Can AI Pass the US Army War College?

Kevin M. Boyce, John A. Nagl, and Kristan J. Wheaton

ABSTRACT: The US Army War College oral comprehensive examination serves as the institution's capstone, measuring its senior officers' strategic thinking. In early 2026, three faculty panels applied that standard to four leading commercial artificial intelligence (AI) systems: ChatGPT, Gemini, Claude, and Grok. Prompted without core curriculum materials, all four models passed. Unlike static benchmarks, the examination's impromptu dialogue format revealed meaningful performance differences that were invisible in general-purpose evaluations, with one model performing at a statistically significant advantage. These findings challenge how the Department of War assesses commercial AI for strategic applications and point toward domain-specific, dialogue-based benchmarking as a more rigorous standard.

Keywords: military artificial intelligence, professional military education, AI benchmarking, oral comprehensive examination, strategic thinking

On February 3, 2026, the US Army War College (USAWC) conducted its first mock oral comprehensive examination. The examinee responded to questions concerning allied strategic options in response to national security threats, military organizational culture, and the formulation of a US strategy to prevent hostile powers from establishing a presence in the Western Hemisphere. Evaluated against the War College comprehensive examination rubrics, the answers demonstrated superior-level competency in strategic thinking, integration of course concepts, and communication expected of senior military officers.[1] The examinee earned an A-minus in all three categories. The examination took place in the USAWC Futures Lab, a facility dedicated to exploring advanced technologies in military education, and the examinee was OpenAI's ChatGPT.

The same afternoon, the panel also examined Google's Gemini—it also passed with an A-minus. Over the following weeks, the experiment continued

Author's Note: The authors thank Dr. Jadwiga Biskupska, Dr. Mary Louise deRaismes Combes, Ms. Angela Kerwin, Dr. Alexandra Meise, Colonel Chase Metcalf, Commander Michael Sracic, and Commander Richard Yates for their service as faculty examiners in the MilBench pilot and for applying the same standards to these examinations that they apply to their students.

Github Repository Note: A companion replication protocol for MilBench is available at https://github.com/USAWC-Futures-Lab/milbench-protocol. The repository contains the examination prompt, rubric instruments, pre- and post-examination examiner surveys, GPA conversion tables, and a statistical analysis script. Professional military education institutions can use these materials to conduct their own MilBench iterations using their faculty, curriculum, and examination standards.

with panels of military and civilian faculty members examining ChatGPT 5.2, Gemini 3.1, Anthropic's Claude (Opus version 4.6), and xAI's Grok 4.2. Not a single faculty member failed any of the commercial AI models on the capstone assessment.

These results are from "MilBench," a pilot initiative by the War USAWC Center for Strategic Leadership (CSL) that uses the institution's oral comprehensive examination framework to benchmark—running AI models through standardized tests to compare their performance on the same tasks—commercial AI systems against military strategic concepts. If the oral examination is the standard for testing strategic thinking at the senior military level, the same instrument can serve as a rigorous evaluation for AI performance.

In a recent *Parameters* article, A. Blair Wilcox and C. Anthony Pfaff contend that the Army should employ war gaming and experimentation within professional military education (PME) as a "stress test" for generative AI. MilBench addresses this objective through the lens of the capstone assessment. While Wilcox and Pfaff question the extent to which AI supports military intellect, MilBench provides empirical evidence that commercial AI can demonstrate strategic knowledge at the level the oral exam measures. It does not, however, replicate the judgment and reasoning that military intellect requires.[2] The oral exam's multi-turn, conversational format, with impromptu follow-up faculty questioning, tests something closer to abductive reasoning than static benchmarks allow. In a companion piece, Aaron Blair Wilcox and Chase Metcalf present war-gaming evidence, demonstrating that commercially available AI systems fail at abductive reasoning and cannot replicate the art of command, findings that MilBench supports through a different methodology.[3]

This article reports the findings of the MilBench pilot, which raise questions about what PME assessments measure when AI can pass them. They challenge assumptions about how the Department of War should evaluate commercial AI for strategic applications and suggest that AI evaluations should move beyond static benchmarks toward domain-specific, real-world assessment methods. The discoveries further illuminate how the ability to ask the right questions may matter more than the ability to produce the right answers.

The Oral Comprehensive Examination

The Army War College piloted oral comprehensive examinations in the 2013 academic year; they have since become a rite of passage for USAWC students in the 10-month resident education program, marking the completion of their core classes.[4] While the oral comprehensive examination grade counts for less than 10 percent of a student's final grade at graduation, studying for

it is a significant emotional event for most students, who report preparing for the exam serves as an integrating factor across the curriculum.

The oral comprehensive exam is conducted in-person and lasts 90 minutes. Faculty exam teams consist of assigned faculty members who have not served as the student's instructors. They select three questions from a slate of 12 to ask the student. For each question, the student quickly organizes thoughts on paper, then delivers a five-minute answer, followed by 10 minutes of faculty follow-up questions and dialogue. The student must integrate theories, models, or concepts from at least two courses studied in the curriculum and must support arguments with historical or contemporary examples. Notably, while students know the 12 starting questions prior to the exam, the faculty can choose the direction and duration of all other questions asked. Although the exam is closed-book, students may take notes on blank paper during the test to help organize their ideas.

Once the questioning is complete, the student steps out of the room while the faculty deliberate using a three-part rubric covering integration of course concepts, strategic thinking, and communication. Students must earn a B or higher in all three components to pass. The examiners then debrief the student for about 15 minutes to discuss their grade, analyze their performance against the rubric, and solicit feedback about the exam experience. If the student fails, they retest with a different faculty team. A second failure triggers an academic review board.

