Paper Presented at the 2026 Artificial Intelligence Workshop: Accelerating Data and Analytics Capabilities for AI Military Operations Research Society (MORS), Carnegie Mellon University

Agentic AI Assessment of Creative Thinking for the U.S. Air Force

Yigal Rosen, PhD[1] and Ilia Rushkin, PhD[2]

[1] Ignis AI: yigal@ignisai.ai [2] Ignis AI: ilia@ignisai.ai

Abstract

This paper proposes psychometrically validated, agentic AI assessment framework for measuring mission-ready creative thinking competencies in the context of talent development and training programs in the United States Air Force (USAF), with particular emphasis on complex and uncertain operational environments. While USAF competency models identify creative thinking as foundational to readiness (Air Force, 2025), prevailing assessment approaches rely largely on self-report instruments or narrowly scoped simulations. Building on the Ignis AI PowerSkillsAssessment™ framework and prior research on large-scale assessment of creative thinking (Rosen, & Ruskin, 2026a; Rosen et al., 2023), this paper introduces an agentic AI–driven, scenario-based method for evaluating creative thinking competencies at scale. Airmen engage with mission-relevant vignettes that require problem reframing, generation of non-obvious alternatives, questioning of assumptions, and adaptive response to new constraints, and explanation of reasoning. Responses are evaluated using a human validated AI-enabled scoring approach grounded in computational psychometrics and learning engineering principles (Rosen, & Rushkin, 2026b) to support reliability, transparency, and fairness. The paper outlines scenario design, rubric construction, and a proposed research design for implementation in USAF training, including pre/post assessment and linkage to performance indicators. The contribution is a practical, scientifically grounded model for assessing creativity competencies at scale, suitable for integration into airmen talent development, mission-readiness analytics, and a demonstration of how AI-enabled assessment can move beyond generic notions of “creative brainstorming" to measure the core competencies required to lead and innovate responsibly in high-stakes environments on the frontlines of national security.

1. Introduction

The accelerating integration of artificial intelligence (AI) into operational, analytic, and decision-making processes is reshaping both the character of modern warfare and the nature of creativity and innovation required to execute missions successfully. In the United States Air Force (USAF), technological superiority has long been a cornerstone of strategic advantage, yet senior leaders increasingly emphasize that mission success depends as much on human judgment, adaptability, and creativity as it does on advanced systems. As operational environments become more complex, contested, and uncertain, the ability of professionals and leaders to inspire others, organize coordinated action, and generate novel responses under pressure has emerged as a defining component of mission readiness.

The USAF competency model explicitly identifies creative thinking as foundational mission-ready capability (Air Force, 2025). Creative thinking is defined as the capacity to develop new insights in novel situations and to question conventional approaches. Importantly, these definitions position creativity not as abstract traits or positional authority, but as enacted capabilities that integrate cognition, communication, ethics, and social coordination in real operational contexts.

Despite the centrality, creative thinking remains difficult to measure rigorously and at scale. Existing approaches within both military and civilian systems rely heavily on self-report surveys, retrospective evaluations, or narrowly scoped simulations. Such methods are limited in their ability to capture how airmen actually diagnose problems, frame tradeoffs, communicate intent, and adapt their actions as situations evolve. From a measurement perspective, these approaches conflate perception with performance and outcomes with underlying capability, reducing their predictive validity in dynamic, high-stakes environments (Rosen, 2015; Rosen et al., 2023).

These limitations have become more consequential with the rapid diffusion of generative AI systems. Outputs that once served as proxies for human creativity—written analyses, strategic plans, or ideation artifacts—can now be partially or fully generated by AI. While such tools can enhance productivity and baseline output quality, they also introduce substantial ambiguity regarding attribution and capability. Recent research demonstrates that generative AI may increase individual creative output while simultaneously reducing collective diversity, a phenomenon associated with convergence toward high-probability solutions and the erosion of differentiated thinking (Doshi & Hauser, 2024; Kleinberg & Raghavan, 2021; Wu et al., 2025). In assessment contexts, this raises a critical challenge: distinguishing underlying human creative judgment from AI-enabled fluency in ways that are fair, scalable, and aligned with real-world high-stakes performance demands.

This paper addresses these challenges by proposing a psychometrically validated, agentic AI assessment framework for measuring mission-ready creative thinking, anchored in the USAF context but designed to generalize across military contexts and talent development systems facing comparable complexity. Building on Ignis AI's Human Power Skills Ontology, advances in creativity measurement under generative AI, and learning engineering principles, the framework operationalizes creative thinking as observable performance rather than inferred traits. Using agentic AI–driven, scenario-based tasks and distribution-aware, competitive scoring, the approach captures how leaders reason, adapt, and differentiate their thinking under evolving constraints (Baker et al., 2022; Craig et al., 2025).

The contribution of this work is threefold. First, it introduces a practical, scientifically grounded model for assessing creative thinking explicitly aligned with USAF definitions of mission readiness (Air Force, 2025). Second, it extends emerging theory on creativity under generative AI into applied assessment contexts, demonstrating how distributional and competitive metrics can mitigate AI-driven homogenization and preserve meaningful signals of human capability (Kleinberg & Raghavan, 2021; Raghavan, 2025). Third, it illustrates how learning engineering can translate insights from cognitive science, psychometrics, and AI research into scalable assessment systems that support decision-making in high-stakes defense and workforce environments.

2. Creative Thinking in the U.S. Air Force and Modern Organizations

Creative thinking plays a complementary and enabling role in effective leadership under such conditions. The Army Talent Attribute Framework broadly defines creative problem solving as the ability to develop and utilize new or novel and useful methods and strategies to accomplish work or achieve goals in both unexpected, unique or infrequent situations and in evolving and new work environments (U.S. Army Research Institute, 2023). The USAF definition emphasizes developing new insights in novel situations and questioning conventional approaches (Air Force, 2025), capabilities that are essential when standard operating procedures prove insufficient or when adversaries exploit predictable patterns. Importantly, creative thinking in this context is not synonymous with unconstrained ideation or artistic expression. Rather, it involves disciplined innovation within constraints, reframing problems to surface non-obvious options, and adapting strategies as operational conditions shift.

