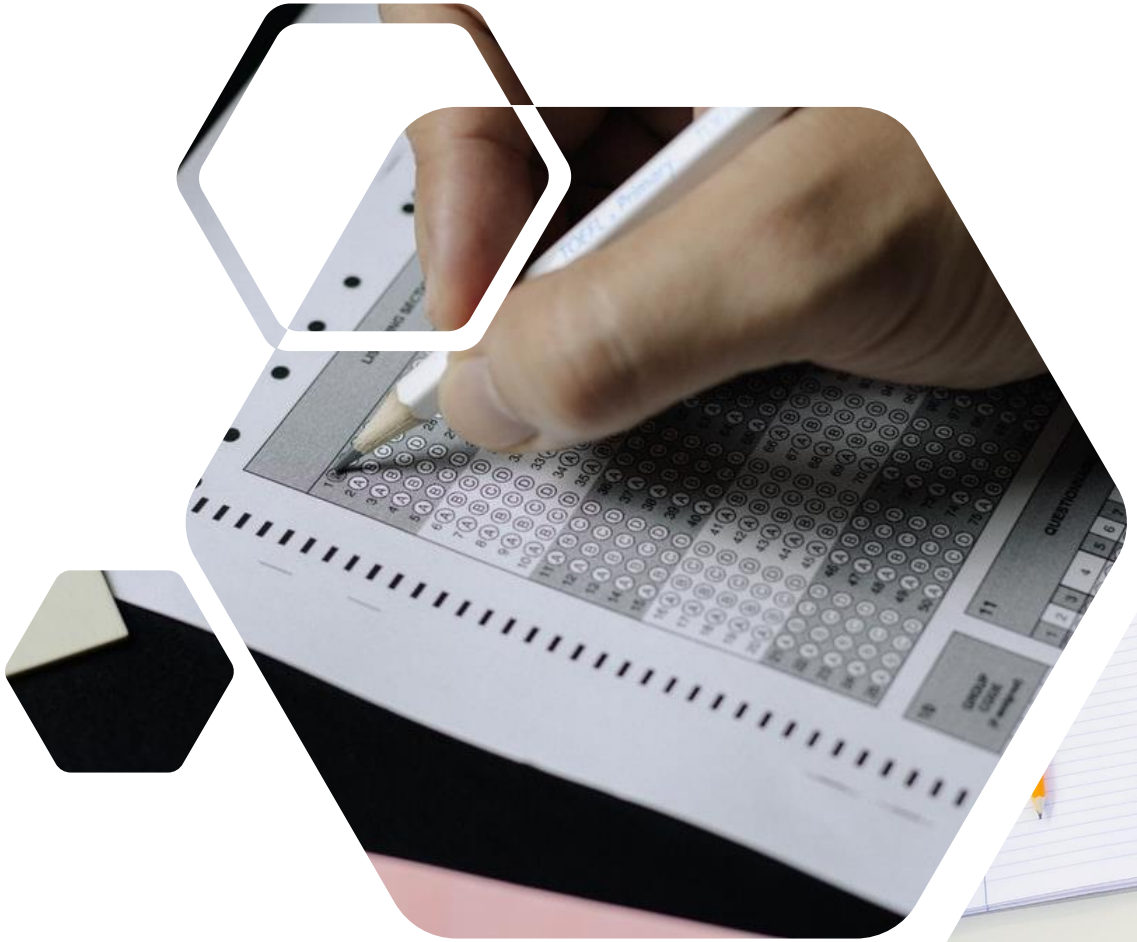
Overcoming Human Error with **AI-Powered** Assessments