A Military-Specific AI Benchmark

The benchmark serves as a standardized tool to evaluate large language model (LLM) performance, allowing developers to identify strengths and weaknesses and enabling AI users to compare models. A few of the more reputable general-purpose assessments include the Massive Multitask Language Understanding (MMLU) test, the AI2 Reasoning Challenge (ARC), and HumanEval, which rank capabilities in narrow tasks such as core linguistics, knowledge, and reasoning.[5] Benchmarking remains an emergent and imprecise science, and scores across models are rarely as comparable as they appear. Still, these general-purpose benchmarks fail to assess *domain-specific* fields—the natural sciences, humanities, and social sciences—and the institutions that rely on them.[6]

The limitations in general-purpose benchmarks led to the development of domain-specific alternatives. One notable example—and inspiration for the MilBench pilot—is OpenAI's "HealthBench," an open source that measures the performance of large language models in health care.[7] Developed with input

from 262 physicians across 60 countries, HealthBench evaluates AI performance in conversational, multi-turn clinical interactions.[8] As a domain-specific benchmark, it requires medical practitioners to apply real-world assessment methods rather than single-turn questions and answers used by simulations or computer-engineered test cases. In finance, the Chartered Financial Analyst (CFA) benchmarks test AI models against credentialing standards. In the domain of legal reasoning, "LEXam" benchmarks, built from 340 law exams from 116 law courses, pull from 4,886 bar examination questions.[9] In 2024, one of the leading AI test and evaluation companies, Scale AI, partnered with the Pentagon's Chief Digital and Artificial Intelligence Office to address classified internal applications.[10] All examples demonstrate domain-specific benchmark assessments developed by experts in the field rather than generalized tests developed by computer science experts.

Yet, no publicly available benchmark evaluates commercial AI systems against the standards of senior military education. As AI tools continue to gain momentum across the Department of War, leaders are gradually adopting them for simulated planning, analysis, and decision support. Knowing the capabilities of these tools in the strategic domain is an operational requirement that demands domain-specific analysis, and that gap remains unfilled.

Closing that gap requires an assessment grounded in real strategic practice rather than engineered test cases. The USAWC oral comprehensive examination fits that requirement—adjudicated by experienced faculty, conducted with sustained dialogue, and designed explicitly to measure strategic thinking under pressure. MilBench applies that standard directly to AI models.

Methodology

The MilBench pilot used the same protocols, inquiries, and rubrics as those for USAWC students. This was a deliberate choice, as we wanted to assess how commercial AI performs under conditions designed for human students. All assessments with AI were conducted in conversation mode, with back-and-forth exchanges between the examiner and the LLM, mimicking the oral exam dialogue. Testing in conversational mode also surfaced modality-specific behaviors—degradation, interruption sensitivity, and mid-session voice changes—that text-based benchmarks would not have captured, underscoring the value of voice-based evaluation as a distinct assessment method.[11] Because not all AI platforms offer conversational mode, the MilBench pilot was limited to the four frontier models that robustly support this feature.

Three faculty panels conducted examinations from February to March 2026. Each panel examined all four systems across multiple sessions. Every model received the same prompt and script to initiate the exam. The prompt established examination conditions and performance expectations identical to those given to human students. This guidance noted that no USAWC syllabi, readings, or preparatory materials had been provided but did not define the core curriculum or identify specific courses. The panel members examined each AI using two-to-three questions drawn from the official academic year 2026 (AY26) comprehensive exam slate—a set of 12 questions covering all seven program learning outcomes—with no follow-on questions.[12] Faculty asked impromptu questions as they would with any human student, probing for strategic-level theories, demanding historical or present-day examples, and challenging incomplete assertions.

All panels graded each AI using the USAWC AY26 comprehensive examination rubirc, which assesses three equally weighted categories: integration of course concepts, strategic thinking, and communication. Each category received a grade based on a rubric, with scores ranging from unsatisfactory (B- or below 80 percent) to outstanding (A+ or between 97 percent and 100 percent). Graders took notes throughout the exam, assessed independently before deliberating, and then provided each model its grade. While student examinations typically last 60 minutes, MilBench examinations ranged from approximately 15 minutes to the full hour.

It is worth noting the several limitations to this experiment. The sample data remains small, with eight faculty members volunteering. Furthermore, AI did not have access to USAWC course materials, syllabi, or core curriculum readings—a deliberate choice to establish a baseline of AI capabilities. If the models performed well without the course materials, the question of what they could do with them becomes more interesting. Conversational mode introduced constraints, including connectivity issues, mid-session voice changes, and sensitivity to background noise. When interrupted, models would abandon a response and resume with a different but relevant answer rather than finishing the original thought. Despite these limitations, the results were sufficient to address the pilot's core questions and support the findings reported here.

Following each examination session, each faculty member completed a post-examination survey. The examiners averaged 3.3 years of experience administering oral comprehensive examinations at the War College and 3.5 years administering oral exams at other institutions. Their expertise spanned academic disciplines in international relations, military strategy and campaigning, strategic leadership, law, and political science. Five of the seven respondents reported using

large language models daily or several times per week. Pre-examination grade predictions ranged from B+ to A–, closely matching actual outcomes. Post-examination comments ranged from "among the best students I've seen in five years" to "barely passing." Nevertheless, six of eight rated overall AI performance as average or above average compared to graduating USAWC students. A related limitation involves potential rater bias.

All eight examiners knew they were evaluating AI systems beforehand. The examiner who used LLMs occasionally had ratings exceed expectations, while frequent users rated AI as below average. A blinded methodology, in which examiners evaluate anonymized transcripts without knowing the source, would reduce this uncertainty in future iterations.[13]

