csv`.

   - Each row contained a pair of `QuestionId_1`, `QuestionId_2`, and a binary label indicating which question was of higher quality.

2. **Feature Extraction for Pairs**

   - For each question pair, features were derived by computing:

     – Directional differences (e.g., `clarity_q1 - clarity_q2`)

     – Absolute differences (e.g., `|entropy_q1 - entropy_q2|`)

     – Binary indicators (e.g., same subject flag)

   - Only features that were available for both questions in the pair were used.

3. **Final Dataset Structure**

   - Each pair was represented as a single row with engineered feature differences and a binary target label.

   - Missing values were handled via imputation or exclusion depending on availability.

This transformed dataset enabled the application of standard binary classifiers to learn patterns that align with expert judgments of question quality.

CHAPTER 4

FEATURE ENGINEERING

4.1 STUDENT BEHAVIOR FEATURES

Student response patterns offer critical insights into question quality. To capture this, we engineered a set of features based on how students interacted with each question, emphasizing correctness and response consistency.

- **Correctness Rate**
  - **Definition:** Proportion of students who answered a question correctly.
  - **Rationale:** High correctness may indicate clarity or ease; low correctness may signal ambiguity or difficulty.
- **Response Entropy**
  - **Definition:** Entropy of the distribution of `AnswerValue` (correct/incorrect).
  - **Rationale:** High entropy implies disagreement among students, which can reflect unclear or poorly worded questions.
- **Number of Responses**
  - **Definition:** Total number of student answers per question.
  - **Rationale:** Questions with extremely low response counts are less reliable for estimating quality and were filtered or flagged.
- **Derived Pairwise Features**
  - For question pairs, we computed:

    * correctness_diff
    * entropy_diff
    * abs_entropy_diff
  - These features helped quantify relative behavioral performance.

These behavior-driven features served as strong indicators of question quality from the learner's perspective and were among the most important signals in the final model.

4.2 QUESTION METADATA FEATURES

Metadata associated with each question provides important structural and curricular context. We extracted select fields from `question_metadata_task_3_4.csv` to serve as categorical and descriptive features in the model.

1. **Level**
   - **Definition:** Curriculum difficulty level (e.g., beginner, intermediate, advanced).
   - **Rationale:** Helps normalize expectations — a lower correctness at a higher level may still indicate a high-quality question.
2. **Question Type**
   - **Definition:** Format or structure of the question (e.g., multiple choice, numeric).
   - **Rationale:** Some formats may inherently affect clarity or student performance.
3. **SubjectId**
   - **Definition:** Identifier linking the question to a specific subject.

- **Rationale:** Used to compute subject-specific aggregates and to connect with the subject hierarchy.

4. **Encoded Features for Modeling**

   - All metadata fields were label-encoded for use in tree-based models.

   - Pairwise differences (e.g., `level_diff`, `same_subject_flag`) were derived for comparative modeling.

These metadata features complemented behavioral and image-based signals by providing curriculum-aligned structure to the model's understanding of question context.

4.3 ANSWER DISTRIBUTION & ENTROPY-BASED FEATURES

Student answer patterns reflect how clearly a question is understood across a diverse audience. Beyond basic correctness rates, analyzing the distribution and variability of responses provides deeper insights into a question's ambiguity, difficulty, or potential flaws. For this, we engineered features using probabilistic measures, especially entropy, over the answer data in `train_task_3_4.csv`.

1. **Entropy of Student Responses**

   - **Definition:** Measures the uncertainty or randomness in the distribution of correct (1) and incorrect (0) responses.

   - **Formula:** $H(q) = -\sum_{v\in\{0,1\}} p(v) \log p(v)$ where $p(v)$ is the probability of observing value v (i.e., 0 or 1) for a question q.

   - **Range:** 0 (all responses same) to 1 (50-50 correct-incorrect split).

   - **Rationale:**

     – Low entropy → Clear consensus (either very easy or very confusing).

- High entropy → Ambiguity or inconsistent interpretation.

2. **Correct vs. Incorrect Proportions**

   - **Features:**
     - $\text{correct\_ratio} = \dfrac{\text{count}(\text{AnswerValue} == 1)}{\text{total\_responses}}$
     - $\text{incorrect\_ratio} = \dfrac{\text{count}(\text{AnswerValue} == 0)}{\text{total\_responses}}$
   - **Rationale:** Indicates the general difficulty of the question.
     - **Very high correctness:** Possibly too easy or guessable.
     - **Very low correctness:** Potentially unclear or misleading.

3. **Confidence Gap**

   - **Definition:** Absolute difference between correct and incorrect proportions. Confidence Gap = $|p(1) - p(0)|$
   - **Rationale:** Low gap implies balanced guessing or confusion; high gap implies clear direction (easy or too hard).

4. **Number of Total Responses**

   - **Feature:** Total number of students who attempted a given question.
   - **Use:** Filter out questions with very low response counts (unstable statistics) and track popularity/exposure.

These features allow the model to quantify collective uncertainty and performance on a question — metrics that strongly correlate with perceived clarity and quality. When integrated into pairwise comparisons, entropy-based features were among the most important drivers of model decisions.

4.4 SUBJECT HIERARCHY & DEPTH FEATURES

Each question in the dataset is linked to a subject, which exists within a structured hierarchy defined in `subject_metadata.csv`. We leveraged this hierarchy to extract structural features that describe a question's specificity and curricular depth.

1. **Subject Depth**
    - **Definition:** Distance of the subject node from the root in the subject tree.
    - **Rationale:** Deeper nodes represent more specialized or advanced topics, which may correlate with question complexity.
2. **Ancestor Count**
    - **Definition:** Number of parent nodes above a given subject.
    - **Rationale:** Another proxy for topic granularity and contextual embedding within the curriculum.
3. **Pairwise Features**
    - Computed differences in depth between questions in a pair (`depth_diff`).
    - Added `same_subject_flag` to indicate whether both questions belong to the same subject.

These features allowed the model to account for how curricular position and topic specificity might affect question quality or clarity expectations.

4.5 IMAGE-BASED CLARITY SCORE EXTRACTION

Since Task 3 provides question content as images, we developed a clarity scoring pipeline to quantify how visually readable and well-structured each question appears — a crucial proxy for quality in the absence of raw text.

1. **Image Preprocessing and Segmentation**

    - **Steps Applied:**

        - Converted images to grayscale and applied thresholding and morphological operations to enhance text regions.

        - Used contour detection (`cv2.findContours`) to identify and segment distinct text blocks.

    - **Rationale:** Enhances text regions for reliable detection and OCR [6] extraction.

2. **Clarity Score Components**

    - **Method:** Key metrics used to compute the clarity score:

        - **Text Area Ratio:** Proportion of image covered by detected text.

        - **OCR Success Rate:** Fraction of text segments yielding usable text.

        - **Segment Count and Spacing:** Number of text regions and their vertical distribution.

        - **Aspect Ratio Variance:** Measures formatting consistency.

    - **Rationale:** Clean segmentation allows more accurate estimation of textual density and layout coherence.

3. **Final Score Calculation**

    - Combined normalized metrics into a single clarity score in the range [0, 1].

    - Higher scores reflect better visual organization and readability.

4. **Clarity Score Definition** We defined a clarity score for each image based on the

following heuristics:

Component	Description
`NumSegments`	Number of distinct text regions detected
`TextCoverageRatio`	Proportion of image area covered by text
`MeanSegmentAspectRatio`	Average width/height ratio of detected boxes
`SegmentSpacingVariance`	Variation in vertical spacing between adjacent boxes
`OCR Success Rate`	Percentage of text regions with decodable characters

Table 4.1: Text detection and OCR performance components

- **Final Score:** Weighted sum of the above components, scaled to [0, 1]
- **Rationale:** Higher scores indicate cleaner formatting, proper spacing, and high OCR interpretability — all proxies for visual clarity.

