Evaluating an LLM-based Course-of-Action-Analysis assistant for simulated tactical decision-making

Dr. Josh Price
CAE (UK) Plc.
Burgess Hill, West Sussex, UK
joshua.price@cae.com

Dr. Aaron Coutino
CAE Inc.
Montréal, Québec, Canada
aaron.coutino@cae.com

Dr. Deniz Yilmaz, Mr. Peter Meyer zu Drewer
CAE GmbH.
Stolberg, Germany
deniz.yilmaz@cae.com, peter.meyerzudrewer@cae.com

Mr. Giles Moore
Dstl
Portsdown West, Fareham, Hants, UK
gmmoore@dstl.gov.uk

ABSTRACT

Military operations require fast-paced decision-making where outcomes are dependent on innate ability, training and the impacts of stressors like workload and fatigue. To overcome human limitations in increasingly complex and dynamic defence environments, Artificial Intelligence (AI) technologies are being considered to enhance capabilities. Large Language Models (LLMs) are a promising area of advancement, demonstrating exceptional abilities in understanding and generating human-like text, which could enhance Course-of-Action Analysis (CoAA).

To evaluate the technology's impacts from technical and human perspectives we integrated a proof-of-concept LLM-based CoAA assistant with our constructive simulation platform, 'CAE GESI', to provide a safe, configurable and repeatable testbed. The assistant responded to the real-time state of a 'protection of a sensitive site' scenario, providing suggestions of the next step to respond to a range of threats, and enabling users to query the suggestions using natural language. A human participation study then considered:

1. Are LLM-generated CoA recommendations credible and is the AI engaged with?
2. Does AI assistance objectively impact task performance?
3. How is the AI perceived and are there effects on subjective experience of the system?

Although the AI responded in real-time with credible suggestions and effective query handling, this did not translate into measurable performance improvements, in terms of threat outcomes or response times. Engagement with the AI was relatively low with its attentional requirements causing distraction and negatively impacting perception of workload, usability and trust, especially when task demand was high. Wider issues of hallucination, disregard for prompted rules and opportunities for misuse were also observed.

The CoAA assistant demonstrated that LLM technology can deliver decision-support capabilities and be integrated with existing simulation platforms effectively. While the approach shows promise, the study highlights the need to consider human performance in future applications to ensure that benefits are realised and limitations addressed. Key areas for improvement include application to more complex scenarios, enhancing simulation integration, refining the user interface design, and implementing output verification & validation to ensure credibility.

ABOUT THE AUTHORS

Dr. Josh Price is an R&D engineer at CAE (UK) plc, with over 10 years of experience in systems and Human Factors engineering. Josh has lead R&D projects concerned with the application of simulation technologies, AI decision-support and biometric monitoring of human performance. Josh holds an EngD in Human Factors engineering.

Dr. Aaron Coutino is a software developer at CAE Inc., with expertise in AI, mathematical modelling and signal processing. Aaron completed a PhD at the University of Waterloo in Applied Mathematics, incorporating skills in applied machine learning, mathematical modelling, signal processing as well as technical communication.

Dr. Deniz Yilmaz is Innovation & Technology Group Lead at CAE GmbH., involved with a range of advanced technologies, including AI, training science, distributed training, and mixed reality simulation. Deniz holds an MSc and PhD in Aerospace Engineering, specialising in flight simulation.

Mr. Peter Meyer zu Drewer is the 'CAE GESI' Product Expert at CAE GmbH., with over 30 years of experience in the modelling and simulation field. He has been involved in a wide range of R&D projects, including human machine interface, interoperability between simulation systems and between simulation systems and Command & Control Information Systems (C2IS).

Mr. Giles Moore is a principal scientist in Dstl's Futures and Innovation Group. Giles acts as a technical partner to organisations across a wide portfolio of research & development projects, including those conducted through Dstl's Human Augmentation project.

Evaluating an LLM-based Course-of-Action-Analysis assistant for simulated tactical decision-making

Dr. Josh Price
CAE (UK) Plc.
Burgess Hill, West Sussex, UK
joshua.price@cae.com

Dr. Aaron Coutino
CAE Inc.
Montréal, Québec, Canada
aaron.coutino@cae.com

Dr. Deniz Yilmaz, Mr. Peter Meyer zu Drewer
CAE GmbH.
Stolberg, Germany
deniz.yilmaz@cae.com, peter.meyerzudrewer@cae.com

Mr. Giles Moore
Dstl
Portsdown West, Fareham, Hants, UK
gmmoore@dstl.gov.uk

INTRODUCTION

Military operations require fast-paced decision-making where outcomes are dependent on innate ability, training and the impacts of stressors like workload and fatigue. To overcome human limitations in increasingly complex and dynamic defence environments, Artificial Intelligence (AI) technologies are being deployed to enhance capabilities. Specific aims include to aid decision-making, improve efficiency, secure operational advantages with new capabilities, and empower personnel to focus on high value functions (see Ministry of Defence (2022)).

There has been rapid progress towards delivering these capabilities, AI applications having been trialled on exercises. The NATO Spring Storm exercise in 2022, for example, utilising machine learning to process complex data and provide instantaneous decision-support for mission planning (Ministry of Defence, 2022). While the NATO Allied Command Transformation Autonomy Programme investigated AI assistance for Air Command and Control (Meeuwissen, 2025), demonstrating significant efficiencies in the creation of air plans.

While quantifiable real-world benefits have been demonstrated, there is increasing awareness and understanding of AI's limitations and risks. These include a lack of interpretability and explainability (Gilpin, et al., 2018), dependence on training data quality and completeness (Shumailov, et al., 2024), and the resource intensiveness of AI applications (Organisation for Economic Co-operation and Development, 2022).

These issues will become more pertinent as AI is applied to more complex tasks such as Course-of-Action Analysis (CoAA) which can be utilised to formulate and evaluate strategies (HM Government, 2021), guiding human decision-making. Advances in AI technologies are enabling new ways to improve decision-making systems. Large Language Models (LLMs) being a promising area of development, having demonstrate exceptional abilities in understanding and generating human-like text, making them ideal for producing CoA recommendations.