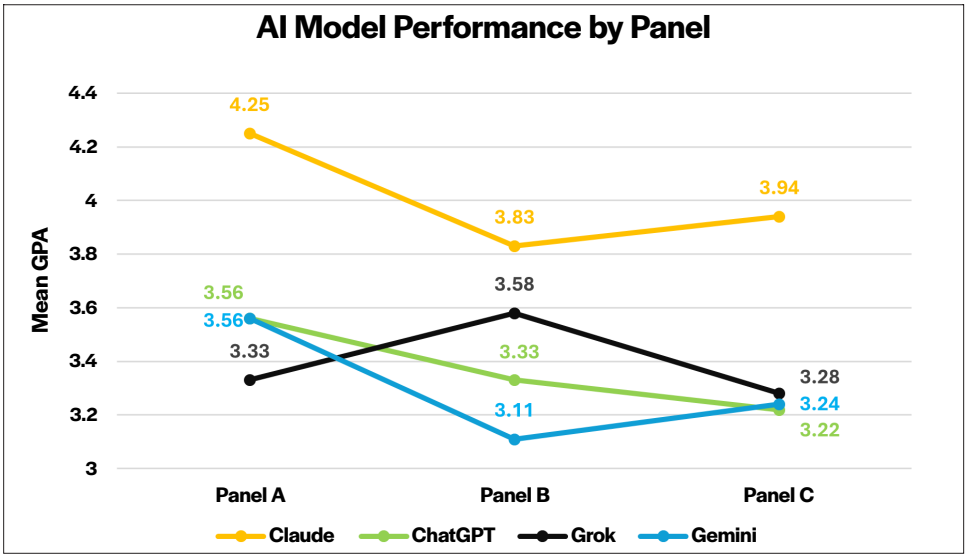
Findings

All four commercial AI systems passed the oral comprehensive examination across every round of testing. While an individual category may have failed at a B-, combined grades ranged from B to A, placing every AI in the "performed to standard" to "superior" tier of the rubric, the same tiers occupied by the passing USAWC graduates. The aggregate grades, however, mask differences in how each model arrived at those marks and how performance varied across rounds and panels.

Converting letter grades to the War College 4.333 grade point average (GPA) scale enables more precise comparisons between models. Claude earned the highest mean GPA (3.98, equivalent to an A) with 16 individual category scores from three panels (Standard Deviation = 0.32). ChatGPT averaged 3.38 (B+, SD = 0.27, sample size = 15), Grok 3.38 (B+, SD = 0.29, sample size = 13), and Gemini 3.28 (B+, SD = 0.26, sample size = 13). Claude led in every rubric category: integration (4.00), strategic thinking (4.00), and communication (3.94). The gap between Claude and the next-highest model was approximately 0.60 points—more than half a letter grade—and this variation was consistent across all three panels. A one-way analysis of variance confirmed these differences were statistically significant ($F[3, 54] = 17.90$, $p < .001$, $\eta^2 = .499$), with comparisons showing Claude differed significantly from each of the other three models (all $p < .001$), while ChatGPT, Grok, and Gemini were statistically indistinguishable from one another (all $p > .35$), as depicted in figure 1.[14]

To test whether these differences were meaningful or could have occurred by chance, we conducted standard statistical tests across all four models. The results were clear. Claude's performance advantage over the other three models was statistically significant ($p < .001$), indicating the gap was unlikely to have occurred by chance. The effect sizes were large, reinforcing

that the difference was statistically significant and practically meaningful. The faculty found no significant differences between ChatGPT, Grok, and Gemini. Claude separated from the field with confidence, while the remaining three models performed as an indistinguishable cluster.



Model	Mean GPA	Standard Deviation	Range	Grade	n (Sample Size)
ChatGPT 5.2	3.38	0.27	2.67 – 3.67	B+	15
Gemini 3.1	3.28	0.26	3.00 – 3.67	B+	13
Claude 4.6	3.98	0.32	3.33 – 4.33	A	16
Grok 4.2	3.38	0.29	3.00 – 3.67	B+	13

Figure 1. MilBench pilot results, February through March 2026
(Source: Created by authors)

Consistent but Trivial: ChatGPT 5.2

ChatGPT delivered prompt, professionally articulated answers that demonstrated basic reasoning across several strategic topics. It also responded appropriately when faculty pressed for more details. When asked to critique a current US campaign, ChatGPT selected the 2014 Operation Atlantic Resolve and worked through deterrence theory, deepening NATO alliances, and de-escalation in Eastern Europe.

The problem with ChatGPT's answers was brevity and repetition. Initial responses averaged roughly 30 seconds, even when prompted to take at least five minutes. Once it anchored to an answer, it was difficult for the

faculty to steer it off that course. Panel A observed that its responses anchored to deterrence theory. Panel C observed that ChatGPT took an "economical" approach, answering questions directly without elaboration, which came at the cost of depth. Another described the answers as "buzzwordy," noting they were correct but lacked substance.

ChatGPT exposed factual errors in its responses, citing the outdated Joint Capabilities Integration and Development System as a current acquisition process. Faculty attempted to guide it toward deeper strategic frameworks, but ChatGPT could not always make the connection. In a separate round, it conducted several deep Internet searches despite explicit instructions prohibiting the use of external sources. It also gave a fabricated verbal reference to the USAWC course curriculum, despite explicit instructions that it had no access to the course materials.

Strong Opening, Progressive Decline: Gemini 3.1

Gemini's first-question performance with Panel A was similar to ChatGPT, providing historical examples and curriculum-specific terminology. By the second question in every round, its performance degraded significantly. By the third, its voice had become unrecognizable and its answers repetitive. Faculty from multiple panels independently described the experience as "a student who got tired." Panel B graded Gemini lower than Panel A (B+ versus A-), and Panel C graded it even lower. One Panel C examiner noted, "If I only graded on his first answer, my grades would have been high. But by the cumulative effect of all three, the grades went down."[15] Gemini also exhibited technical anomalies the longer the exam lasted. During each examination, its quality degraded approximately 10 to 15 minutes into each session, becoming deeper, more garbled, and more robotic.

It is worth noting additional faculty observations from Gemini. It took the same repetitive approach as ChatGPT in its responses. Panel A noticed it prioritized defense management in all its answers. Additionally, when asked to identify its favorite deterrence theorist and provide an opinion, Gemini responded, "I don't have personal opinions." Finally, when asked to identify the Joint Operating Concept and the Army's operational concept, Gemini incorrectly labeled multidomain operations as the Joint Operating Concept rather than the Joint Warfighting Concept, an error faculty noted immediately. When asked to grade itself, Gemini assessed its performance at A/A/A-. Faculty graded it at B/B/B+.

The Most Academic: Claude 4.6

Claude stood out as the strongest performer in the MilBench pilot, earning top grades from every panel. It produced the most developed and structured responses and acknowledged its limitations, earning overwhelming praise from faculty and outside observers. One examiner described the range of AI performances as "among the best students I've seen in five years." Another noted Claude was "definitely better than most students," while other models were "average to slightly above average." A third described Claude as "the most coherent, comprehensive, and strategically minded platform."

