Automating Training Analysis through Retrieval Augmented Generation and Hierarchical Reasoning

Taja Hillier

Chief Data and AI Officer, Mission Decisions

Bristol, UK

taja@missiondecision.com

Sally Powling MBE

Head of Product Innovation & Development, Aquila Learning

Oxford, UK

spowling@aquilalearning.com

ABSTRACT

The US Systems Approach to Training (SAT), and its equivalent in the UK (Defence SAT (DSAT)) place a high demand on Training Analysts. The process is highly manual, resulting in thousands of hours of analyst work. This paper describes our approach to increase the use of automation for Training Analysis that can deliver consistent, high quality results as well as a mechanism for validation and verification.

Working with a Learning Lifecycle Management System (LLMS) in service with the Royal Air Force, we have developed a process using securely hosted Large Language Models (LLMs) within a bespoke, agentic, Retrieval Augmented Generation (RAG) architecture to automate parts of the Training Analysis process.

Each step of the automated process is controlled, including text extraction, chunking strategy, retrieval, prompt generation, style and tone, and the output format. Using one model's output as the input for the next model (a hierarchical structure of LLMs), we can generate results that are not possible to achieve with a single step process. This approach allows us to handle higher-level output and focus on more specialised tasks - a methodology known as Hierarchical Reasoning.

The human in the loop is still key to this process. The LLM generated results are fed directly to the LLMS application via an API, along with source references for validation. This allows the analyst to accept the outputs, modify, or reject it altogether - all within the LLMS application. This feedback is then fed back to the Secure Data platform via an API, leading to continuous model adaptation and improvement over time. Using a real world source document of over 2,000 pages, our Agentic AI generated 4,000 duties, tasks and standards - compliant with UK Defence training policy and standards (articulated in JSP822). This work would take a full-time human analyst around 3 months to complete. Our process required 15 minutes of compute time to reduce the analyst workload by 60%-80%.

ABOUT THE AUTHORS

Taja Hillier, Chief Data and AI Officer, co-founder of Mission Decisions. Taja has broad experience in all aspects of Data Intelligence. This has been gained from roles across UK Government, industry consultancies and startups over the past 20 years. Taja has extensive, hands-on expertise of Gen AI, analytics and data science techniques, data mining and best practices to deliver strategic advice in order to drive business improvements in line with data security, compliance, and governance practices aligning it with business objectives, decision making and future growth plans. In addition, Taja brings a deep understanding of digital data and analytics infrastructure transformation for optimal performance. Along with extensive knowledge of data architecture, infrastructure

management and development of scalable and robust analytics solutions, she is able to build strong relationships with clients and stakeholders by communicating technical concepts effectively to non-technical stakeholders.

Sally Powling MBE, Head of Product, Aquila Learning**.** Sally has a Degree in Human Psychology, a Post Graduate Certificate in Education (PGCE) and a Masters in Lifelong Learning. Sally served in the British Army for 23 years in the Educational and Training Services Branch, as one of the Army's Learning and Development SMEs. During this time she worked across all aspects of the Defence Systems Approach to Training process, including training policy, assurance and governance roles and as the Army's lead for learning technology. She received an MBE for her work as the training advisor at the Royal Military Academy Sandhurst (RMAS). She left the Army and joined Aquila Learning in 2022 as the Head of Product for ALaRMS. Her responsibilities include defining the product vision and strategy, determining the product roadmap, exploiting emerging technologies and determining how ALaRMS will align with various international training standards.

INTRODUCTION

In the Defence landscape, the operational readiness of personnel is intrinsically tied to the precision, responsiveness, and adaptability of training systems. The Defence Systems Approach to Training (DSAT) has long served as the cornerstone methodology for aligning Defence training with job, role and mission-specific requirements. However, as Defence organisations grapple with a surge in equipment procurement, increasing complexity of that equipment and associated concept of employment (CONEMP),, opportunities emerge to enhance legacy approaches to training needs analysis (TNA). Current systems employed across military institutions predominantly utilise software that support role analysis with varying amounts of automation, but application of DSAT in a manual and time-intensive manner remains challenging. These constraints hinder the agility required to adapt training strategies compromising speed and placing an unsustainable burden on human analysts.

This paper seeks to demonstrate AI applications, showcasing the design and deployment of hierarchical, Agentic AI architectures that can operate effectively in real-world Defence contexts. Specifically, the paper introduces a bespoke AI-powered system for ALaRMS (a Learning Lifecycle Management Software), although the methodology can augment any software.