Course-of-Action Analysis using LLMs

Traditionally, Reinforcement Learning (RL) has been the foundational approach for decision-making problems in machine learning, with applications across diverse domains, including robotics, gaming and autonomous systems (e.g. Sutton & Barto (2018), Mnih, et al. (2015)). In RL agents are trained to make optimal decisions by providing rewards for desirable actions and penalties for undesirable ones. Through this reward-driven interaction with the environment, agents learn to navigate complex decision spaces, aiming to maximise cumulative rewards over time.

Implementing RL requires precisely mapping out all possible actions and assigning appropriate rewards, which can be computationally intensive and complex, especially in dynamic or high-dimensional environments. This raises challenges in scalability and adaptability, particularly in scenarios requiring rapid decision-making and contextual understanding. The exhaustive exploration of action spaces and the need for substantial training data can limit the practicality of RL in time-sensitive or resource-constrained situations.

Recent advancements suggest that LLMs enhanced with human feedback offer a promising alternative for decision-making tasks. Instead of relying solely on predefined action mappings and reward systems, LLMs leverage their extensive language understanding capabilities to interpret context and generate actions accordingly. This approach capitalises on the models' ability to process and generate human-like text, enabling them to comprehend complex scenarios and provide contextually relevant suggestions.

Research by Goecks & Waytowich (2024) has demonstrated that LLMs perform equivalently to RL, and with human-in-the-loop feedback can outperform traditional RL methods in generating CoAs. Unlike RL agents that require specific training on the problem domain, LLMs utilise their general language understanding and contextual awareness to suggest actions. While LLMs are not inherently trained to produce 'good' solutions in the reward-maximising sense, their ability to comprehend nuance and handle complex scenarios allows them to generate solutions that can surpass those derived from RL strategies. LLMs having demonstrated more effective and adaptable CoAs within simulated wargames (see Figure 7 in Goecks & Waytowich (2024)).

Impacts on Human Performance

Critically, human oversight, as well as a fundamental ethical and legal requirement within military applications, offers significant performance benefits over solely AI-generated solutions. It is therefore critical that the impacts on human performance are understood (Passerini, Aryo, Pasquale, Burcu, & Tentori, 2025). Concerns include the presentation of incorrect or fabricated information, known as hallucinations (Huang, et al., 2023), sensitivity to phrasing (Pezeshkpour & Hruschka, 2023), and lack of critical assessment of human inputs (Sharma, et al., 2023).

Additionally, there are implementation challenges to ensure that user needs are met, with a desire to ensure the technology is human-centric and explainable (Jenia, Maathuis, & Sent, 2024). To achieve this, it is critical that outputs are trusted, transparent, understandable, usable and fair (Haque, Islam, & Mikalef, 2023), with an increasing body of work evaluating AI technology from a Human Factors and cognitive science perspective (e.g. Bertrand et al. (2022), Liao & Vershney (2021) and Ferreira & Monteiro (2020)).

Future military decision-support systems could benefit from incorporating LLM technology, however there is a need to evaluate the impacts from technical and human perspectives. Constructive simulations offer a safe, configurable and repeatable testbed where AI can be assessed against representative scenarios. We integrated a proof-of-concept LLM-based CoAA assistant with our ‘CAE GESI’ constructive simulation and conducted a human participation study to evaluate its impacts in a tactical decision-making scenario, considering:

1. Are LLM-generated CoA recommendations credible and is the AI engaged with?
2. Does AI assistance objectively impact task performance?
3. How is the AI perceived and are there effects on subjective experience of the system?

METHODOLOGY

Simulation & Scenario Overview

A ‘protection of a sensitive site’ scenario was modelled using the ‘CAE GESI’ constructive simulation with an aim for participants to identify and resolve a range of threats. The mission area was a generic secure site with a security perimeter and several entry points, located in an industrial area with waterway access (see Figure 1). Friendly forces consisted of a drone, patrol boat, security guard, armed response unit and emergency services vehicle. We selected ‘CAE GESI’ for its ability to model the scenario and because we had access to its developers to support AI integration.

Military advisors from the UK’s Defence Science and Technology Laboratory (Dstl) helped identify representative threats (including intruders, hostile attacks, false alarms, and environmental hazards) which we then modelled as neutral or hostile entities that appeared at set times and remained active until resolved. Detailed threat information was revealed via a textual ‘event’ message on the display when a friendly unit was in proximity to it, and output in real-time to an event log for use by the CoAA assistant.

Threat investigations could be conducted by any unit, for example the drone might return “*Reports of a suspicious individual loitering close to the site, individual identified, appears to be a lost civilian*”. Participants could then use

this information to decide whether to send additional or different units to deal with the threat, in the example potentially sending a security guard to intercept the trespasser.

When a designated resolution unit arrived at the scene, the system either provided additional information, such as the need for another unit, like an emergency vehicle for medical support, or marked the threat as resolved if all necessary units were present. For example, in the case of a lost civilian, the security guard might report, *"Confirmed lost civilian, provided assistance and they have left the area".* At that point, the system would consider the threat resolved and remove it from the display.

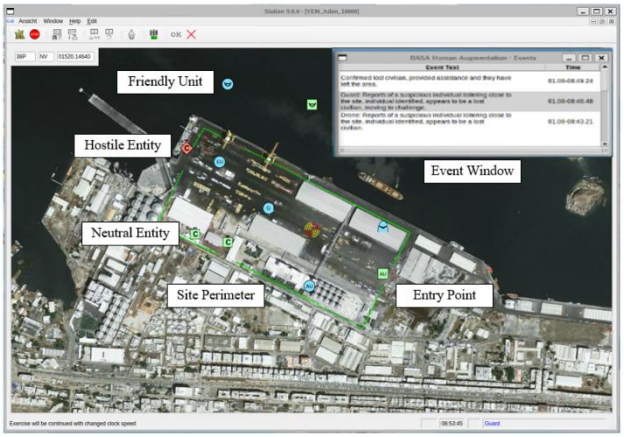

Figure 1: Experimental Software

Course-of-Action-Analysis AI Development

A review of the current LLM landscape was conducted to identify the model on which to develop the CoAA assistant. LLMs are broadly categorised into closed-source and open-source models, with the primary distinction between them being the accessibility of their underlying architecture and model weights:

- **Closed-Source Models:** such as OpenAI's GPT-4 are accessed exclusively through an API with the model weights proprietary and not publicly available. Users cannot host the model locally or modify its internal workings, however they offer state-of-the-art performance in language understanding and generation, benefiting from continuous updates and optimisations. The computational burden is offloaded to the provider reducing local hardware requirements, though incurring ongoing compute costs based on usage.
- **Open-Source Models:** such as Meta's Llama have publicly available model weights and architecture, allowing users to modify and deploy them on local hardware. This facilitates customisation and fine-tuning for specific tasks and enables control over the deployment environment, improving data privacy and security. Although significant computation resources are required to run these models, barriers are reducing as commercial hardware is increasingly targeted to support AI applications.

