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.
{njchoma, sreeis, rccheruk}@amazon.com

Abstract

Effectively assessing AI systems, particularly those operating in specialized domains or producing dynamic outputs, requires translating nuanced human expertise into scalable, quantitative measures. Traditional metrics often fall short in capturing qualitative requirements that domain experts intuitively grasp. This paper presents a novel framework that systematically transforms qualitative expert feedback into quantitative metrics for assessing the output quality of AI systems. Our methodology leverages Large Language Models (LLMs), first to help articulate and formalize these metrics from expert input, and subsequently as "judges" to apply them in an automated fashion. As validation, we present initial results from calibration against expert ratings, demonstrating that automated assessments align with human judgment and can evolve with changing requirements. Learning content creation serves as our illustrative specialized domain. Its reliance on learning design frameworks, coupled with the need for nuanced expert evaluation of pedagogical quality, makes it an ideal test case for our framework. Results confirm that our LLM-generated, expert-calibrated metrics achieve promising alignment with expert evaluations, enabling robust, scalable, and adaptable assessment.

1 Introduction

A fundamental challenge in deploying sophisticated AI systems, especially in specialized domains, is bridging the gap between expert conceptualizations of quality and measurable, automated evaluation. Domain experts often possess an intuitive, rich understanding of what constitutes a "good" system output or behavior, yet this understanding can be difficult to articulate in precise, quantitative terms suitable for scalable AI assessment and iterative improvement. Existing methodologies for qualitative analysis, such as the manual creation of validated "codebooks" and "coding schemes," demand significant expert effort and adherence to strict processes [Kawulich, 2017; Skjott Linneberg and Korsgaard, 2019], even when partially automated [Sankaranarayanan *et al.*, 2025; Simon *et al.*, 2025]. Furthermore, applying these schemes, even post-validation, remains an expert-driven, non-scalable task, particularly problematic for rapidly evolving AI systems or those producing dynamic outputs.

This paper introduces a systematic methodology to address this challenge by transforming qualitative, often ill-defined, quality requirements into quantitative metrics that maintain strong fidelity to expert judgment. Our core contribution is a framework leveraging Large Language Models (LLMs) in two key stages: first, assisting in the generation and formalization of assessment rubrics from patterns in expert feedback; and second, acting as "judges" to apply these rubrics for automated evaluation, a concept that has gained traction in various evaluation scenarios [Zheng *et al.*, 2023; Kocmi and Federmann, 2023].

To ensure genuine alignment and address divergence risks, we present a rigorous expert calibration protocol where human experts validate and refine the LLM-driven assessment. This iterative loop allows metrics and their application to adapt, reflecting evolving expert understanding or system requirements, akin to principles in learning from human preferences [Christiano *et al.*, 2017].

We validate our approach in a generative AI system creating workplace learning content. In this domain, essential qualities like "content alignment with learning objectives" or "scenario authenticity," while grounded in learning design frameworks such as Backwards Design [McTighe and Thomas, 2003], ADDIE [Branch and Varank, 2009], and SAM [Allen, 2012], are ultimately nuanced and context-dependent. Our framework systematically quantifies these attributes using expert feedback on generated content, bypassing time-intensive qualitative coding. This enables scalable assessment and offers a pathway for system improvement analogous to reinforcement learning from human feedback (RLHF) [Ouyang *et al.*, 2022] when there is much less data to draw from.

Overall, this work contributes to the broader need for assessment paradigms that are not only efficient but rooted in user (expert) values, vital as AI systems become more integrated into complex, human-centric tasks.

1.1 Contributions

Our primary contributions are: a **Qualitative-to-Quantitative Transformation Framework** that systematically leverages expert feedback to convert ill-defined quality concepts into measurable metrics; an **LLM-Powered Metric Generation and Application** approach where LLMs help define evaluation criteria from qualitative input and automate their application as an "LLM-as-a-Judge"; a rigorous **Expert Calibration Protocol for Alignment** to calibrate LLM-generated assessments against domain expert judgment, ensuring sustained alignment and enabling metric evolution; and an **Empirical Validation in Learning Design** demonstrating the framework's effectiveness in quantifying pedagogical quality for a generative AI, showing measurable improvements in human-AI assessment agreement.