These competencies are not unique to military contexts. Civilian organizations operating in sectors such as energy, healthcare, finance, and technology face analogous challenges as they navigate regulatory complexity, technological disruption, competitive pressure, and ethical tradeoffs. Leaders in these domains must similarly integrate data, human judgment, and social coordination across complex systems. As in defense, success

depends less on technical optimization alone than on the ability to align people, interpret uncertainty, and adapt action in real time.

Despite this alignment across sectors, assessment practices in both military and civilian settings have struggled to keep pace with evolving demands (National Academies of Sciences, Engineering, and Medicine, 2024). Self-report instruments capture dispositions or self-perceptions rather than enacted behavior and are vulnerable to social desirability and reference bias. Retrospective evaluations often conflate outcomes with underlying skill, failing to account for context, chance, or structural constraints. Narrow task-based tests may isolate specific cognitive functions, but rarely capture the integrative nature of creative thinking as they unfold in practice (Rosen, 2015; Rosen et al, 2020; Rosen et al., 2023).

Scenario-based assessment offers a promising alternative by situating individuals in contexts that elicit the behaviors of interest. When designed carefully, scenarios can reveal how leaders frame problems, communicate intent, adapt to new constraints, and organize collective action. However, traditional scenario-based assessments often rely on static prompts and manual scoring, limiting scalability and sensitivity to process-level evidence. From a learning engineering perspective, this constrains the ability to iteratively improve assessment design based on data and to link observed behaviors to development pathways (Baker et al., 2022; Craig et al., 2025).

3. Creativity Measurement in the Age of Generative AI

Creativity has long been recognized as a driver of innovation, adaptability, and strategic advantage, yet its measurement has historically been indirect and contested. Traditional approaches have treated creativity either as a stable trait inferred from personality measures or as a property of isolated outputs evaluated against subjective rubrics. Such approaches were already limited in predictive validity and are increasingly inadequate in AI-mediated environments, where output quality alone is no longer a reliable signal of human capability.

Recent empirical work highlights a paradox introduced by generative AI (e.g., Runco, 2023; Vinchon et al., 2023; Wingström et al., 2024). While AI tools can enhance individual productivity and apparent creativity, they may simultaneously reduce collective diversity by encouraging convergence around high-probability ideas. Doshi and Hauser (2024) demonstrate that generative AI can increase individual creative performance while diminishing population-level variation, a dynamic that undermines long-term innovation. Related work characterizes this phenomenon as algorithmic monoculture, in which widespread reliance on similar models and defaults leads to

homogenization of behavior and outcomes (Kleinberg & Raghavan, 2021; Wu et al., 2025).

From a computational perspective, this convergence reflects well-documented properties of generative models. Neural text generation systems tend to collapse toward dominant modes unless explicitly constrained or incentivized to diversify, a pattern observed across language modeling research (Holtzman et al., 2020; Li et al., 2024). Even when prompts request originality, AI-generated outputs often cluster around high-probability regions of the idea space, producing responses that appear novel in isolation but are highly redundant at scale.

In response, emerging theory reconceptualizes creativity as a distributional property rather than a binary attribute or isolated outcome (Acar, 2025; Amabile, & Pratt, 2016). Under this view, creativity is reflected in how ideas are distributed across a conceptual space under shared constraints and incentives (Kleinberg & Raghavan, 2021; Raghavan, 2025). Key dimensions include distinctiveness relative to peers, within-agent variability, sensitivity to contextual changes, and robustness when novelty is explicitly rewarded. Distribution-aware metrics operationalize these dimensions by examining dispersion, divergence, and redundancy across responses rather than relying solely on surface-level quality judgments (Srivastava et al., 2023).

This framing aligns closely with real-world contexts, where creativity is evaluated implicitly through competition for scarce resources, roles, or opportunities. Individuals are rarely assessed on absolute idea quality alone; instead, value is determined by whether ideas offer differentiated advantage relative to alternatives. Crucially, this distributional perspective also provides a principled basis for distinguishing human creative capability from AI-generated or AI-assisted output, which tends to exhibit characteristic clustering patterns even when superficially polished.

The implications for assessment are significant. Quality-only scoring systems—including many rubric-based approaches—are increasingly brittle under generative AI, as they reward fluency and coherence without accounting for convergence or redundancy. Distribution-aware metrics, by contrast, capture properties of ideation and reasoning that remain informative even when surface-level quality is inflated by AI assistance.

This perspective suggests that creative thinking should be evaluated not as a static trait or one-off performance, but as a pattern of behavior across evolving situations. Leaders demonstrate creativity by reframing problems, generating non-obvious options, adapting to new constraints, and sustaining differentiated judgment over time. Capturing these dynamics requires assessment systems that are interactive, comparative, and sensitive to process-level evidence—capabilities that agentic, AI-enabled assessment architectures are uniquely positioned to support.

4. The Ignis AI PowerSkillsAssessment™ Framework

To operationalize creative thinking as mission-ready capabilities, this work builds on the Ignis AI PowerSkillsAssessment™ framework (Rosen, & Rushkin, 2025), a performance-based, AI-enabled system designed to make complex human capabilities visible, measurable, and developable at scale. The framework is grounded in a Human Power Skills Ontology that defines creativity not as latent traits inferred from self-report, but as patterns of observable behavior enacted across realistic, high-stakes contexts.

At its core, the framework reflects a learning engineering approach to assessment design, integrating theory, task construction, scoring, and iteration within a coherent system (Rosen, & Rushkin, 2026; Baker et al., 2022; Craig et al., 2025). Rather than treating assessment as a static instrument, the framework is designed as an adaptive measurement system in which design decisions are continuously informed by empirical evidence and aligned with operational use cases, including selection, development, and readiness analytics.

