EduSathi

AI-Powered Handwritten Answer Sheet Grading

https://edusathi.qzz.io

Initial Benchmarking Report

AI Evaluation vs. Human Examiner

AS1020 Fluid Mechanics — Undergraduate Quiz

Prepared by: Siddhant Gujar
Role: Founder, EduSathi

Date: April 21, 2026
Dataset: 9 Handwritten Answer Sheets
Course: AS1020 Fluid Mechanics Quiz (April 10, 2026)
Max Marks: 30

Key Results at a Glance

0.92	1.44	4.8%	−0.11
Pearson r	MAE (pts)	Error Rate	Bias (pts)
Correlation	Avg. Error	of 30 marks	Near-Zero

This is an initial benchmark on a small sample. Results are indicative and will be validated with larger datasets.

Abstract

This report presents the initial benchmarking of **EduSathi**, an AI-powered evaluation platform for handwritten academic answer sheets. The objective was to measure how closely EduSathi's Vision-Language Model (VLM) pipeline matches human examiner grades on real undergraduate answer sheets.

Nine handwritten answer sheets from a Fluid Mechanics undergraduate quiz (Course Code: AS1020, conducted April 10, 2026) were processed. Results show a **Pearson Correlation Coefficient of** $r = 0.92$, a **Mean Absolute Error of 1.44 marks** on a 30-mark scale (4.8%), and a **near-zero Mean Bias Error of** -0.11, indicating no systemic tendency toward leniency or harshness. These results demonstrate strong viability for AI-assisted grading in academic environments.

Contents

1 Introduction **2**

2 Methodology **2**
2.1 Dataset Collection 2
2.2 Data Pre-processing and Anonymisation 2
2.3 AI Configuration and Evaluation Pipeline 2

3 Results **3**
3.1 Raw Data 3
3.2 Visualisation 3

4 Statistical Analysis **4**
4.1 Mean Absolute Error (MAE): 1.44 points 4
4.2 Mean Bias Error (MBE): −0.11 points 4
4.3 Root Mean Square Error (RMSE): 1.58 points 4
4.4 Pearson Correlation Coefficient (r): 0.92 5
4.5 Summary of Statistical Metrics 5

5 Limitations and Scope **5**

6 Conclusion and Future Work **6**

Introduction

Manual evaluation of handwritten answer sheets is one of the most time-consuming tasks in academic assessment. A single teacher may spend 4–6 hours grading a batch of 40 papers, with inconsistency increasing as fatigue sets in. For large institutions or competitive exam coaching centres handling hundreds of papers per cycle, this becomes operationally unsustainable.

EduSathi addresses this problem by using Vision-Language Models (VLMs) to evaluate handwritten answer sheets directly from scanned images or photographs. Unlike traditional Optical Character Recognition (OCR) approaches which first convert handwriting to text (introducing errors), EduSathi's vision-based pipeline reads the answer sheet holistically — including diagrams, equations, and messy handwriting — and evaluates it against a teacher-calibrated, step-wise marking rubric.

This report documents the first real-world benchmark of EduSathi's grading pipeline against human examiner scores, with the goal of quantifying accuracy, consistency, and bias.

Methodology

Dataset Collection

The dataset consisted of **9 handwritten answer sheets** from a Fluid Mechanics Quiz (Course Code: **AS1020**) conducted on **April 10, 2026**. The maximum achievable score was **30 marks**.

Answer sheets were collected on a **voluntary basis** from undergraduate students via a WhatsApp message explaining the purpose of the data collection. No student was pressured or compensated.

Data Pre-processing and Anonymisation

To ensure unbiased evaluation and strict data privacy compliance, the following preprocessing steps were undertaken:

1. **Anonymisation:** All Personally Identifiable Information (PII), including student names and roll numbers, was fully removed from all scanned copies before processing.

2. **Masking of Human Marks:** All physical annotations by the human examiner — ticks, crosses, and numerical scores — were digitally masked from the PDF files. This ensured the AI could not "read" the existing grades and was evaluated independently.

AI Configuration and Evaluation Pipeline

The evaluation used **Google Gemini Vision-Language Models** on default hyperparameters with no custom fine-tuning. The pipeline operated as follows:

1. **Rubric Generation:** The raw Question Paper was input into the AI, which autonomously generated a preliminary step-wise model answer and marking scheme.

2. **Human-in-the-Loop Calibration:** The generated rubric was manually reviewed and edited by the human operator (the founder) to ensure academic accuracy and alignment with course standards.

3. **Evaluation:** The anonymised, masked answer sheets were processed through the EduSathi engine against the finalised rubric. For this specific benchmark, no external textbook context was supplied; the model relied on the question paper and calibrated rubric alone.

4. **Score Extraction:** EduSathi's output was a structured PDF report for each paper containing question-wise marks and an overall total, which was compared against the original human examiner score.

Results

Raw Data

Table 1 presents the complete set of human examiner scores alongside EduSathi AI scores for all 9 papers.

Table 1: Raw Evaluation Data: Human Examiner vs. EduSathi AI (Max Marks: 30)

Paper ID	Human Score (/30)	AI Score (/30)	Deviation (AI − Human)
P1	27	25.0	−2.0
P2	20	19.5	−0.5
P3	19	17.5	−1.5
P4	17	19.5	+2.5
P5	14	14.5	+0.5
P6	26	27.0	+1.0
P7	17	15.5	−1.5
P8	22	24.0	+2.0
P9	22	20.5	−1.5
Average	**20.4**	**20.3**	−0.11

Visualisation

Figure 1 presents a scatter plot of human vs. AI scores alongside a deviation chart. The scatter plot illustrates the tight clustering of data points around the perfect agreement line (red dashed), confirming high correlation. The deviation chart shows that fluctuations are balanced around zero with no systemic direction.

