Use of Large Language Models in Assessing Training Performance

Brian VanVoorst, Charlie Meissner, Nick Walczak
RTX BBN Technologies
Minneapolis, MN
brian.vanvoorst@rtx.com,
charles.meissner@rtx.com,
nicholas.walczak@rtx.com

Jack Norfleet, Matthew Hackett
DEVCOM Soldier Center STTC
Orlando, FL
jack.e.norfleet.civ@army.mil, xxx

ABSTRACT

The Army has long invested in automated techniques to advance the process of teaching Tactical Combat Casualty Care (TCCC) medical procedures. This investment includes the use of computer vision and AI technologies to create specialized systems that recognize and analyze student performance of TCCC work. Recent advances in generalized AI systems, including Large Language Models (LLMs) have gained considerable attention. LLMs such as OpenAI's ChatGPT-4b and Google's Gemini have learned, through their training data, various TCCC medical procedures, including the proper steps to perform the skills, the reasons for those steps, and what those steps look like in a photograph. The systems can now take video input, it is expected that soon they will be able to reason about real-time video input.

Our paper asks the question: "how can LLM capabilities can be utilized by the simulation and training community?" We have identified several areas and present preliminary work that utilize these concepts. Our exploration includes: 1) using LLMs to assess a student's performance of a TCCC procedure in video (this relates to ongoing work by the Army DEVCOM STTC CITADEL program, and DARPA's PTG program); 2) using LLMs to make accurate assessments of the start and end of medical skill steps in a video (also relates to CITADEL and PTG); 3) using LLM's to answer questions students may have in-the-moment while training on a skill (relates to PTG and Army Applications Lab's ARTEMIS program); and 4) using LLM's to annotate images to improve object recognition models (of interest to the all programs).

We specifically report on the capabilities of the ChatGPT-4b and Google Gemini Advanced LLM's, and speculate on future capabilities. We report lessons learned and provide insights into how this technology can be used in the future for the concept of operations described and other uses in simulation and training.

ABOUT THE AUTHORS

Brian VanVoorst is a Lead Scientist and Raytheon Fellow in the Intelligent Software and Systems Business Unit at RTX BBN Technologies. He has contributed to research and technology development in computer vision, machine learning, autonomy, and Artificial Intelligence (AI) during his 14 years with BBN. Mr. VanVoorst is currently the Principal Investigator for RTX BBN's work under the CASTLE Improved Tasks to Advance the Delivery of Education Lessons (CITADEL) for training and evaluation with DEVCOM Soldier Center STTC, as well as the DARPA Perceptually-enabled Task Guidance for Medical Assistance, Guidance, Instruction, and Correction (PTG-MAGIC) program. Prior to CITADEL Mr. VanVoorst has worked a number of related combat medical training programs using computer vision for STTC. Outside of computer vision for training, Mr. VanVoorst has also made contributions to research in the fields of applications of LIDAR, computer vision for maintenance and inspection, autonomy, and benchmarking high-performance computers. He holds a B.S. and M.S. in Computer Science from Michigan Technological University.

Charles Meissner joined the Intelligent Software and Systems Business Unit at RTX BBN Technologies at the start of 2024. He is currently working as a Research Software Engineer on BBN's DARPA PTG contract and a combat medical program called ARTEMIS for the U.S. Army Applications Lab. He holds a B.S. and M.S. in Computer Science and a B.A. in Political Science from Michigan State University.

Dr. Nicholas Walczak is a Scientist in the Intelligent Software and Systems Business Unit at RTX BBN Technologies, where he has been for over 8 years. He received his Ph.D. in Computer Science from the University of Minnesota in 2017. His research interests lie in the area of computer vision, particularly in object detection and object tracking. He has worked on and/or lead numerous projects at BBN Technologies in this area, including Camera Analytic System for Teaching, Learning and Education (CASTLE), CASTLE Improved Tasks to Advance the Delivery of Education Lessons (CITADEL), Terrain Learning During Drop/ Learning Useful Models of Information in Networks to Advance TAK Ecosystems (TLDD/LUMINATE), and Perceptually-enabled Task Guidance for Medical Assistance, Guidance, Instruction, and Correction (PTG-MAGIC).

Dr. Jack Norfleet is the Chief Engineer for Medical Simulation Research at the Simulation and Training Technology Center (STTC) of the DEVCOM Soldier Center under Army Futures Command. He executes medical simulation R&D while managing a multidisciplinary team of engineers and scientists. His current research seeks to improve medical training effectiveness by applying cutting edge simulation technologies. Interests include tissue characterization, advanced airways, synthetic environments, gender disparities, computer vision, AI, brain measures of knowledge, multi-modal automated skill measures, and military working dog medical simulations. Dr. Norfleet started his federal career as a GS-1 student trainee in 1984. He has 38 years of experience in developing and fielding modeling, simulation and training technologies for the U.S. Army and U.S. Navy. His degrees include a Ph.D. in Modeling and Simulation, 2018, a Master of Business Administration, 2001, and a Bachelor of Science in electronics Engineering, 1990.

