AI-Powered Rubric Feedback: Lessons from Purdue's 'Charlie' Writing Assistant

Jason Dufair (jase@purdue.edu)
Purdue University

Introduction

Purdue's "Charlie" is a rubric-based pre-flight writing assistant included in the "Circuit" peer review application that was developed by Purdue. Charlie has evolved from a quantitative, predictive scoring model to a qualitative, actionable,` LLM-based assistant.

Providing timely, meaningful feedback on student writing is one of the most challenging tasks in teaching at scale. This poster discusses practical lessons from Purdue building and evaluating Charlie, Purdue's AI-powered rubric feedback tool. This includes design strategies, pitfalls to avoid, and frameworks for integrating AI writing support into assignments – with a focus on preserving instructional quality and supporting student learning.

Objectives

- Identify key considerations and challenges in using AI tools for rubric-based writing feedback
  - **Alignment and validity**: LLM-generated feedback must accurately use rubric criteria and apply them consistently—models may misjudge nuance rather than intended learning outcomes
  - **Transparency and trust**: Explaining how the model came up with the feedback remains challenging; opaque reasoning, not to mention hallucinations, can reduce instructor and student confidence in fairness and objectivity
  - **Bias and scalability trade-offs**: While LLMs enable scalable feedback, they risk amplifying biases embedded in training
- Evaluate strategies for integrating AI writing support tools into assignments while maintaining instructional integrity
  - **Avoid high stakes**: Charlie is and has always been a "pre-flight" assistant. Language models are not sophisticated enough for high-stakes (or even low-stakes) evaluation.
  - **Give the learner control**: Charlie encourages submitting early and often. This has led to, according to some instructors, some students using better critical thinking skills based on current feedback
- Apply a framework for selecting, testing, and refining AI feedback tools in your own teaching or design context. Purdue has evaluated Charlie using:
  - Survey instruments
  - Learning outcome vs. a control section of the course

Quantitative Charlie

Charlie 1.0 – 2019 (Quantitative)

- Initially trained on a corpus of over 600 essays, hand graded by instructors and teaching assistants
- Used an in-house developed neural network to offer a predicted score on the entire essay
- Implemented as a pre-flight assistant where learners could revise and resubmit, using the predicted score to determine magnitude and directional quality changes.
- For Purdue's Sociology 220 course, the model training converged successfully. Charlie's **QWK** (inter-rater agreement) vs. either grader met or exceeded the **QWK** between the two graders.
- Learners typically submitted several drafts during the submission window, one 32 times, to increase their scores on the final drafts of their essays over the two-week period it was available to them.

Successes

- Statistically significant improvement in learning outcomes over previous semesters where Charlie was not available

Challenges

- While demonstrably accurate, Charlie's feedback did not provide anything actionable for the student.

Charlie 1.5 – 2021 (Quantitative, Rubric-based)

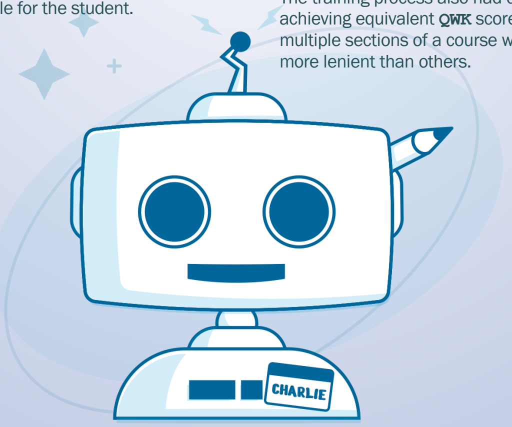
- Charlie was trained on an additional 500 hand graded essays
- The essays were graded on a rubric, so Charlie was trained on a model-per-rubric strategy.
- Learners were given a predicted score per rubric criterion
- Subsequent submissions would indicate directional and magnitude quality changes based on specific rubric criterion

Successes

- This version led to learners submitting earlier and more often, frequently seeing improved score prediction upon resubmission

Challenges

- While feedback was more specific, related to rubric criteria provided by the instructor, it still was not actionable.
- The training process was unable to converge on several rubric criterion
- The training process also had difficulty converging or achieving equivalent **QWK** scores when there were multiple sections of a course where some instructors were more lenient than others.



Qualitative Charlie

Charlie 2.0 – 2023 (Quantitative)

- To provide actionable feedback, the neural network-based models were retired in place of Large Language Models
- A custom prompt was developed that includes the rubric, its criteria, and the scoring levels and descriptions
- Results provide feedback that is actionable including
  - Areas for improvement
  - Suggestions
- The custom prompt prevents Charlie from giving learners specific verbiage to use in their essays, encouraging engagement with the feedback vs. copy/pasting

Successes

- Available to all students at Purdue
- Works with any rubric – no training necessary

Challenges

- Feedback in certain domains (i.e. Pharmacy drug monograph reviews) was not suitable for use in the course. Too much domain-specific knowledge is required and not found in current language models
- Feedback is not yet specific to the course materials
- Token costs are manageable but not negligible. The university is currently absorbing the cost of the tokens.

Future Work

- Charlie currently uses the OpenAI API for generating feedback
  - We want to expand to using more models, including open-source models hosted on campus
- Purdue plans on dynamically building a RAG from course materials so the student can get course-specific feedback including links into the reference material

PURDUE
UNIVERSITY®