2 Problem Formulation

2.1 The Qualitative-Quantitative Gap

Specialized domains feature quality standards that experts recognize but struggle to formalize. This gap arises from several factors: experts often rely on *implicit knowledge* developed through experience; quality judgments are frequently *context-dependent*, varying with subtle factors; *subjective criteria* can lead to multiple valid interpretations of quality; and there's an inherent *articulation difficulty* in converting tacit knowledge into explicit, measurable rules.

2.2 Requirements for Quantification

To be effective, the quantification of qualitative metrics must satisfy several requirements. Metrics must maintain *fidelity*, capturing the essence of expert judgment. They need *consistency*, producing reliable results across similar contexts. *Scalability* is crucial, ensuring automated evaluation is computationally feasible. Scores must have *interpretability* for domain experts. Critically, metrics require *adaptability*, evolving as expert understanding and requirements change.

2.3 The Learning Design Challenge

Learning design exemplifies this qualitative-quantitative challenge with concepts such as pedagogical soundness (alignment with learning science), learner engagement, authenticity (relevance to real-world contexts), clarity of communication, and appropriate progressive complexity or scaffolding. While readily recognized by learning design experts, these concepts resist direct, straightforward measurement.

3 Methodology

Our framework transforms qualitative requirements into quantitative metrics through four key phases.

Phase 1: Qualitative Requirement Elicitation This initial phase focuses on understanding and capturing expert notions of quality. It involves structured *expert interviews* to identify salient quality dimensions, a systematic *feedback analysis* of historical expert evaluations on similar content or systems, and a *prioritization* step where identified quality dimensions are ranked by their importance and perceived measurability.

Phase 2: Metric Generation and Formalization With LLM assistance, qualitative insights from Phase 1 are translated into formal metrics. Each metric is generated individually in four steps, where each step corresponds to one LLM call. These steps include creating an *operational definition* for each abstract concept to describe observable behaviors, developing a detailed *scoring rubrics* with graduated scales for quality assessment, collecting *exemplars* of content that represent different quality levels to anchor the rubrics, and establishing *validation criteria* to define standards for subsequent metric effectiveness. When generating each metric, the LLM receives relevant expert feedback and available background information about the AI system as context. Example generated metrics are provided in the Appendix.

Phase 3: Automated Implementation The formalized metrics are then implemented for automated assessment, using an *LLM-as-a-Judge architecture* where a language model applies the scoring rubrics. This phase involves careful *prompt engineering* to craft instructions that enable the LLM to replicate expert reasoning, implementing *consistency mechanisms* (discussed later) to ensure reliable evaluation across samples, and *efficiency optimization* to balance assessment accuracy with computational cost.

Phase 4: Expert Calibration and Refinement This human-in-the-loop phase ensures the automated assessments align with expert judgment. It involves *parallel evaluation*, where both automated systems and human experts score a common set of samples. An *alignment analysis* statistically assesses human-AI agreement. *Bias identification* detects systematic differences in evaluation patterns, and an *iterative refinement* process adjusts metric definitions, rubrics, or LLM prompts based on calibration results to improve alignment.

3.1 LLM-assisted Metric Specification

Feedback Theme Extraction

The input in Phase 1 is a collection of expert feedback on AI-generated content. This feedback can be provided at various granularities: overall content, selected sections, sentences, or even phrases. Experts are typically solicited to provide the reason for their feedback, a critique, substantiation for the critique, and, if applicable, a suggested rewrite. While comprehensive feedback is beneficial, the system can work with varying levels of detail, though richer, targeted feedback tends to reduce the iterations needed for convergence. The output from Phase 2 is a structured taxonomy of quality dimensions or themes that emerge from the feedback.

Metric Specification Generation

For each theme identified from expert feedback, the system, with LLM assistance, generates a full metric specification. This typically includes a concise *Metric Name*, a clear *Definition* of what the metric measures, detailed *Evaluation Guidelines* for assessment, a graduated *Scoring Rubric* with specific criteria for each level, and *Example Anchors* using sample content to illustrate different score levels.