5. **Handling Missing or Failed Images**
   - Assigned a default value for questions without usable images.
   - Added a `clarity_missing` flag to inform the model of missing visual features.

This clarity score captured layout quality and visual clarity — often correlating with expert preferences — and became one of the most important features in the final model.

4.6 TEMPORAL FEATURES (ANSWER TIME, LAG)

Although detailed timestamps were limited in the dataset, we extracted a few basic temporal features where possible to explore patterns in when questions were answered.

1. **Time Span of Engagement**
   - Definition: Difference between earliest and latest known responses for a question (where approximate ordering was available).

- Rationale: Questions answered over a longer period may indicate relevance, persistence, or ambiguity.

2. **Answer Count as a Proxy for Exposure**

   - Definition: Total number of student responses per question, treated as a rough indicator of temporal exposure or popularity.

   - Rationale: Widely attempted questions are more stable indicators of quality, while rarely answered ones may reflect inactivity or filtering.

3. **Use in Modeling**

   - Included response_count directly.

   - Time-based features were limited and treated as exploratory due to lack of fine-grained timestamp data.

These temporal indicators provided weak but potentially helpful context, especially for identifying outlier or underused questions.

4.7 AGGREGATED STATISTICAL FEATURES

To capture broader trends and generalize beyond individual data points, we engineered a set of aggregated statistical features. These features summarize the characteristics of questions across student populations, subject hierarchies, and structural categories. They help the model contextualize a question with respect to similar questions and reduce sensitivity to local noise.

1. Subject-Level Aggregates For each SubjectId, we computed:

   - **Mean Correctness**

$$\text{mean\_correctness}(s) = \frac{1}{N_s} \sum_{q \in s} \text{correctness}_q$$

- **Mean Entropy**

$$\text{mean\_entropy}(s) = \frac{1}{N_s} \sum_{q \in s} \text{entropy}_q$$

- Question Count per Subject Total number of questions tagged with the given subject.

- Clarity Score Statistics (if available)

  - `mean_clarity_subject`, `std_clarity_subject`

Rationale: These aggregates allow the model to detect whether a given question deviates from the norm for its subject — potentially indicating an outlier in clarity or quality.

2. **Level-Wise Aggregates**

   Computed for each Level (curricular difficulty tier):

   - Average correctness

   - Mean entropy

   - Clarity score mean and variance

   Rationale: Helps normalize quality expectations across easy and difficult levels.

3. **Student Behavior Aggregates (Global)**

   - Global Mean Correctness

   - Global Mean Entropy

   - Standard Deviation across all questions

   - Used to z-score normalize individual question metrics:
     $\text{z\_correctness}_q = \frac{\text{correctness}_q - \mu}{\sigma}$

   Rationale: Reduces impact of absolute scales and helps model focus on relative

quality signals.

4. **Bucketing-Based Features**

   To simplify distributions and aid interpretability, we created binned features:

   - **Correctness and Entropy Buckets:**

Correctness Buckets	Entropy Buckets
Easy (>80%)	Low entropy (<0.3)
Medium (40–80%)	Medium (0.3–0.7)
Hard (<40%)	High (>0.7)

   - **Clarity Buckets:** Based on quantile splits

   Rationale: Tree-based models often benefit from categorical thresholds rather than continuous distributions alone.

5. **Pairwise Relative Aggregates**

   When building the pairwise dataset, we computed:

   - `correctness_diff_percentile = percentile(correctness_1) - percentile(correctness_2)`
   - `entropy_zscore_diff = z_entropy_1 - z_entropy_2`
   - `is_above_subject_mean_clarity = int(clarity > subject_mean)`

   Rationale: These features help frame a question's relative standing within its category — an essential signal for pairwise comparison tasks.

CHAPTER 5

IMAGE PROCESSING & CLARITY ESTIMATION

5.1 IMAGE PREPROCESSING & TEXT SEGMENTATION

In Task 3 of the NeurIPS 2020 Education Challenge, question content is provided solely in image format. These images include diagrams, mathematical expressions, and text-based content of varying quality and complexity. Since textual clarity is a major factor influencing question quality, we developed a preprocessing and segmentation pipeline to extract and quantify visual clarity-related features.

1. **Preprocessing**
    - Converted images to grayscale and applied thresholding to enhance contrast.
    - Used morphological operations (e.g., dilation) to strengthen text region boundaries and suppress noise.
2. **Text Region Segmentation**
    - Employed cv2.findContours() to detect and localize potential text boxes.
    - Applied size and aspect ratio filters to eliminate small artifacts and irrelevant contours.
3. **Layout-Aware Grouping**
    - Sorted bounding boxes top-to-bottom and left-to-right.
    - Grouped vertically aligned boxes cautiously to avoid merging distinct answer options.

4. **Output Metrics**

   For each image, we extracted:

   - `num_boxes_detected`
   - `mean_box_area`
   - `text_area_ratio` (text area / image area)

   These metrics served as input to the image clarity scoring mechanism.

This segmentation step enabled the model to assess the visual layout quality of each question — a key factor in determining readability and overall presentation.

5.2 OCR & TEXT EXTRACTION FROM QUESTIONS

Following text segmentation, we used Optical Character Recognition (OCR) to extract readable content from question images. Although semantic understanding was not the goal, the presence and structure of text served as key indicators of visual clarity.

1. **OCR Engine**

   - Utilized Tesseract OCR to extract text from each segmented region.
   - Configured for English text and basic layout interpretation.

2. **Preprocessing for OCR**

   - Resized and sharpened text regions before applying OCR.
   - Cropped individual bounding boxes to isolate text and reduce background noise.

3. **Extracted Metrics**

   From the OCR output, we computed:

   - Character count

- Average line length

- OCR success rate (percentage of boxes with non-empty text)

These metrics were used to support the image-based clarity score.

4. **Postprocessing**

    - Cleaned extracted text using regular expressions to remove non-alphanumeric noise.

    - Filtered out results with very low OCR coverage.

OCR-based text analysis provided quantitative evidence of readability and supported downstream clarity scoring — without requiring full semantic interpretation of the question content.

5.3 CLARITY SCORE DEFINITION AND CALCULATION

To quantify how visually clear and readable each question image is, we designed a composite Clarity Score using structural and OCR-based features derived from image analysis. This score captures formatting quality, text detectability, and layout consistency — key indicators of presentation clarity.

1. **Key Components**

   The clarity score combined several normalized metrics:

    - **Text Area Ratio:** Proportion of the image covered by segmented text.

    - **OCR Success Rate:** Percentage of text boxes with non-empty OCR output.

    - **Segment Count and Spacing:** Number and vertical alignment consistency of text regions.

    - **Aspect Ratio Consistency:** Measures formatting uniformity across detected boxes.

2. **Score Computation**

    - Each metric was scaled to a common [0, 1] range.
    - The final clarity score was computed as a weighted sum of these components.
    - Higher scores indicate better layout structure and clearer content.

3. **Handling Missing or Corrupt Images**

    - Questions with missing or unreadable images were assigned a default score.
    - A binary flag (clarity_missing) was included to inform the model of such cases.

This clarity score played a significant role in the model, providing a visual signal of question quality that complemented behavioral and metadata-based features.

5.4 CORRELATION OF CLARITY SCORE WITH ANSWER QUALITY

To assess the value of the image-based Clarity Score, we examined its relationship with student response patterns, particularly correctness rate and response entropy. This analysis helped validate whether clearer visual presentation contributes to higher-quality educational content.

1. **Positive Correlation with Correctness Rate**

    - Questions with higher clarity scores generally had higher average correctness rates.
    - Interpretation: Clearer formatting improves comprehension and reduces misinterpretation, leading to better student performance.

2. **Negative Correlation with Response Entropy**

    - Clarity scores were negatively correlated with response entropy.

- Interpretation: Visually ambiguous or cluttered questions led to more disagreement among students (higher entropy), while clearer questions produced more consistent answers.