4.1 Human Power Skills Ontology and Creative Thinking Construct

The Ignis AI Human Power Skills Ontology specifies creative thinking as a multi-dimensional construct composed of interrelated cognitive, social, and ethical capacities. In alignment with the USAF competency model (Air Force, 2025), creative thinking is operationalized as the capacity to develop new insights, question conventional approaches, and adapt solutions within constraints.

These capacities are intentionally defined at a level of abstraction that supports both defense and civilian workforce applications, while remaining anchored in USAF doctrine and operational realities.

Figure 1 below illustrates Ignis AI PowerSkillsPrint™, a structured visualization designed to make human power skills visible, interpretable, and actionable. The central "talent flower" represents an individual's multi-dimensional performance profile across seven human power skills assessed by the Ignis AI PowerSkillsAssessment™: Creative Thinking, Leadership, Communication, Collaboration, Analytical Thinking, Productivity, and AI Fluency. Each petal corresponds to a distinct skill domain, with its relative size and numerical proficiency estimate reflecting performance inferred from scenario-based, agentic AI interactions rather than self-report.

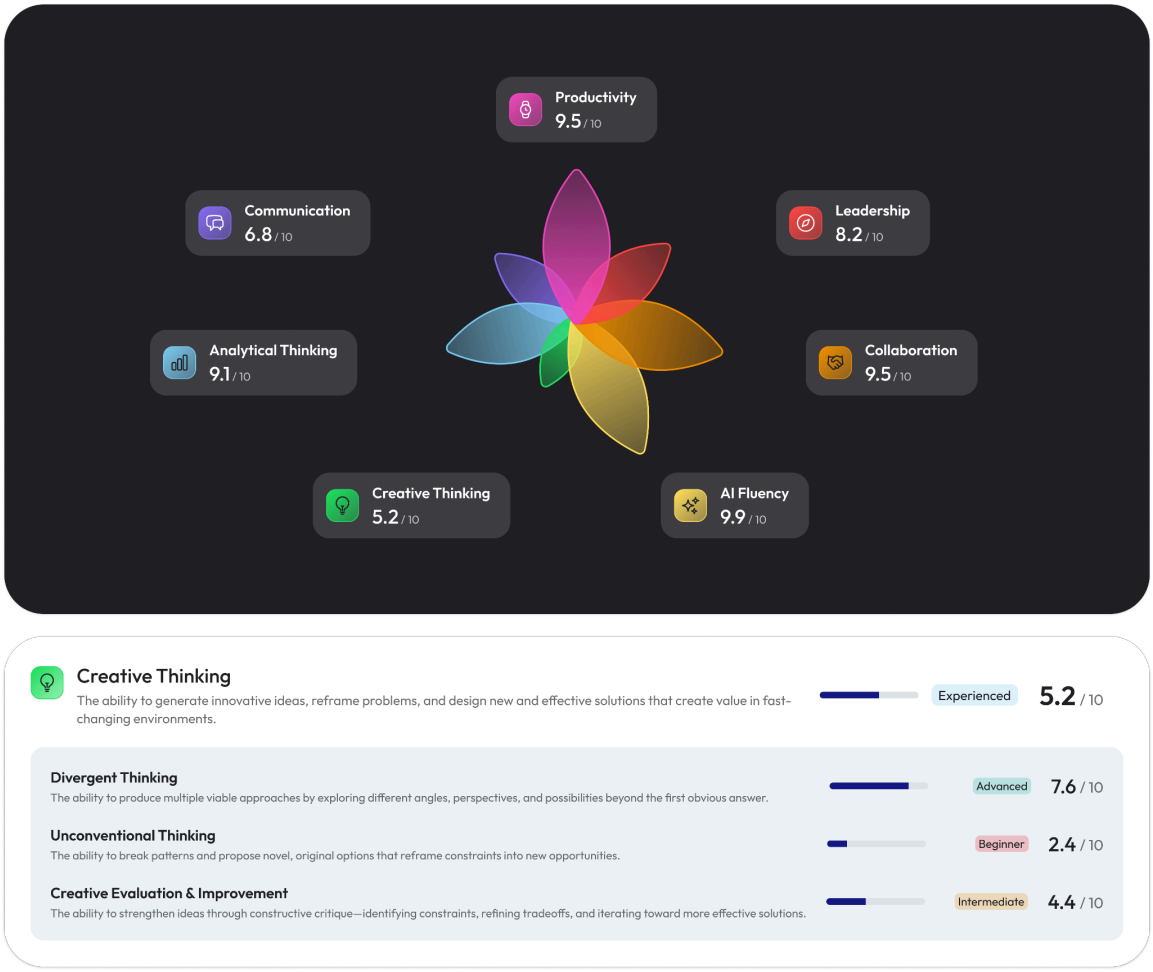

Figure 1. Ignis AI PowerSkillsPrint™: Human Power Skills Talent Flower

Proficiency scores are expressed on a calibrated scale and accompanied by uncertainty-aware performance bands, enabling both developmental insight and responsible interpretation for high-stakes contexts. Surrounding the talent flower, the PowerSkillsPrint™ provides deeper diagnostic views for selected competencies, including Creative Thinking, which are central to USAF mission readiness. These expanded views decompose each competency into theoretically grounded sub-skills aligned with operational definitions and learning science. For Creative Thinking, sub-skills include Divergent Thinking (the ability to generate multiple viable approaches), Unconventional Thinking (the capacity to break patterns and propose original options), and Creative Evaluation and Improvement (the ability to refine ideas through critique, tradeoff analysis, and iteration). These dimensions capture creativity as disciplined innovation under constraints, consistent with USAF requirements for developing new insights and questioning conventional approaches in uncertain environments.

