Transforming Expert Insight into Scalable AI Assessment: A Framework for LLM-Generated Metrics and User-Calibrated Evaluation

Nicholas Choma*, Sreecharan Sankaranarayanan and Rajesh Cherukuri
Amazon.com, Inc.

Introduction and Motivation

Assessing sophisticated AI systems, especially in specialized domains, is hindered by a persistent qualitative-quantitative gap: experts intuitively recognize “good” output, yet their nuanced judgments resist translation into scalable, automated measures.

We address this challenge with **a two-stage framework that first leverages Large Language Models (LLMs) to distill expert feedback into formal evaluation rubrics, and then deploys LLMs as “judges” to apply those rubrics rapidly and consistently**, preserving expert intent. A human-in-the-loop calibration protocol iteratively aligns LLM scores with expert ratings, ensuring metrics and evaluators co-evolve with shifting requirements.

We validate the approach on a generative-AI system that produces workplace learning content; automated assessments correlate directionally with expert judgments while eliminating the bottleneck of manual coding. By rooting scalable evaluation in expert values, our framework advances trustworthy, adaptable assessment for next-generation AI systems.

Methodology

Our framework converts qualitative expert insight into scalable, quantitative assessment through a rigorously staged process that blends large-language-model (LLM) assistance with human oversight. At a high level, it transforms subject matter experts’ requirements into evaluation metrics across four sequential phases, ensuring that both the metrics themselves and the automated evaluator remain aligned with evolving expert expectations.

Scalable AI Assessment Phases

1. **Qualitative requirement elicitation** – structured interviews and analysis of historical expert feedback isolate and prioritize the domain-specific quality dimensions to be measured.
2. **Metric generation & formalization** – using the elicited insights as context, an LLM iteratively drafts operational definitions, scoring rubrics, exemplar anchors, and validation criteria for each abstract concept.
3. **Automated implementation (LLM-as-Judge)** – the formalized rubrics are embedded in carefully engineered prompts so that the LLM can score new artifacts consistently; reliability mechanisms (e.g., multi-pass low-temperature evaluation) balance accuracy with cost.
4. **Expert calibration & refinement** – human experts score a stratified sample in parallel with the LLM; statistical alignment metrics guide iterative updates to definitions, rubrics, and prompts until predefined agreement thresholds are met.

Together, these phases establish a feedback loop in which expert knowledge seeds the metrics, LLMs apply them at scale, and continued expert calibration sustains fidelity as both the AI system and domain standards evolve.

Case Study: Learning Design

A generative-AI system that produces workplace learning content was selected, as evaluating pedagogical quality demands nuanced expert judgment.

- **Dataset**: 13 content items spanning diverse topics and expected quality levels.
- **Expert baseline**: 60 independent ratings from learning-design experts on those items.
- **Automated Scoring:** The same artefacts were then scored by the LLM-as-Judge using the metrics generated in earlier phases.

Algorithm 1 Calibration Feedback Loop Algorithm.

```
while alignment $<$ predefined_threshold do
    Identify discrepancy patterns between LLM and expert scores.
    Analyze expert reasoning for these disagreements.
    Refine metric definitions, rubrics, or exemplars.
    Update automated evaluation prompts for LLM-as-a-Judge.
    Re-evaluate a calibration sample set.
    Measure new alignment scores.
end while
```

Alignment between expert and model scores was computed with Pearson, Spearman, ICC, MAE and RMSE statistics; the entire calibration cycle followed **Algorithm 1**, iteratively analyzing discrepancies and refining metric definitions, rubrics and prompts until alignment met a predefined threshold.

Results

Initial alignment showed strong interclass correlation, indicating high agreement in ranking order of preference between experts and automated scores:

- Interclass Correlation: **ICC = 0.9376**.
- Pearson **r = 0.3089** (p = 0.0163) and Spearman **ρ = 0.3731** (p = 0.0033).
- Overall error: **MAE ≈ 0.40; RMSE ≈ 0.49**.

Although though the model tended to assign slightly higher scores on average, experts and the LLM agreed directionally on content quality, suggesting our framework can capture core aspects of pedagogical quality.

Conclusions

This study **presents a framework that transforms qualitative expert insights into quantitative, scalable metrics** by harnessing LLMs both to draft evaluation rubrics and to act as automated “judges.” An initial learning-design case study showed directional correlation between LLM and expert scores, demonstrating a practical path toward trustworthy, expert-grounded AI assessment.

Acknowledgements

We would like to thank the learning design experts who provided valuable feedback during the calibration process, as well as our internal collaborators for their support throughout this research. Their insights and contributions were instrumental in shaping and validating this work.

AIA Workshop IJCAI 2025

*Corresponding author: njchoma@amazon.com