We decided to use a closed-source LLM, specifically the OpenAI family of models, for several reasons:

- **Performance:** GPT-4 has demonstrated exceptional capabilities in generating high-quality CoAs, as evidenced by the performance in Goecks and Waytowich (2024).
- **Resource Constraints:** the lack of sufficient local hardware to effectively run large open-source models.
- **Accessibility and Licensing:** an enterprise license with OpenAI provided prioritised access to their models, support services, and compliance with enterprise-level security and privacy standards.

The framework used to create, deploy, and host the models was 'Microsoft Azure', with the 'Azure AI Foundry' service, and 'Prompt Flow' interface used to create and deploy the CoAA assistant's AI models, by utilising:

- **Customised System Prompts:** to tailor model's behaviour to specific contexts.
- **Integration with Databases for Retrieval-Augmented Generation (RAG):** to facilitate connection between LLM agents and external databases and knowledge bases. This enabled the models to access and retrieve relevant information from connected data sources in real-time, enhancing the generated responses.

The CoAA assistant used three specialised LLM agents, each responsible for a different part of the decision-support process, with the overall architecture illustrated in Figure 2.

- **CoA Generation (LLM A):** created a full CoA for the entire scenario using a custom prompt that included the mission goals, environment, available resources, unit capabilities, and operational rules. It also received a threat list from GESI at scenario startup and accessed doctrine information via a RAG system linked to open-source U.S. Army training manuals (2025). Although users didn't see this full CoA directly, LLM A shared it with the other agents. We used the o1 model for this task due to its strong reasoning capabilities, accepting its slower response time (~2 minutes) since it only ran once at scenario initialisation.

- **Next Step Generation (LLM B):** translated LLM A's overall plan into specific next steps based on the current scenario state. It monitored GESI's event log, which updated in real time as users interacted with the simulation. Whenever the log changed LLM B generated a new recommendation output (e.g. a suggestion of which unit to send) via a command line interface, with a lockout period used to prevent excessive updates. To ensure fast responses (within ~10 seconds), we selected the lighter o1-mini model for this role.
- **User Interaction (LLM C):** was a conversational agent handling user queries about the CoA. It synthesised responses using information from both LLM A and LLM B, along with the ongoing conversation history. We chose the GPT-4o model for this task, as it allowed us to fine-tune the response consistency by adjusting the model's temperature.

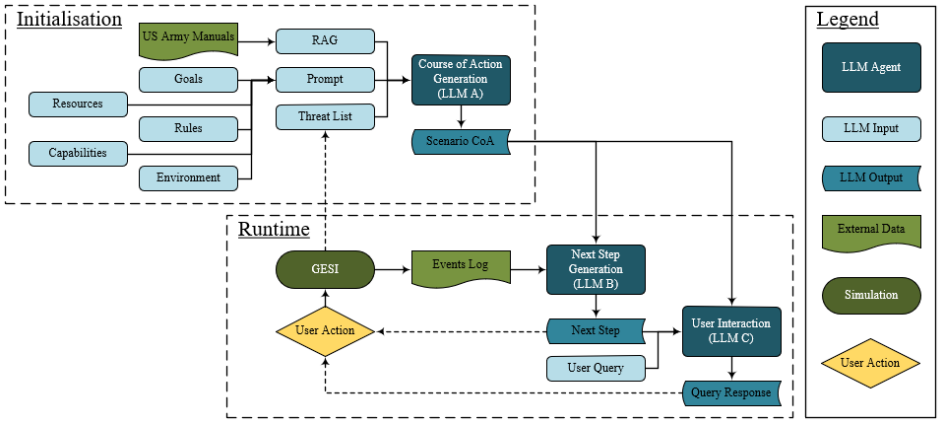

Figure 2: CoAA Assistant Architecture

Research Design

The experiment utilised a 2x2 repeated measures design where the independent variables demand and automation were varied. Demand was adjusted through the frequency, tempo and complexity of threats, low demand consisting of five sequential threat events over a ten-minute period, while high demand consisted of fourteen threats which could occur concurrently requiring a greater degree of resource prioritisation. Automation was determined as whether support from AI was provided or not. This design resulted in four experimental conditions; A. Low Demand (LD); B. High Demand (HD); C. Low Demand AI (LD CoA); D. High Demand AI (HD CoA).

Dependent variables consisted of objective performance based on threat investigation and resolution, attention across Areas of Interest (AoIs) utilising eye-tracking data, subjective experience utilising the NASA-TLX (Hart & Staveland, 1988) and System Usability Scale (SUS; Brooke, 1996) questionnaires, and AI engagement (based on interactions), credibility and trust (utilising a questionnaire adapted from Körber, 2019).

Lab Setup

A meeting room provided a consistent environment and ensured participants weren't disturbed. Participants were seated in front of a 21.5" computer monitor and interacted with the system using a standard keyboard and mouse. The experimental software (see Figure 3) consisted of:

- **GESI:** the constructive simulation used to model each scenario and on which participants could direct friendly units.
- **Events:** displayed threat information as provided by the simulation when units were sent to investigate or resolve threats.
- **CoAA:** the AI which suggested CoAs in automated conditions. In manual conditions the window was still present however the assistant was not active and so did not update.

