GYMGOV.ORG · CERTIFICATION GOVERNANCE ARCHITECTURE

STORY PAPER · CGA SERIES

The Story Paper: Scalable Certification of Reasoning in the Age of AI

Extending Triple-Jump / OSCE-Style Assessment Through AI-Mediated Governance

Waleed S. Nema · April 2026 · gymgov.org/docs/whitepapers/the-story-paper.html

CONTENTS

EXECUTIVE SUMMARY

Advances in artificial intelligence are rapidly expanding how learning occurs. Individuals can now access tools that assist with explanation, synthesis, and problem-solving across a wide range of domains. As a result, learning is becoming more scalable, adaptive, and widely accessible.

This shift has important implications for professional certification.

In high-stakes fields such as medicine, law, engineering, and aviation, competence depends not only on knowledge, but on the ability to reason under conditions of uncertainty. Established assessment models—including oral examinations, Objective Structured Clinical Examinations (OSCEs), and related formats—are designed to evaluate this capability directly. They rely on contextual scenarios, interactive probing, and expert judgment to assess how candidates think.

These approaches are effective. However, they do not scale easily.

As participation in professional pathways grows—and as learning increasingly occurs in AI-assisted environments—a gap is emerging. While learning can scale, the certification of reasoning remains constrained by resource-intensive methods. At the same time, the presence of AI-assisted tools introduces new ambiguity in assessment, making it more difficult to distinguish between assisted output and independently generated reasoning.

Addressing this gap requires more than incremental adjustments to existing methods. A potential direction involves re-examining how assessment is structured and governed—separating learning environments from certification processes, representing training in terms of structured exposure to reasoning challenges, and verifying whether candidates can reason consistently across variations of those challenges.

Importantly, this direction is not tied to a single system or platform. It reflects an architectural and governance perspective that can support multiple, independent implementations across institutions. Progress requires coordinated collaboration among certification bodies, academic

institutions, technical participants, and governance stakeholders—and can occur incrementally.

SECTION 1

The Inflection Point

In recent years, advances in artificial intelligence have begun to reshape how knowledge is accessed, acquired, and applied. AI-assisted learning environments are no longer experimental; they are rapidly becoming integrated into educational workflows across disciplines. In many contexts, learners now have continuous access to tools that can retrieve, synthesize, and even generate domain-relevant information in real time.

This shift has profound implications for professional education and certification. Historically, assessment systems have evolved around the assumption that access to information is limited and that knowledge must be internalized to be applied. As a result, a significant portion of assessment design has focused on evaluating recall, recognition, and structured problem-solving within controlled conditions.

That assumption is no longer stable.

As access to information becomes effectively ubiquitous, the distinction between knowing and reasoning becomes more consequential. The ability to retrieve information is increasingly commoditized; the ability to interpret, prioritize, and act on that information in context remains a core professional competency. In high-stakes domains such as medicine, law, engineering, and aviation, this distinction is not academic—it directly affects decision quality, safety, and outcomes.

Existing assessment models have long recognized this. Oral examinations, Triple-Jump, Objective Structured Clinical Examinations (OSCEs), and related formats were designed specifically to evaluate reasoning under conditions of uncertainty. They rely on interactive, often adversarial engagement to probe how a candidate thinks, not simply what they know. These approaches remain among the most trusted methods for assessing professional competence.

However, they do not scale.

They depend on expert time, are logistically complex, and are difficult to standardize across large populations. As demand for professional certification grows—and as learning pathways diversify globally—these constraints become more pronounced.

The result is a growing tension between what is known to be the most meaningful form of assessment and what is feasible to deliver at scale.

> *We are entering a phase where learning can scale globally, but the certification of reasoning cannot—at least not using existing models alone.*

Addressing this gap does not require abandoning established assessment practices. Rather, it requires examining how their core strengths—contextual probing, adversarial dialogue, and evaluation of reasoning—might be extended through new forms of infrastructure that preserve trust while enabling scale.

SECTION 2

What Already Works

Established models of professional assessment have long recognized that competence is not defined solely by the possession of knowledge, but by the ability to apply that knowledge under conditions of uncertainty. Oral examinations, OSCEs, and multi-stage formats such as the "triple jump" share several essential characteristics.

First, they are interactive. This interaction allows for probing beyond initial answers, revealing how a candidate interprets information, adjusts to new inputs, and navigates ambiguity.

Second, they are contextual. Candidates are situated within realistic scenarios that require synthesis across multiple concepts, requiring them to prioritize, make trade-offs, and justify decisions that reflect real-world practice.

Third, they are adaptive and, at times, adversarial. Examiners introduce variations, challenge assumptions, and explore edge cases to test the robustness of a candidate's reasoning—distinguishing surface-level familiarity from deeper conceptual understanding.

Finally, they rely on expert judgment. Trained evaluators recognize not only correct conclusions, but also the quality and structure of the reasoning that led to those conclusions.

These characteristics—interaction, context, adaptability, and expert judgment—are central to why such formats are trusted. At the same time, because they depend on expert time and individualized engagement, they are resource-intensive and difficult to scale.

> *The problem is not that reasoning cannot be assessed. It is that it cannot be assessed at scale.*

The question, therefore, is not whether existing methods are effective. They are. The question is whether their core strengths can be preserved and extended in a way that allows reasoning to be evaluated more consistently and at greater scale.