Automated Application of Generated Metrics for Evaluation: LLM-as-a-Judge Architecture

The automated evaluation leverages an LLM (Claude Sonnet 4 in this case) to apply the defined metrics and rubrics. A structured prompt guides the LLM. The conceptual Python code below is aggressively formatted with very short lines and reduced font size to fit a narrow two-column display:

```
def eval_content(
    content_to_eval,
    metric_definition_text,
    rubric_details_text
):
    # Build prompt for narrow display
    # All strings are kept very short
    prompt_lines = [
        "EVALUATE THIS CONTENT:",
        f"METRIC: {metric_definition_text}",
        f"RUBRIC: {rubric_details_text}",
        f"INPUT: {content_to_eval}",
        "SCORE (1-5) AND PROVIDE",
        "A DETAILED JUSTIFICATION."
    ]
    prompt = "\n".join(prompt_lines)

    # Conceptual LLM interaction (pseudo):
    # response = llm_service.call(prompt)
    # score, analysis_text = parse(response)
    # return score, analysis_text

    # Example placeholder return:
    return 0, "Output not generated"
```

(Actual LLM call and response parsing are omitted for brevity in this example.)

Consistency Mechanisms

To ensure reliability in the LLM's evaluations, several consistency mechanisms are employed. These include averaging scores from *multiple evaluation runs*, using *temperature control* (e.g., lower temperature settings) to reduce randomness in LLM outputs, adhering to strict *prompt standardization* for consistent instruction formatting, and incorporating *calibration anchors* (reference examples) within evaluation prompts to guide the LLM.

Dynamic Metric Evolution

The framework is designed for metrics to adapt over time. This evolution is facilitated through *Feedback Integration*, where new expert input continually refines existing metrics; *Performance Monitoring*, which tracks metric effectiveness in predicting expert satisfaction or specific outcomes; *Concept Drift Detection* to identify shifts in expert preferences or understanding of quality; and *Automatic Rebalancing*, which can adjust metric weights or components based on observed usage patterns or changing importance. Together, this constitutes the last phase: Phase 4 of the methodology.

3.2 Expert Calibration Protocol

In Phase 4, the expert calibration protocol ensures ongoing alignment between automated scores and human expert judgment.