Importantly, the PowerSkillsPrint™ integrates performance-level descriptors (e.g., Beginner, Intermediate, Experienced, Advanced) to translate quantitative estimates into interpretable developmental signals. This design supports formative feedback for individual development while maintaining the rigor required for selection and readiness analytics. By grounding all scores in observable behavior elicited through mission-relevant scenarios, the Ignis AI PowerSkillsPrint™ demonstrates how AI-enabled assessment can move beyond abstract potential toward a concrete, evidence-based representation of mission-ready creative thinking.

4.2 Scenario-Based, Agentic AI Task Design

Assessment within the Ignis AI framework is conducted through scenario-based tasks that simulate mission-relevant decision environments. Scenarios are designed to elicit the behaviors specified in the ontology by placing leaders in situations characterized by ambiguity, competing objectives, resource constraints, and ethical tension. Rather than selecting from predefined options, participants engage in open-ended interaction with agentic AI systems that dynamically respond to their inputs.

This agentic design enables the capture of process-level evidence, including how leaders frame problems, generate options, revise assumptions, and communicate rationale over time. Prior research demonstrates that such human–agent interaction can validly elicit complex skills such as collaboration, communication, and creative problem solving in ways that traditional static tasks cannot (Rosen, 2015; Rosen et al., 2023).

Importantly, scenarios are structured to evolve based on participant responses, introducing new information, constraints, or stakeholder perspectives. This dynamic structure allows assessment of adaptability and creative thinking as situated behavior rather than isolated output, consistent with USAF definition of creativity as enacted capabilities (Air Force, 2025).

4.3 Distribution-Aware and Competitive Scoring

To address the challenges posed by generative AI, the framework employs distribution-aware scoring methods that evaluate responses relative to one another rather than in isolation. Drawing on emerging theory in generative AI and creativity measurement, creativity is treated as a distributional property that emerges under shared constraints and competitive conditions (Kleinberg & Raghavan, 2021; Raghavan, 2025). These conditions are informed by computational work on diversity and mode collapse in language models (Holtzman et al., 2020; Li et al., 2024; Wu et al., 2025) and by distributional measures of generated text diversity (Srivastava et al., 2023). By focusing on patterns of ideation rather than surface-level quality alone, the framework

distinguishes human creative judgment from AI-generated or AI-assisted fluency, which tends to cluster around high-probability solutions even when prompts request originality.

4.4 Human–AI Hybrid Scoring and Fairness

Scoring within the Ignis AI framework is implemented through a human–AI hybrid approach. AI models are used to extract features, model distributions, and generate preliminary scores, while human expert oversight is retained for rubric validation, bias monitoring, and adjudication of edge cases. This design supports reliability and scalability while maintaining transparency and alignment with ethical and fairness considerations.

From a learning engineering perspective, this hybrid architecture enables iterative refinement of scoring models based on empirical evidence, supporting continuous improvement and alignment with operational needs (Baker et al., 2022; Craig et al., 2025). It also mitigates risks associated with over-automation, particularly in high-stakes defense and talent decisions where accountability and explainability are essential.

Together, these design elements establish a measurement system capable of capturing creative thinking as dynamic, context-sensitive performance. The next section outlines the proposed research and pilot design for evaluating this framework within USAF training pipelines, including validation strategies and linkage to performance indicators.

5. Task Design, Agentic AI, and the Science of Measurement

Assessing mission-ready creative thinking competencies requires tasks that reflect how individuals actually operate in real environments rather than simplified testing conditions. Leaders in the USAF and comparable civilian organizations rarely encounter well-structured problems with clear objectives and complete information. Instead, they operate under uncertainty, competing priorities, ethical constraints, and dynamic adversary or market behavior (Air Force, 2025). To capture these realities, the Ignis AI PowerSkillsAssessment™ relies on scenario-based, multi-turn tasks delivered through agentic AI systems that elicit creative thinking as enacted performance.

From a learning engineering perspective, this approach treats assessment as a designed system in which task structure, interaction dynamics, and scoring models are intentionally aligned with the constructs of interest (Baker et al., 2022; Craig et al., 2025). The goal is not to approximate operational complexity perfectly, but to elicit the cognitive, social, and ethical behaviors that underlie creativity in mission-relevant contexts.

5.1 Scenario-Based Task Design

Each assessment scenario is designed to reflect authentic challenges, such as allocating limited resources, responding to unexpected failures, balancing mission objectives with safety and ethics, or communicating intent across hierarchical and functional boundaries. Scenarios unfold over multiple turns, with new information, constraints, or stakeholder perspectives introduced as the participant responds. This structure enables the assessment to capture not only what decisions are made, but how leaders reason, adapt, and communicate as situations evolve.

From an assessment science perspective, scenarios are explicitly aligned with a construct model grounded in the Human Power Skills Ontology and USAF competency definitions (Air Force, 2025). his alignment ensures that observed performance can be interpreted meaningfully rather than treated as an opaque outcome, consistent with prior work on performance-based assessment of complex skills (Rosen, 2015; Rosen et al., 2023).

5.2 Agentic AI as Assessment Orchestrator

A defining feature of the framework is the use of agentic AI to orchestrate assessment interactions. Unlike static assessments, agentic systems dynamically adapt prompts and follow-up questions based on participant responses. When a leader proposes a solution, the system may introduce a resource shock, ethical dilemma, or stakeholder conflict, requiring the participant to reassess tradeoffs, reframe the problem, and communicate revised intent.

This interactional design mirrors how performance unfolds in practice and enables the assessment to probe deeper into reasoning processes rather than capturing surface-level outputs alone. Prior research demonstrates that agent-based and conversational assessment approaches can validly elicit higher-order skills such as creative problem solving when tasks are designed to capture process-level evidence (Rosen, 2015).

Importantly, agentic AI also supports controlled variation across participants. While all participants engage with scenarios drawn from a common template, specific details may differ to reduce rote responding and limit the effectiveness of AI-only generation. This design choice strengthens fairness and mitigates gaming in AI-mediated assessment contexts, where static prompts are increasingly vulnerable to automation.