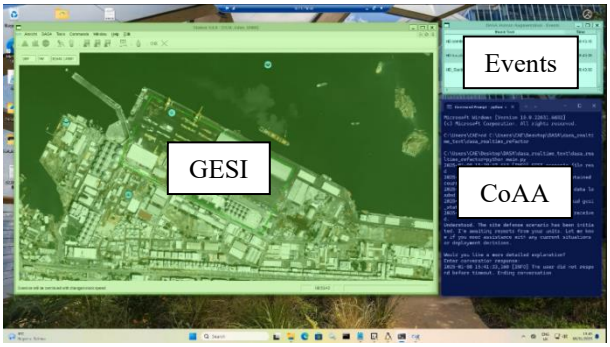

Figure 3: Experimental Software

A researcher station perpendicular to the participant consisted of a laptop running 'Tobii Pro Lab' software to record the participants' screen, control the 'Tobii Pro Fusion' eye tracker mounted below the participant workstation, and determine attention across each AoI.

Participants
Ten CAE (UK) plc Defence & Security employees (5 male, 5 female; mean age 45.2±12.1 years) participated in the study, none reporting having prior military experience.

Procedure
Participants were briefed on the research aims and process and signed a consent form. Age, gender and years of regular or reserve military experience were recorded and the experimental workstation set up. A familiarisation session instructed participants how to direct friendly entities, investigate and resolve threats, and use the CoAA assistant.

The four experimental conditions followed an identical procedure in which the eye tracker was calibrated, the threat scenario undertaken, followed by completion of three subjective questionnaires (NASA-TLX, SUS and automation trust, for AI conditions). The conditions were conducted in a predetermined counterbalanced order using a Latin square to account for learning effects.

After each condition, simulation and AI output logs were taken, with a debrief discussion held on completion. The entire session was completed within 90 minutes for each participant. As a human participation study the experimental protocol was reviewed by Dstl's Scientific Advisory Committee, and a favourable opinion for the protocol obtained from the Ministry of Defence Research Ethics Committee (2327/MODREC/24).

Data Analysis
For each quantitative measure a two-way repeated-measures Analysis of Variance (ANOVA) was conducted to analyse the effect of the independent variables demand (Low & High) and automation (no CoA & CoA). Where significant effects were found the effect size was calculated using omega-squared (ω^2), and two-tailed paired samples t-tests conducted to assess specific differences between conditions, effect size being calculated using Cohen's D (d). For multiple comparisons a Bonferroni correction was applied to the reported p values to determine significance.

RESULTS & DISCUSSION

AI Engagement & Credibility

The technical capabilities of LLM-based AI within simulation have been successfully demonstrated (e.g. Goecks & Waytowich (2024)), however a key question from a human performance perspective is how well users engage with the AI. Our CoAA assistant provided:

1. a suggested next step based on the current scenario state (e.g. *"A suspicious individual has been identified as a lost civilian **Next step:** Dispatch a security guard to ensure their safety")*,
2. the reasoning if prompted to (by describing the source, unit capabilities and objectives etc.),
3. a detailed action plan if requested (e.g. *"Deploy the nearest guard > monitor the situation with the drone until the guard arrives > engage with and assist the individual"*),
4. responses to direct questions typed into the assistant's command line interface (e.g. *"What if no security guards are available? > Deploy the armed guard instead as they have similar capabilities"*).

Users were encouraged to engage with the assistant but were responsible for implementing their own strategies to resolve threats and could ignore the AI if they wished. Engagement was objectively assessed by interrogating the AI chat logs to determine how many responses from LLM C occurred throughout each condition, which only occurred in response to a direct question. Simple yes/no answers to questions posed by the CoAA assistant (e.g. *"Would you like a detailed explanation?"*) and complex queries (i.e. specific questions asked by participants) were considered.

On average participants conducted four simple and one complex interaction per condition. There was no significant difference ($t(9) = -0.18$, $p = .863$) in the number of simple interactions between low demand ($\mu_{LD\ CoA} = 3.8±5.4$) and high demand ($\mu_{HD\ CoA} = 4.0±3.9$) conditions. Similarly, no significant difference ($t(9) = 0.42$, $p = .685$) was observed in the number of complex interactions ($\mu_{LD\ CoA} = 1.0±1.9$; $\mu_{HD\ CoA} = 0.7±1.1$).

This is a relatively low degree of engagement, especially with regard to complex interactions. Regardless of the specific threat, the investigation and resolution strategy remained similar (e.g. send a relevant unit and leave it in place). While this simplification enabled inexperienced participants to undertake the study, it could also have resulted in a task not complex enough to justify interacting with the AI beyond the initial CoA next step provided.

To determine the level of attention given to the CoAA assistant, fixation duration from the eye tracker was assessed (see Figure 4). This provided the amount of time participant's gaze was focused on the CoAA assistant, events window and GESI. Overall participants spent ~20% of their time observing the CoAA assistant which had the effect of reducing time spent monitoring the simulation by ~25% compared to when the CoAA assistant was not available.

In automated conditions the CoAA assistant was observed for a comparable amount of time regardless of the level of demand ($\mu_{LD\ CoA}$ = 108.9±50.2secs, $\mu_{HD\ CoA}$ = 91.8±39.8secs) with the difference not significant ($t(9) = 1.37$; $p = .205$). This is aligned with the comparable level of engagement discussed above. With respect to the 'Events' and 'GESI' AoIs no significant interaction was found between demand and automation, however main effects analysis showed that automation significantly reduced fixation duration when the CoAA assistant was available in both 'Events' ($F(1, 36) = 0.37$, $p_{automation} = .001$, $\omega^2_{automation} = .53$) and 'GESI' AoIs ($F(1, 36) = 0.36$, $p_{automation} = <.001$, $\omega^2_{automation} = .72$).

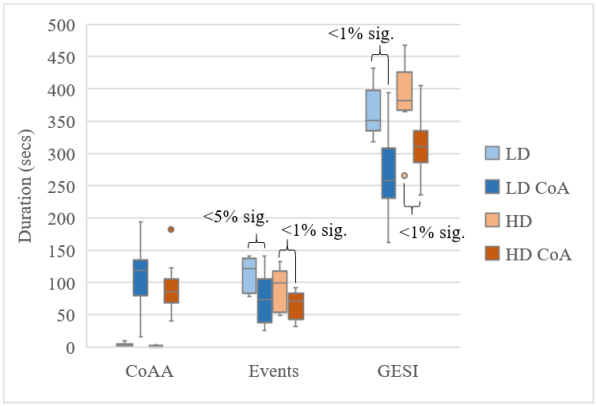

Figure 4: Total fixation duration by AoI

Paired samples t-tests (corrected for four comparisons; see Table 1) showed more time was spent observing the events window when the CoAA assistant wasn't available compared to when it was regardless of whether demand was low or high. Availability of the CoAA assistant also reduced the time spent observing GESI compared to manual conditions when demand was both low and high.