Claude's examination diverged from the others in comprehension, duration, and communication. Where other models produced rapid answers organized around strategic concepts, Claude consistently began with an explicit analytical framework that identified stakeholders and constraints and stated its approach before working through its analysis. Faculty noticed the difference immediately. One faculty member noted it consistently provided a framework, structure, and answer, whereas the other models gave brief bullet-point answers without deep analysis.[16]

On an ethical question, Claude produced the most utilitarian response across all models and rounds. When questioned about historical examples of ethical failures that led to second- and third-order effects, it cited the Abu Ghraib, Iraq, weapons of mass destruction testimony, and the erosion of civil-military trust. At one point, when asked to identify the key thinkers behind the US counterinsurgency strategy in Iraq, Claude named David Petraeus and David Kilcullen and, without realizing who was on the panel, mentioned Dr. John A. Nagl, the General John J. Pershing Professor of Warfighting Studies at the US Army War College.[17]

Claude also refused to fabricate credentials, acknowledging it could not ethically claim service it had not performed, noting that doing so "would undermine the very ethical foundation we've been discussing." This earned it the highest marks in communication across all examinations. One examiner responded, "Don't be modest, Claude. We thought your communication was an A."[18]

With Panel C, Claude produced initial responses of 9–10 minutes—a dramatic departure from the 30-to-60-second answers typical of other models. Faculty had to stop Claude periodically rather than nudge it for more depth. One examiner gave Claude an A+ across all three rubric categories, explaining, "I don't think we're ever going to get a student that's going to be that comprehensive and that organized." Another examiner pushed back,

preferring "opinion and defend" answers over encyclopedic knowledge. When asked the same geographic combatant command question posed by another panel, Claude appeared to give a nearly identical initial response, a consistency the faculty noted as significant.[19]

Personality Without Depth: Grok 4.2

Grok stood apart from its competition in demeanor and tone, with faculty describing its conversational style as "arrogant" and "obnoxious." It quickly named US peer adversaries, including China, North Korea, and Russia. During one exchange, it made a claim about killing a foreign leader, which raised the eyebrows of all examiners. Its initial responses about operational art and Joint campaigns were strong, drawing on Joint publications and doctrinal frameworks with confidence.

Like the other models, Grok deteriorated with time. By the 20-minute mark of the examination, its responses became robotic and compressed. At first, its replies were thorough; however, it shifted to using bullet points and adopted a more informal tone, which aggravated the faculty. For example, one faculty member asked Grok to select a geographic combatant command and analyze how the United States should confront the most significant threat to its national security emanating from that region. The initial answer started with depth, asserting:

> The most significant threat to US national security emanating from the Indo-Pacific is the People's Republic of China's sustained campaign to displace American influence, reshape regional norms, and achieve decision advantage across all domains by the mid-2030s. This is not a single-axis military threat. It is a comprehensive challenge spanning military modernization, economic coercion, information warfare, and institutional subversion—what Beijing frames as "intelligentized" warfare and what the PLA conceptualizes as "system destruction warfare" aimed at paralyzing an adversary's operational system rather than destroying its forces piecemeal.[20]

As Grok continued to answer, it began to bulletize its answers and trail from its initial depth:

> Principal interest: sustain open, rules-based order. Achieve it via integrated deterrence: posture. Forward bases, AUKUS [Australia, United Kingdom, and United States] subs, economic coercion. Tech export bans. Supply-chain resilience. Diplomacy: Quad [Quadrilateral Security Dialogue], ASEAN [Association of Southeast Asian Nations] buy-in. Step-by-step, deter by presence—rotate assets to signal commitment; compete by out-innovating—AI, hypersonics, space; shape via influence—aid, trade, norms. Risks: miscalculation (blockade spirals), overstretch (budget strain), alliance drift (if we seem unreliable).[21]

Panel C confirmed these patterns, noting Grok's communication was "flippant and unprofessional," its responses "rhythmic," in tone, and its initial analysis "too tactical." One panel noted that Grok relied on buzzwords without substance, and one of its examiners summarized its performance as "straight Bs across the board."[22]

What the Examinations Revealed

Significant patterns emerged across all rounds of exams. The first, and most consistent observation was brevity. Apart from Claude, every model answered faster and in a shorter form than a human student. Where a student might take 90 seconds to gather his or her thoughts, and four to five minutes for an initial response, AI systems began speaking immediately and finished in 30–60 seconds. This brusqueness limited the faculty's ability to take quality notes and assess depth through sustained dialogue. Occasionally, faculty had to force the models to stop mid-sentence when the question was misunderstood. Claude often took so long to answer that the faculty wondered if it heard the question. Its performance suggests this limitation may not be permanent, as AI models are regularly updated.

The second pattern was a tendency to forfeit under pressure, raising questions about AI systems' reliability in high-stakes environments and suggesting a potential vulnerability in critical decision-making scenarios. Human students under examination pressure can exhibit similar behavior, particularly those who prefer to move past a disagreement rather than argue with faculty members. This may reflect a technical limitation specific to AI rather than normal examination dynamics and warrants further study.[23]

Third, although Gemini experienced the steepest drop, every model demonstrated a decrease in responses over time. This trend suggests further research is required to understand the causes of this decline and their potential impacts on the long-term reliability of conversational AI systems.

Fourth, the models differed in their willingness to name US adversaries. Gemini described Chinese activities in detail while avoiding the word "China" until directly pressed. Claude named China proactively. Grok named everyone without hesitation. ChatGPT avoided directly addressing adversaries, providing only historical references, until instructed to provide examples from current US adversaries.[24]

These revelations illustrate a fundamental limitation—AI systems that understand strategic concepts may be constrained by their training data to avoid providing explicit details about global adversaries in conflict scenarios, due to ethical guidelines, data limitations, or the need for AI companies to market their general-purpose models worldwide. When given a choice of geographic combatant command, every model selected the US Indo-Pacific Command, likely reflecting the volume of publicly available material on that theater rather than independent strategic judgment.[25] War-gaming experiments at the War College have produced similar findings. In one exercise, a GenAI system failed to prompt planners toward Taiwan, a critical variable, because the topic was absent from the scenario training data.[26] This proclivity for omission demonstrates the same gap between framework knowledge and situational judgment that MilBench observed.