SECTION 3

The Emerging Gap

AI-assisted environments are increasingly integrated into both formal and informal education. Learners can now access systems that provide explanations, suggest approaches, generate examples, and assist in problem-solving across a wide range of domains. This shift is expanding access to learning in meaningful ways—but it also introduces new ambiguity into the assessment process.

When assistance is readily available, the boundary between independently generated reasoning and externally supported output becomes less clear. A response may be correct and well-structured, but it may not fully reflect the candidate's own reasoning process. This ambiguity is not easily resolved through restriction: limiting tools during assessment does not address the broader shift in how individuals learn and prepare, nor does it reflect the environments in which professional reasoning is increasingly exercised.

As a result, the focus of assessment is beginning to shift. The central question is no longer whether a candidate can arrive at a correct answer in isolation, but whether they can demonstrate consistent, coherent reasoning across variations of a problem, including in settings where information and support are present.

> *As learning becomes AI-mediated, the boundary between assisted performance and genuine reasoning becomes increasingly difficult to define and verify.*

This gap is not theoretical. It is already beginning to surface in discussions around assessment integrity, the role of AI in education, and the future of credentialing. The question is not whether change will occur, but how it can be guided in a way that preserves trust while adapting to new conditions.

SECTION 4

IV Why Incremental Adjustments Are Not Sufficient

A range of incremental approaches has emerged in response to evolving assessment challenges—expanding standardized testing, strengthening proctoring, and restricting access to external tools. Each can play a role in maintaining assessment integrity, but they are not designed to address the core shift described in the preceding section.

Standardized testing provides limited visibility into how a candidate reasons through complex, context-dependent problems. Enhanced proctoring reduces certain forms of unauthorized assistance, but does not resolve the underlying ambiguity introduced by AI-assisted learning. Restricting tool access approximates traditional testing conditions, but does not reflect the environments in which professional reasoning is increasingly exercised.

Taken together, these approaches operate within the existing assessment paradigm—adjusting parameters without fundamentally changing what is being observed or how it is evaluated.

> *The challenge is not preventing access to information, but verifying reasoning in its presence.*

Addressing that gap requires a shift in perspective—from restricting access to information toward designing mechanisms that can reliably observe reasoning in its presence.

SECTION 5

V A Direction, Not a Product

Addressing the gap between scalable assessment and meaningful evaluation of reasoning does not begin with a single system or platform. It begins with a shift in how assessment is structured and governed—specifically, by separating functions that have historically been coupled.

One such boundary is between learning and certification. Training environments can be optimized for guidance and iterative improvement, including AI-assisted tools. Certification environments can be structured to evaluate reasoning independently, using methods designed specifically for verification rather than instruction.

A second boundary concerns identity and performance data. Structuring assessment in a way that minimizes dependence on identity-linked training data can support more neutral and portable forms of certification.

A third element involves how exposure to reasoning scenarios is represented. Rather than relying on static question banks, a candidate's training can be characterized in terms of structured exposure to classes of reasoning challenges—enabling evaluation that verifies whether a candidate can reason consistently across variations of those conceptual dimensions.

> *The objective is not to build a system, but to define a governance architecture within which independent systems can operate.*

These elements describe a set of architectural considerations that can be realized in different ways, across institutions and systems. Framed in this way, the problem becomes one of governance as much as of engineering.

SECTION 6

VI Why This Requires Collaboration

The considerations outlined above point toward a form of assessment that cannot be defined or implemented by a single entity alone. They involve questions of educational design, technical infrastructure, certification authority, and governance —each of which sits within a different institutional domain.

Certification bodies establish standards and maintain public trust. **Academic institutions and educators** contribute domain expertise and pedagogical structure. **Technical participants** bring capabilities that enable new forms of interaction, data handling, and evaluation. **Policy and governance stakeholders** provide oversight related to privacy, data use, and regulatory alignment.

These roles are interdependent but not interchangeable. No single participant can fully define the problem or its solution in isolation.

> *This is inherently a multi-party problem requiring coordinated governance, not centralized control.*

A collaborative model allows for the development of shared specifications and interoperable components, while preserving institutional autonomy. It enables iterative validation, avoids premature standardization, and creates space for consensus to emerge around what aspects of reasoning should be assessed, how they should be represented, and how outcomes should be interpreted.

SECTION 7

VII An Incremental Path Forward

Any evolution in assessment models must proceed carefully. Progress is more likely to occur through a series of incremental steps that allow new approaches to be explored, evaluated, and refined over time.

A first step involves conceptual validation—clarifying definitions, identifying key assumptions, and engaging stakeholders in structured discussion. From there, **limited pilot efforts can be introduced** alongside existing assessment processes, focusing on specific aspects of reasoning-based evaluation. As insights are gathered, they can **inform the development of shared specifications** that define how different components interact within a governed framework.

Throughout this process, transparency is essential. Assumptions, methods, and outcomes should be documented and made available for review. Equally important is continuity with existing systems: established assessment formats embody principles that have been validated over time, and an incremental approach seeks to extend those principles rather than replace them.

> *Progress does not require replacement of existing systems, but structured extension and validation.*

As AI continues to expand the scale and accessibility of learning, the question of how to verify reasoning will become increasingly central. The aim of this work is to contribute to that progress by providing a structured basis for discussion, experimentation, and collaboration.

GymGov.org · waleed.nema@gymgov.org