Dr. Matthew Hackett is a science and technology manager for the Combat Capabilities and Development Command Soldier Center. Dr. Hackett had led a variety of medical simulation and training research efforts, including holographic display research, serious gaming, training effectiveness, and computer vision. Dr. Hackett led the effort to modernize tactical combat casualty care training, resulting in a standardized curriculum and the Deployed Medicine platform. Dr. Hackett also serves as the technology committee chair for the Society for Simulation in Healthcare. Dr. Hackett holds a PhD in Modeling and Simulation, an MS in Biomedical Engineering, and a BS in Computer Engineering.

Use of Large Language Models in Assessing Training Performance

Brian VanVoorst, Charlie Meissner, Nick Walczak
RTX BBN Technologies
Minneapolis, MN
brian.vanvoorst@rtx.com, charles.meissner@rtx.com, nicholas.walczak@rtx.com

Jack Norfleet, Matthew Hackett
DEVCOM Soldier Center STTC
Orlando, FL
jack.e.norfleet.civ@army.mil, xxx

INTRODUCTION

Artificial Intelligence (AI) can have a profound impact on training in several different ways, including scenario generation, improved synthetic behaviors, adaptive instruction, and more. Within this manuscript, we focus on enhancing combat medicine training by providing special purpose AI tools to improve after-action review and coaching as well as creating more realistic training environments. Prior research experimented with various sensor systems (VanVoorst et al., 2015) and ultimately selected helmet-mounted cameras to provide portable first-person views of trainees performing combat medicine (Walczak et al., 2022). While helmet-mounted video footage is useful on its own, the creation of specialized AI tools has allowed us to streamline the whole-exercise video to its essential educational parts and take measurements within the video to assess user performance. Complementary research within the DARPA PTG program has advanced this concept by developing specialized AI systems that can watch videos and verify that all necessary steps have been completed.

Generalized AI systems, such as Large Language Models (LLMs) (Brown et al., 2020) and image generation systems (Ho et al., 2020), are trained on extensive corpuses of text or images covering a wide range of topics. LLMs, such as OpenAI's GPT-4 (OpenAI, 2023) and Google's Gemini (Reid, 2024), have learned various Tactical Combat Casualty Care (TCCC) medical procedures through their training data, including procedure steps, rationales, and visual representations. For years, the I/ITSEC community has developed tools to assist in simulation and training for various specialized job functions, all of which have been specialized efforts requiring novel underlying AI technologies often customized to a specific training domain. The introduction of generative AI systems, like OpenAI's GPT-4 and Google's Gemini, represents a disruptive technology with significant potential to impact this work. Initially, this impact may be seen in generative imagery from systems like DALL-E 3 (Ramesh, 2021) and Midjourney (Holz, 2024) being used to create new training imagery, but the broader potential of these foundational models extends much further.

In this paper, we explore the capability of LLMs to perform or assist in some of these AI tasks that the training community has been invested in creating. If successful we can leverage the cloud-based LLMs[1] from companies such as Google and OpenAI which, in their pursuit of generalized intelligence, leverage several orders of magnitude more funding, staff, data, and computing resources. Some colleagues have questioned why we would highlight that our research might be eclipsed by third-party products. Our response is twofold. First, there is no reason to ignore the incredible advances being made by LLMs. In their pursuit of generalized intelligence, these products will inevitably subsume many specialized AI efforts. As researchers and technologists, we must remain agile and adapt to technological advances, leveraging them to achieve new and more exciting tasks. Second, we have been genuinely surprised by the areas in which LLMs excel and those in which they struggle. Some tasks we deemed nearly impossible are handled well by LLMs, while seemingly simpler tasks remain beyond their current capabilities. Consequently, customized AI solutions or human intelligence are still required for certain tasks at this time.

We have focused our exploration of LLMs on the following challenging and resource-intensive areas: 1) The assessment of a medical skill in a video for the presence of an error; 2) The timestamping of moments of medical skills (Figure 1); 3) The answering of open-ended questions a student may have about medical skills; and 4) The labeling of training data for object recognition. We will explore each topic in turn, explaining the core challenge, how an LLM can help, the success of an LLM in the process, and the implications for the simulation and training community. Finally, we will summarize our results and make some predictions about the future use of LLMs in this area.

Figure 1. Gemini is given steps of a chest seal application and is asked to find the moments in the video when they occur.

BACKGROUND WORK

To provide context, we present background information on past special-purpose AI projects and details on how publicly-available cloud-based AI systems function.

Recent projects focused on applying computer vision and AI for point-of-injury medical care.

Hundreds of thousands of individuals are trained in point-of-injury care, such as TCCC or civilian emergency response. Several research efforts aim to support this training through various approaches, including the identification of key moments for after-action review, providing automated performance assessment, and offering automatic guidance in performing medical interventions. We briefly discuss three projects to highlight the need for and investment in special-purpose AI projects, thereby motivating the interest in this paper.

The Army's DEVCOM Soldier Center STTC has been working towards the frictionless use of video in after-action reviews as documented in (VanVoorst et al., 2023), referred to as CASTLE. By utilizing helmet-worn egocentric cameras and artificial intelligence software, CASTLE eliminates the need for a separate videographer. CASTLE videos have several use cases, including educational review, instructor workload reduction, automated coaching, experimental ground truth, and serving as a source for training material. CASTLE was used at Joint Base Lewis-McChord to record performance during training and was presented at IITSEC in 2022 (Walczak et al., 2022).