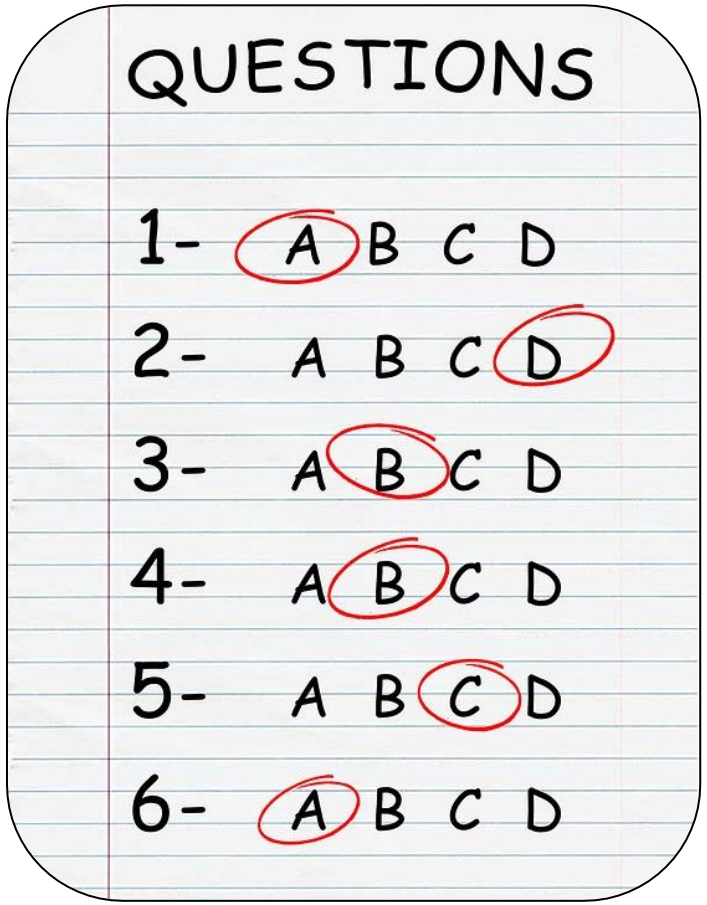
Okan Bulut

Measurement, Evaluation, and Data Science
University of Alberta

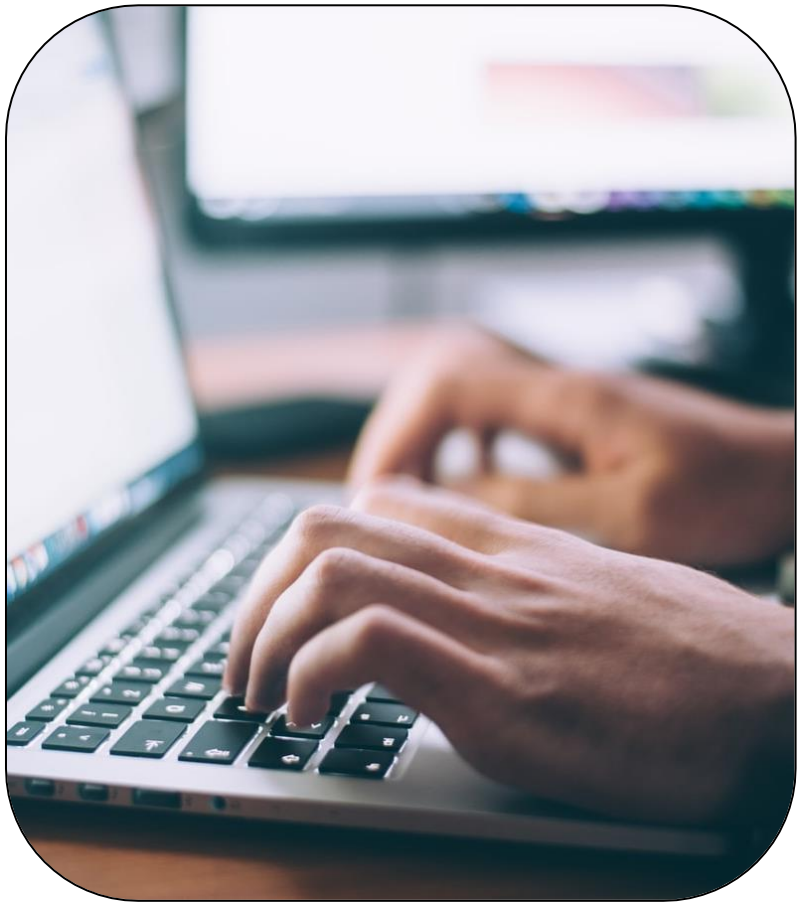
bulut@ualberta.ca www.okanbulut.com @drokanbulut



- Technology-enhanced items
- Test administration features
- Flexible test scheduling
- Adaptive testing
- “Objective” and “quick” scoring