5.3 Human–AI Hybrid Scoring and Bayesian Proficiency Estimation

Scoring follows a human–AI hybrid model designed to balance scalability, transparency, and accountability. AI models generate preliminary scores, feature representations, and uncertainty estimates based on trained classifiers and probabilistic models. Human

experts oversee rubric calibration, audit scoring behavior, and adjudicate ambiguous cases. This structure aligns with responsible AI principles for high-stakes assessment and mitigates risks associated with fully automated judgment.

Proficiency estimation is conducted using Bayesian modeling, integrating evidence across multiple tasks and interactions. Rather than producing point estimates alone, the system generates posterior distributions that explicitly represent uncertainty. As participants complete additional tasks, uncertainty bands narrow, increasing confidence in proficiency estimates. This approach supports both summative decisions and formative feedback while avoiding overinterpretation of sparse or noisy data, consistent with best practices in performance-based assessment and learning engineering (Rosen et al, 2018; Baker et al., 2022; Craig et al., 2025).

Below we present an evaluation of AI model scoring assessments on a human-labeled dataset, for a subset of skills: this is a direct comparison of the scores produced by the AI model and those produced by a human expert ("labels"). The scores are normalized to lie on the 0-to-1 scale. We measure the mean absolute error of the score, and the Pearson correlation between the AI score and the label score.

Table 1: A direct comparison of the scores produced by the AI model and those produced by a human expert ("labels"), for a subset of skills.

Power Skill	Mean Absolute Error (MAE)	Pearson Correlation with Label	Label standard deviation	AI score standard deviation	Quadratic Weighted Kappa (QWK)	Accuracy	n
AI Fluency	0.082	0.737	0.262	0.231	0.731	0.775	285
Analytical Thinking	0.068	0.760	0.221	0.210	0.758	0.797	286
Collaboration	0.077	0.730	0.239	0.217	0.726	0.782	321
Communication	0.067	0.850	0.273	0.271	0.849	0.800	210
Leadership	0.103	0.718	0.257	0.232	0.712	0.692	120
Overall	**0.077**	**0.774**	**0.256**	**0.238**	**0.772**	**0.778**	**1229**

Both the labels and the AI scores are scaled to be on the 0-to-1 scale. Column "Accuracy" is made possible by the fact that in this implementation the model spectrum (all the possible produced score values) was the same as in the labels (0, 1/3, 2/3, 1), so it could be viewed as a classification model as well as a regression model.

We can also compare the distributions of assessment responses by score, using the human-label scores ("label") and the scores from the AI model ("AI"). These histograms show that the distribution is not heavily skewed, and that difference between the label distribution and the AI score distribution is small.

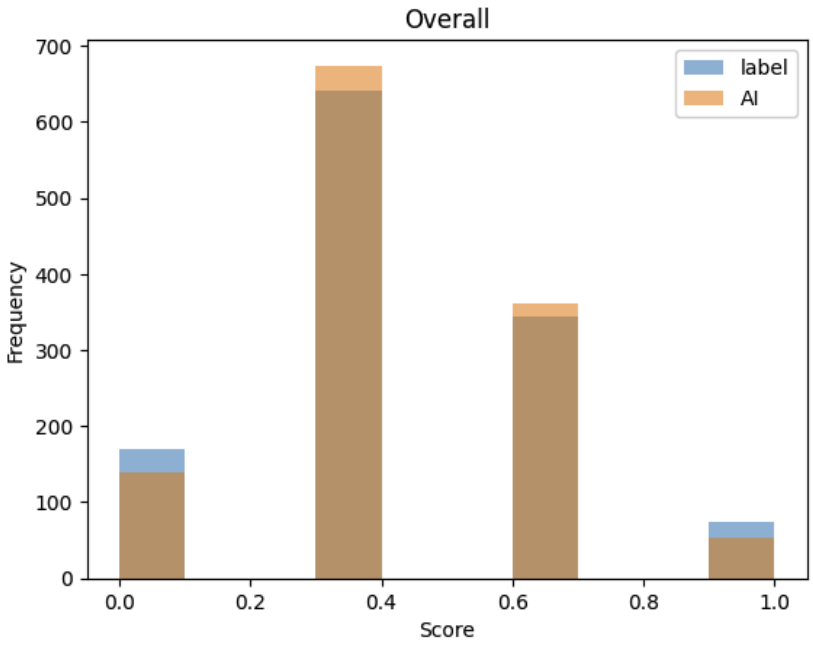

Figure 2. The distributions of assessment responses by score, using the human-label scores ("label") and the scores from the AI model ("AI").

Assessment of the skill *Creative Thinking* can follow the same general framework as other skills; however, key aspects of creativity require specialized measurement approaches. One such approach involves asking the test-taker to generate one or more ideas or solutions (e.g., "Suggest a solution to the following problem in mission-readiness…"). The resulting concepts that meet pre-requisite criteria of appropriateness (e.g., on topic, correctness, feasibility) are classified into a predefined set of topics, which are derived from a human analysis of historical responses (e.g., Rosen et al, 2023). Based on this classification, a score is computed deterministically using measures such as the number of qualitatively distinct topics present in the response and the relative frequency of each topic in the historical data. For instance, the presence of lower-frequency topics is interpreted as a signal of unconventionality and therefore contributes to a higher score for unconventional thinking.

The topic classification itself shows near-perfect agreement with human reviewers when modern reasoning-oriented AI models are used. By conventional accuracy metrics, this component of the system can therefore be considered essentially error-free. The primary source of uncertainty instead arises from the inherent subjectivity involved in constructing the topic taxonomy. Several sources of uncertainty can be identified:

1. **Topic definition and granularity.** There may be genuine disagreements among experts regarding which topics should be included and how many topics are

appropriate. In such cases, no single topic list can be considered objectively "correct."