In related work, the primary goal of The Trauma THOMPSON Challenge (TCGC, 2034) is to identify the best algorithms for automatic action recognition and prediction using computer vision from a first-person perspective in the medical domain. This challenge utilized an egocentric-view dataset of life-saving intervention procedures with detailed annotations by medical professionals. The Trauma THOMPSON Challenge has collected over 200 procedure videos with environmental simulator and type variability. The Challenge's tasks include action recognition, action anticipation, procedure recognition, and visual question answering (VQA) algorithm development. This work is funded by the U.S Army Medical Research Acquisition Activity (USAMRAA).

DARPA's Perceptually-enabled Task Guidance (PTG) program (Marge, 2024) aims to develop artificial intelligence technologies to assist users in performing complex physical tasks. This assistance aims to expand a user's skillset and reduce errors, making them more versatile and proficient. One of the selected domains for PTG is combat medicine. PTG seeks to develop methods, techniques, and technology for AI assistants that provide just-in-time visual and audio feedback during task execution. The goal is to equip users with an augmented reality (AR) headset that includes cameras and microphones enabling the AI assistant to see and hear what the user does. Inside the AR display the AI assistant will provide feedback through speech and aligned visual graphics. These AI assistants will combine task knowledge with a perceptual model of the environment to support mixed-initiative and task-focused user dialogs. These dialogs will help users complete tasks, identify and correct errors, and provide guidance through new tasks, considering the user's level of expertise.

Common threads to all three of these projects are: 1) that they seek to answer questions about what is happening in the videos; 2) that they need to recognize medical devices; 3) that they depend on a considerable amount of labeled imagery; and 4) that one of their end goals is to help the user, in the context of learning in a simulated medical scenario, become better at performing the end skills through AI techniques.

A short description of Large Language Models**.**
In recent years, Large Language Models (LLMs) have become a significant focal point within the realm of artificial intelligence, revolutionizing how we interact with and leverage machine intelligence. LLMs, such as OpenAI's GPT-4 and Google's Gemini 1.5 Pro are advanced neural networks trained on vast amounts of text data, enabling them to generate, understand, and respond to human language with remarkable fluency and coherence. Unlike traditional rule-based AI systems, which rely on predefined instructions, LLMs learn patterns and nuances from data, making them exceptionally versatile in a variety of tasks ranging from natural language processing to creative writing. Their ability to comprehend context and generate human-like text has opened new possibilities for applications in customer service, content creation, and even complex problem-solving.

The past two years have witnessed rapid advancements in LLMs, particularly with the development of multi-modal models that can process and generate text, images, and other types of data. This evolution marks a significant departure from earlier AI models, which were typically specialized for specific tasks such as image recognition or speech processing. Multi-modal LLMs integrate diverse data types, enabling more comprehensive and context-aware interactions. For instance, the models used in this paper (provided as a paid-for cloud service) can analyze a combination of text, images, and in some cases, video, to provide richer and more nuanced responses. This capability enhances their utility in fields like education, healthcare, and research, where understanding and synthesizing information across different media is crucial. The convergence of these capabilities highlights the unique position of LLMs within the broader AI landscape, showcasing their potential to transform various industries by providing more dynamic and human-like interactions.

OUR PROCESS AND THE REMAINDER OF THE PAPER

The remainder of the paper will discuss four different experiments using LLMs and image generation AIs to augment or replicate the work done by special-purpose AI efforts in tools for combat medical training. For each challenge, we briefly provide the context, how the generalized AI can help, the impact if successful, and quantitative results with analysis. We conclude with a discussion on the future impact of AI in this work.

In each section we will make note of what AI system we worked with. All data was collected between May and June of 2024. We note the period because all AI vendors are continually developing new releases and features and we have no control over what is included in those capabilities.

Part of our work involved experimenting with prompts to solicit proper responses. Due to space constraints, we did not share our prompts but will provide them upon request. For each medical skill (application of a tourniquet, a pressure bandage, and an XSTAT 12), we prompted with a prescribed set of steps expected from the user. These steps are also available upon request.

CHALLENGE 1: DETECTING THE PRESENCE OF AN ERROR IN A MEDICAL SKILL IN A VIDEO

A "holy grail" task for AI in TCCC training videos is detecting if someone in training performs a medical skill correctly or has made a mistake in performing that skill. Achieving such a feat would offer significant automated advantages in assessing human performance in TCCC medical skills. It would enable more rigorous assessment and better allocation of human instructors' time towards one-on-one instruction rather than performance assessment. The notion of an error can vary widely, especially with novice users, and the environment in the video can be highly variable, posing enormous challenges for computer vision experts utilizing traditional techniques. However, recent multi-modal models have shown promising abilities when focused on video. This study aims to explore what these models can achieve with minimal effort.