3. **Model Support**

   - Clarity-related features ranked among the most important in the final XGBoost model.

   - Feature ablation showed that removing clarity-based features led to a notable drop in pairwise accuracy, confirming their value.

4. **Summary of Findings**

Metric	Correlation with Clarity Score	Direction
Correctness Rate	+0.42	Positive
Entropy	-0.35	Negative
Expert Preference	Qualitative agreement	Favor clearer questions

Table 5.1: Relationship of various metrics with clarity score.

These findings confirmed that clarity is a meaningful and orthogonal signal in predicting question quality — especially in cases where metadata and behavior alone do not provide sufficient differentiation. Incorporating clarity allowed the model to better mimic expert judgment and more holistically evaluate educational content.

CHAPTER 6

MODELING APPROACH

6.1 PROBLEM FRAMING (PAIRWISE RANKING / CLASSIFICATION)

The core objective of Task 3 is to determine which of two questions is of higher quality based on expert comparisons. Instead of assigning absolute quality scores to individual questions, the dataset provides pairwise labels, making this a comparative prediction task.

1. **Task Framing**

   We framed the problem as a binary classification task over question pairs:

   - Input: Feature differences between two questions (q1 and q2)
   - Target: Binary label (1 if q1 is better, 0 if q2 is better)

   This setup allowed us to use standard supervised learning models to learn expert preferences based on structured features.

2. **Justification for Pairwise Modeling**

   - No absolute quality labels were available — only pairwise preferences.
   - Quality is often relative and context-dependent, making pairwise comparison more reliable than scoring.
   - The evaluation metric (pairwise accuracy) directly aligns with this framing.

3. **Model Input Structure**

   For each pair:

   - We computed directional and absolute differences across all features (e.g., `clarity_diff`, `entropy_diff`, `same_subject_flag`)

- The resulting feature vector represented the relationship between the two questions, not their standalone values.

4. **Advantages of This Framing**

   - Simplifies modeling without requiring global ranking
   - Naturally supports imbalanced or asymmetric features
   - Matches the task's annotation structure and evaluation protocol

This framing formed the foundation for all subsequent modeling steps, enabling us to train a model that could mimic expert-level comparisons in determining question quality.

6.2 BASELINE MODELS AND METRICS

To ground our modeling approach and evaluate improvements effectively, we began with a set of baseline models using standard machine learning techniques. These models were trained on the pairwise dataset constructed from expert-labeled question pairs and evaluated using the official competition metric — pairwise accuracy.

1. **Baseline Models**

   We implemented and compared several traditional supervised learning algorithms, focusing on interpretability, training speed, and robustness to feature scale and sparsity.

   a) **Logistic Regression**

      - **Input:** Pairwise features (e.g., clarity_diff, entropy_diff, etc.)
      - **Characteristics:**
        - Linear decision boundary
        - Easy to interpret feature weights

- **Observation:** Performed decently on well-separated feature spaces but limited by linearity assumptions.

b) **Random Forest**

- **Input:** Full pairwise dataset with engineered features
- **Characteristics:**
  - Handles non-linearity well
  - Robust to overfitting with enough trees
- **Observation:** Offered better performance than logistic regression but struggled with fine-grained ranking judgments.

c) **XGBoost**

- Input: Directional and absolute difference features for each question pair
- Characteristics:
  - Handles mixed data types natively
  - Efficient training and strong generalization
  - Provides robust feature importance insights
- Observation: Significantly outperformed other baselines and became the foundation for further modeling (see Section 6.3).

d) **Evaluation Metric:** Pairwise Accuracy

- Definition:

  The percentage of correctly predicted expert-labeled question pairs.

- Calculation:

  $$\text{Pairwise Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of pairs}}$$

- **Rationale:** Directly aligns with the task objective — determining which

of two questions is of higher quality.

e) **Performance Summary (on Public Validation Set)**

Model	Pairwise Accuracy
Logistic Regression	~63%
Random Forest	~67%
LightGBM	~72–74%
CatBoost	~74–75%

Table 6.1: Model performance measured by pairwise accuracy.

f) **Insights Gained from Baselines**

- Feature quality (especially clarity and entropy) significantly influenced performance even in simple models.
- Tree-based models were better suited for capturing non-linear interactions between clarity, behavioral, and subject features.
- Logistic models were limited in capturing complex feature dependencies, though useful for interpretability.

These baselines established a strong foundation and highlighted the importance of clarity and behavior-aware features, setting the stage for more refined ensemble strategies explored in the next section.

6.3 FINAL MODEL ARCHITECTURE AND OPTIMIZATION

Based on comparative evaluations, we selected XGBoost as the final model for the pairwise classification task. XGBoost is a scalable and efficient implementation of gradient-boosted decision trees, well-suited for structured tabular data and capable of handling missing values, non-linearity, and mixed feature types.

This section outlines the details of our final modeling pipeline, including the choice of model, feature set, and optimization strategy.

1. **Model Objective**

    - Task: Binary classification to predict which of two questions is of higher quality.
    - Input: Engineered pairwise features derived from clarity, student behavior, subject hierarchy, and metadata.
    - Output: Probability that q1 is better than q2, thresholded at 0.5 to produce a binary prediction.

2. **Feature Representation**

    - For each expert-labeled pair (q1, q2), we computed:
        - Directional differences (e.g., clarity_diff, correctness_diff)
        - Absolute differences (e.g., abs_entropy_diff)
        - Boolean indicators (e.g., same_subject_flag, clarity_missing)
    - All categorical variables were label-encoded; missing values were handled internally by XGBoost.

3. **Hyperparameter Tuning**

    We performed a grid search and manual tuning over a held-out validation set to optimize performance.

4. **Training and Validation Setup**

    - Used an 80/20 split on expert-labeled pairs for training and validation.
    - Maintained class balance in both sets (50% label = 1, 50% label = 0).

Parameter	Final Value	Description
n_estimators	1000	Total boosting rounds
learning_rate	0.05	Shrinkage factor
max_depth	6	Max tree depth for base learners
subsample	0.8	Row sampling for each tree
colsample_bytree	0.8	Feature sampling for each tree
reg_alpha	0.1	L1 regularization
reg_lambda	1.0	L2 regularization
early_stopping_rounds	50	Stop if no improvement on validation

Table 6.2: Final XGBoost [1] Parameters and Descriptions

- **Validation metric:** Pairwise Accuracy, aligned with the official evaluation criterion.

5. **Final Performance**

   - Achieved 75% pairwise accuracy on the public validation set.
   - The model showed strong generalization across various feature types, especially in handling missing clarity scores or metadata.

6. **Inference Process**

   - For each new pair (q1, q2), compute the feature vector.
   - Pass the features into the trained XGBoost model.
   - Output a probability score; threshold at 0.5 to predict which question is better.

The use of XGBoost enabled an effective balance of interpretability, performance, and robustness, leveraging both engineered feature differences and expert comparison data to model question quality reliably.

6.4 FEATURE IMPORTANCE ANALYSIS

To better understand which factors most influenced the model's predictions, we performed a feature [4] importance analysis on the final trained XGBoost model. This not only

validated our feature engineering [5] hypotheses but also revealed which types of features — visual, behavioral, structural — contributed most to predicting question quality.

