**[Week 4] Team Project: Evaluation Team**

**Setup the Hypothesis for a goodmodel + define the Evaluation Dataset Structure.**

**a) AI detection** (KHUSHI)

**B) Feedback AI (Van Hieu Nguyen)**

**- Hypothesis:** evaluate whether fine-tuning modell on learnersourced student-written programming feedback produces AI-generated feedback that is more accurate and closer in style to human student feedback compared to baseline models using basic or engineered prompts.

*(1) Feedback-Writing task:* where students take on the role of

tutors, providing feedback on buggy code. Students receive a problem description, a buggy

implementation, and an instruction asking them to provide feedback from a tutor’s perspective.

*(2) Student-Written feedback task:* the student is given a problem and a buggy program and asked to act in the flipped role of a tutor to write feedback for that buggy program.

*(3) AI-Generated Feedback Data:* which involves leveraging a problem description and a buggy code, along with additional symbolic information (failing test case and a fixed version of the code) to generate feedback.

1. **Identify the key evaluation metrics:** use two complementary categories of metrics, as recommended in the research and aligned with the programming feedback generation task
2. *Automatic text similarity metrics:* measure lexical and semantic similarity between AI-generated feedback and student-written feedback.

* BLUE: captures extract wording similarity
* ROUGE-L: captures recall and coverage of reference content
* METEOR: measures unigram matches using stemming, synonyms, and paraphrase matching.

1. *Rubric-based alignment metrics:* from the research paper for programming feedback generation scene:

* Correctness (binary): does the feedback help fix the bug?
* Gives fix (binary): Does the feedback explicitly provide a solution path?
* Mentions variables (binary): Does the feedback refer to specific variables?
* Mentions lines (binary): Does the feedback refer to specific code lines?
* Word count & sentence count: Measures conciseness and alignment with student style.

1. Human Evaluation Metrics: independent expert raters (larger than 2) will score each feedback instance on a 5-point Likert scale for:

* Usefulness: actionable for debugging.
* Clarity: easy to understand.
* Tone appropriateness: supportive and non-punitive.
* Specificity: concrete and precise advice.
* Overall preference: if comparing two model outputs (A/B testing).

      Reliability target: Cohen's kappa of 0.6.

1. **Define Success Criteria:** meet or exceed the success threshold for both automatic text similarity and rubric based alignment
2. *Automatic Text Similarity metrics*

* BLUE: greater or equal 0.25. For short feedback (between 40 and 70 words), BLEU above 0.25 indicates reasonable n-gram overlap while allowing for paraphrasing.
* ROUGE-L: greater or equal 0.35. It captures recall of important phrases; threshold set to reflect retention of key concepts in reference feedback.
* METTOR: greater or equal 0.30. Accounts for synonym and stem matches.

1. *Rubric based on research paper:*

* Correctness greather than 85%. Fine-tuned models in study reached between 86 and 88%, higher than baseline engineered prompts.
* Num of words around 40–70. Matches student mean (46 words) while allowing in range of 20% flexibility.
* Num of sentences: 2 or 3 sentences. Matches student average (2.7 sentences).
* Gives fix: between 45 and 70%. Students gave fixes around 46% of the time; fine-tuned models exceeded this (71–98%) which cap upper bound to keep style balanced.
* Mentions variables: greater than 36% percent. Matches student rate (36.3%).
* Mentions lines : less than 12 percent. Matches student rate (11.3%), which prevents overly rigid line number feedback.
* Therefore, a model is successful if:
* Automatic metrics: Meets all three thresholds for BLEU, ROUGE-L, and METEOR, and shows significant improvement over baseline prompts.
* Rubric metrics: Meets 5 out of 6 rubric thresholds above, with Correctness mandatory.

1. **Define the evaluation Dataset Structure:** we will work with data team to make sure we follow the correct format for evaluation. Assume that
2. Dataset format and content hypothesis:

* student\_id: string (unique identifier).
* Input\_prompt: text (The problem description, buggy code, and other content).
* generated\_feedback: text (Feedback produced by the evaluated model).
* ground\_truth\_feedback: text (Reference feedback, such as student-written or expert-written used for comparison).
* source\_type: categorical, such as "student", "AI\_model", or "baseline".

1. Metrics used hypothesis (based on research paper):

* Automatic Text Similarity Metrics:
* BLEU: n-gram precision and ground truth.
* ROUGE-L: longest common subsequence recall.
* METEOR: semantic match with synonym/stemming support.
* Rubric-Based Alignment Metric:
* Correctness (binary, expert-annotated).
* Num of words (automatic count).
* Num. sentences (automatic count).
* Gives fix (binary, annotated).
* Mentions variables (binary, annotated).
* )Mentions lines (binary, annotated).

**Reference:**

1. Humanizing Automated Programming Feedback: Fine-Tuning Generative Models with Student-written feedback.

<https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.short-papers.35/2025.EDM.short-papers.35.pdf>