Table 1: Gemini's Performance in Detecting Errors in Training Videos (Green: Correct, Yellow: Partially Correct)

Skill	Ground Truth About Error	Trial 1	Trial 2	Trial 3
Tourniquet	Out of 8 steps, skipped steps 4-7.	Flagged error at step 4	Flagged error at step 3	Flagged error at step 5
Pressure Bandage 1	Sets a book on the wound to apply pressure instead of using hands.	The user did not apply direct hand pressure before applying the pressure dressing. The user starts by placing a book on the wound instead of applying direct hand pressure.	They did not apply direct hand pressure to the wound before opening the dressing package. This is a critical step to help control bleeding.	The error is that they did not apply direct hand pressure to the wound before applying the pressure dressing.
Pressure Bandage 2	After wrapping the pressure bandage, the user does not use clips to secure the bandage.	User failed to secure dressing with securing device.	The user failed to secure the dressing with a securing device.	The user failed to secure the dressing with a securing device.
Pressure Bandage 3	Compound error. User does not apply direct hand pressure for Step 1.	The user in the second video failed to apply direct hand pressure to the wound before applying the pressure dressing.	Yes, the user made an error in the second video. The error was that they did not apply direct hand pressure to the wound before opening and applying the pressure dressing.	They did not apply direct hand pressure to the wound site before opening the dressing packaging.

In this effort we used Google's Gemini 1.5 Pro for error detection of four medical skills: the application of a tourniquet, the application of a pressure bandage, the application of a chest seal, and the use of a X-STAT 12. In all cases we prompted with a single video of the skill being performed correctly, and the steps we took in performing the skill. We then showed Gemini, in separate runs, different videos of the skill being performed (some of which had correct applications, others which had a mistake performed in the video). The response of the AI is shown in Table 1 (videos with errors present) and Table 2 (videos with no errors present). We note that at the time of our experiments Gemini would take as input a whole video and examine only one frame per second, effectively throwing away 29 of the 30 other frames of information. We do not know the algorithm for selecting the single frame, and can only speculate on how performance may have increased had it used additional information.

Table 2. Gemini's Performance in Identifying Errors in Error-Free Training Videos (Dark Green: Correct, Light Green: True but Not an Error, Yellow: Matters of Interpretation, Red: Wrong)

Skill	Ground Truth	Trial 1	Trial 2	Trial 3
Tourniquet	No Error	Step 4 Error: The user does not appear to sufficiently tighten the tourniquet before moving to Step 5. They turn the windlass only a couple of times.	No Error	Step 7 Error: Not performed – omitted the use of the windlass securing device (Velcro strap).
Result:		*True, but not an error.*	*100% Correct*	*Wrong.*

Skill	Ground Truth	Trial 1	Trial 2	Trial 3
Pressure Bandage 1	No Error	Step 3 Error: The user does not maintain adequate pressure while applying the dressing to the wound site. They let go of the dressing entirely before wrapping it.	Step 3 Error: Variation of Step 3, dressing was not applied directly over wound site on first attempt.	Step 3 Error: Did not apply pressure point directly over the wound.
Result:		*True, but not an error.*	*Wrong.*	*Wrong.*
Pressure Bandage 2	No Error	Step 4 Error: The user wraps the bandage directly on top of the pressure dressing. Wrapping directly on top of the pressure dressing could displace it and compromise pressure on the wound.	Step 3 Error: Failed to maintain continuous direct pressure on the wound site.	Step 3 Error: Did not immediately place the dressing on the wound site after opening the package.
Result:		*AI seems confused as to what it sees.*	*True, but not an error.*	*Opinion, we disagree that it wasn't immediately.*
Pressure Bandage 3	No Error	No Error	No Error	Step 3 Error: User does not apply the dressing with the pressure point directly over the wound site.
***Result*:**		*100% Correct*	*100% Correct*	*Wrong.*
Chest Seal	No Error	Step 1 Error: The user removes their hand before the chest seal is fully applied and sealed to the mannequin's chest.	Steps 1 & 2 Error: The user failed to maintain hand pressure on the wound site after initially applying it in Step 1. The user removed their hand completely to open the chest seal package.	Steps 3 & 4 Error: The user failed to cover the wound site with their hand after wiping the area and before applying the chest seal.
Result:		*True, but not an error.*	*True, but not an error.*	*True, but not an error.*

The results shown in Table 1 and 2 capture how Gemini did at watching five videos with errors (Table 1) and five videos without errors (Table 2). Our prompts included: a) a list of expected steps per skill, b) an example of the skill being performed correctly in a video, and c) instructions to examine the videos we showed and see if there were any errors. Repetitions were conducted (Trials 1-3) to see if each outcome was repeatable.

As shown in Table 1, Gemini was remarkably accurate in identifying errors. Given that Gemini is only observing one frame per second, we believe that performance in the tourniquet trials 2 and 3 could improve with additional frames. However, Table 2 suggests that this capability comes with a degree of false-positive error detection. Many responses indicate a stricter interpretation of our instructions than intended. We believe this could be improved with better prompting. We note that chest seal performed better than the others, but can offer no explanation as to why. We also note that repeated trails of different videos were important in uncovering the variability of the performance of the system. If this capability can be made robust, it could significantly impact training programs across various domains.

CHALLENGE 2: TIMING THE STEPS OF A MEDICAL SKILL

Timely medical care is critical to achieving ideal casualty outcomes; as such, during training we seek to identify when a medical skill starts and stops to analyze student proficiency within a margin of +/- 2 seconds. To achieve this, we created a rule-based system that monitors hand interactions with medical objects. Similarly, in the DARPA PTG program, a key metric is knowing when each step of a medical skill is completed to notify the user of their progress against a series of prescribed steps. This step recognition is performed by neural network models trained on labeled video data, where video frames are annotated with the steps being performed. This required a great deal of effort in both collecting the video examples and having them manually labelled for ground truth. The desired timing accuracy per step is also +/- 2 seconds, and real-time streaming video results are required for both applications.