2. **Dependence on topic frequencies.** Scores for unconventional thinking rely on topic frequency estimates, which may change as the labeled dataset is expanded or updated.

3. **Topic overlap.** Despite efforts by human authors to define topics as distinctly as possible, some responses may plausibly be assigned to more than one topic.

A secondary, though smaller, source of uncertainty concerns the extraction of concepts from the response text. For example, a response such as "I would provide mission-ready training or make changes to the personnel, or something like that, and maybe incentives folks really care about" may be interpreted as containing two or three distinct concepts, depending on whether "mission-ready training" and "personnel" are treated jointly or separately. This variation can affect the resulting score even when "mission-ready training" and "personnel" fall within the same topic. In principle, this source of uncertainty could be eliminated by providing test-takers with a structured interface that allocates a separate input field for each concept.

Taken together, these considerations motivate a systematic analysis of item performance following standard processes in assessment development. Items showing excessive uncertainty can be revised or removed, and the topic taxonomy can be refined.

Another approach to measuring creativity, which we are currently developing (Rosen & Rushkin, 2026a), relies on modeling creativity as novelty and entropy in synthesis of ideas. Given a set of premise statements (abstract ideas, disparate facts, concepts from different domains, etc.), an inference statement is produced by the test subject in response.

This formulation builds on established theories of creativity, including divergent thinking as the expansion of the idea space (Runco & Acar, 2012) and associative recombination as the basis of creative insight (Mednick, 1962). It also aligns with research on ambidextrous leadership, which emphasizes the interplay between exploratory and exploitative processes in innovation (Rosing et al., 2011).

We implement this framework using a geometric representation in embedding space, where premises and responses, as well as their subelements, are represented as high-dimensional vectors. Considering the projection of a response (or a response subelement) onto a cone spanned by the premises (or their subelements), novelty is quantified by the norm of the complement to the projection, while transformation is

quantified using entropy-based measures that capture the breadth and balance of contributions across premises. The two measures—novelty and entropy—are then combined into a single creativity score by means of a mathematical model that allows tunable meta-parameters.

Importantly, this approach does not rely on opaque generative processes or subjective scoring. Instead, it provides a transparent and consistent numerical measure grounded in first principles. While it does not claim to capture all aspects of creativity, it offers a valid and operationalizable definition that can be applied across diverse contexts.

We evaluate the proposed framework using a synthetic AI-generated dataset of activities (sets of premises) and responses of varying levels of creativity. This provides us with a labeled set of activities and responses. Generation is done with a general-purpose generative model with straightforward prompts, so that it serves as an approximation to human judgments of what is more or less creative. Thus, while we compare our model scores to the labels, the labels cannot be regarded as absolute truth, only an approximation to human-perceived creativity. Hence we are evaluating not only literal closeness (mean absolute error), but also correlation: Pearson rho and Kendall tau (a measure of ordinal agreement).

Figure 3. Preliminary results demonstrate that creativity can be measured reliably across assessment activities.

Mean absolute errors of 40 activities are typically low (with both the model outputs and the labels being on 0-to-1 scale): 0.20 average across activities. Kendall and Pearson correlation coefficients are typically high (0.61 and 0.76 averages across activities).

Score distributions are provided to illustrate the relationship: on the x-axis are the labels (converted to integers 0-4). The box-and-whiskers plots show the sets of model outputs for those responses. We see that they go steadily up with X and cover a large portion of the theoretically defined 0-1 range on the y-axis.

Results indicate low mean absolute error and substantial correlation between model scores and intuitive labels, suggesting that the framework aligns with common notions of perceived creativity. Moreover, the variability of metrics across activities provides us with further insights and serves as the beginning of an iterative convergent process: analyzing differences between activities with better and worse agreement metrics (like #33 vs. #13), we will iterate on the activity-generation process in order to further stabilize it and reduce the scatter. Tests on human-labeled data will also be done.

6. Implications for Mission Readiness and Talent Development

The assessment framework described in this paper has direct implications for how mission-ready performance is identified, developed, and sustained within the U.S. Air Force. More broadly, it offers a generalizable model for organizations operating in complex, high-stakes environments where performance effectiveness depends on creative judgment, ethical reasoning, and coordinated action under uncertainty.

Within the USAF, mission readiness depends not only on technical proficiency and procedural compliance, but on leaders' ability to adapt when plans encounter friction, ambiguity, or adversarial disruption. By operationalizing creative thinking as observable performance aligned with Air Force competency definitions (Air Force, 2025), the proposed framework enables a more precise understanding of readiness at the individual and unit levels.

The scenario-based, agentic design supports early identification of strengths and gaps in capability before they manifest in operational settings. Because proficiency estimates are probabilistic and updated as additional evidence is collected, the framework supports longitudinal tracking of development rather than one-time certification. This is particularly relevant for professional military education and command preparation pipelines, where leaders must demonstrate growth across increasingly complex roles.

Importantly, the framework positions assessment as a developmental asset rather than a gatekeeping mechanism. By capturing how leaders reason, reframe problems, and communicate intent, assessment outputs can be used to tailor feedback, coaching, and experiential learning. This aligns with USAF emphasis on continuous development and adaptive performance rather than static qualification.

6.1 Navigating AI-Enabled Decision Environments

As AI systems become embedded in operational planning, data analysis, and logistics, leadership increasingly involves managing human–AI interaction. Leaders must determine when to rely on algorithmic recommendations, when to question them, and how to integrate machine-generated insights with human judgment and values.

The proposed assessment framework explicitly incorporates this reality by distinguishing between human creative judgment and AI-mediated fluency. Distribution-aware scoring mitigates the risk that leaders who rely uncritically on AI-generated outputs are misclassified as highly creative, while those who demonstrate disciplined innovation and ethical reasoning are undervalued. This distinction is essential for maintaining operational resilience in environments where algorithmic monoculture and convergence can undermine adaptability (Kleinberg & Raghavan, 2021; Wu et al., 2025).