This shows that despite low levels of direct engagement with the CoAA assistant, a significant proportion of participant's attention was given to monitoring it. This was likely impacted by LLM outputs being text-dense, requiring full sentence reading, unrefined user interface design, and indirect simulation integration. Improvements in these areas could minimise the degree of attentional demand required. A key concern is that by distracting users from the tactical situation important information could be delayed or missed. A more focused approach to the definition of AoIs could be taken to quantify whether there is a detrimental effect on situational awareness. This insight could ultimately be incorporated into the AI's capability to enable it to prioritise areas or events which the user is not focused on.

Table 1: Total fixation duration by AoI paired samples t-test significant results

AoI	Demand	μ (No CoA)	μ (CoA)	*t*(9)	*p(corrected)*	*d*
Events	Low	112.7±25.5secs	73.6±38.4secs	4.07	.011	1.22
	High	92.3±31.0secs	64.7±22.4secs	4.35	.007	1.03
GESI	Low	362.5±37.6secs	269.1±65.8secs	5.44	.002	1.81
	High	386.3±53.7secs	312.3±45.7secs	4.42	.007	1.49

Credibility

The CoAA assistant's required attentional demand is not inherently detrimental to performance; so long as it provides credible and actionable responses which aid the user's decision-making. An objective assessment of the LLM's accuracy was not conducted, however the AI logs were subjectively assessed for credibility by the research team. The majority of proposed CoAs appeared to be credible with an appropriate resolution unit and strategy. In response to a hostile intruder threat, for example, the CoAA assistant proposed to:

1. *"Deploy the Armed Unit to engage and neutralise the hostile individual"*
2. *"Utilise the Drone to provide real-time intelligence on the individual's movements and additional threats"*

The assistant demonstrated capability to provide CoAs for multiple ongoing threats concurrently, and also to provide reasonable responses to direct user queries. For example, in response to a question on whether a security guard should

be used instead of an armed unit for a hostile threat it returned; *"Given the severity of the situation with a vehicle crashing through the gate, deploying the Armed Unit is the most appropriate response to intercept and neutralise the intruder. The Security Guard may not be equipped to handle a potentially hostile threat effectively."*

While the majority of the CoAA assistant's responses appeared to be credible, several hallucinations, where incorrect or nonsensical information was presented as accurate, were noted. These are a known issue of LLMs (e.g. Huang, et al. (2023)) and can be detrimental to trust in the systems and impact the validity of any decision made based on the erroneous output. This highlights the need for future systems to incorporate effective verification & validation of CoA outputs (e.g. Bearss (2024)), for example by checking consistency against scenario rules, doctrine and physical constraints (e.g. unit positions). Two notable hallucination examples were:

1. **Fabrication of the location of the main entrance:** *"The main entrance is located on the north side of the site, adjacent to the primary access road."* The CoAA assistant was only prompted with a general description of the site, and it can be seen from Figure 1 that there is a body of water to the north of the site. The AI did however sometimes demonstrate an awareness that it did not have this information, highlighting how LLMs are sensitive to phrasing (Pezeshkpour & Hruschka, 2023) and provide variable responses, representing challenges where consistency is required.
2. **Confirmed additional non-existent resources:** *"Yes, we do have additional Armed Units available."* The AI was directly prompted with the available units (one of each type), however when asked if more could be applied to a threat the AI gave a positive response, demonstrating how LLMs can support human opinions regardless of the context (see Sharma, et al. (2023)) and highlighting the need for users to exhibit a level of critical thinking in reviewing LLM outputs.

Participants were asked to use the AI however they wished to help identify the technology's limitations. This resulted in potential misuses (i.e. employing technology in a manner it hasn't been designed for; see Parasuraman & Riley (1997)) being identified. Specifically, the AI could be used to find out about future scenario events by:

- detailing what an event was before it was investigated based on the trigger message in the event file,
- listing all events in the condition if the user requested more detailed information at the start of a scenario,
- returning what the next event would be if asked explicitly by a user.

This was possible because LLM A utilised the scenario configuration files to generate its initial CoA on system initialisation. The issue was identified during development and its prompt updated to forbid this however the success of this seemed dependent on the underlying GPT model, newer models better understanding temporal factors.

While a real system would not be pre-loaded with scenarios this does highlight the problem that LLMs are capable of inadvertently breaking domain rules despite their underlying prompting if explicitly asked to do so. This could be a serious issue if, for example, a CoAA AI integrated into a constructive training system inadvertently provided trainees with information that they shouldn't have access to. Future systems may need to consider how the prompting is made more robust, or the underlying data federated in order to prevent this (e.g. Ahmadi et al. (2024)).

Task Performance

The effectiveness of the CoAA assistant was evaluated objectively by considering threat outcomes and response times. At the end of each condition threats could have been resolved (i.e. all required units sent, and resolution actions completed), investigated (i.e. at least one unit sent to identify the threat) or unidentified (i.e. no units sent to the threat), the proportion of investigated and resolved threats being shown in Figure 5.

For investigated threats the interaction between demand and automation was significant ($F(1, 36) = 5.34$, $p = .027$, $\omega^2 = .30$), main effects analysis showing that both automation and demand significantly reduced the proportion of threats investigated ($p_{automation} = <.001$, $\omega^2_{automation} = .57$; $p_{demand} = <.001$, $\omega^2_{demand} = .87$).

Paired samples t-tests (corrected for four comparisons; see Table 2) showed that a greater proportion of threats were investigated when demand was low compared to high regardless of whether the CoAA assistant was available or not. In high demand conditions the proportion of threats investigated when the CoAA assistant was available was also significantly lower, however the difference for low demand conditions was not.