In our second challenge, we aimed to determine if Gemini could assess the start and stop times of steps in a skill application by being shown an example in a prompt. Due to time and space constraints, we focused on the 8-step tourniquet application process, conducting three trials to understand repeatability and repeating the experiment over three different tourniquet application videos. Step 6 (securing extra strap) was omitted because there was no additional strap to secure[1]. Since the API does not support streaming video and receiving replies, we modified our experiment to give the AI the entire video to consider at once, which admittedly gave the AI a considerable advantage. The results are shown in Tables 3, 4, and 5. The ground truth indicates the start and stop times for each step. The trial entries (1-3) indicate each attempt by Gemini to answer the question of the start and end times per step. The delta is computed as Ground Truth – Trial. Entries within 2 seconds are considered comparable to the state of the art. To better understand this task, consider timing the start and stop of each step depicted in Figure 1.

Table 3. Gemini attempting to delineate the eight steps of a tourniquet application, Video 1. Average performance (last row) is equivalent to the state of the art.

Steps	**Ground Truth (s)**	**Trial 1 (s)**	**Delta**	**Trial 2 (s)**	**Delta**	**Trial 3 (s)**	**Delta**	
S1 Start	3.13	5.00	1.87	4.00	0.87	4.00	0.87	
S1 End	9.20	7.00	-2.20	7.00	-2.20	8.00	-1.20	
S2 Start	9.27	7.00	-2.27	7.00	-2.27	8.00	-1.27	
S2 End	11.53	13.00	1.47	15.00	3.47	13.00	1.47	
S3 Start	11.60	8.00	-3.60	9.00	-2.60	9.00	-2.60	
S3 End	15.33	15.00	-0.33	17.00	1.67	14.00	-1.33	
S4 Start	19.93	20.00	0.07	20.00	0.07	20.00	0.07	
S4 End	25.53	27.00	1.47	27.00	1.47	27.00	1.47	
S5 Start	25.60	27.00	1.40	27.00	1.40	27.00	1.40	
S5 End	27.53	28.00	0.47	29.00	1.47	28.00	0.47	
S6 Start	NA	NA	NA	NA	NA	NA	NA	
S6 End	NA	NA	NA	NA	NA	NA	NA	
S7 Start	27.73	28.00	0.27	29.00	1.27	28.00	0.27	
S7 End	29.80	29.00	-0.80	30.00	0.20	29.00	-0.80	
S8 Start	31.67	31.00	-0.67	31.00	-0.67	31.00	-0.67	
S8 End	38.20	36.00	-2.20	37.00	-1.20	36.00	-2.20	**Avg**
Mean Abs Dev			1.36		1.49		1.15	1.33
Root Mean Sq Dev			2.79		3.01		1.77	2.53

Gemini provided excellent results for Video 1. However, for Videos 2 and 3, it failed to find the starting and ending times for some steps in certain trials (though not in *all* trials). These results are on par with the best-in-class solutions that operate in real-time. Due to the closed nature of the cloud-based AI services, we cannot determine the extent of video review being performed to generate the answers, but each response was returned within a few seconds. Considering the ground truth at a resolution of 1/100th of a second and the median duration of these steps being slightly less than 3 seconds, we believe that the system might significantly improve if it considered more frames when answering the question.

[1] Gemini correctly identified this was a skipped step.

Table 4. Gemini attempting to delineate the eight steps of a tourniquet application, Video 2. Average performance (last row) is equivalent to the state of the art.

Steps	Ground Truth (s)	Trial 1 (s)	Delta	Trial 2 (s)	Delta	Trial 3 (s)	Delta	
S1 Start	1.67	1.00	-0.67	1.00	-0.67	1.00	-0.67	
S1 End	5.13	4.00	-1.13	4.00	-1.13	4.00	-1.13	
S2 Start	5.20	4.00	-1.20	4.00	-1.20	4.00	-1.20	
S2 End	7.40	8.00	0.60	8.00	0.60	9.00	1.60	
S3 Start	7.47	5.00	-2.47	5.00	-2.47	7.00	-0.47	
S3 End	12.00	11.00	-1.00	11.00	-1.00	12.00	0.00	
S4 Start	15.73		-15.73	12.00	-3.73		-15.73	
S4 End	20.00		-20.00	19.00	-1.00		-20.00	
S5 Start	20.07	11.00	-9.07	19.00	-1.07	12.00	-8.07	
S5 End	22.47	12.00	-10.47	22.00	-0.47	15.00	-7.47	
S6 Start	NA	NA	NA	NA	NA	NA	NA	
S6 End	NA	NA	NA	NA	NA	NA	NA	
S7 Start	22.73	12.00	-10.73	22.00	-0.73	15.00	-7.73	
S7 End	24.53	25.00	0.47	25.00	0.47	24.00	-0.53	
S8 Start	26.00	25.00	-1.00	25.00	-1.00	25.00	-1.00	
S8 End	32.93	32.00	-0.93	32.00	-0.93	32.00	-0.93	**Avg**
Mean Abs Dev			3.31		1.18		2.57	2.35
Root Mean Sq Dev			26.64		2.11		15.73	14.83

Table 5. Gemini attempting to delineate the eight steps of a tourniquet application, Video 3. Average performance (last row) is equivalent to the state of the art.