6.2 Transferability to Civilian Workforce and Talent Systems

Although anchored in the USAF context, the framework is intentionally designed to generalize to civilian workforce systems facing similar complexity. Leaders in energy, healthcare, finance, technology, and critical infrastructure must navigate regulatory constraints, rapid technological change, and ethical tradeoffs while coordinating action across diverse stakeholders.

In these settings, traditional hiring and development practices are increasingly misaligned with AI-mediated work. Self-report measures and artifact reviews struggle to distinguish underlying capability from AI-enabled output. The distribution-aware, scenario-based approach described here provides a principled alternative, enabling organizations to measure creativity in ways that remain robust under generative AI (Doshi & Hauser, 2024; Raghavan, 2025).

This has direct implications for skills-forward leadership development, and succession planning. By separating AI fluency from creative judgment, organizations can design clearer development pathways and reduce the risk of workforce homogenization.

6.3 Learning Engineering as an Enabler of Scalable Impact

A central contribution of this work is demonstrating how learning engineering can translate theory into operational systems that scale. Rather than treating assessment, training, and analytics as separate functions, the framework integrates them within a coherent design cycle in which data from assessment informs both individual development and system-level improvement (Baker et al., 2022; Craig et al., 2025).

This integration is particularly important in defense contexts, where resources are constrained and the cost of misalignment is high. By embedding assessment within training pipelines and readiness analytics, the framework supports evidence-based decision-making without over-reliance on intuition or retrospective judgment.

6.4 Limitations and Future Directions

While promising, the framework has limitations. Scenario-based assessment necessarily abstracts from real-world complexity, and continued validation is required to strengthen links between assessment performance and operational outcomes. Additionally, governance mechanisms must evolve alongside the technology to ensure transparency, fairness, and appropriate use.

Future research will focus on expanding scenario libraries, refining distribution-aware metrics, and examining long-term predictive validity across diverse operational contexts. Further work is also needed to explore how leaders learn to use AI as a creative scaffold rather than a substitute, a distinction that has important implications for both readiness and workforce resilience (Zhang et al., 2024).

7. Discussion

This paper advances a performance-based approach to measuring mission-ready leadership that responds to three converging forces: the operational demands of modern defense environments, the limitations of traditional assessments, and the growing influence of generative AI on how human capability is expressed and evaluated. By anchoring creative thinking in USAF competency definitions while leveraging advances in learning engineering and AI-enabled assessment, the framework reframes what it means to measure readiness in an era of technological acceleration.

7.1 Reframing Creativity as Enacted Capability

A central contribution of this work is conceptual rather than purely technical.Creative thinking are treated not as latent traits inferred from self-report or reputation, but as enacted capabilities observable through behavior in context. This framing aligns closely with USAF doctrine, which emphasizes inspiring others, organizing coordinated action,

and generating new insights in novel situations (Air Force, 2025). By using scenario-based, multi-turn tasks, the framework captures creative thinking as a dynamic process that unfolds over time. This enables assessment of judgment, adaptability, and reasoning—dimensions that are critical to mission success but difficult to observe through traditional instruments. Prior work in performance-based assessment demonstrates the feasibility and value of this approach for complex skills, including collaboration and creative problem solving (Rosen, 2015; Rosen et al., 2023). The present work extends that foundation into creative thinking assessment under generative AI conditions.

7.2 Implications of Generative AI for Leadership Assessment

Generative AI fundamentally alters the signal environment for leadership and creativity assessment. As AI systems increasingly generate fluent analyses, plans, and recommendations, output quality alone becomes an unreliable indicator of human capability. Without careful design, assessment systems risk conflating AI-enabled fluency with underlying judgment, thereby misclassifying leaders and reinforcing convergence.

This paper contributes to emerging scholarship by operationalizing creativity as a distributional property rather than a one-off outcome. Distribution-aware, competitive scoring addresses the well-documented tendency of generative models to collapse toward dominant modes (Holtzman et al., 2020; Li et al., 2024; Wu et al., 2025) and aligns with recent evidence that AI can enhance individual output while reducing collective diversity (Doshi & Hauser, 2024). By evaluating distinctiveness, variability, and robustness under novelty incentives, the framework preserves meaningful signals of human creative judgment even in AI-mediated contexts (Kleinberg & Raghavan, 2021; Raghavan, 2025).

For the USAF, this distinction is not merely academic. Leaders who uncritically accept algorithmic recommendations may appear effective in routine contexts while failing under adversarial adaptation. Conversely, leaders who demonstrate disciplined creativity—questioning assumptions, reframing problems, and integrating ethical considerations—are better positioned to sustain mission advantage.

7.3 Learning Engineering as a Unifying Framework

Another key contribution is methodological. The assessment framework exemplifies how learning engineering can integrate theory, design, measurement, and iteration within a coherent system (Baker et al., 2022). Rather than optimizing isolated components, the framework treats assessment as part of a broader socio-technical system that includes training pipelines, development pathways, and readiness analytics.

The use of a nested learning engineering cycle enables continuous refinement based on empirical evidence while maintaining alignment with operational goals (Craig et al., 2025). This is particularly important in defense contexts, where assessment systems must balance rigor, scalability, and accountability. The human–AI hybrid scoring architecture further reinforces this balance, ensuring that automation supports rather than replaces expert judgment.

7.4 Limitations and Open Questions

Several limitations warrant discussion. Scenario-based assessments necessarily abstract from real-world complexity, and continued validation is required to strengthen links between assessment performance and operational outcomes. Additionally, while distribution-aware metrics mitigate AI-driven homogenization, they do not eliminate all risks associated with misuse or overreliance on automated systems.