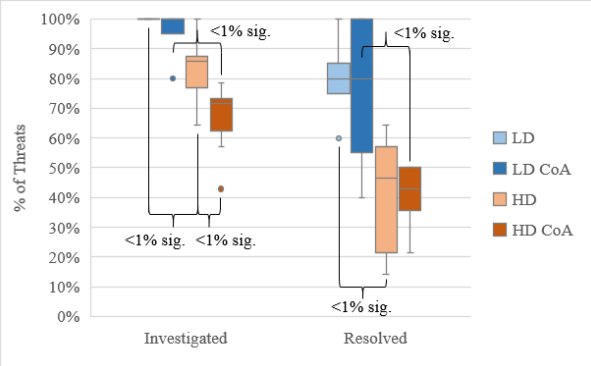

Figure 5: Threat Outcome Percentage

For resolved threats interaction between demand and automation wasn't significant ($F(1, 36) = 0.10$, $p = .753$), main effects analysis showing automation didn't significantly impact the proportion of threats resolved ($p_{automation} = .651$), however demand level did cause a significant impact ($p_{demand} = <.001$, $\omega^2_{demand} = .83$).

Paired samples t-tests showed that a greater proportion of threats were resolved in low compared to high demand conditions regardless of whether the CoAA assistant was available or not.

This shows that demand was the key driver of performance with a lower proportion of threats investigated and resolved when task demand was high regardless of whether AI was available. Having access to CoAs did not translate into improved threat resolutions, suggesting that the recommendations were either unhelpful, not implemented, or task complexity was not sufficient to elicit any benefits from the AI as was observed regarding AI engagement. The reduced investigation rate in automated conditions suggests that the distraction from the simulation, revealed by the eye tracking data, translated to a meaningful performance impact. By requiring participants to monitor an additional text-based system less attention was given to observing threats and implementing responses.

Response times were calculated by interrogating GESI's event log to determine the time between a threat being displayed in the environment to when it was first investigated, and from then until resolved by friendly units. The threat investigation and resolution times are shown in Figure 6.

There was no significant interaction between demand and automation ($F(1, 36) = 0.19$, $p = .663$) for investigation time. Main effects analysis showed that automation had no impact ($p_{automation} = .681$), however demand did cause a significant impact ($p_{demand} = <.001$, $\omega^2_{demand} = .55$).

Paired samples t-tests showed investigation times were faster in low compared to high demand conditions when the CoAA assistant wasn't available, however the difference in automated conditions was not significant.

For resolution time there was no significant interaction between demand and automation ($F(1, 36) = 2.41$, $p = .128$). Main effects analysis showed that neither automation nor demand caused a significant impact ($p_{automation} = .750$; $p_{demand} = .051$).

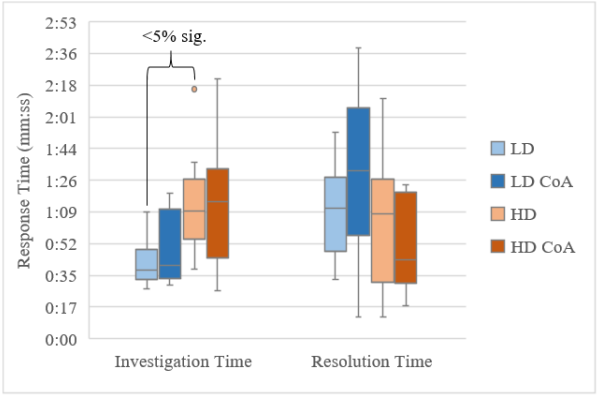

Figure 6: Threat Response Times

Table 2: Task performance paired samples t-test results

Metric	CoAA	μ (Low Demand)	μ (High Demand)	*t*(9)	*p(corrected)*	*d*
Threats Investigated	No	100±0%	83.6±10.1%	6.57	<.001	3.01
	Yes	96±8.4%	67.1±10.8%	5.13	<.001	3.24
	LD vs LD CoA			1.50	.168	N/A
	HD vs HD CoA			4.44	.006	1.57
Threats Resolved	No	80.0±13.3%	42.1±17.3%	7.06	<.001	2.47
	Yes	76.0±22.7%	41.4±8.8%	4.04	.012	2.20
Investigation Time	No	40.7±12.4secs	73.3±27.7secs	-3.46	.029	1.62
	Yes	47.5±18.8secs	73.1±35.4secs	-2.41	.158	N/A
Resolution Time	No	69.1±25.6secs	64.1±35.5secs	N/A	N/A	N/A
	Yes	89.6±46.9secs	50.6±25.4secs	N/A	N/A	N/A

Again, demand was the key performance driver with threats taking longer to investigate in high demand conditions, which given threat density was higher, necessitating prioritisation of friendly resources, was not surprising. The AI did not appear to provide a tangible benefit over manual decision-making, though it didn't inhibit performance. The CoAA assistant didn't have access to entity's positional data, therefore any suggested prioritisation was based on threat descriptions, incorporating this information into the CoAs could allow for more efficient response strategies which could lead to response time improvements, especially where there are concurrent threat events.

Subjective Experience

Workload

Subjective workload was assessed using the NASA-TLX questionnaire (Hart & Staveland, 1988) evaluating perceived mental (1), physical (6) and temporal demand (2), performance (3), effort (4) and frustration (5), see Figure 7. Participants ranked each subscale in order of importance (shown in brackets above, 1 = most important) to determine a weighted average of overall workload.