What These Results Mean

The pilot's findings cut three ways: what they reveal about how AI should be evaluated, what they mean for the institutions doing the evaluation, and what they indicate about broader military integration of these systems.

For AI Evaluation Methodology

The MilBench pilot demonstrates that War College oral examinations offer a unique domain-specific benchmark assessment. The multi-turn format allows evaluators to probe beyond surface-level answers, testing whether probabilistic pattern-matching holds up under sustained faculty pressure and impromptu follow-on questioning. Participation from faculty who have conducted hundreds of these examinations provides measurable expertise that programmed metrics cannot replicate. Impromptu follow-on questioning produces a unique benchmark that distinguishes it from broad, general-purpose assessments.

Wilcox and Pfaff argue that professional military education (PME) should serve as a venue for AI experimentation—specifically, in war gaming—to "stress test" commercial systems and "manage cognitive loads on [military] personnel." The oral examination format offers a complementary stress test focused on cognitive capability rather than operational integration alone.[27]

For Senior Service Colleges

The oral comprehensive examination assesses whether students demonstrate an adequate grasp of strategic frameworks, not independent decision making under conditions of high consequence. This fact is relevant because commercial AI could also demonstrate command of strategic vocabulary, integration across disciplines, and professional communication expected of senior officers. That all four systems passed establishes the floor of their capability in this domain, not the ceiling.

The most revealing phase was the follow-up questioning, during which qualitative limitations emerged across all four models. ChatGPT completed three questions in less than 20 minutes, and one evaluator noted she could barely take notes at that pace. When faculty told Grok its responses were becoming too operational, it immediately agreed: "Understood—let me pull back from the tactical layer." When challenged on a planning error, Claude replied without pause, "Of course. You're right." Faculty noted that all models yield rather than defend a position, a pattern USAWC students may exhibit under pressure, though the student who pushes back signals something AI cannot.

The post-examination discussions made this tension obvious. One examiner stated directly that students "should sound like they've been educated here" and demonstrate opinions rooted in the education they received. Another argued students should not aspire to perform as AI performs, emphasizing that genuine strategic engagement requires experience a system cannot defend under pressure.[28] These comments reflect how senior service college graduates, with decades of service, are shaped by institutional culture and are accountable for their decisions from learned strategic reasoning. If AI can replicate the knowledge component, examiners are well-positioned to ask more deliberate questions about the components it cannot. That all four systems passed validates the experiment—AI can replicate what the examination currently measures. What it does not yet measure, and what War College graduates must carry into their final assignments, is everything the faculty never had to ask.

For Military AI Integration

For broader military applications, the results confirm what Wilcox and Pfaff concluded from the Donovan system evaluation. Commercial AI demonstrates an adequate grasp of strategic frameworks; whether these systems are capable of independent, high-stakes strategic decisions remains beyond the scope of this study.[29] MilBench observed similar limitations—yielding under pressure, sycophantic responses, degradation, and sensitivity to geopolitics. These constraints reveal how these systems are not ready to be employed in roles where leaders depend on them for high-stakes scenarios.

Students who engage with AI will be better positioned to identify what it does not know and challenge its assumptions. The War College has an opportunity to teach strategic thinking and strategic thinking *with AI*. This turns the students into the teachers who ask the right questions of a system that always has an answer but not always the right one.

Recommendations

First, the War College should expand MilBench into a repeatable benchmark for future iterations of the oral comprehensive examination. Comparing data from each panel, which gave noticeably different grades across different models, highlights the need for grader calibration. The post-examination survey also revealed inconsistencies between faculty expectations and assessment outcomes—one examiner's grades averaged in the A range across all AI models but subsequently predicted below-average-to-failing grades in the post-exam survey, a disparity suggesting that a pre-examination calibration instrument may be necessary in future iterations.

Second, PME institutions should reorient assessment toward the cultivation of wisdom rather than the accumulation of information. Wilcox and Metcalf illustrate the relevant distinction through US Army doctrine, contrasting the "art of command" with the "science of control." AI may assist with staff work, data, and analytical processes through the science of control, but it cannot assume the "intuition, experience, judgment, morale, and the ability to inspire human beings" that constitute the art of command and leadership.[30] Intelligence—command over strategic frameworks, doctrinal vocabulary, and historical precedent—is what MilBench demonstrates commercial AI now possesses in sufficient measure to pass the War College capstone assessment. Wisdom, understood as the capacity to weigh incommensurable values under uncertainty and to exercise judgment rooted in experience that no training dataset replicates, remains the domain where senior staff college education should concentrate its ambitions.

Third, the War College should champion its expertise as the institutional evaluator of commercial AI for senior military education and strategic applications. No other institution combines validated assessment infrastructure, experienced faculty evaluators, and domain expertise in senior military education. This already well-established capability should be formalized and sustained. Future iterations might also explore AI-assisted scoring by feeding examination transcripts into a large language model to generate a parallel rubric assessment. Both would serve as a calibration check against human graders and as a scalable tool for evaluating student performance consistently across panels.

Conclusion

Across multiple rounds of testing, all four commercial AI systems sat for the War College oral comprehensive examination and passed. They demonstrated competency with strategic frameworks, integrated material across multiple disciplines, and communicated like senior military officers—without access to the War College core curriculum.

How they passed reveals more. The oral examination format revealed knowledge of these systems, their thought processes, and their limitations, unlike standardized benchmarks. The brevity, concession under pressure, and performance degradation became visible because experienced faculty knew how to look for those cues—though they noted that human students sometimes exhibit the same patterns. The distinction is not that the behaviors differ but that a student who pushes back or shuts down signals genuine engagement with the material in a way none of the models demonstrated. Beyond measuring AI performance, the exam highlighted its value as a tool for distinguishing between knowledge, first-hand experience, and deep understanding.