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
```

Calibration Study Design

The design of calibration studies incorporates several elements to ensure rigor: *Sample Selection* involves stratified sampling of content across various quality levels and types to ensure diverse coverage. A user-friendly *Evaluation Interface* (prototype created in Python Dash) is provided for expert assessment. *Blind Evaluation* is employed, where experts are unaware of automated scores during their assessment to prevent bias. The involvement of *Multiple Raters* allows for inter-rater reliability analysis among expert evaluators, helping to establish a human expert baseline.

Alignment Measurement

Human-AI agreement is quantified using various statistical measures. These include Pearson Correlation to assess the linear relationship between human and automated scores, Spearman Rank Correlation for monotonic relationship preservation, Cohen's Kappa to measure agreement while accounting for chance, and Mean Absolute Error to determine the average magnitude of score differences.

Calibration Feedback Loop

The calibration process is iterative, following a loop aimed at improving alignment until a satisfactory threshold is met, as depicted in Algorithm 1. Note that this threshold must be computed taking the variance of expert scores into account; low variance among experts necessitates stronger alignment than does high variance.

4 Case Study: Learning Design

Our approach is demonstrated through a generative AI system designed to create workplace learning content at Amazon. This domain was chosen due to the inherent subjectivity and complexity of evaluating pedagogical quality.

4.1 Domain-Specific Quality Dimensions

Through expert interviews with learning designers and analysis of their feedback on AI-generated content, our system identified key qualitative dimensions. Pedagogical dimensions included *Learning Outcome Alignment* (content supporting stated objectives), *Cognitive Load Management* (appropriate information density), *Knowledge Transfer* (facilitating application to new contexts), and *Scaffolding Quality* (progressive skill building). Content quality dimensions included *Scenario Authenticity* (realistic workplace situations), *Action Specificity* (concrete, implementable guidance), *Self-Containment* (complete explanations without external dependencies), and *Engagement Factors* (elements that maintain learner interest). These dimensions formed the basis for metric development.

4.2 Quantification Process

For each identified qualitative dimension, we developed a detailed quantitative metric by supplying the aforementioned inputs to the Claude Sonnet 4 model with conservative sampling parameters (temperature = 0.1, top-k = 250) to ensure consistent outputs while maintaining sufficient flexibility to capture nuanced expert feedback patterns. This involved defining the metric, creating evaluation guidelines, and establishing a scoring rubric with illustrative examples. Below are summarized examples for two such metrics.

Learning Outcome Alignment Example

The qualitative concept for this metric is that "Content should clearly support the learning objectives." The metric evaluates how effectively learning content aligns stated learning objectives with the actual cognitive processes required by the content and assessments, often using a framework like Bloom's Taxonomy. It ensures that if higher-order thinking is claimed, the content indeed provides opportunities for such engagement. The full metric description, guidelines, and scoring rubric are provided in Appendix A.

Scenario Authenticity Quantification Example

The qualitative concept here is that "Scenarios should feel realistic and relevant." This metric assesses how well learning scenarios reflect genuine workplace situations while appropriately balancing realism with pedagogical needs, such as by anonymizing confidential elements while maintaining authentic processes. The full metric description, guidelines, and scoring rubric are provided in Appendix B.

5 Evaluation and Results

This section presents the setup and findings from our initial expert calibration process, designed to validate the alignment of the LLM-generated metrics with human expert assessments.

5.1 Calibration Setup and Initial Findings

For the initial calibration study, we used a set of 13 distinct content samples, representing different topics and expected quality levels within the learning design domain. These samples were evaluated by a panel of learning design experts, resulting in 60 unique expert ratings across the samples for key metrics. The same samples were also evaluated using our LLM-as-a-Judge system with the initially generated metrics.

Quantitative analysis of alignment between expert scores and LLM scores yielded the following initial results: Pearson Correlation was 0.3089 ($p = 0.0163$), and Spearman Rank Correlation was 0.3731 ($p = 0.0033$). The Intraclass Correlation Coefficient (ICC) indicated a high degree of consistency for certain metrics when looking at patterns, with one variant reaching 0.9376. The Mean Absolute Error (MAE) was 0.3972 on a normalized scale, and the Root Mean Square Error (RMSE) was 0.4864.

5.2 Qualitative Observations

The main qualitative takeaway from this initial calibration phase was that, directionally, the experts and the LLM agreed on the quality of content. However, a tendency was observed for the LLM to provide slightly higher scores on average compared to the human experts. These initial findings, both quantitative and qualitative, serve as the baseline for the iterative refinement process described in Section 3.2 (Calibration Feedback Loop), guiding adjustments to metric definitions, rubrics, and LLM prompts to improve alignment. Further iterations of this loop are expected to enhance these correlation figures and reduce error margins, as suggested by improvements typically seen in such calibration processes [Ouyang *et al.*, 2022].

6 Discussion

6.1 Implications of Expert-Driven, LLM-Powered Assessment

Bridging Subjective Expertise and Objective Measurement