Steps	Ground Truth (s)	Trial 1 (s)	Delta	Trial 2 (s)	Delta	Trial 3 (s)	Delta	
S1 Start	0.47	1.00	0.53	1.00	0.53		-0.47	
S1 End	5.40	10.00	4.60	4.00	-1.40		-5.40	
S2 Start	5.47	9.00	3.53		-5.47	0.00	-5.47	
S2 End	7.33	14.00	6.67		-7.33	5.00	-2.33	
S3 Start	7.40	2.00	-5.40	4.00	-3.40	3.00	-4.40	
S3 End	11.67	17.00	5.33	17.00	5.33	17.00	5.33	
S4 Start	14.93	18.00	3.07		-14.93	17.00	2.07	
S4 End	19.13	22.00	2.87		-19.13	19.00	-0.13	
S5 Start	19.20	23.00	3.80		-19.20	19.00	-0.20	
S5 End	21.47	24.00	2.53		-21.47	19.00	-2.47	
S6 Start	NA	NA	NA	NA	NA	NA	NA	
S6 End	NA	NA	NA	NA	NA	NA	NA	
S7 Start	21.53	24.00	2.47	17.00	-4.53	19.00	-2.53	
S7 End	23.33	25.00	1.67	23.00	-0.33	23.00	-0.33	
S8 Start	23.87	25.00	1.13	23.00	-0.87	24.00	0.13	
S8 End	30.80	29.00	-1.80	30.00	-0.80	30.00	-0.80	**Avg**
Mean Abs Dev			3.24		1.43		2.18	2.29
Root Mean Sq Dev			13.42		8.03		8.39	9.95

CHALLENGE 3: OPEN-ENDED QUESTIONS AND ANSWERS ABOUT MEDICAL SKILLS

Another "holy grail" capability in AI assisted training would be a system that could answer arbitrary student questions about the skills they were learning. While such a system would have seemed science fiction two years ago, those who have interacted with systems like ChatGPT would not be surprised that a LLM can provide an answer to nearly any question posed to it. However, the key question is, how accurate of a response would that answer be?

Q. Should the tourniquet be applied directly on the wound?
A. *No, place tourniquet 2-3 inches above wound site, not directly on it.*

Q. Should I remove the tourniquet once the bleeding has stopped?
A. *No, keep the tourniquet in place until medical professionals can assess and provide further treatment.*

Q. Where exactly should I place the chest seal on the chest?
A. *Center the chest seal over the wound, ensuring the circle of vents is directly above the injury site.*

Q. Does a chest seal work if the patient is sweating heavily?
A. *Yes, but ensure skin is as dry as possible before application for better adhesion.*

Figure 2. Representative questions and answers.

RUBRIC FOR SCORING AI ANSWERS

5: As good of an answer in one sentence as you could expect.
4: A human might have done better and included some other important bits
3: Incomplete answer even for one sentence
2: Wrong answer
1: Terrible and/or dangerous answer

Figure 3. Visualization of scoring of answers per medical topic for Gemini.

\

We generated 80 one-sentence questions, 20 for each of our target medical skills. In this scenario we imagined a "simple" question asked in the moment of performing a skill and instructed the AI to respond in kind with a simple one-sentence response. Without limiting the answer to a single sentence, the LLM's tended to produce lengthy and very complete answers that would not necessarily be actionable in the moment. The challenge of a single sentence (targeting 12 words or less) response made the AI get to the core of the question. We then took these answers and asked a professional TCCC instructor to evaluate the quality of each answer on a 1-5 scale. Samples from the list of questions and responses are shown in Figure 2. The rubric given for evaluation is shown in Figure 3. An evaluation of the responses is shown in Table 6 with summary statistics shown in Figure 4.

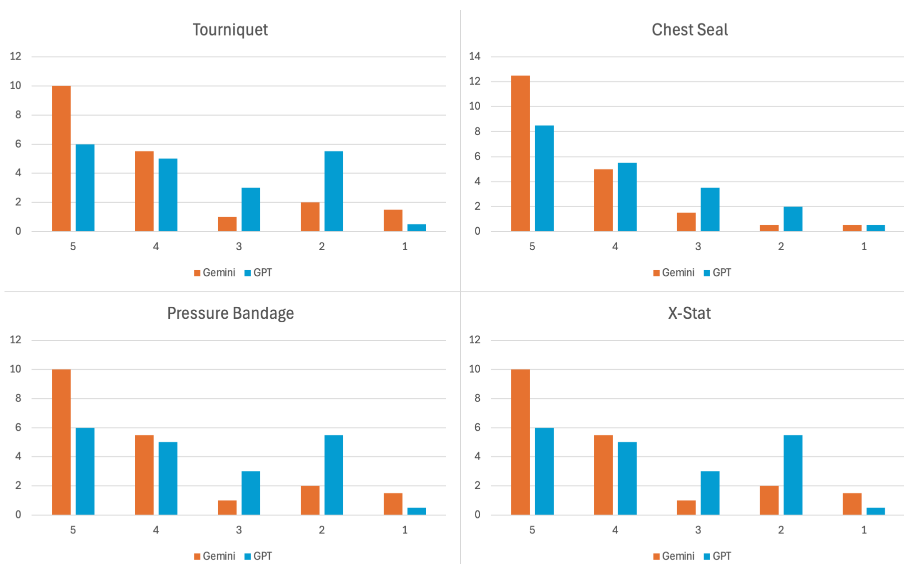

Figure 4. Visualization of scoring of answers per medical averaged across

Table 6. Professionals scores of the answers of Gemini 1.5 Pro and GPT-4o on 80 medical related questions.

