QA Automation of Canvas Courses

Natalia Echeverry • University of Pittsburgh •

1 Problem

Online education growth demands scalable quality assessment:

- → Tight development timelines
- → QA standards and rubrics are underutilized
- → Difficulty connecting learning sciences with instructional design practice
- → Lack of systematic identification of improvement opportunities

2 Solution

QA Bot with three innovations:

- → Hybrid architecture (rule-based + LLM prompting)
- → Resource efficient (8B parameters)
- → Open source and free access
- → Actionable feedback based on course quality standards

3 Research

- → **RQ1:** Small models vs experts?
- → **RQ2:** Agreement among models?
- → **RQ3:** Small vs large models?

Validation: 7 Canvas courses, 20 OLC Essential Design standards, Human review benchmark.

4 System Architecture & Rating Agreement Results

1 Parse .imscc → 2 Map Criteria → 3 Extract Content → 4 LLM Analysis → 5 Generate Report

OLC EVALUATION - HUMAN BENCHMARK

65%

Rated as "Exemplary"

OLC EVALUATION - ALL LLM MODELS

0%

Rated as "Exemplary"

Model	% Agreement	Cohen's κ	Bias
Llama 3.1 8B	48.6%	**0.06**	-0.60
DeepSeek R1 8B	30.0%	**0.04**	-1.13
Claude Sonnet 4	15-17%	**~0.01**	-1.32
GPT-4o	17-20%	**~0.00**	-1.17

κ < 0.20 = poor agreement (Landis & Koch, 1977). Negative bias = systematic underrating.

5 Frameworks

Quality Matters
28 criteria

Online Learning Consortium
20 objectives

Universal Design for Learning
5 principles

Customized Rubric
Future work

6 Models

Open-Source Models:

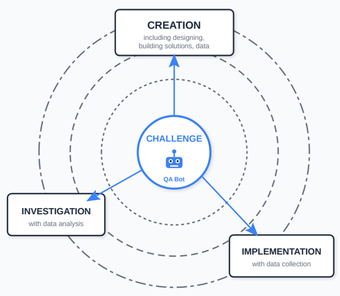
Llama 3.1 8B **DeepSeek-R1 8B**
Llama 3.1 70B **DeepSeek-R1 70B**

Commercial Models:

Claude Sonnet 4 **GPT-4o**

8B models run on standard laptop (8GB RAM)

7 Learning Engineering



Adapted from: Learning Engineering Process by Aaron Kessler, Jim Goodell, Sae Schatz (CC BY)

QA Bot supports nested improvement cycles:

- → **Creation:** Validate course design against standards
- → **Implementation:** Identify improvement opportunities
- → **Investigation:** Targeted analysis of design elements

8 Semantic Similarity Analysis

LLMs generate semantically similar justifications to humans but assign systematically different numeric ratings.

Comparison	Justification Similarity	Rating Agreement	Gap
Humans vs. DeepSeek	**0.41**	**0.30**	+0.11
Humans vs. Llama	**0.43**	**0.49**	-0.06

Interpretation: Positive gap = justification quality exceeds rating accuracy. The reasoning module works; classification is poorly calibrated.

9 Impact & Future

Open-source for learning engineering teams

Future: Classification recalibration • Hybrid human-bot workflows • Larger model evaluation • Custom rubric integration • Multi-LMS support

Acknowledgments: Rae Mancilla, Stephen Butler. Claude Sonnet 4.5 for writing assistance.
Keywords: Instructional design • Quality Assurance • Learning Engineering • LLMs • Canvas LMS • Automated QA