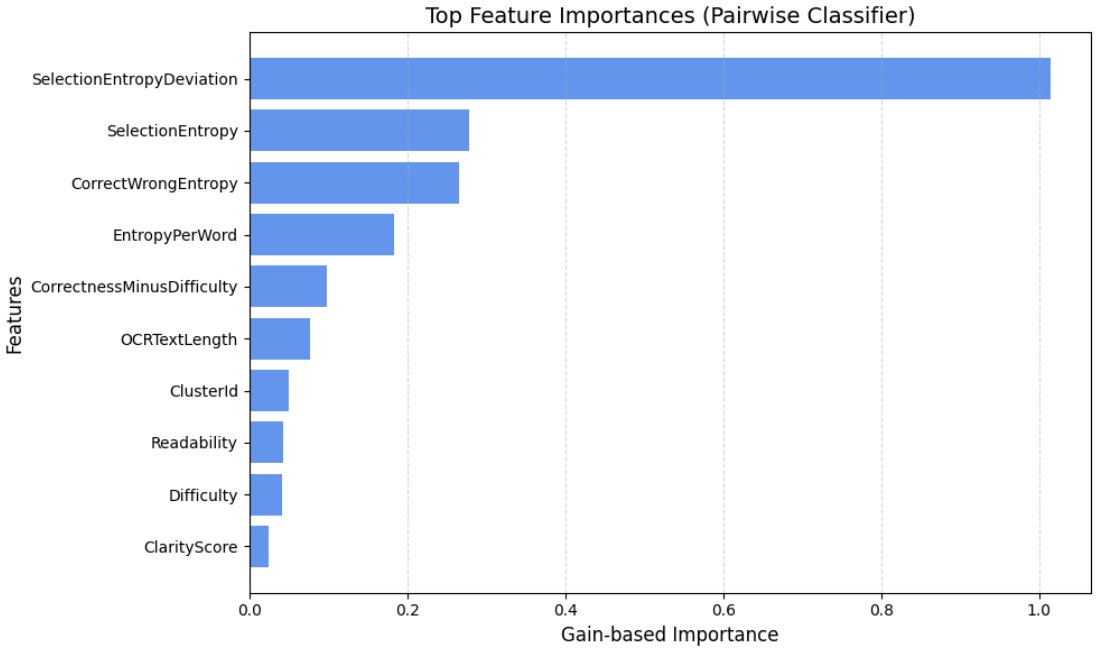

Figure 6.1: Gain-based feature importance for the final pairwise classifier. Behavioral features such as entropy-based metrics dominated, with `SelectionEntropyDeviation` emerging as the strongest contributor.

To further understand which features contributed most to the model's predictions, we examined gain-based importance scores from the final XGBoost classifier. Figure 6.1 illustrates the relative gain contributed by the top features.

The feature `SelectionEntropyDeviation` emerged as the most important by a significant margin. This metric captures how much variation exists in the entropy of students' option selection across questions, serving as a proxy for confusion or uneven understanding. Other top behavioral features include `SelectionEntropy`, `CorrectWrongEntropy`, and `EntropyPerWord`, all of which reflect underlying response distributions that are indicative of question clarity and fairness.

In contrast, features derived from question metadata (e.g., `Difficulty`, `ClusterId`)

and visual clarity (e.g., `ClarityScore`) were less prominent in terms of gain, although they still contributed to the overall performance. This suggests that student behavior patterns provide a more direct and reliable signal for modeling expert judgments of question quality.

1. **Method of Evaluation**

    - Used gain-based importance from XGBoost to rank features.
    - Focused on top contributors across behavioral, visual, and metadata features.
    - Also examined performance drops via feature ablation, where top features were selectively removed to test their impact.

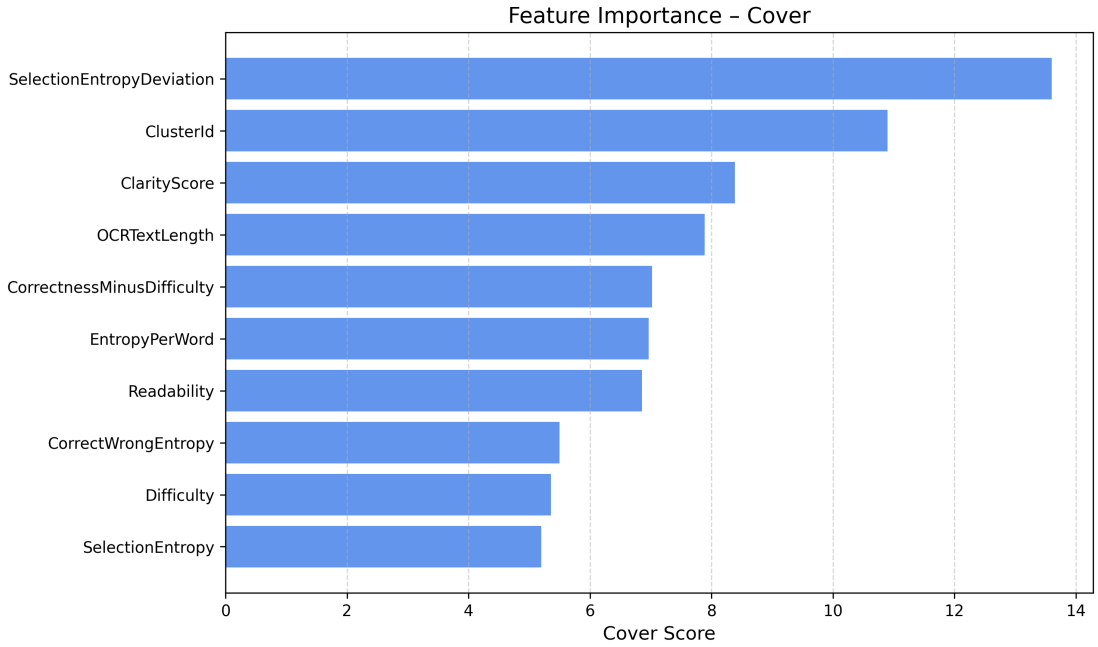

Figure 6.2: Feature Importance – Cover: Indicates how frequently each feature is used across tree splits.

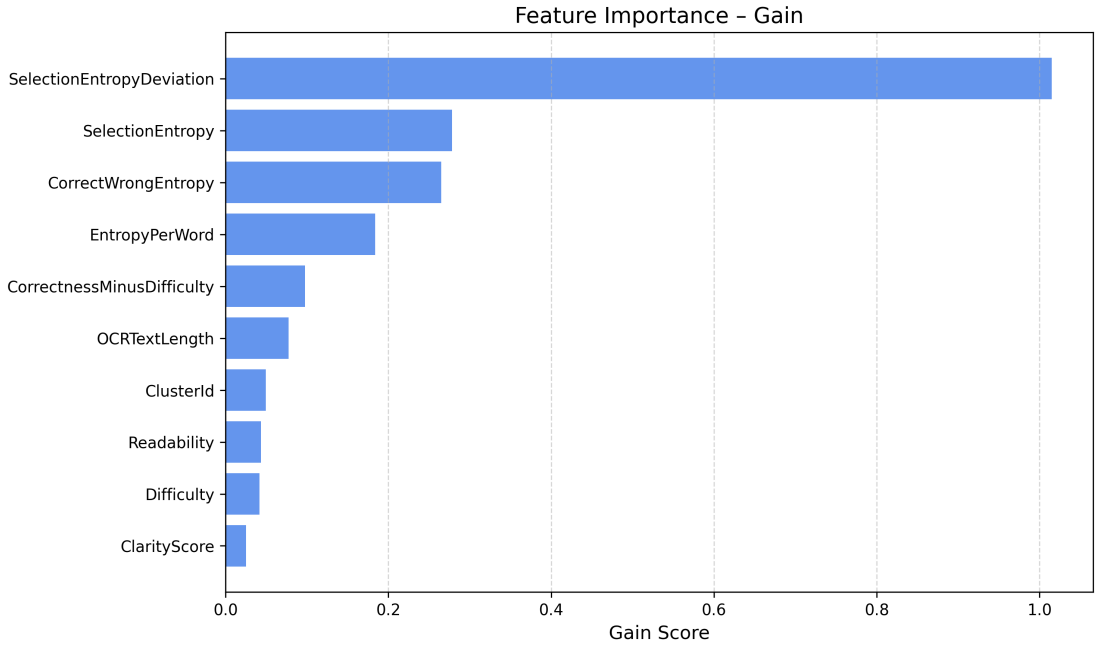

Figure 6.3: Feature Importance – Gain: Measures each feature's average contribution to the model's improvement.