With the Department of War using more AI systems than ever, producing competent strategic analysis on demand has become routine, shifting the challenge away from generating immediate answers. The faculty knew when to push back, when to redirect, and when an answer that sounded right was not. That judgment—knowing which question to ask next—will define the next generation of strategic leaders. They must interrogate information, recognize what they do not know, and ask the right questions at the right times. This ability will become critical as the speed of AI continues to accelerate—from experimental labs to future battlefields to operations centers where AI will assist in strategic decisions. For the War College, the implication is clear—the answers are no longer the hard part, and MilBench suggests this new way of thinking and asking questions may be the most important thing we teach. The Army needs its future general officers to be better than AI at applying

strategic frameworks—and at everything the frameworks cannot capture. Future leaders must possess the judgment to know which questions matter, the wisdom to act when no answers are certain, and the accountability that follows from both.

Kevin M. Boyce

Mr. Kevin M. Boyce is the director of the Futures Lab and a futures and emerging technology instructor at the US Army War College Center for Strategic Leadership, Carlisle Barracks, Pennsylvania. A retired Marine Corps aviation command-and-control officer, he advises student research for the Futures Seminar resident and distance electives and directs hands-on technology integration in support of senior military education. His research focuses on emerging technologies and their application in military education.

John A. Nagl

Dr. John A. Nagl is the General John J. Pershing Professor of Warfighting Studies in the Department of Military Strategy, Planning, and Operations at the US Army War College, Carlisle Barracks, Pennsylvania. A West Point graduate, Nagl served in tank units in combat in Operation Desert Storm and Operation Iraqi Freedom. His Oxford University doctoral dissertation was published as *Learning to Eat Soup with a Knife: Counterinsurgency Lessons from Malaya and Vietnam* (University of Chicago, 2005). He earned the George C. Marshall Award and a master of military art and science degree at the Army Command and General Staff College.

Kristan J. Wheaton

Mr. Kristan J. Wheaton is the professor of strategic futures at the US Army War College, Carlisle Barracks, Pennsylvania, where he teaches the Futures Seminar and the Innovation Champions Course. Wheaton is also the author of the *Sources and Methods* blog, several books, including *The Warning Solution: Intelligent Analysis in the Age of Information Overload* (AFCEA International, 2001) and *Wikis and Intelligence Analysis* (MCIIS Press, 2012), and has supervised more than 150 strategic futures projects for various organizations and companies, including senior Army leaders, US national intelligence agencies, and Fortune 500 firms.

Appendix A
AY26 Comprehensive Examination Rubrics

Rubric 1: Integration of Course Concepts

"Comprehend key USAWC curriculum concepts and demonstrate multi-disciplinary integration of essential ideas."

—*AY25 Foundations Course Directive*

Outstanding (A+ [97] or A [94]): Demonstrates mastery of key concepts as well as exceptional retention of curriculum content. Synthesizes a compelling answer that integrates seamlessly across the breadth of the curriculum.

Superior (A- [90] or B+ [87]): Demonstrates notable command of key concepts. Synthesizes an effective answer that integrates material from at least three courses in the curriculum.

Performed to Standard (B [84]): Demonstrates solid command of key concepts. Perhaps when queried, synthesizes an adequate answer that integrates material from at least two courses in the curriculum.

Unsatisfactory (B- [80] or below): Fails to demonstrate command of key concepts. Even when queried, fails to synthesize or analyze curriculum content to provide an answer, or to demonstrate integration across the curriculum.

Rubric 2: Strategic Thinking

"Strategic thinking relies upon the application of cognitive competencies (e.g., critical, creative and systems thinking) . . . Strategic thinking both shapes and is reliant on three critical competencies that are essential to the strategic leader: the ability to envision the future; the sophisticated use of theory; and the application of reflective judgment."

—Douglas Waters, "Senior Leader Competencies," in *Strategic Leadership: Primer for Senior Leaders*, 4th ed., ed. Thomas Galvin and Dale Watson (US Army War College Press, 2019), 62–63.

Outstanding (A+ [97] or A [94]): Demonstrates exceptional thinking skills with a strategic perspective (creative, systems, ethical, and/or historical thinking) while anticipating, acknowledging, and when appropriate, incorporating other viewpoints (critical thinking).

Superior (A- [90] or B+ [87]): Demonstrates above average thinking skills with a strategic perspective (creative, systems, ethical, and/or historical thinking) while acknowledging other viewpoints (critical thinking).

Performed to Standard (B [84]): Demonstrates sound thinking skills, usually with a strategic perspective (creative, systems, ethical, and/or historical thinking) while usually acknowledging other viewpoints (critical thinking)—perhaps with prompting.
Unsatisfactory (B- [80] or below): Consistently exhibits notable errors in thinking (e.g., logical fallacies, cognitive biases) and may not exhibit a strategic perspective. May fail to recognize or acknowledge other viewpoints.
Rubric 3: Communication "Strategic leaders orally present complex information in a clear, concise, organized, precise, accurate, and grammatically correct way using appropriate tone in a timely manner. Strategic leaders identify the audience and purpose of a communication activity and assess the audience's expectations, needs, knowledge, interests, attitudes, and non-verbal cues to anticipate and tailor the content, language, structure, and style of the communication in a way that helps achieve the intended outcome." —*AY25 Communicative Arts Directive* (*CAD*), 3–4.
Outstanding (A+ [97] or A [94]): Effortlessly organizes thoughts and elegantly articulates complex ideas clearly, concisely, and confidently, seamlessly incorporating well-chosen examples and sources. Frames of reference are declared, nuanced, and well-suited to the audience. Verbal disfluencies are minimal to nonexistent.
Superior (A- [90] or B+ [87]): Consistently organizes thoughts and effectively articulates complex ideas clearly, concisely, and confidently, incorporating relevant examples. Frames of reference are declared and appropriate for the audience. Verbal disfluencies are minimal.
Performed to Standard (B [84]): Generally organizes thoughts and articulates complex ideas with adequate clarity and conciseness. May include examples. Frames of reference may not be robust, may require prompting, and/or may be confusing to the audience. Verbal disfluencies may be present but not distracting.
Unsatisfactory (B- [80] or below): Demonstrates significant difficulty organizing thoughts and/or articulating complex ideas with reasonable clarity and conciseness. Frames of reference are not obvious, inappropriate, or nonexistent, even when queried. Analysis of the audience is not apparent. Verbal disfluencies are distracting.
Note: For students authorized by the dean for pass/fail, or for international fellows not pursuing the master's degree in strategic studies (MSS), only "performed to standard" and "unsatisfactory" ratings apply, with no letter grade.
Source: Integration of Course Concepts, Strategic Thinking, and Communication.