Our framework, centered on transforming expert feedback into LLM-generated and calibrated metrics, demonstrates that qualitative domain expertise can be systematically quantified. This offers a path to scalable application of expert knowledge by codifying tacit expertise for consistent, automated assessment. It also leads to transparent and evolvable criteria, making implicit quality standards explicit and allowing them to adapt through ongoing expert calibration. Furthermore, it enables data-driven system improvement, allowing for the refinement of AI systems based on metrics grounded in human expertise.

A Collaborative Model for Human-AI Assessment

The calibration process fosters a collaborative dynamic where human insight guides AI judgment; experts define and refine "quality," which LLMs then apply at scale. This builds trust in automated evaluation through transparent alignment metrics and iterative refinement. The assessment system itself can learn and improve via expert interaction, making it suitable for evaluating evolving AI systems. This human-AI collaboration mirrors principles from interpretable machine learning, where understanding model behavior is key [Doshi-Velez and Kim, 2017].

6.2 Methodological Contributions

The primary methodological contributions are the LLM-assisted metric generation from expert feedback and the rigorous calibration for sustained alignment. Using LLMs to help draft metric specifications from qualitative feedback accelerates a traditionally labor-intensive process and aids in capturing nuances. The systematic expert calibration protocol provides a robust mechanism to ensure the LLM-as-a-Judge's evaluations remain faithful to human expert judgment over time, crucial for maintaining assessment validity as systems or expert understanding evolves.

6.3 Challenges and Limitations

Several challenges and limitations are inherent in this approach. First, regarding the *boundaries of quantification*, not all qualitative nuances, especially those requiring deep situational knowledge not easily captured in rubrics, may be fully measurable. Second, there's a *risk of bias amplification*, where biases present in initial expert feedback could be encoded into the metrics if not carefully managed during elicitation and subsequent calibration phases. This is mainly mitigated by broadening the group of experts we gather feedback from as a part of the initial input. Third, the *LLM reliability as judges* is contingent on prompt quality, rubric clarity, and the evolving capabilities of the LLMs themselves, necessitating ongoing monitoring and potential re-calibration. A related issue is one of entirely novel AI outputs not foreseen during initial metric design.

We discuss some ways to address these challenges in the next section.

7 Future Work

To address model drift and related issues with the reliability of LLMs as judges, we expect to use multiple LLMs and a voting/concordance procedure for updates to the metrics and their definitions. This reduces the likelihood of idiosyncrasies with a single model leading to large changes in the metrics and their definitions.

While we face a small data issue, we expect to kick-off model fine-tuning runs when sufficient feedback items have been collected. Our use-case is closest to instruction-following and prior work shows improved fine-tuning performance with approximately 1000 carefully crafted examples [Zhou *et al.*, 2023; Dong *et al.*, 2023].

Likewise, a granular evaluation of the framework, assessing the LLM's ability to extract qualitative requirements, the quality of metric conversion, and the fidelity of rubric application will reveal areas for further development that offer the highest impact. Tracing quantitative gains across multiple iterations will further strengthen the analysis and guide future enhancements to the framework. Specifically with respect to the metrics, we aim to explore methods for LLMs to automatically identify areas where current metrics are insufficient or expert disagreement is high, thereby prompting targeted calibration or soliciting additional expert feedback.

To streamline expert calibration, developing active learning techniques to select the most informative samples for expert review could reduce expert workload further. Additionally, exploring few-shot or zero-shot calibration methods, where LLMs can better generalize from minimal expert input once robust initial metrics are established, is a promising direction. From the user-alignment standpoint, the key is to demonstrate consistent scoring by the LLM-as-a-judge and little drift in several iterations before we can gain the trust of experts to let the system independently work end-to-end without the human experts in the loop. Once that is unlocked, however, we can run automated prompt engineering cycles to update the content generation prompt to score highly on the metrics. Assuming the metrics suite is comprehensive, the content should not overfit to the metric and their definitions.

Finally, systematically applying and evaluating this framework (expert feedback → LLM-generated metrics → LLM-judge → expert calibration) across diverse domains beyond learning design, such as healthcare (clinical decision quality), legal (document compliance), and creative industries (artistic assessment), will be crucial for establishing its generalizability.

8 Conclusion

This paper presented a comprehensive framework for transforming qualitative expert insights into quantitative, scalable metrics for AI system assessment. Our core methodology emphasizes the synergistic use of domain expert feedback to guide LLM-assisted generation of evaluation criteria, the application of these criteria by LLMs acting as judges, and a continuous expert calibration loop to ensure enduring alignment. The initial application of this framework to a learning design use case, achieving promising directional correlation with expert judgment, validates our approach and sets the stage for further refinement.

The ability to systematically translate human expertise into reliable, automated assessment mechanisms that can adapt over time is increasingly important. Our work provides a practical pathway to achieve this, offering a robust method for user-aligned assessment relevant for complex AI systems, including those that are adaptive or LLM-based. This approach contributes to building more trustworthy AI systems by ensuring their evaluation is grounded in the nuanced understanding of human experts.

Acknowledgments

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.