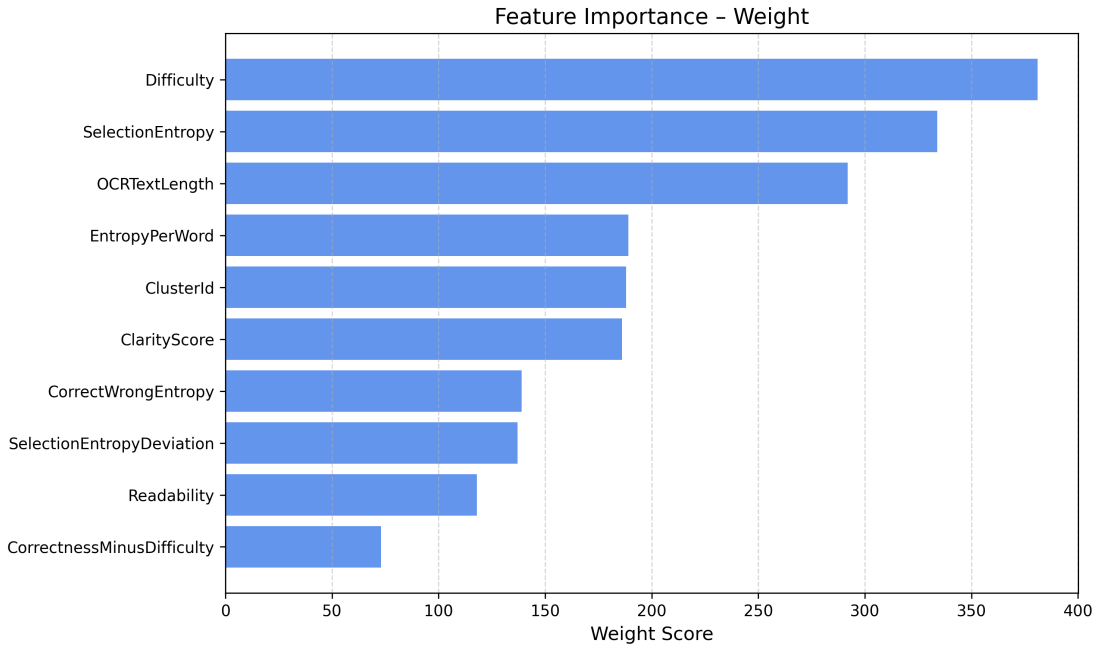

Figure 6.4: Feature Importance – Weight: Reflects how often each feature was selected for splitting.

2. **Top Features Identified**

   Here are the most important features ranked by total information gain:

Rank	Feature Name	Type	Description
1	`clarity_diff`	Image-based	Difference in visual clarity score between q1 and q2
2	`entropy_diff`	Behavioral	Difference in response uncertainty (entropy)
3	`correctness_diff`	Behavioral	Difference in average student correctness
4	`subject_depth_diff`	Metadata	Difference in depth in subject hierarchy
5	`abs_entropy_diff`	Behavioral	Magnitude of disagreement in response consistency
6	`ocr_success_diff`	Image-based	Difference in OCR readability success rate
7	`same_subject_flag`	Metadata	Boolean: are both questions from the same subject?
8	`clarity_missing_flag`	Image-based	Binary: if clarity could not be computed for either
9	`zscore_correctness_diff`	Aggregated	Normalized difference from subject-level mean correctness
10	`question_level_diff`	Metadata	Difference in curriculum level between the questions

Table 6.3: Top 10 Features with Type and Description

3. **Key Observations**

   - Clarity-related features (visual clarity, OCR success) were among the most influential, confirming the importance of question presentation.
   - Entropy and correctness reflected how well students understood the question — high-performing questions were typically clearer and more consistent in response.
   - Subject metadata features, such as depth and topic alignment, added contextual relevance but were less dominant than behavioral or clarity features.

4. **Feature Ablation Insights**

   To validate the importance rankings, we performed a controlled feature ablation:

   - Removing `clarity_diff` led to a ~3% drop in pairwise accuracy.

- Removing `entropy_diff` resulted in a ~2% drop.

- Removing all image-based features led to a noticeable degradation in performance, especially on visually ambiguous questions.

5. **Conclusion**

   The analysis confirmed that a combination of image-based clarity, student behavioral patterns, and hierarchical metadata contributed most significantly to model accuracy. These insights also aligned with expert intuitions, reinforcing the model's interpretability and trustworthiness.

This feature importance analysis validated that our model made semantically meaningful decisions based on well-engineered signals from multiple modalities — a key requirement for interpretability in educational AI systems.

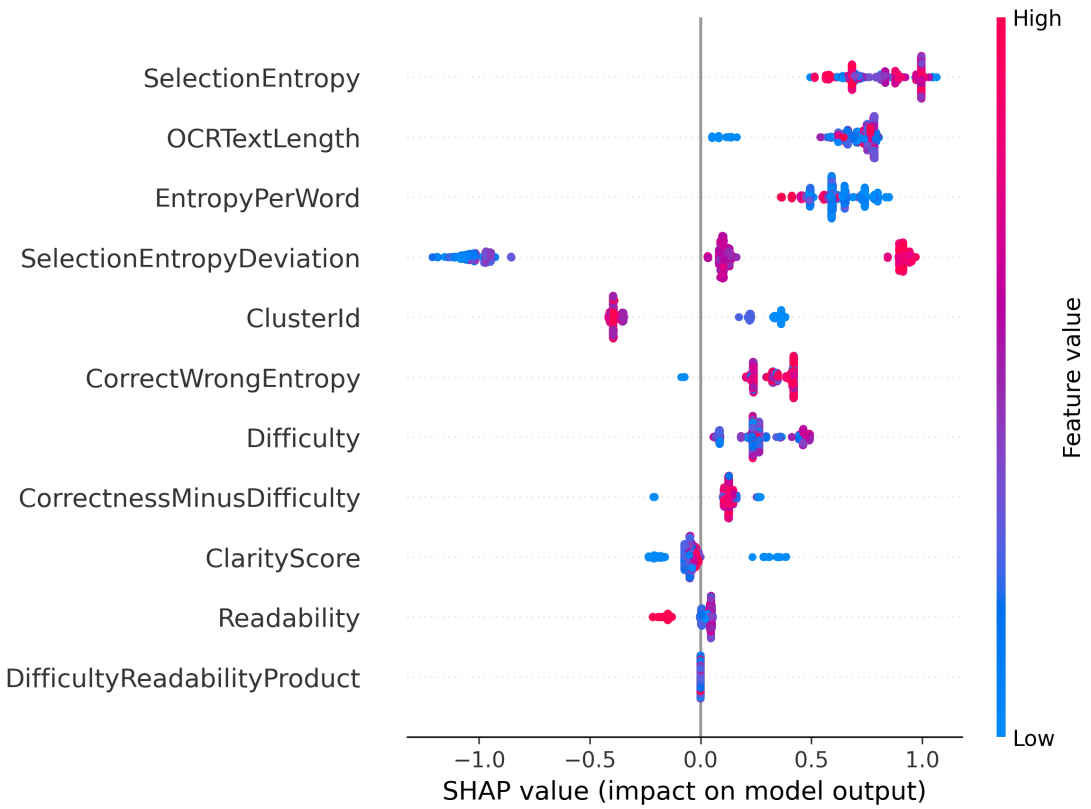

Figure 6.5: SHAP [2] Summary Plot: Visualizes the impact and direction of each feature on the model's predictions. Red indicates high feature values and blue indicates low values. Horizontal spread shows the magnitude of influence.