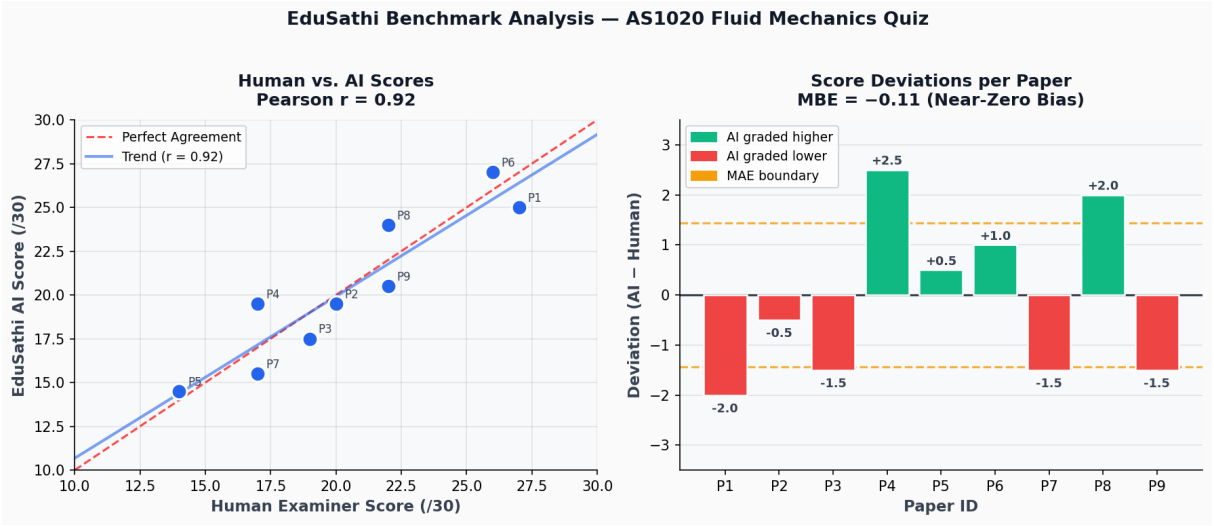


Figure 1: Left: Human vs. AI Score scatter plot with trend line ($r = 0.92$). Right: Per-paper score deviations showing near-zero bias.

Statistical Analysis

Mean Absolute Error (MAE): 1.44 points

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = 1.44$$

MAE measures the average magnitude of prediction errors. An MAE of **1.44 on a 30-mark scale represents an average error margin of only 4.8%**. This is an operationally excellent margin for a subjective, handwritten evaluation task.

Mean Bias Error (MBE): −0.11 points

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = -0.11$$

MBE captures the directional tendency of the AI — whether it systematically over-grades or under-grades. An MBE of −0.11 **indicates near-perfect neutrality**. The AI does not exhibit systemic leniency or harshness relative to the human benchmark.

Root Mean Square Error (RMSE): 1.58 points

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = 1.58$$

RMSE penalises larger errors more heavily. The fact that **RMSE (1.58) is very close to MAE (1.44)** confirms high consistency in the AI's grading. There are no wild, erratic outliers that would undermine trust in the system.

Pearson Correlation Coefficient (r): 0.92

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} = 0.92$$

The Pearson coefficient measures the linear relationship between human and AI scores. A value of $r = 0.92$ **is exceptionally strong**, indicating that EduSathi correctly identifies which papers are excellent, average, and poor — maintaining the same grading "curve" as a human examiner.

Summary of Statistical Metrics

Table 2: Summary of Benchmark Statistical Metrics

Metric	Value	Scale Reference	Interpretation
MAE	1.44 pts	out of 30	Excellent (<5% error)
MBE	−0.11 pts	out of 30	Near-zero bias
RMSE	1.58 pts	out of 30	Highly consistent
Pearson r	0.92	−1 to +1	Very strong correlation

Limitations and Scope

This benchmark has the following known limitations which will be addressed in subsequent studies:

1. **Small Sample Size ($n = 9$):** The current dataset is statistically small. While the results are promising, they cannot be considered statistically robust until validated on $n \geq 100$ papers.
2. **Single Subject and Single Exam Type:** All papers are from one Fluid Mechanics quiz. Performance may vary across subjects (especially humanities, languages, or diagram-heavy papers) and exam formats.
3. **Volunteer Bias:** Answer sheets were collected voluntarily, which may not represent the full distribution of handwriting quality and academic performance in the class.
4. **Single Grader Reference:** Human scores were from one teaching assistant. Inter-rater variability between multiple human graders was not measured.
5. **No Textbook Context:** This benchmark did not use EduSathi's Textbook Lock-in feature. Results with full textbook context are expected to improve further.

Conclusion and Future Work

The initial benchmark demonstrates that EduSathi's vision-based evaluation pipeline is highly capable of replicating human grading standards on handwritten undergraduate answer sheets.

Key findings:

- 92% correlation with human examiner scores
- Less than 5% average error margin
- Near-zero systemic bias
- Consistent performance with no erratic outliers

These results indicate strong viability for deployment in academic environments. A teacher using EduSathi can expect AI-generated scores that align closely with their own grading judgment, while retaining full authority to review and modify any result.

Future Work

1. Scale benchmark to $n \geq 100$ papers across multiple subjects
2. Test with Textbook Lock-in feature enabled
3. Benchmark across diverse handwriting styles and regional language scripts (Hindi, Tamil, Marathi, etc.)
4. Conduct inter-rater reliability study comparing multiple human graders
5. Publish findings as a formal research paper