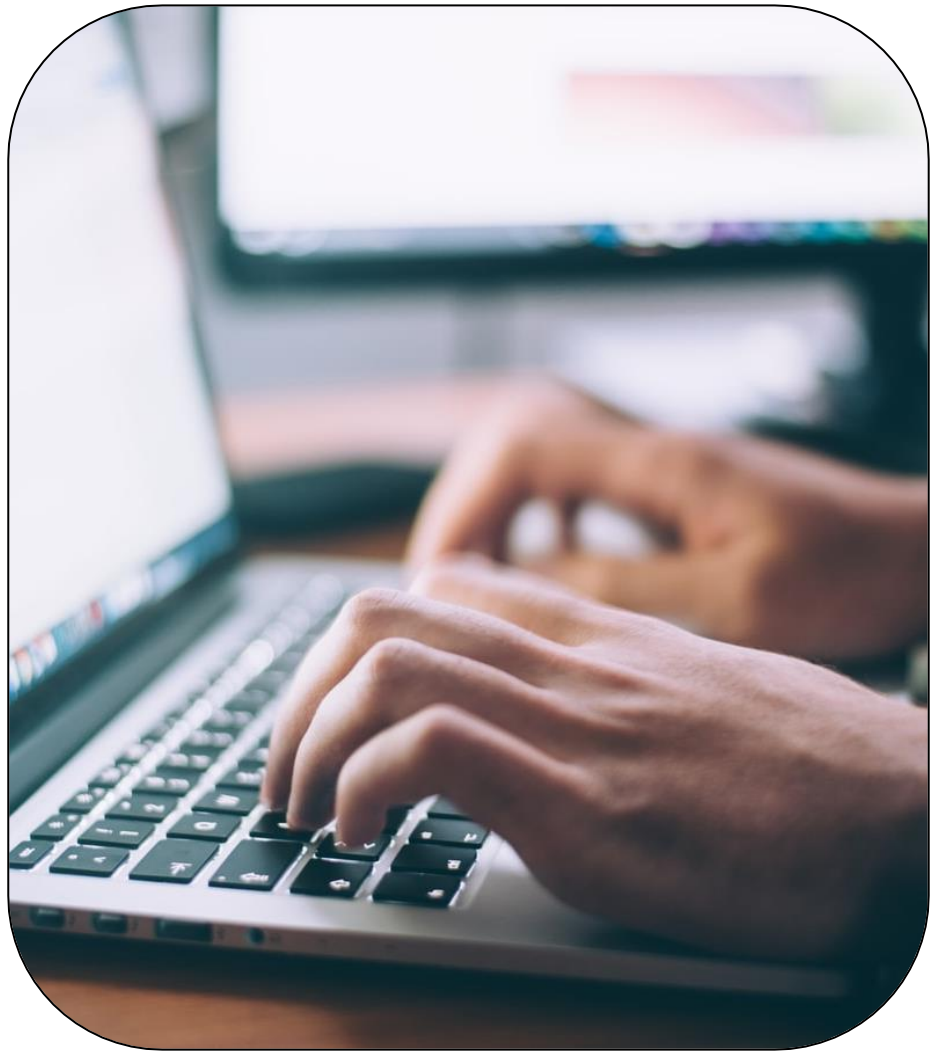


VS



Selected-Response
(Multiple-Choice, T/F)

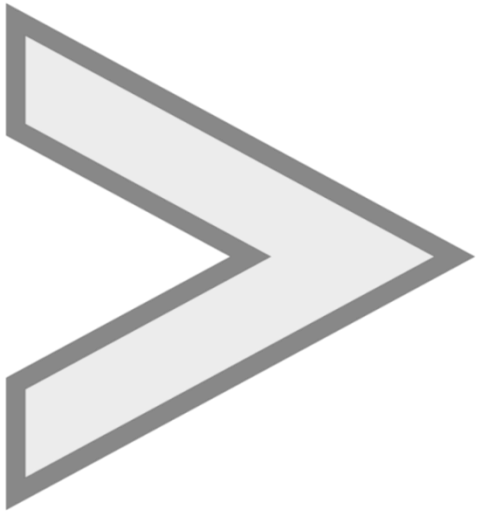
Constructed-Response
(Open-Answer)



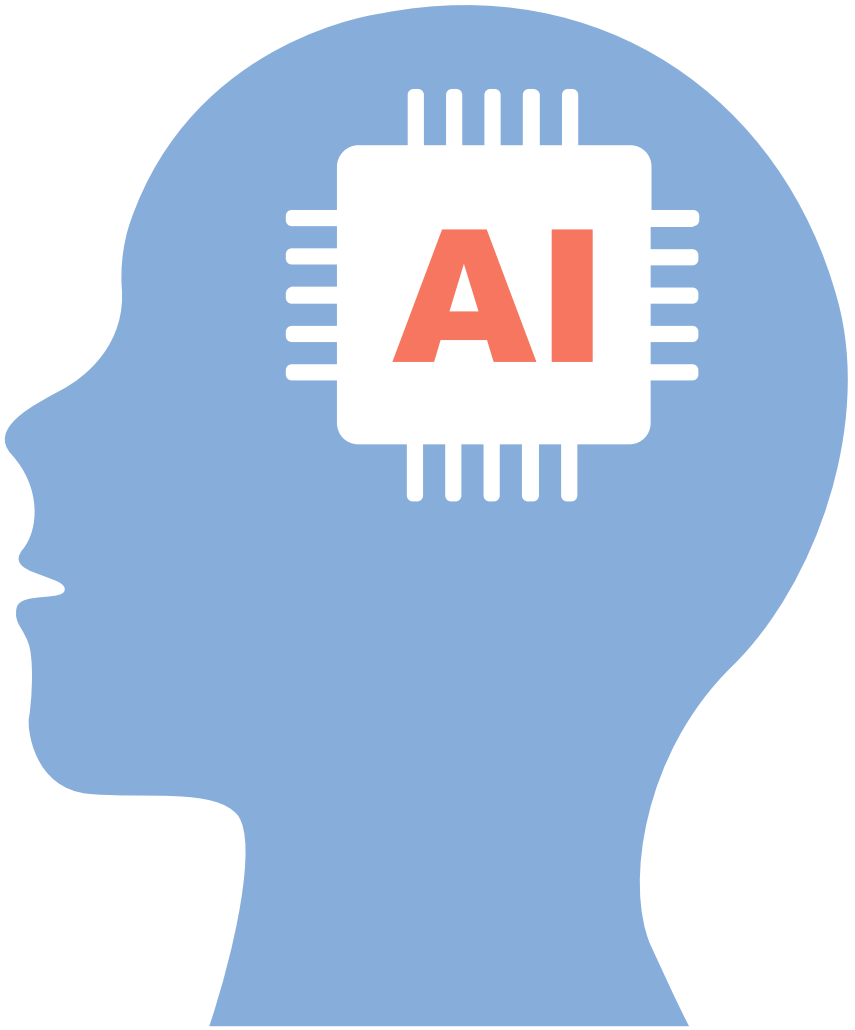
- Easier to develop
- Capability to assess complex skills
- Ease of partial credit scoring
- Less prone to cheating