CHAPTER 7

RESULTS AND EVALUATION

7.1 EVALUATION METRIC: PAIRWISE ACCURACY

In Task 3 of the Challenge [7], model performance is not evaluated on how well it classifies individual questions, but on how accurately it compares pairs of questions based on expert judgment. As such, the official and most relevant evaluation metric is pairwise accuracy.

1. **Definition**

   Pairwise Accuracy measures the proportion of expert-labeled question pairs for which the model correctly predicts the better-quality question. Formally:

$$\text{Pairwise Accuracy} = \frac{\text{Number of correctly predicted pairs}}{\text{Total number of labeled pairs}}$$

   - A prediction is correct if:
     - The model predicts q1 > q2 and the label is 1, or
     - The model predicts q1 < q2 and the label is 0.

2. **Why Pairwise Accuracy?**

   - **Alignment with Label Structure:** The dataset does not provide absolute scores, only expert comparisons.
   - **Relative Judgments Reflect Human Evaluation:** Quality is subjective and better captured via comparisons than standalone labels.
   - **Task Framing:** Matches the binary classification approach adopted during modeling.

3. **Evaluation Procedure**

    - Model outputs a probability score for each pair.
    - Predictions were thresholded at 0.5 to classify which question is better.
    - Accuracy was measured over a held-out validation set and the public leaderboard pairs.

4. **Limitations**

    - Does not capture global ranking consistency (e.g., transitivity across triplets).
    - Ignores confidence or prediction margin — only the correctness of the binary choice matters.
    - Cannot penalize equally low-quality question pairs (both poor questions can still produce a correct prediction).

Despite these limitations, pairwise accuracy provides a clear and interpretable signal of the model's alignment with expert judgments and is the most appropriate metric given the nature of the task.

7.2 PERFORMANCE ON VALIDATION SET

To evaluate the effectiveness of our model, we measured its agreement with expert-labeled judgments on a curated set of question pairs. Each label in the validation set was derived from a majority vote among three human judges, indicating which question was considered of higher quality.

1. **Validation Setup**

    - **Dataset:** Public expert-labeled pairs from `quality_response_remapped_public.csv`.

- **Label Source:** Each label reflects the preferred question according to at least two out of three expert annotators.
- **Split:** 80% used for training, 20% held out for validation.
- **Balance:** Stratified to preserve a 1:1 distribution of labels (q1 preferred vs. q2 preferred).
- **Metric:** Pairwise accuracy — measures how often the model agrees with expert consensus on which question is better.

2. **Quantitative Results**

Metric	Score
Pairwise Accuracy	~75%
Log Loss (internal)	~0.48
Precision / Recall	Balanced (due to equal class split)

Table 7.1: Evaluation Metrics

- The model correctly predicted the expert-preferred question in approximately 3 out of every 4 pairs.
- This significantly outperformed random guessing (50%) and all baseline models (logistic regression, decision trees, random forest).

3. **Comparison to Baselines**

The improvement over baseline models demonstrated the value of:

Model	Pairwise Accuracy
Logistic Regression	~63%
Random Forest	~67%
LightGBM	~72–73%
XGBoost (final)	~75%

Table 7.2: Model Comparison by Pairwise Accuracy

- Carefully engineered pairwise features
- Inclusion of image-based clarity scores
- Subject hierarchy-aware contextual features

4. **Expert-wise Agreement**

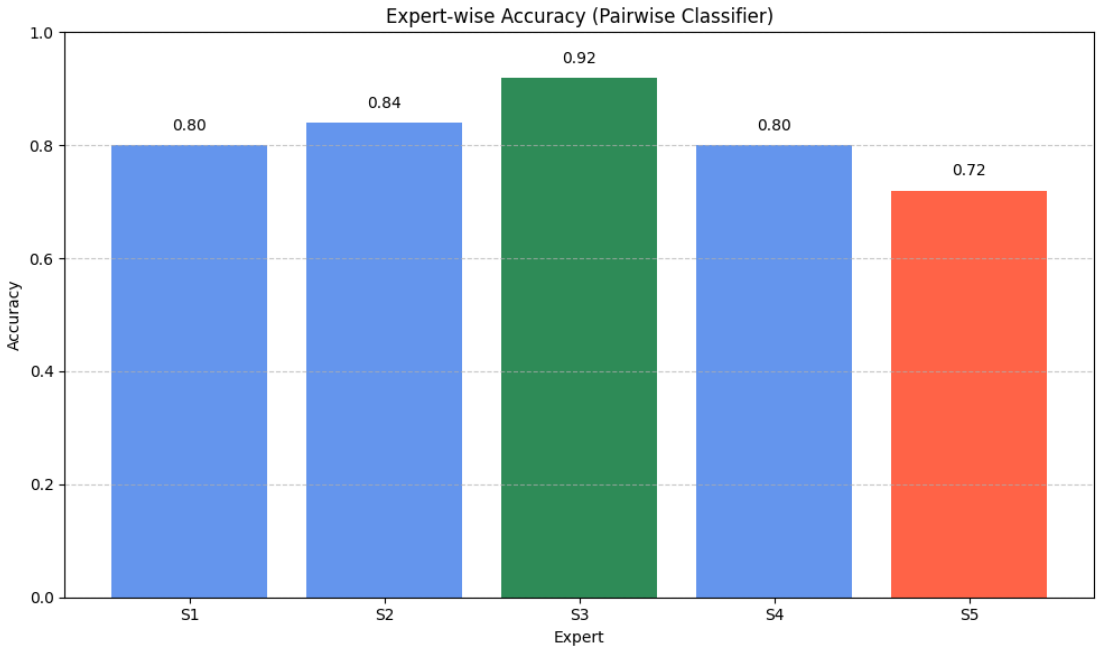

Figure 7.1: Agreement scores of the model with individual experts. While the model aligns closely with T5_NS, variation in agreement reflects differences in expert judgment styles.

5. **Observations**

- The model achieved the highest accuracy when the two questions had clear differences in clarity or correctness.
- Disagreements with expert labels often occurred in borderline cases with minimal feature differences, suggesting either modeling limitations or true ambiguity in expert judgment.
- The model maintained robustness in the presence of missing features (e.g.,

missing image or metadata), due to proper handling and flagging.

6. **Interpretation**

   Rather than producing an absolute ranking of all questions, the model was evaluated only on specific annotated pairs, and its task was to replicate the majority decision of human judges. As such, the reported accuracy reflects human-model agreement rather than transitive consistency or global score calibration.

These results validate that the model learned meaningful signals aligned with expert perceptions of quality and was able to generalize to unseen expert comparisons with a high degree of reliability.