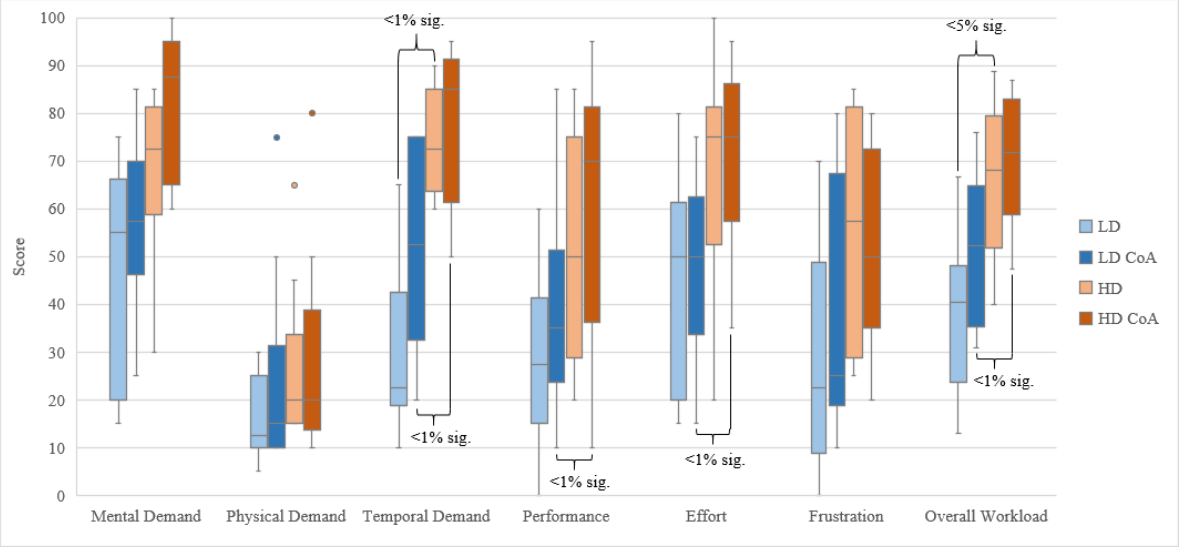

Figure 7: NASA-TLX Scores

Indicatively a similar trend is observed across each subscale and the overall workload, with participants reporting lower scores when demand was low compared to high and without the CoAA assistant versus when it was available at a similar demand level. In terms of the overall workload there was no significant interaction between demand and automation ($F(1, 36) = 0.60$, $p = .445$). Main effects analysis showed automation didn't cause an impact, ($p_{automation} = .087$), however demand increased it significantly ($p_{demand} = <.001$, $\omega^2_{demand} = .69$). Paired samples t-tests (corrected for four comparisons; see Table 3) showed that overall workload was perceived to be lower in low demand conditions compared to high demand regardless of whether the CoAA assistant was available or not.

Table 3: NASA-TLX paired samples t-test significant results

Subscale	Automation	μ (Low Demand)	μ (High Demand)	*t*(9)	*p(corrected)*	*d*
Overall Workload	No	38.6±16.1	65.9±15.9	-3.68	.020	1.71
	Yes	50.9±15.8	70.73±13.4	-6.56	<.001	1.36
Temporal Demand	No	30.5±17.6	73.5±11.1	-5.52	.001	3.00
	Yes	53.0±21.0	77.5±17.0	-3.30	.037	1.29
Performance	Yes	39.5±22.8	60.0±27.8	-3.83	.016	0.81
Effort	Yes	49.0±18.2	72.0±19.0	-9.22	<.001	1.23

There was no interaction between automation and demand for any subscale, with main effects (see Table 4) found for automation in relation to temporal demand, and demand in relation to mental demand, temporal demand, performance, effort and frustration. Paired samples t-tests (see Table 3) revealed that only demand-based comparisons were significant in relation to temporal demand, performance (lower scores indicating better perception) and effort.

Table 4: NASA-TLX subscale significant interactions

Factor	Subscale	$F(1,36)$	p	ω^2
Automation	Temporal Demand	0.39	.019	.34
Demand	Mental Demand	0.09	<.001	.59
	Temporal Demand	0.39	<.001	.79
	Performance	0.03	.006	.43
	Effort	0.04	.002	.49
	Frustration	0.79	.019	.34

A key automation goal is to reduce user's workload, enabling focus on higher value tasks (e.g. Parasuraman & Riley (1997), Ministry of Defence (2022)). Ideally, AI would help users manage greater demand, however these results showed no performance gains with the CoAA assistant. The general trend was in fact that cognitive demand was higher in automated conditions compared to when relying on manual decision-making. Combined with reduced objective performance, this highlights the need to apply LLMs cautiously. Of course, the entire system and the task itself was unfamiliar to participants (i.e. they were novices) therefore the automated conditions imposed additional learning demands and an additional system to monitor which could have limited any potential benefits.

Usability

The System Usability Scale (SUS) was used to assess how usable the system was perceived to be. Participants responding to ten standard questions (e.g. *"I felt confident using the system"*) using a Likert scale with a quantitative measure of usability then calculated from the responses (see Brooke (1996)).

There was no significant interaction between demand and automation ($F(1, 36) = 0.10$, $p = .760$) for system usability. Main effects analysis showed that automation did not have a significant impact ($p_{automation} = .302$), but demand did ($p_{demand} = .029$, $\omega^2_{demand} = .30$). The comparisons between low and high demand were not significant when the Bonferroni correction was applied to paired samples t-tests for both automated ($\mu_{LD\ CoA} = 72.0 \pm 11.0$, $\mu_{HD\ CoA} = 61.5 \pm 14.7$; $t(9) = 2.6$, $p_{corrected} = .116$) and manual conditions ($\mu_{LD} = 75.0 \pm 10.7$, $\mu_{HD} = 67.0 \pm 14.3$, $t(9) = 1.77$, $p = .110$).

The system's experimental nature makes usability assessment academic, especially given the limited attention to interface design and simulation integration. While automation didn't reduce usability, there is a clear need to ensure future applications are properly designed to avoid introducing problems. Future studies may consider using experienced participants and more representative scenarios to better isolate LLM technology's impacts from issues caused by domain unfamiliarity or unrealistic scenario design.

Automation Perception

Participants' perception of the CoAA assistant was assessed using an automation questionnaire (adapted from Körber (2019)). Participants answered eleven standard questions (e.g. *"The system interpreted situations correctly"*) on reliability & competence, understanding & predictability, and trust in automation using a Likert scale. A quantitative measure was calculated similarly to the SUS (see Table 5). Participants rated the CoAA assistant more reliable, predictable and trustworthy in low demand versus high demand conditions, which was reflected in the overall score.

Table 5: Automation perception questionnaire paired samples t-test significant results

Category	μ (Low Demand)	μ (High Demand)	$t(9)$	p	d
Reliability & Competence	79.5±9.0	62.0±15.1	4.34	.002	1.45
Understanding & Predictability	68.8±13.8	54.4±15.3	2.27	.049	0.99
Trust	70.0±12.1	52.5±14.2	2.58	.029	1.33
Overall	60.5±6.0	51.9±7.1	3.39	.008	1.32

While the scores seem quite high across each area it is worth noting that participants did not have any experience of a baseline system against which the CoAA assistant could be compared, therefore the results are best described as a 'gut feel' for how the system is perceived. The results do show a reduction across all scores when demand is increased, which indicates limitations with the CoAA assistant's design. Operational systems will need to be perceived well regardless of the external situation therefore further work is required to refine the AI's implementation. This also highlights the need to understand the system's limitations and the contexts in which it is best used, for example if the AI should be confined to less temporally demanding activities such as slower than real-time training.