- Inefficient (scoring + testing time)
- Subjective human scoring → bias

Rater fatigue
Rater's personal beliefs
Off-topic elements

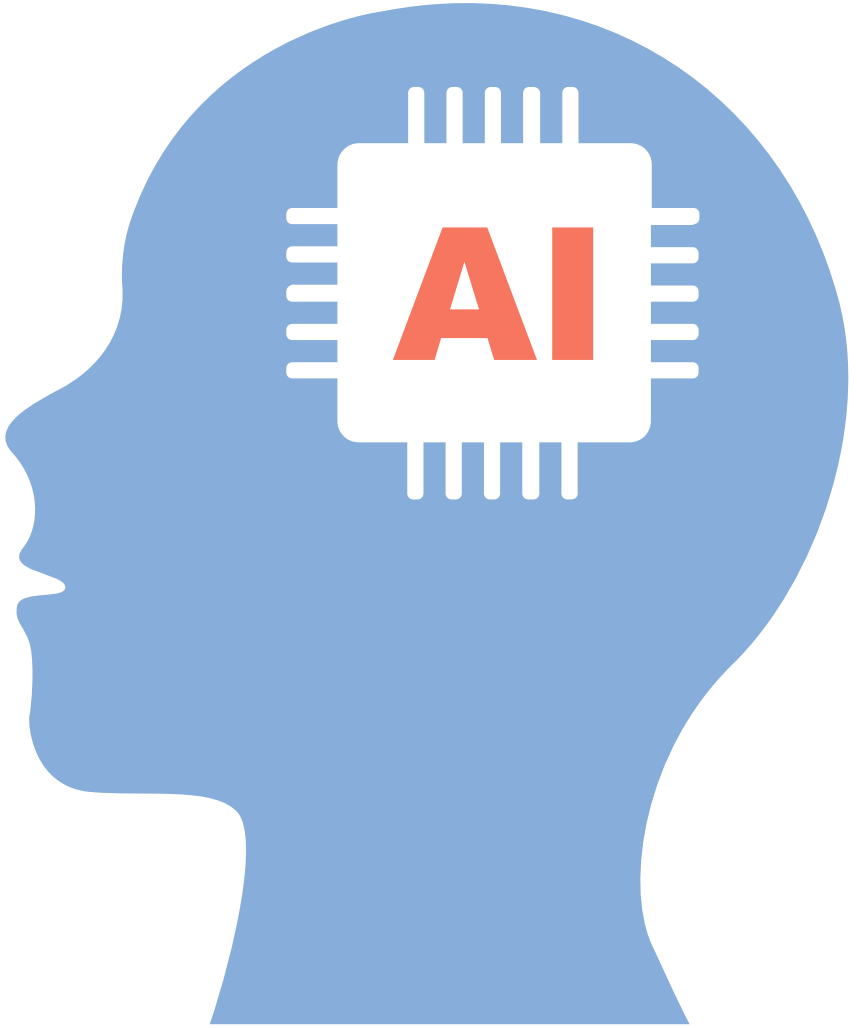


Response accuracy
Completeness of response