	Tourniquet				Chest Seal				Pressure Dressing				X-STAT			
Question	Gemini Score		GPT-4o Score		Gemini Score		GPT-4o Score		Gemini Score		GPT-4o Score		Gemini Score		GPT-4o Score	
1	5	4	5	2	5	5	5	5	5	5	5	5	4	5	4	5
2	1	5	4	2	1	4	4	3	4	NA	NA	3	1	5	4	5
3	5	4	5	2	5	5	5	5	4	5	5	5	5	5	4	5
4	4	3	4	2	5	5	4	5	4	5	4	5	4	5	4	5
5	5	5	4	2	4	5	3	5	4	5	4	5	3	5	4	5
6	2	4	2	4	4	5	5	5	3	5	3	5	5	5	4	5
7	5	5	4	2	2	3	3	2	5	5	4	5	3	5	4	5
8	5	5	5	2	4	4	3	1	4	5	4	5	1	5	3	5
9	5	5	5	4	4	5	2	5	3	5	4	5	4	5	4	5
10	5	3	3	3	4	5	5	2	5	5	4	5	NA	4	2	4
11	2	5	4	5	5	5	4	5	3	3	4	3	3	5	2	1
12	4	5	3	5	5	5	4	5	2	5	4	5	NA	4	4	3
13	1	2	5	2	5	4	4	3	4	4	3	3	2	5	4	1
14	1	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5
15	4	5	5	5	5	5	4	5	1	5	5	5	4	5	4	3
16	4	5	2	1	5	5	5	5	4	5	4	4	1	5	4	5
17	5	2	4	2	4	3	4	3	3	5	4	4	4	5	5	5
18	4	4	3	4	3	5	2	4	5	5	5	5	5	5	4	5
19	4	5	3	5	5	5	4	3	3	3	4	3	2	5	4	5
20	4	5	4	3	4	5	4	4	2	5	3	5	4	5	4	2

Of the overall answers Gemini scored 79% as a 4 or 5, indicating they were suitable, whereas 10% of the answers were deemed dangerously incomplete or inaccurate. GPT scored 73% as a 4 or 5 indicating they were suitable, with 14% deemed dangerously incomplete or inaccurate. We believe that if the LLM were specifically trained with TCCC medical knowledge, the overall quality could be improved considerably. This capability, if improved, could be an exceptional resource for students who are either unable to get instructor time to ask questions, or unwilling to ask questions in a training setting in front of their peers.

CHALLENGE 4: LABELING TRAINING DATA FOR OBJECT RECOGNITION

The goal of this study was to employ automated methods for labeling medical object data, which is crucial for medical object recognition. Many AI efforts in this domain require reasoning about medical objects, necessitating object detectors trained with labeled imagery. Typically, object recognition tasks require a minimum of 1,000 labeled instances, with 10,000 or more often being desirable. Moreover, when an object classifier encounters additional objects in the scene (e.g., hands in a tourniquet dataset), and those objects later become of interest, the images need to be re-examined and additional labels created.

Given the success of LLM's in analyzing video, we hypothesized that they might also excel in analyzing still images and labeling target objects. The advantage of using simulation and training with AI is the potential to significantly increase the number of labeled objects, thereby improving object detection under various conditions. This enhancement would ultimately result in better simulation and training tools that rely on accurate object detection.

To assist the model, we augmented the images with a coordinate frame. By providing the AI with a prompt and an example of the desired output, we tasked Gemini 1.5 Pro and GPT-4o with generating bounding box coordinates for new, unseen images. These AI-generated labels were compared to human-labeled ground truth using the Intersection Over Union (IOU) metric. The IOU is defined as the area of the intersection of the bounding boxes divided by the area of their union. An IOU of 1.0 indicates identical boxes (ideal), while a score of 0 indicates no overlap.

The results of the study are presented in Table 7, with a representative example shown in Figure 5. To consider replacing human labelers to create training data, we aimed for IOUs of 90% or higher. However, the AI's performance on this task was particularly poor. It is intriguing to consider how the AI can successfully reason about video tasks but fail to accurately locate the main object of interest in still images. However, this question is beyond the scope of this paper.

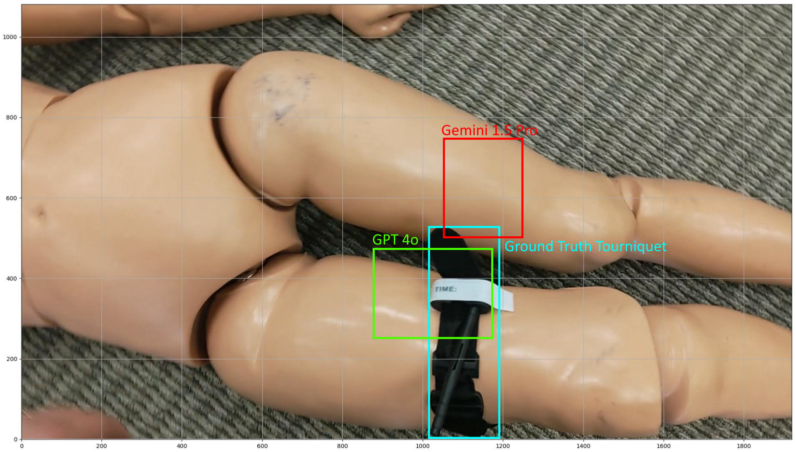

Figure 5. Ground Truth (blue) alongside GPT4o's (green) and Gemini's (red) attempts to label the tourniquet.