A Detailed Metric: Learning Outcome Alignment

Qualitative Concept: "Content should clearly support the learning objectives."

Description: This metric evaluates how effectively learning content maintains alignment between stated learning objectives and the actual cognitive processes required by the content and assessments, using Bloom's Taxonomy as the foundational framework. The metric specifically focuses on whether the cognitive level specified in learning objectives (e.g., "analyze," "evaluate," "create") is matched by the cognitive demands of both instructional content and assessment items. At its core, this metric examines cognitive process alignment across three key components:

1. The stated learning objective's cognitive level (e.g., "analyze trade-offs between options")
2. The cognitive processes required by instructional content and activities

3. The cognitive level at which assessments test learning

For example, if a learning objective states learners will "analyze trade-offs," but the content only presents descriptions and the assessment only tests recall, this represents misalignment. The metric identifies such gaps between intended and actual cognitive processing levels. This metric is particularly crucial for adaptive learning systems where content generation must maintain consistent cognitive alignment across variations. It ensures that when an AI generates content claiming to develop higher-order thinking skills (analysis, evaluation, creation), it actually provides opportunities for learners to engage in those cognitive processes rather than defaulting to lower-level activities (remembering, understanding). This alignment is essential for learning effectiveness. The metric focuses solely on cognitive process alignment and does not attempt to evaluate knowledge dimensions, content progression, or general instructional quality, as these are covered by other metrics. Its specific purpose is to ensure that what learners are asked to do mentally matches what the learning objectives claim they will learn to do.

Quantified Metric Definition: Degree to which content elements directly address stated learning outcomes by aligning required cognitive processes with those specified in the objectives, based on Bloom's Taxonomy.

Scoring Scale (Example): 1-5 point scale.

- 5: All content directly supports objectives with clear, demonstrable cognitive alignment at the specified Bloom's level across objectives, content, and assessments.
- 4: Most content supports objectives with strong cognitive alignment; minor elements may be slightly off-level but do not detract significantly.
- 3: Content generally supports objectives, but there are noticeable inconsistencies in cognitive alignment in some areas; some activities or assessments may operate at a lower level than specified.
- 2: Content partially supports objectives, with significant sections showing cognitive misalignment; higher-order objectives are often addressed with lower-order content/assessment.
- 1: Content fails to address stated objectives in terms of cognitive alignment; fundamental mismatch between claimed cognitive level and actual demands.

Evaluation Guidelines:

1. Each learning objective must explicitly state a Bloom's Taxonomy cognitive process level.
2. Instructional content must require learners to engage in cognitive processes at the same level specified in the learning objective.
3. Assessment items must test learners at the cognitive process level stated in the learning objective.
4. If a learning objective specifies higher-order thinking (analyze, evaluate, create), the content must provide explicit opportunities for learners to practice these complex cognitive processes.
5. Learning activities must match the cognitive level of the learning objective through appropriate task design.
6. Examples and scenarios must engage learners at the stated cognitive process level.
7. Content variations and adaptations must maintain consistent cognitive process alignment with the original learning objective.
8. When multiple cognitive processes are present in a learning objective, the content must address each process level explicitly.
9. Feedback and remediation must operate at the same cognitive process level as the learning objective.
10. The cognitive process level should remain consistent across all components (objectives, content, activities, assessments) unless there is an explicit instructional reason for variation.

Scoring Rubric:

- *Comprehensive Alignment (Corresponds to original score 3):* Content demonstrates comprehensive cognitive alignment across all components. Learning objectives specify clear Bloom's levels, instructional content requires matching cognitive processes, and assessments test at the stated levels. All activities and examples engage learners at the intended cognitive level, with consistent alignment maintained across variations.
- *General Alignment (Corresponds to original score 2):* Content shows general cognitive alignment with occasional minor mismatches. Most components operate at the stated Bloom's level, but some activities or assessments may default to slightly lower cognitive processes than specified in objectives. Core alignment is maintained but with room for improvement.
- *Significant Misalignment (Corresponds to original score 1):* Content exhibits significant cognitive misalignment in multiple areas. While learning objectives may specify higher-order thinking, many content elements and assessments operate at lower cognitive levels. Some alignment exists but fails to consistently engage learners at intended cognitive levels.
- *Fundamental Misalignment (Corresponds to original score 0):* Content shows fundamental cognitive misalignment throughout. Learning objectives lack clear Bloom's levels, content operates at basic recall regardless of stated objectives, and assessments fail to test at appropriate cognitive levels. No consistent alignment between intended and actual cognitive processes.

B Detailed Metric: Scenario Authenticity

Qualitative Concept: "Scenarios should feel realistic and relevant."