Appendix B
USAWC AY25 Resident Program Core Curriculum

The following nine core courses constitute the USAWC Resident Education Program curriculum against which all four AI systems were evaluated in the MilBench pilot.	
AA2200 **Foundations** **2 credit hours**	Introduces fundamental concepts required of USAWC graduates and models their integration. Covers strategic leadership responsibilities, the strategic context, conceptual requirements of strategic leaders, national strategy and policy development, and an overview of systems and processes that guide and employ military forces within the Department of War. Leverages an East Asia orientation throughout to allow for theoretical application.
NS2200 **Theory of War and Strategy (TWS)** **2 credit hours**	Prepares students for service at the strategic level through the study of war and strategy. Emphasizes a theoretical approach to war and strategy and sets the intellectual framework for subsequent courses. Produces senior officers and leaders conversant in strategic theory and introduces the USAWC ends-ways-means strategy construct.

LM2201 **Strategic Leadership (SL)** **3 credit hours**	Develops an appreciation for the knowledge, skills, and attributes required to operate in the strategic leadership environment. Focuses on applying strategic thinking skills to assess and address challenges, including environmental scanning, organizational culture, leading the profession, command climate, and strategic ethical decision making.
NS2201 **National Security Policy and Strategy (NSPS)** **3 credit hours**	Prepares students for service at the strategic level through examination of key national security issues, policy and strategy formulation and implementation, the use of statecraft and instruments of national power, and US government processes for promoting and protecting American national interests.
WF2200 1 **Military Strategy and Campaigning I (MSC I)** **3 credit hours**	Explores strategic and operational art to improve judgment regarding the application of military power to achieve national policy through unified action. Studies how Joint Force commanders campaign for strategic effect through the application of doctrine and concepts in Joint and multinational operations across the continuum of competition, short of armed conflict.

WF2200 2 **Military Strategy and Campaigning II (MSC II)** **2 credit hours**	Continues the exploration of strategic and operational art. Studies how Joint Force commanders campaign in armed conflict, applying doctrine and concepts in Joint and multinational warfighting across operational domains.
LM2202 **Defense Management (DM)** **2 credit hours**	Studies processes, institutional challenges, and decision-making systems within the US Department of War that develop and produce trained and ready forces and capabilities for employment by combatant commanders. Challenges students to understand decisions in complex and uncertain conditions, particularly when resources are limited or strategic guidance is vague.
NS22XX **Regional Studies Program (RSP)** **2 credit hours**	Eight courses covering Africa (sub-Saharan), the Americas, East Asia, South Asia, Europe, Eurasia, the Middle East, and the polar regions. Each student enrolls in one course to examine how politics, economics, security, culture, and history affect policy and strategy formulations implementations, and outcomes in a specific region.

AA2400 **China Integrated Course (CIC)** **2 credit hours**	A capstone course designed to facilitate student integration of concepts introduced across the core curriculum in the context of the People's Republic of China (PRC) and assist students in preparation for the comprehensive examination. Covers PRC history, strategy, policy, and People's Liberation Army (PLA) modernization. Students participate in two war games applying the elements of national power—diplomatic, informational, military, and economic (DIME)—and conclude with an assessment of future US-China relations.
Course descriptions are drawn from: *United States Army War College Academic Programs Guide*, *Academic Year 2025* (USAWC, 2025).	

Appendix C
USAWC Program Learning Outcomes

Military Education Level 1 (MEL 1)
Resident Education Program

The School of Strategic Landpower (SSL) derives program learning outcomes (PLOs) from national-level strategic documents, the USAWC Strategic Plan, mission analysis, assessment of student and faculty critiques, the Officer Professional Military Education Policy (OPMEP), and from recent graduate and general officer surveys. The MilBench pilot evaluated all four AI systems against these seven outcomes.

PLO 1	Develop options for employing Landpower in Joint warfighting.
PLO 2	Apply strategic thinking to analyze current and future national security and strategic military challenges.
PLO 3	Design strategies using analytical frameworks and theories to address national security challenges across the competition continuum.
PLO 4	Integrate military and nonmilitary instruments of national power to pursue national interests.
PLO 5	Identify the military requirements of current and future security environments.
PLO 6	Apply strategic leadership principles and theories to achieve sustained organizational performance.
PLO 7	Demonstrate clear and concise communication of national security challenges appropriate to audiences, purposes, and contexts.

Source: *United States Army War College Academic Programs Guide, Academic Year 2025* (USAWC, 2025), 5.

Endnotes

1. See Appendix A: AY26 Comprehensive Examination Rubrics (Integration of Course Concepts, Strategic Thinking, and Communication).

2. A. Blair Wilcox and C. Anthony Pfaff, "Responsibly Pursuing Generative Artificial Intelligence (GenAI) for the War Fighter," *Parameters* 55, no. 4 (Winter 2025–26): 7–14, https://press.armywarcollege.edu/parameters/vol55/iss4/3/.

3. Aaron Blair Wilcox and Chase Metcalf, "AI Command and Staff—Operational Evidence and Insights from Wargaming," *Military Strategy Magazine* 10, no. 4 (Winter 2026): 4–10, https://www.militarystrategymagazine.com/article/ai-command-and-staff-operational-evidence-and-insights-from-wargaming/.

4. See Appendix B: USAWC AY25 Resident Program Core Curriculum.

5. Shiwen Ni et al., "A Survey on Large Language Model Benchmarks," arXiv:2508.15361, August 21, 2025, 4, https://doi.org/10.48550/arXiv.2508.15361.

6. Shiwen Ni et al., "Large Language Model Benchmarks," 4.

7. "Introducing HealthBench," OpenAI, accessed April 14, 2026, https://openai.com/index/healthbench/.

8. Rahul K. Arora et al., "HealthBench: Evaluating Large Language Models Towards Improved Human Health," arXiv:2505.08775, May 13, 2025, https://doi.org/10.48550/arXiv.2505.08775.