Table 7. Gemini attempting to locate target objects in a scene as a surrogate for a human labeler.

Object	Gemini IOU	GPT4o IOU
Tourniquet	0.67	0.34
Tourniquet	0.22	0.71
Tourniquet	0.05	0.28
Package	0.34	0.35
Package	0.00	0.37
Package	0.01	0.00
Plunger	0.17	0.00
Plunger	0.00	0.00
Plunger	0.00	0.08
Package	0.00	0.08
Package	1.00	0.00
Package	0.38	0.12

SUMMARY

Training is a labor-intensive process for which the military community has invested in special purpose computer vision AI systems for automation. When this automation is successful it can give considerable benefit to the trainee and instructor in several ways, ranging from bookmarking moments of a training session to review, to performing automated assessment, or ultimately an always-available computerized instructor.

Recently, AI systems that purport to have broad capabilities over a wide set of tasks have become available as online services. In this paper we have examined the use of these generalized AI models (LLM's and diffusion image generation systems) as methods to augment or replace the work done by specialized AI systems on five challenge problems. Our findings show that while not ready for deployment, in many cases they operate at a technology readiness level that is on par with specialized systems. Due to the considerable investment these systems are receiving we can expect their capabilities to improve.

CONCLUSIONS

In this paper we have explored several AI challenges that we thought cloud-based AI services may be able to assist with to benefit TCCC Training. As these AI services continue to advance their capabilities with a new announcement every month, our evaluation is a snapshot in an ever-changing landscape of next generation AI services. Based on the experience of this paper, Table 8 shows the authors belief of where we are today and where we will be in the near future. Speaking in generalities, we optimistically believe that these services can be applied with care to offer great benefit to simulation and training in many different capacities. The next generation of work will be the proper way to harness this intelligence to our benefit, understanding that 1) the capabilities will be ever-changing and 2) these techniques will always be heuristics without a 100% guarantee of correctness.

Table 8. Author's opinions of the state of the AI's at the time of submission.

Challenge	Performer	Comparison to traditional methods	Ready for deployment today?	Expectation for near future
Error detection	Google Gemini	NA	Yes / TRL 6	Will surpass SOA
Skill timing	Google Gemini	As good as specialized existing methods	Yes / TRL 6	Will surpass SOA
Open Q&A	GPT4o & Gemini	NA	No	Probably will soon reach a TRL for deployment
Image labeling	GPT4o & Gemini	Human labeling	No	Unknown

ACKNOWLEDGEMENTS

We would like to thank Shane Storks at the University of Michigan for his assistance in formulating the prompting and collecting the data in Challenge 3. We'd also like to thank Aaron Jamison of Valkyries Austere Medical Solutions & Consultants and Steve Murphy (RTX BBN Technologies) for their assistance in evaluating the answers provided by the AI systems in Challenge 3.

The research reported in this document/presentation was performed in connection with contract number W912CG-19-C-0001 with the U.S. Army Contracting Command - Aberdeen Proving Ground (ACC-APG). The views and conclusions contained in this document/presentation are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, CCDC-SC STTC or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

In addition, this material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No HR001122C0010.

REFERENCES

VanVoorst, B. R., Walczak, N. R., Hackett, M. G., Norfleet, J. E., Schewe, J. P., & Fasching, J. S. (2023). Automated Video Debriefing Using Computer Vision Techniques. Simulation in Healthcare, 18(5), 326-332.

Walczak, N., VanVoorst, B., Noyes, E., Mazzeo, M., Norfleet, J.. Automating Video After Action Reviews for Military Medical Training Exercises. I/ITSEC 2022.

VanVoorst, B. R., Hackett, M., Strayhorn, C., Norfleet, J., Honold, E., Walczak, N., & Schewe, J. (2015, September). Fusion of LIDAR and video cameras to augment medical training and assessment. In 2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI) (pp. 345-350). IEEE.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. Retrieved from https://arxiv.org/abs/2005.14165

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems, 33, 6840-6851. Retrieved from https://arxiv.org/abs/2006.11239

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. Proceedings of the 38th International Conference on Machine Learning, 139, 8821-8831. Retrieved from https://arxiv.org/abs/2102.12092

Holz, D. (2021). Midjourney: AI-generated art exploration. Retrieved from https://www.midjourney.com

Marge, Matthew. (2024). Perceptually-enabled Task Guidance (PTG). Retrieved from https://www.darpa.mil/program/perceptually-enabled-task-guidance

OpenAI. (2023). ChatGPT: Language model trained on GPT-4 architecture. Retrieved from https://www.openai.com/research/chatgpt

Reid, Machel et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." ArXiv abs/2403.05530 (2024)

TCGC. (2024). Thompson Challenge Grand Challenge. Retrieved from https://thompson-challenge.grand-challenge.org