Description: This metric evaluates how well learning scenarios reflect genuine work situations (e.g., within a specific company like Amazon, as per the original detailed context) while appropriately balancing realism with pedagogical needs. This metric focuses specifically on the authenticity of workplace scenarios used in learning content, recognizing that scenarios may use anonymized or fictionalized elements (like project names) while maintaining authentic company processes, roles, and organizational dynamics. The metric evaluates scenarios across a progression of complexity – from simplified introductory scenarios that may use "magic wand" solutions to teach basic concepts, to more nuanced scenarios that reflect the full complexity of the working environment. This progression should match the learning objectives and the learner's developing understanding. A key aspect is the appropriate use of the organizational context. While scenarios may use fictional characters or project names, they must accurately represent how work gets done – including the use of specific mechanisms or practices (e.g., "working backwards documents" in an Amazon context). The metric specifically evaluates whether scenarios demonstrate authentic problem-solving approaches, even when teaching about poor practices or common mistakes as learning moments. This metric is distinct from factual groundedness, source coverage, or writing style. Instead, scenario authenticity focuses specifically on how well the learning scenarios serve as authentic vehicles for teaching specific ways of working, while maintaining appropriate pedagogical scaffolding.

Quantified Metric (example, weights can be adjusted):

- Workplace Realism (40% weight): Accuracy of organizational context, roles, interactions, and cultural elements.
- Process & Technical Precision (30% weight): Correctness of domain-specific processes, tools, and technical details, appropriately anonymized.
- Situational Plausibility & Pedagogical Appropriateness (30% weight): Likelihood of scenario occurrence and its suitability for the learning objectives and learner progression.

Evaluation Guidelines:

1. Verify scenarios use appropriate organizational mechanisms, documents, and practices, allowing for fictionalized elements to protect confidentiality.
2. Assess whether the scenario's complexity level matches learning objectives and learner progression.
3. Confirm scenarios demonstrate authentic problem-solving approaches, even when illustrating mistakes.
4. Check that organizational dynamics accurately represent the unique culture and practices.
5. Evaluate whether technical and process details align with current practices, with appropriate anonymization.
6. Verify scenario challenges reflect genuine business situations at an appropriate scale, using realistic constraints.
7. Ensure scenario resolutions demonstrate authentic approaches while maintaining pedagogical scaffolding.
8. Check that cross-team dependencies and partnerships reflect actual working relationships and collaboration mechanisms.
9. Assess whether scenario progression across sections builds understanding while maintaining consistent context.
10. Verify customer impacts and business outcomes reflect actual scale and scope, protecting confidential information.

Scoring Rubric:

- *High Authenticity (Corresponds to original score 3):* Scenario demonstrates authentic organizational practices while appropriately scaffolding learning progression. Uses accurate mechanisms and organizational dynamics with appropriate fictionalization. Complexity increases logically across sections while maintaining consistent context. Cross-team interactions and problem-solving approaches reflect genuine practices at the right level for learning objectives.
- *Good Authenticity (Corresponds to original score 2):* Scenario uses appropriate organizational terminology and practices with proper fictionalization, but shows minor issues in progression or organizational dynamics. May oversimplify some dependencies or compress timelines unrealistically for later-stage learning. Core context remains accurate but some nuances are imprecise.
- *Basic Authenticity (Corresponds to original score 1):* Scenario maintains basic organizational terminology and context but lacks appropriate progression of complexity. May use overly simplified solutions in advanced sections. Organizational dynamics may be oversimplified beyond pedagogical necessity. Cross-team interactions may not reflect actual mechanisms.
- *Lacks Authenticity (Corresponds to original score 0):* Scenario fails to use appropriate organizational practices or shows fundamental misunderstanding of ways of working. May expose sensitive information or fail to properly fictionalize details. Learning progression is absent or inappropriate. Dynamics and problem-solving are unrealistic.

References

[Allen, 2012] Michael Allen. The successive approximation model (sam). In *Trends and Issues in Instructional Design and Technology*, pages 67–81. Routledge, 2012.

[Branch and Varank, 2009] Robert Maribe Branch and İlhan Varank. *Instructional design: The ADDIE approach*, volume 722. Springer, 2009.

[Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[Dong *et al.*, 2023] Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*, 2023.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Kawulich, 2017] Barbara B Kawulich. Coding and analyzing qualitative data. *The BERA/SAGE handbook of educational research*, 2:769–790, 2017.

[Kocmi and Federmann, 2023] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.

[McTighe and Thomas, 2003] Jay McTighe and Ronald S Thomas. Backward design for forward action. *Educational leadership*, 60(5):52–55, 2003.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Sankaranarayanan *et al.*, 2025] Sreecharan Sankaranarayanan, Conrad Borchers, Sebastian Simon, Elham Tajik, Amine Hatun Ataş, Berkan Celik, Francesco Balzan, et al. Automating thematic analysis with multi-agent llm systems. 2025.

[Simon *et al.*, 2025] Sebastian Simon, Sreecharan Sankaranarayanan, Elham Tajik, Conrad Borchers, Francesco Balzan, Sebastian Strauß, Sree Viswanathan, Amine Ataş, Mia Čarapina, Li Liang, et al. Comparing a human's and a multi-agent system's thematic analysis: Assessing qualitative coding consistency. 2025.

[Skjott Linneberg and Korsgaard, 2019] Mai Skjott Linneberg and Steffen Korsgaard. Coding qualitative data: A synthesis guiding the novice. *Qualitative research journal*, 19(3):259–270, 2019.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[Zhou *et al.*, 2023] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.