Open questions remain regarding how leaders learn to integrate AI as a creative scaffold rather than a substitute, and how assessment feedback can best support that development. Longitudinal research is needed to examine how creative thinking evolves over time and how assessment-informed interventions influence readiness at scale (Zhang et al., 2024).

8. Conclusion

Anchored in the U.S. Air Force competency definitions (Air Force, 2025), the proposed AI-enabled assessment framework operationalizes mission-ready creative thinking as enacted behavior rather than inferred potential. By combining scenario-based, agentic AI task design with distribution-aware scoring and human–AI hybrid evaluation, the framework captures how individuals reason, adapt, communicate, and exercise judgment in conditions that mirror operational reality. In doing so, it addresses a growing challenge in the generative AI era: distinguishing human creative judgment from AI-enabled fluency in ways that remain fair, scalable, and predictive of real-world performance.

The framework also demonstrates how learning engineering can serve as a unifying methodology for designing assessment systems that integrate theory, measurement, and application (Baker et al., 2022; Craig et al., 2025). Rather than treating assessment as a static instrument, the approach embeds measurement within iterative design cycles that support development, readiness analytics, and organizational learning. This is particularly critical in defense contexts, where accountability, transparency, and adaptability are paramount.

Beyond the USAF, the implications of this work extend to civilian workforce and talent systems confronting similar dynamics. As generative AI becomes embedded in

knowledge work, organizations across sectors face rising risks of homogenization, misattribution of capability, and erosion of differentiated judgment (Doshi & Hauser, 2024; Kleinberg & Raghavan, 2021; Wu et al., 2025). The distribution-aware, performance-based approach articulated here offers a principled path forward for measuring and developing creativity in AI-mediated environments.

In closing, this paper contributes a practical, scientifically grounded model for assessing creative thinking at scale—one that aligns with mission readiness requirements while remaining adaptable to broader workforce challenges. By making human capability visible, measurable, and developable, AI-enabled assessment can strengthen both national security and organizational resilience, ensuring that innovation on the frontlines remains guided by human judgment, values, and purpose.

References

Acar, S. (2025). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal, 37*(2), 181-187.

Air Force (2025). *Air Force Handbook 36-2647, Competency Modeling*. Secretary of the Air Force.

Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior*, *36*, 157–183.

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. Technology, Mind, and Behavior, 3(1).

Craig, S. D., Avancha, K., Malhotra, P., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a nested learning engineering methodology to develop a team dynamic measurement framework for a virtual training environment. In *ICICLE 2024 Conference Proceedings* (pp. 115–132).

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces collective diversity. *Science Advances, 10*(28).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *Proceedings of ICLR 2020*.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22).

Li, J., Zheng, Y., Zhou, D., et al. (2024). On the diversity collapse of large language models. In *Proceedings of NeurIPS 2024*.

National Academies of Sciences, Engineering, and Medicine (2024). *Adult Learning in the Military Context*. Washington, DC: The National Academies Press.

Raghavan, M. (2025). Competition and diversity in generative AI. *arXiv preprint arXiv:2412.08610v2*.

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of a human-to-agent approach. *International Journal of Artificial Intelligence in Education, 25*(3), 380–406.

Rosen, Y., Jaeger, G., Newstadt, M., Bakken, S., Rushkin, I., Dawood, M., & Purifoy, C. (2023). A multidimensional approach for enhancing and measuring creative thinking and cognitive skills. *International Journal of Information and Learning Technology, 40*(4), 334–352.

Rosen, Y., & Rushkin, I. (2025). *Ignis AI PowerSkillsAssessment™: Advances in Science and AI Technology Enable Measurement of Human Power Skills*. Chestnut Hill, MA.

Rosen, Y., & Rushkin, I. (2026a). *Measuring Creativity in the Age of Generative AI: Distinguishing Human and AI-Generated Creative Performance in Hiring and Talent Systems*. Paper Presented at the 2026 BIG.AI@MIT Conference, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

Rosen, Y., & Rushkin, I. (2026b). *From Measurement to Action: A Learning Engineering Approach to AI-Powered Assessment for Human Power Skills Development*. Paper Presented at the 2026 Learning Engineering Research Network Convening, Learning Engineering Institute, Arizona State University, Tempe, AZ.

Rosen, Y., Rushkin, I., Ang, A., Munson, L., Lopez, G., & Tingley, D. (2018). *The effects of adaptive learning in a massive open online course on learners' skill development*. Proceedings of the Fifth ACM Conference on Learning @ Scale. London, UK.

Rosen, Y., Stoeffler, K., & Simmering, V. (2020). Imagine: Design for creative thinking, learning, and assessment in schools. *Journal of Intelligence, 8*(2), 18.

Rosing, K., Frese, M., & Bausch, A. (2011). Explaining the heterogeneity of the leadership-innovation relationship: Ambidextrous leadership. *The Leadership Quarterly, 22*(5), 956-974.

Runco, M. A. (2023). AI can only produce artificial creativity. Journal of Creativity, 33(3).

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75.

Srivastava, A., Li, C., & Hashimoto, T. (2023). Measuring diversity in generated text using topic and distributional metrics. *Transactions of the Association for Computational Linguistics, 11*, 1234–1249.

U.S. Army Research Institute (2023). *Army Talent Attribute Framework – FY24 Annual Update Using a Mixed Methods Research Design*. United States Army Research Institute for the Behavioral and Social Sciences.

Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., & Gaggioli, A. (2023). Artificial intelligence & creativity: A manifesto for collaboration. *The Journal of Creative Behavior, 57*(4), 472–484.

Wingström, R., Hautala, J., & Lundman, R. (2024). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity Research Journal, 36*(2), 177-193.

Wu, F., Black, E., & Chandrasekaran, V. (2025). Generative monoculture in large language models. In *Proceedings of ICLR 2025*.

Zhang, Y., Müller, M., Wang, D., et al. (2024). Beyond the prompt: How humans strategically use AI systems for creative work. In *Proceedings of CHI '24*.