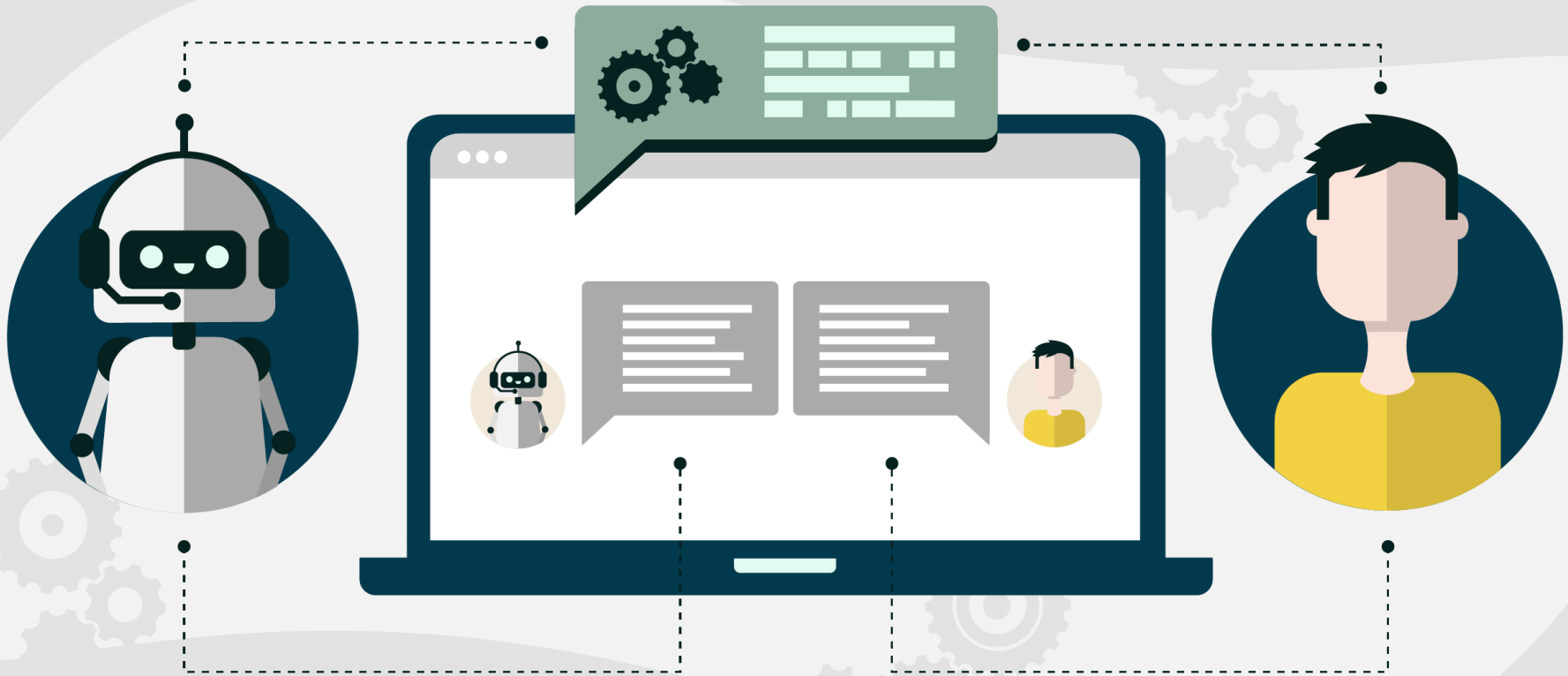

AI

ARTIFICIAL INTELLIGENCE

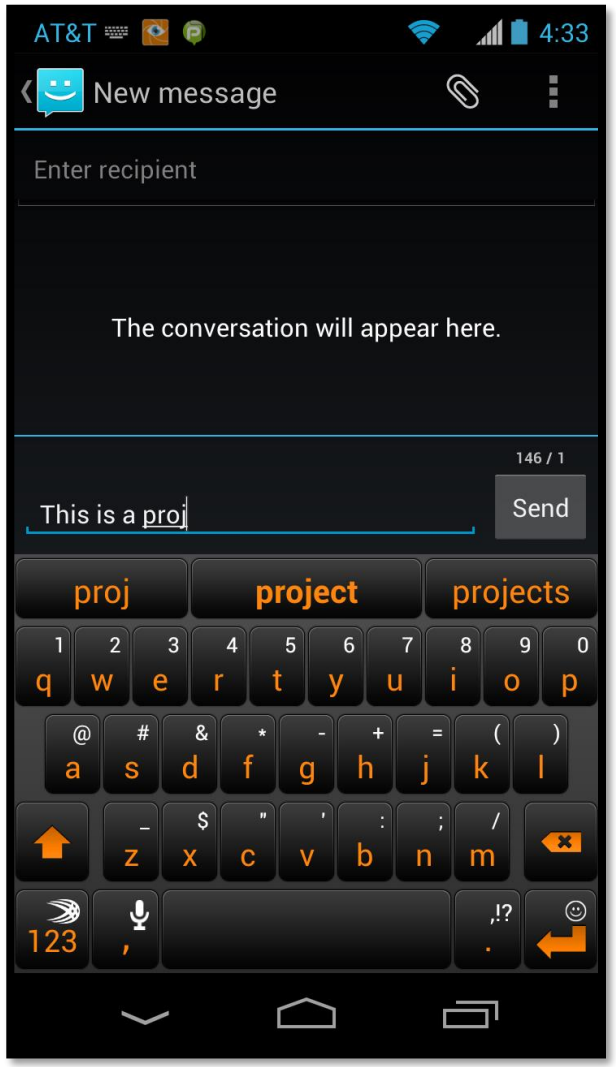
“Having a computer do something that we assume(d) only entities having human-like intelligence could do.”