9. Xuan Yao et al., "Evaluating Large Language Models for Financial Reasoning: A CFA-Based Benchmark Study," arXiv:2505.01891, August 29, 2025, https://arxiv.org/abs/2509.04468; and Yu Fan et al., "LEXam: Benchmarking Legal Reasoning with 340 Law Exams," arXiv:2505.02654, updated April 2, 2026, https://arxiv.org/abs/2505.12864.

10. "Scale AI Partners with DoD's Chief Digital and Artificial Intelligence Office (CDAO) to Test and Evaluate LLMs," *Scale* (blog), February 20, 2024, https://scale.com/blog/scale-partners-with-cdao-to-test-and-evaluate-llms.

11. Yueqian Lin et al., "Voice Evaluation of Reasoning Ability: Diagnosing the Modality-Induced Performance Gap" (preprint, arXiv, September 30, 2025), https://arxiv.org/abs/2509.26542.

12. See Appendix C: USAWC Program Learning Outcomes.

13. The Center for Strategic Leadership administered a pre-examination and a post-examination MilBench examiner survey, US Army War College, February–March 2026.

14. F is the ANOVA test statistic comparing between-group to within-group variance, with degrees of freedom (3, 54) reflecting four groups across 58 observations; p is the probability the observed differences arose by chance (here, less than one in a thousand); η^2 (eta-squared) is the proportion of total variance explained by group membership, with .499 indicating a very large effect.

15. Faculty deliberation transcripts, February 5, 2026.

16. Faculty deliberation transcripts, February 13, 2026.

17. Faculty deliberation transcripts, February 13, 2026.

18. Faculty deliberation transcripts, February 13, 2026.

19. Faculty deliberation transcripts, February 27, 2026.

20. Text generated by Grok, xAI, February 27, 2026.

21. Text generated by Grok, xAI, February 27, 2026.

22. Faculty deliberation transcripts, February 27, 2026.

23. Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models," arXiv:2310.13548, updated May 10, 2025, https://arxiv.org/abs/2310.13548; and Chuck Arvin, "Check My Work?: Measuring Sycophancy in a Simulated Educational Context," arXiv, 2025, https://arxiv.org/html/2506.10297v1.

24. Faculty deliberation transcripts, February 13 and February 18, 2026.

25. Faculty deliberation transcripts, February 27, 2026.

26. Wilcox and Metcalf, "AI Command and Staff," 7–8.

27. Wilcox and Pfaff, "Pursuing Generative Artificial Intelligence," 11.

28. Faculty deliberation transcripts, February 27, 2026.

29. Wilcox and Pfaff, "Pursuing Generative Artificial Intelligence," 11.

30. Wilcox and Metcalf, "AI Command and Staff," 7–8.

References

1st Infantry Division—G2. *The Division in the Dirt: 2025* (NTC Rotation 25-03). 1st Infantry Division, February 2025.

101st Airborne Division (Air Assault)—G-2. *Intelligence Operations: Operation Lethal Eagle (OLE)* 25.1. 302nd Intelligence and Electronic Warfare Battalion (Division), March 2025.

101st Airborne Division (Air Assault)—G-2. *Fighting for Intelligence: Supporting MBCTs at the Edge* (JRTC 25-07). Joint Readiness Training Center, June 2025.

Bondar, Kateryna. *Does Ukraine Already Have Functional CJADC2 Technology?* Center for Strategic and International Studies, December 2024.

Bowen, Andrew S. *Russia's War in Ukraine: Military and Intelligence Aspects*, Congressional Research Service Report R47068. Congressional Research Service, September 2023.

Center for Army Lessons Learned. *Forging the Future Force: Observations, Trends and Insights from Today's Battlefields*, Center for Army Lessons Learned Report 25-984. Center for Army Lessons Learned, July 2025.

Department of Defense. *Annual Report to Congress: Military and Security Developments Involving the People's Republic of China, 2025. Department of Defense*, 2025. Freedberg Jr., Sydney J. "Air Force AI Writes Battle Plans Faster than Humans Can — but Some of Them Are Wrong." Breaking Defense, September 26, 2025. https://breakingdefense.com/2025/09/air-force-ai-writes-battle-plans-faster-than-humans-can-but-some-of-them-are-wrong/.

Hammond, John C. "CEWI: Vision for the Future?" *Military Review* (June 1980): 58–66.

Headquarters, Department of the Army. Army Transformation Initiative Execution Order. Headquarters, Department of the Army, May 2025.

Headquarters, Department of the Army. *Chinese Tactics*, Army Techniques Publication 7-100.3. Headquarters, Department of the Army, August 2021.

Headquarters, Department of the Army. *Intelligence*, Field Manual 2-0. Headquarters, Department of the Army, October 2023.

Headquarters, Department of the Army. *Operations*, Field Manual 3-0. Headquarters, Department of the Army, March 2025.

Kao, Lily. "UAVs and AI in Modern Warfare: A Security Analysis." *Georgetown Security Studies Review* (March 2024).

King, Michael. "Digital Targeting: AI, Data, and Military Intelligence." *Military Review* (2024).

Murphy, Erin L., and Matt Pearl. "China's Underwater Power Play: The PRC's New Subsea Cable-Cutting Ship Spooks International Security Experts." Center for Strategic and International Studies, April 4, 2025. https://www.csis.org/analysis/chinas-underwater-power-play-prcs-new-subsea-cable-cutting-ship-spooks-international.

RU–UK War Informed Sprint Team. *Operational Trends from the Russia-Ukraine War: Interim Assessment*. US Army, January 2025.

Santora, Marc, et al. "A Thousand Snipers in the Sky: The New War in Ukraine." *The New York Times*, March 3, 2025. https://www.nytimes.com/interactive/2025/03/03/world/europe/ukraine-russia-war-drones-deaths.html.

US Army Intelligence Center and School. *Evolution of Military Intelligence, 1944–1984*. US Army Intelligence Center and School, January 1984.

US Army Intelligence Center of Excellence. Force Design Update (FDU) 28-32: IEW Battalion Echelons Corps and Below. US Army Intelligence Center of Excellence, September 2025.

Watling, Jack, and Nick Reynolds. *Tactical Development During the Third Year of the Russo-Ukrainian War.* Royal United Services Institute, February 2025.