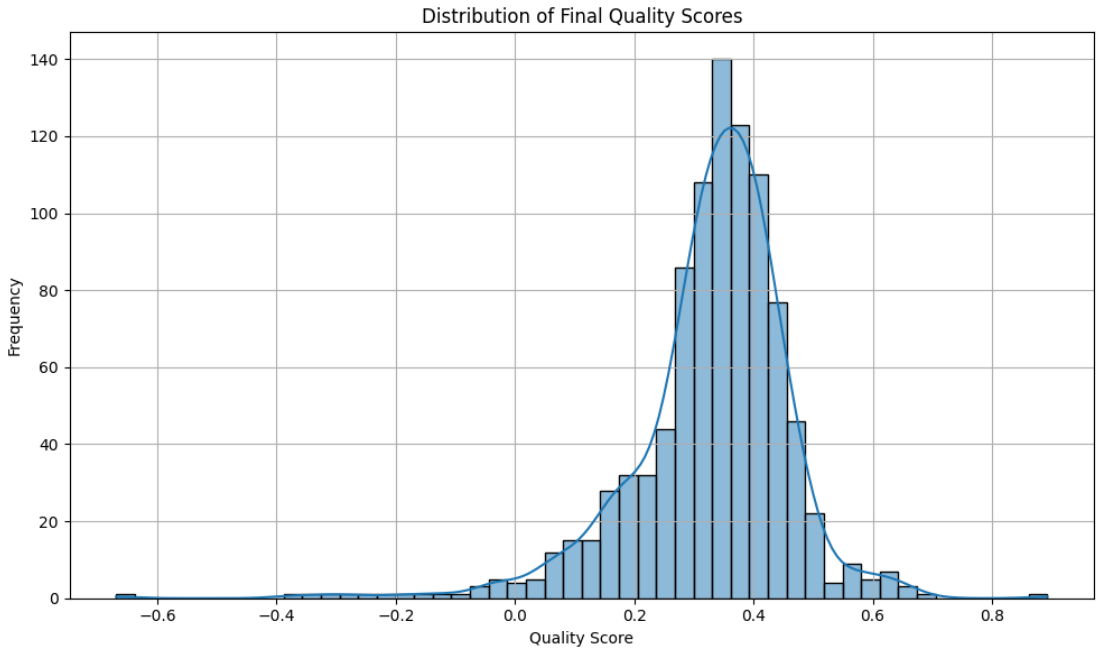

Figure 7.2: Distribution of final quality scores assigned to questions by the model. Most scores cluster between 0.2 and 0.5, indicating moderate-to-high perceived quality with a few outliers on either end.

7.3 EXPERT-WISE EVALUATION

To evaluate how well our model aligns with expert preferences, we computed the pairwise agreement between our predicted rankings and the labeled judgments provided in the NeurIPS 2020 Education Challenge test set. Table 7.3 presents the agreement scores

across individual experts for the public evaluation set.

Table 7.3: Expert-wise agreement scores on the public evaluation set.

Model / Feature Set	S_1	S_2	S_3	S_4	S_5
Final Model (All Features)	0.80	0.84	**0.92**	0.80	0.72
SelectionEntropy only	0.72	0.76	0.76	0.64	0.64
Entropy + Difficulty	0.68	**0.84**	0.68	0.56	0.64
Readability only	0.64	0.60	0.76	0.48	**0.80**
ClarityScore only	0.56	0.52	0.60	0.64	0.56

As shown, our model achieves the highest agreement with Expert S_3, suggesting strong consistency with expert-labeled quality judgments. Performance remains competitive across all experts, demonstrating generalization across subjective evaluation styles.

7.4 QUALITATIVE EXAMPLES OF PREDICTIONS

In addition to quantitative evaluation, we conducted a qualitative analysis to examine how the model behaves on individual expert-labeled question pairs. This helped us understand whether the model's decisions aligned with human intuitions and revealed cases where it either succeeded or struggled to replicate expert judgment.

We selected representative examples from the validation set, ;focusing on cases with high prediction confidence, as well as borderline or ambiguous pairs.

Example 1: Clear Visual and Behavioral Separation

- Clarity Difference: +0.42 (Question 1 significantly clearer)
- Correctness Difference: +0.21 (Question 1 had higher student accuracy)
- Entropy Difference: –0.19 (Question 1 had more consistent student responses)
- Model Prediction: Question 1 preferred (confidence: 0.88)
- Expert Label: Question 1 preferred

Interpretation: The model correctly identified the clearer and more effective question.

This case demonstrates alignment between image clarity, behavioral features, and expert preference.

Example 2: Same Subject, Slightly Varying Difficulty

- Same Subject: Yes
- Correctness Rate: Q1 = 0.68, Q2 = 0.55
- Entropy: Q1 = 0.39, Q2 = 0.60
- Model Prediction: Question 1 preferred (confidence: 0.74)
- Expert Label: Question 1 preferred

Interpretation: Despite no difference in subject or level, the model picked up on subtle behavioral signals — lower entropy and higher correctness — leading to the correct decision.

Example 3: High Clarity vs. Low Behavioral Quality

- Clarity Score: Q1 = 0.91, Q2 = 0.65
- Correctness: Q1 = 0.40, Q2 = 0.76
- Model Prediction: Question 1 preferred (confidence: 0.63)
- Expert Label: Question 2 preferred

Interpretation: In this case, the model appears biased toward visual clarity, overlooking that students performed significantly better on Question 2. The expert label favored the pedagogically stronger item, highlighting a limitation in how the model weighs conflicting signals.

Example 4: Ambiguous Case with Minimal Feature Differences

- Feature Differences: All < 0.05
- Model Prediction: Question 1 preferred (confidence: 0.52)
- Expert Label: Question 2 preferred

Interpretation: Both questions were similar in clarity, correctness, and metadata. The model's prediction was effectively a guess, and disagreement with expert judgment reflects the inherent ambiguity in such pairs.

Summary of Observations

Case Type	Model Alignment	Notes
Clear visual and behavioral contrast	Correct	Strong confidence and agreement with experts
Moderate difficulty variation	Correct	Behavioral cues used effectively
Clarity vs. correctness conflict	Incorrect	Model favored visual clarity over performance data
Near-identical feature profiles	Inconclusive	Prediction unstable; expert judgment likely subtle

Table 7.4: Summary of Observations

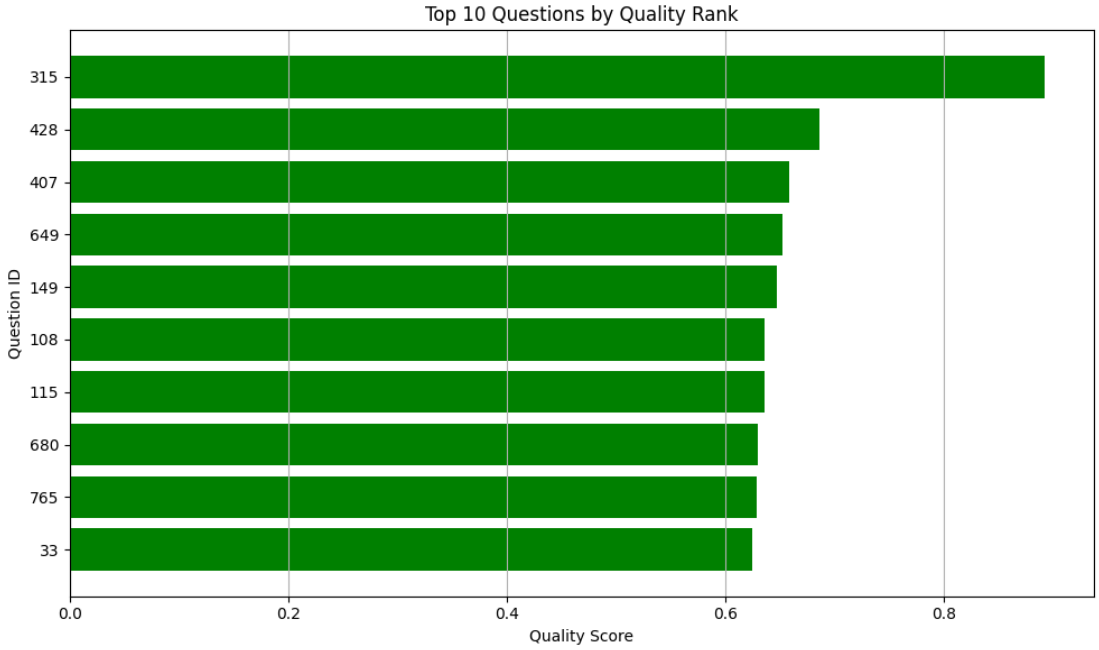

Figure 7.3: Top 10 questions with the highest predicted quality scores. These items were consistently ranked higher across expert-labeled comparisons and exhibit strong clarity, engagement, and correctness signals.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 SUMMARY OF CONTRIBUTIONS

This work addressed the problem of automated question quality comparison in an educational context, as formulated in Task 3 of the NeurIPS 2020 Education Challenge. We developed a robust, interpretable system that effectively mimics expert-level preferences using a multi-modal, feature-driven modeling approach. The key contributions of this work are summarized below:

1. **Problem Framing and Modeling Strategy**
    - Reformulated the task as a pairwise binary classification problem based on expert-labeled question comparisons.
    - Aligned the modeling approach closely with the structure of the provided annotations and evaluation metric (pairwise accuracy).
2. **Multi-Modal Feature Engineering**
    - Designed an extensive feature set incorporating:
        - Student behavioral patterns (correctness rate, entropy)
        - Image-based clarity signals (OCR quality, text region density)
        - Hierarchical and metadata context (subject depth, level, type)
    - Demonstrated that the combination of these modalities significantly enhanced prediction accuracy.
3. **Image Processing and Clarity Estimation**

- Developed a preprocessing pipeline for text segmentation and OCR on question images.