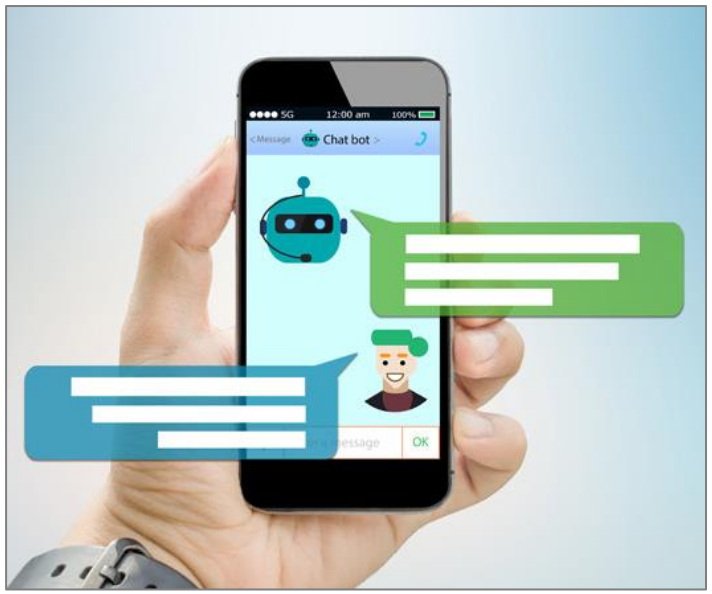
NATURAL LANGUAGE PROCESSING



Source: https://aliz.ai/natural-language-processing-a-short-introduction-to-get-you-started/



amazon alexa



Source: https://pub.towardsai.net/ai-can-bring-fairness-to-assessments-but-are-we-ready-for-it-4bd2b039ee8d

1

Storing students' written responses digitally

2

Human scoring

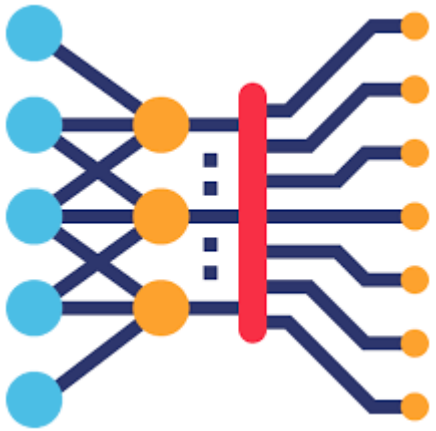
3

Processing responses & scores

6

Building a scoring model

5



Applying **deep learning** models

4

Extracting linguistic features

Short-Answer Scoring

- 10 short-answer items
- 2,230 students (Grade 10)
- 2 human raters to score each item
- Scores: 0, 1, 2, or 3 points

Some information about a science experiment

Prompt—Acid Rain

A group of students wrote the following procedure for their investigation.

Procedure:

1. Determine the mass of four different samples.
2. Pour vinegar in each of four separates, but identical, containers.
3. Place a sample of one material into one container and label. Repeat with remaining samples, placing a single sample into a single container.
4. After 24 hours, remove the samples from the containers and rinse each sample with distilled water.
5. Allow the samples to sit and dry for 30 minutes.
6. Determine the mass of each sample.