CONCLUSION

Rapid AI progress promises significant benefits to defence, with technologies like LLMs enabling partial automation of higher-level cognitive processes such as CoAA. It is hoped that this will lead to operational advantages and address human limitations tested by the complexity and demands of contemporary warfare. While potentially transformative, the technology's limitations and risks must be thoroughly explored before application to a domain where the consequences of errors can be severe.

Integration of a proof-of-concept LLM-based CoAA assistant within the 'CAE GESI' constructive simulation enabled testing within a safe, configurable and repeatable environment. The assistant responded to the real-time state of a 'protection of a sensitive site' scenario, suggested next steps to address threats, and enabled natural language querying of proposed CoAs. A human participation study evaluated research questions relating to AI engagement & credibility, objective impacts on task performance, and effects on user's subjective experience.

The CoAA assistant was able to generate reasonable actions based on its prompting, real-time simulation state and doctrine information. Explicit interaction with the CoAA assistant was low but attentional demand was significant, potentially distracting from the tactical situation. While responses were generally credible, there were instances of hallucination, where incorrect information was generated, and cases where the AI broke prompt instructions, such as by referencing future events, raising concerns about its reliability, especially for safety-critical applications.

Objectively, task performance was comparable across conditions with the AI not eliciting any benefits to threat outcomes or response times, however the relatively low complexity of the experimental task may have limited the need for AI support. Subjectively, task demand drove participants' experience of the system, high task demand being associated with the perception of greater workload, regardless of AI availability, as well as lower system usability and trust in the AI, suggesting its design was not robust enough to bring benefits in the more demanding scenarios.

The CoAA assistant demonstrated that LLMs can support decision-making and effectively integrate with existing simulation platforms. The study highlights how human performance must be considered to realise benefits and address limitations, with a need to evaluate the technology through more realistic use cases, and a larger, more representative, user base. In particular, developing support for complex military or sub-threshold scenarios warranting AI assistance, and potentially leveraging additional types of AI agent to increase functionality. Consideration should also be given to tighter simulation integration, with provision of dynamic data such as unit positions, user interface improvements to limit distraction and improve usability, and verification & validation of outputs to ensure credibility is maintained.

ACKNOWLEDGEMENTS

Thanks to the UK Nuclear Decommissioning Authority for funding the research, Dstl's Human Augmentation team for technical guidance and review, and the wider team including Alexander Vollmer, Andy Ripley, Christian Lagarde, Dave Sexton, Daniel Pabst, Jean-François Delisle, Jonas Leuchtenberger, Máté Koch, Nico Helie and Timo Raff.

REFERENCES

Ahmadi, E., Green, C., Russell, K., Marx, W., & Hill, T. (2024). Are LLMs Too Smart for Their Own Good? *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2024 Proceedings.*

Bearss, M. E. (2024). Evaluating the Trustworthiness of Large Language Models. *2024 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Proceedings.*

Bertrand, A., Belloum, R., Eagan, J. R., & and Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: a systematic review. *roceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 78-91). Oxford, UK: Association for Computing Machinery.

Brooke, J. (1996). SUS: A 'Quick and Dirty' Usability Scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), *Usability Evaluation In Industry.*

Ferreira, J. J., & Monteiro, M. S. (2020). What are people doing about XAI user experience? A survey on AI explainability research and practice. *Design, User Experience, and Usability. Design for Contemporary Interactive Environments: 9th International Conference, DUXU 2020*, (pp. 56-73). Copenhagen, Denmark.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA).* Turin: IEEE.

Goecks, V. G., & Waytowich, N. (2024). COA-GPT: Generative Pre-trained Transformers for Accelerated Course of Action Development in Military Operations. *2024 International Conference on Military Communication and Information Systems (ICMCIS)*, (pp. 1-10). Koblenz, Germany.

Haque, A. B., Islam, A. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change, 186*. doi:10.1016/j.techfore.2022.122120

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-189). North-Holland.

HM Government. (2021). *National AI Strategy.* London: HM Government.

Huang, L., Weijiang, Y., Weitao, M., Weihong, Z., Zhangyin, F., Haotian, W., . . . Ting, L. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems, 43*, 1-55.

Jenia, K., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence, 7*. doi:10.3389/frai.2024.1456486

Körber, M. (2019). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. *Proceedings of the 20th Congress of the Int. Ergonomics Assoc. (IEA 2018).* Springer.

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): from algorithms to user experiences. *arXiv [Preprint]*. doi:10.48550/arXiv.2110.10790

Meeuwissen, M. (2025). Navigating the Realm of Artificial Intelligence in AirC2, Education, Training, Exercise, and Evaluation. *The Journal of the Joint Air Power Competence Centre, 39*.

Ministry of Defence. (2022). *Defence Artificial Intelligence Strategy.* London: HM Government.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 529-533.

Organisation for Economic Co-operation and Development. (2022). *Measuring the environmental impacts of artifical intelligence compute and applications: The AI Footprint.* OECD Publishing.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39*(2), 230-253.

Passerini, A., Aryo, G., Pasquale, M., Burcu, S., & Tentori, K. (2025). Fostering effective hybrid human-LLM reasoning and decision making. *Frontiers in Artificial Intelligence, 7*. doi:10.3389/frai.2024.1464690

Pezeshkpour, P., & Hruschka, E. (2023). LLM Sensitivity to the Order of Operations in Multiple-Choice Questions. *Annual Conf. of the North American Chapter of the Assoc. for Computational Linguistics.*

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., . . . & Perez, E. (2023). Towards Understanding Sycophancy in Language Models. *Int. Conf. on Learning Representations.* Kigali, Rwanda.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*, 755-759.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

United States Army. (2025). *Field Manuals*. Retrieved from Army Publishing Directorate: https://armypubs.army.mil/ProductMaps/PubForm/FM.aspx