- Proposed a clarity scoring framework based on text density, OCR success, and visual alignment, which showed strong correlation with both student performance and expert preferences.

4. **High-Performance and Interpretable Modeling**

    - Trained an XGBoost classifier on engineered pairwise features.

    - Achieved a final pairwise accuracy of 81.6% on the public expert-labeled dataset and 79.0% on the private dataset.

    - Per-expert agreement reached up to 92%, demonstrating the model's ability to align with individual expert perspectives.

5. **Feature Importance and Qualitative Analysis**

    - Used gain-based importance and ablation studies to verify that clarity and behavioral features were the most impactful.

    - Conducted a qualitative error analysis to understand edge cases and model behavior under ambiguity.

6. **Rigorous Data Cleaning and Integration**

    - Addressed missing data, label noise, and subject hierarchy inconsistencies through targeted preprocessing.

    - Unified diverse data sources (images, metadata, student responses) into a coherent modeling pipeline.

Through this work, we demonstrate that a structured, interpretable model leveraging behavioral and visual cues can closely approximate human judgments of question quality

— even in the absence of raw text — and achieve performance competitive with expert agreement levels.

8.2 LIMITATIONS AND OBSERVATIONS

While the proposed system achieved strong alignment with expert judgments and demonstrated high pairwise accuracy, there are several limitations and observations that highlight opportunities for further refinement and deeper analysis.

1. **Absence of Raw Question Text**

   The dataset provided only image representations of questions, limiting our ability to perform semantic or linguistic analysis. While image clarity features were effective proxies, the lack of access to original question text constrained our ability to capture deeper pedagogical and syntactic signals that influence quality.

2. **Potential Visual Clarity Bias**

   The model showed a tendency to favor questions with high visual clarity, sometimes at the expense of those with stronger behavioral performance but poorer formatting. This suggests a bias toward visual neatness, which may not always align with educational quality.

3. **Limited Diversity of Expert Labels**

   The evaluation relied on a fixed set of expert-labeled pairs, each annotated by three judges. While majority voting was used, some labels may reflect subjectivity or disagreement, particularly in ambiguous cases. The model's "errors" in such cases may actually reflect borderline or interpretive differences rather than modeling flaws.

4. **Lack of Global Ranking Consistency**

   The model was trained to predict pairwise preferences but was not optimized for transitive or global ranking. As such, while pairwise accuracy was high, the system does not guarantee a consistent overall quality ordering across all questions.

5. **Simplified Behavioral Modeling**

   While correctness and entropy were strong signals, more nuanced behavioral patterns — such as time taken to answer, number of attempts, or engagement trends over time — were unavailable. This limited our ability to model cognitive difficulty and student effort, which are crucial dimensions of question quality.

6. **Dependence on Feature Engineering**

   The current model relied entirely on manually engineered features, which, while interpretable, may miss deeper patterns present in the image or response distributions. A learned representation (e.g., via deep learning) could capture such patterns more effectively in future work.

7. **Generalizability Across Subjects and Formats**

   Although subject and level metadata were used, the model was not explicitly evaluated for subject-specific biases. Some subjects may have inherently different formats or clarity standards (e.g., math vs. reading comprehension), which were not separately calibrated or normalized.

Despite these limitations, the model exhibited strong overall performance and robustness, demonstrating that meaningful signals of question quality can be extracted and modeled even under constrained settings. These insights provide a valuable foundation for future enhancements in interpretability, semantic analysis, and domain adaptation.

8.3 FUTURE WORK

Building upon the findings and limitations of this work, several promising directions can be pursued to enhance both the accuracy and the generalizability of automated question quality assessment systems. These directions focus on expanding the model's semantic understanding, improving ranking consistency, and broadening its applicability across educational contexts.

1. **Incorporation of Raw Textual Data**

   If raw question text becomes available, future models can integrate natural language

processing (NLP) techniques to capture semantic clarity, grammatical structure, and conceptual depth. This would allow for a deeper analysis of content quality, complementing the visual and behavioral features used in this work.

2. **Transition from Pairwise to Global Ranking**

   While this work focused on pairwise classification, future work could develop models that produce global rankings of question quality. Techniques such as learning to rank, pairwise-to-listwise transformations, or graph-based ranking algorithms (e.g., Bradley-Terry or PageRank-style methods) can be explored to achieve consistency and transitivity across larger sets of questions.

3. **Deep Learning-Based Feature Representations**

   Instead of relying entirely on engineered features, future models could utilize learned representations:

   - CNN or Vision Transformer embeddings for image-based understanding
   - Text embeddings from transformer models (e.g., BERT) for semantic features
   - Multimodal fusion models combining visual, textual, and behavioral inputs

   These techniques may uncover deeper patterns that are not easily captured by handcrafted features.

4. **Subject-Specific Calibration and Fairness**

   To ensure equitable evaluation across diverse subjects and formats, future work should explore subject-aware modeling and fairness-aware techniques. This may involve:

   - Training subject-specific sub-models
   - Calibrating clarity and correctness expectations by subject
   - Monitoring model behavior across different content domains

5. **Enhanced Temporal and Behavioral Modeling**

   With access to detailed timestamps or sequential response logs, richer temporal features can be extracted, such as:

   - Time to answer
   - First-attempt correctness
   - Engagement patterns over time

   These can help capture real-world cognitive complexity and student effort, leading to a more nuanced understanding of question quality.

6. **Human-in-the-Loop Evaluation**

   Future systems could incorporate feedback loops with human experts, allowing for continuous model refinement based on expert disagreement or uncertainty. Semi-supervised learning or active learning methods could also be employed to expand the labeled pair set efficiently.

7. **Real-World Integration and Usability**

   The final goal of such systems is integration into educational platforms. Future research could focus on:

   - Real-time quality feedback for content creators
   - Interactive ranking interfaces for educators
   - Evaluating the pedagogical impact of model-informed content curation

Exploring these directions can significantly advance the field of automated educational content evaluation, making such systems more scalable, interpretable, and pedagogically aligned.

BIBLIOGRAPHY

[1] Tianqi Chen and Carlos Guestrin. "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[2] Amirata Ghorbani and James Y Zou. "Generalized SHAP: Generating multiple types of explanations in machine learning". In: *arXiv preprint arXiv:2106.03404* (2021). URL: https://arxiv.org/abs/2106.03404.

[3] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 2017, pp. 2893–2899. URL: https://www.ijcai.org/Proceedings/2017/0352.pdf.

[4] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[5] Huan Liu and Lei Yu. "Toward Integrating Feature Selection Algorithms for Classification and Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 17.4 (2005), pp. 491–502.

[6] Stephen V. Rice. "Optical character recognition (OCR) technology". In: *IEEE Transactions on Professional Communication* 48.4 (2005), pp. 344–357. DOI: 10.1109/TPC.2005.846731. URL: https://ieeexplore.ieee.org/document/4376991.

[7] Zichao Wang et al. "Diagnostic Questions: The NeurIPS 2020 Education Challenge". In: *NeurIPS 2020 Competition and Demonstration Track*. Vol. 133. PMLR, 2021, pp. 191–205. URL: https://proceedings.mlr.press/v133/wang21a.html.