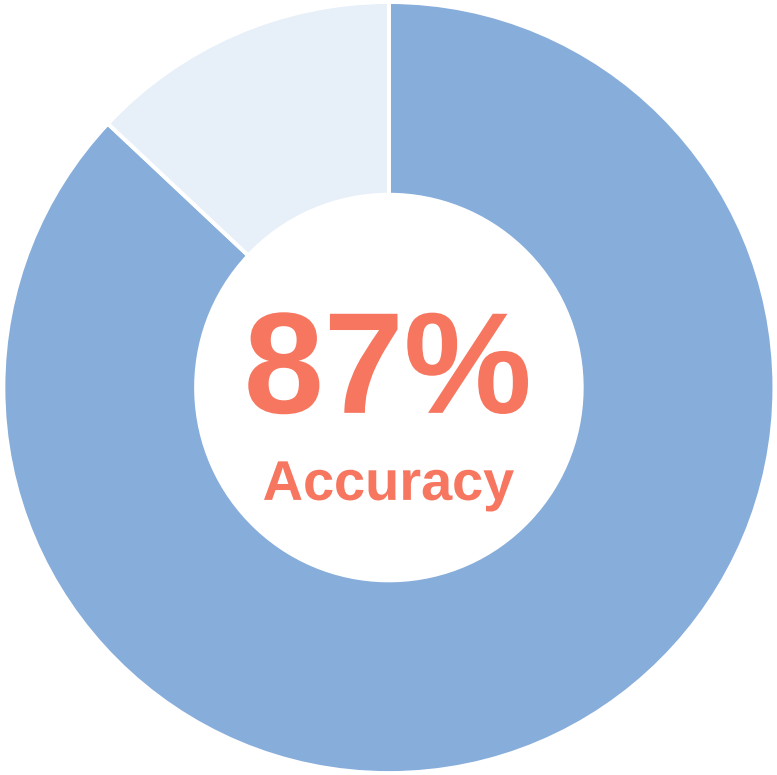
The students' data are recorded in the table below.

Sample	Starting Mass (g)	Ending Mass (g)	Difference in Mass (g)
Marble	9.8	9.4	–0.4
Limestone	10.4	9.1	–1.3
Wood	11.2	11.2	0.0
Plastic	7.2	7.1	–0.1

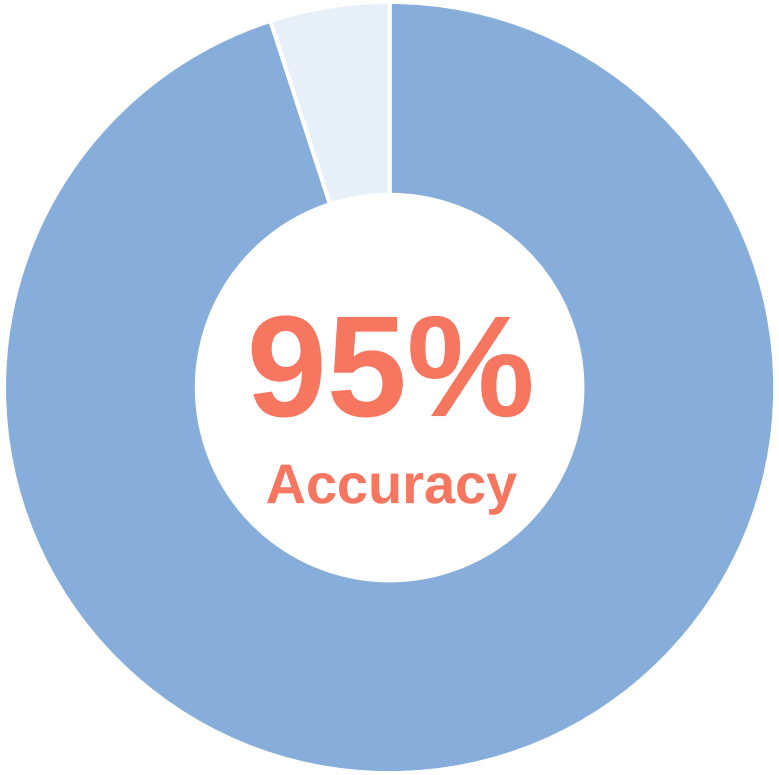
Question

After reading the group's procedure, describe what additional information you would need in order to replicate the experiment. Make sure to include at least three pieces of information.

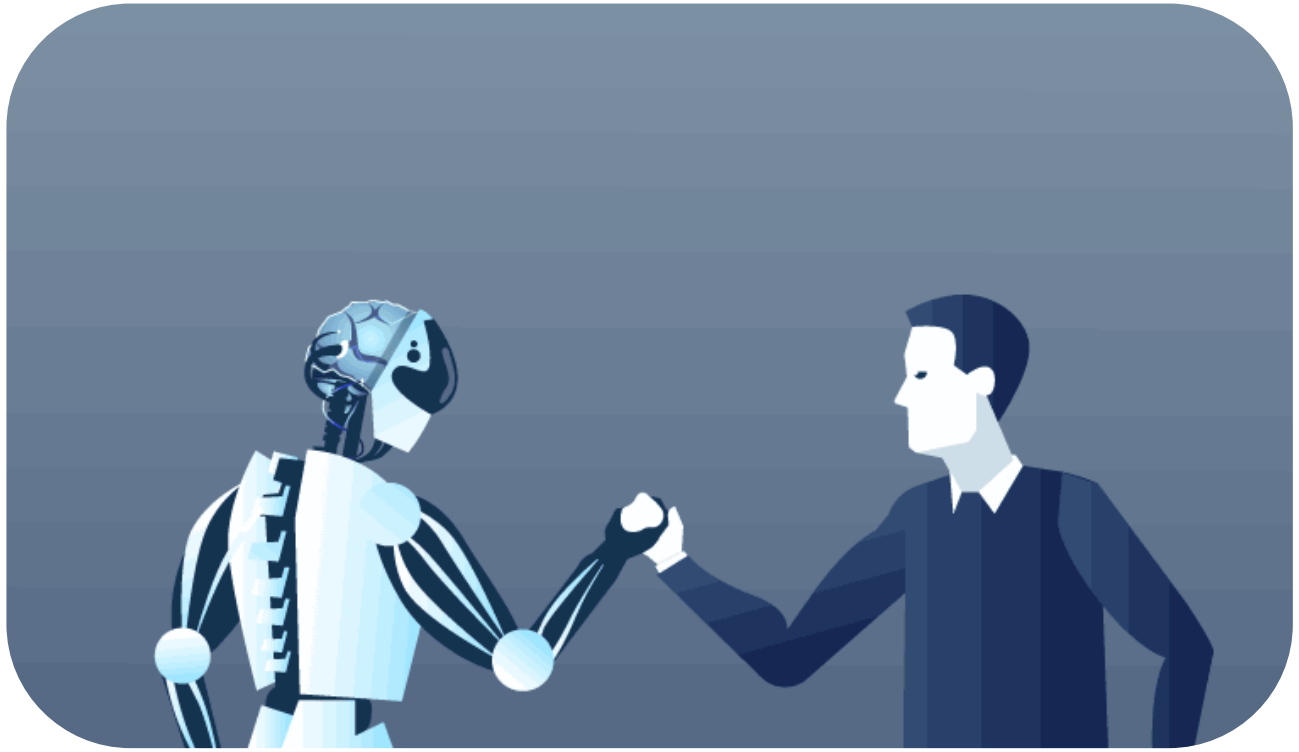
A scoring algorithm based on the Long Short-Term Memory (LSTM) model yielded:



1 item

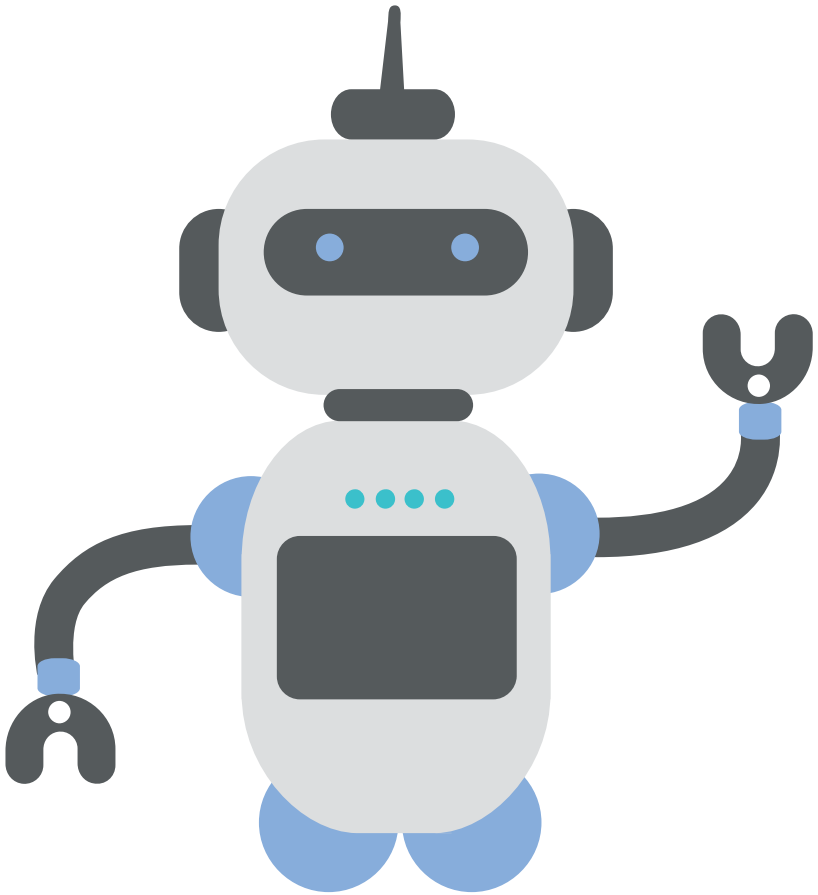


10 items



~~REPLACING~~ HUMAN SCORING

IMPROVING



Thank You

bulut@ualberta.ca