The research draws upon recent advancements in General Generative Artificial Intelligence (Gen AI), with a technical focus on Retrieval-Augmented Generation (RAG). To support this, the methodology pays special attention to foundational AI techniques such as chunking and embedding, which are critical for segmenting and encoding large volumes of data in ways that preserve context and relevance.

In moving toward a real-world implementation, the paper proposes a bespoke methodology tailored for Defence TNA. This includes the introduction of hierarchical chunking strategies that reflect the structured nature of DSAT documentation, and the deployment of advanced embedding models alongside a secure data cloud provider’s hybrid search framework to optimise information retrieval. The proposed system is not merely conceptual; it is validated through a systematic implementation process that demonstrates both feasibility and effectiveness in operational contexts. These components are exposed via a secure, scalable API, allowing seamless integration into ALaRMS and facilitating a continuous feedback loop between AI-generated insights and end-user validation.

Subsequent sections discuss the underlying AI principles guiding the design of ALaRMS, including model architecture, decision transparency, and the balance between autonomy and human oversight. The results of implementation are then analysed, showing marked productivity gains, streamlined QA processes, and a significant acceleration in training insights - from an initial output accuracy of 80–90%, improving to 96–99% with iterative fine-tuning.

Finally, the paper explores the wider operational benefits of the system, including its modularity in responding to changes in source documentation and its capacity to incorporate new data sources with minimal disruption. By embedding these capabilities into the DSAT lifecycle, ALaRMS provides a scalable, future-proof solution that enhances the strategic agility of Defence training infrastructures.

In doing so, this work contributes a practical, field-tested roadmap for deploying Agentic AI systems in high-stakes domains, pushing the boundary of what is achievable in military training automation and demonstrating the full potential of AI beyond controlled demos and into mission-critical, real-world deployment.

THE PROBLEM

The landscape of Defence training demands precision, adaptability, and efficiency - yet, the processes that underpin this training often fall short of meeting the pace and complexity of modern military needs. At the heart of the UK Ministry of Defence's (MoD) training development lies the Defence Systems Approach to Training (DSAT), a structured and systematic framework designed to ensure that all Defence training is aligned with operational goals and executed in a timely, appropriate, effective, and efficient manner.

DSAT is robust, offering a comprehensive methodology for analysing, designing, delivering, and assuring training. However, the reality on the ground reveals several pain points in relation to the analysis element which include:

- Labour intensive and time consuming. The conduct of training analysis remains highly manual, labour-intensive, and protracted, consuming significant people-hours which could be spent on other high value activities.
- Human variation. Training Analysts require significant training themselves in the application of DSAT and often vary greatly in their experience, skillset and availability. This results in analysis outputs of varying quality and natural bias.
- Complex high volume data analysis. Training Analysts must navigate reams of data including extensive evolving technical documentation and complex stakeholder feedback.
- Multiple unconnected data entry points. If they don't have a specialist Learning Lifecycle Management Tool, Training Analysts and designers often work in isolation on independent spreadsheets and output documents, requiring duplicitous data entry. They are required to iteratively assure and refine training documentation, often without cohesive digital support or automation.
- Speed of change. The speed and agility with which training is able to change in response to evolving operational needs is limited, leading to potential delays in readiness and reduced efficiency across training pipelines.

The ALaRMS software was developed in response to many of these inefficiencies. Rooted in the same systems approach to training that DSAT champions, ALaRMS enables organisations to manage the entire training lifecycle through a modular and highly configurable platform. It automates many aspects of the DSAT elements (Training Analysis, Design, Delivery, and Evaluation), provides a handrail for Training Analysts and designers, and ensures a single traceable golden data thread runs throughout all four elements. It can be integrated within a wider learning technology ecosystem, e.g. comprising learning management systems (LMS), content management systems (LCMS), content creation and authoring tools and learning experience platforms (LXP). Despite these capabilities, its effectiveness is constrained by the limited decision-making autonomy embedded within the software - placing the cognitive load of analysis, design judgement, and assurance strategies squarely on human operators.

The dependency on human expertise for every decision node slows down the DSAT process and also restricts scalability. The integration of agentic AI into the ALaRMS ecosystem represents a transformative opportunity. By embedding hierarchical AI within the system, it becomes possible to automate complex DSAT tasks such as needs analysis, training gap identification, and curriculum optimisation. The AI can function at various layers - strategic, operational, and tactical - mirroring the layered nature of Defence decision-making itself.

In essence, the core problem is not with DSAT or ALaRMS as frameworks or tools, but with their executional constraints. Both DSAT and ALaRMS are structured, comprehensive, and theoretically sound - but they would significantly benefit from a greater level of intelligent automation. Without Agentic AI integration, Defence training risks stagnation in an era where agility, adaptability, and speed are paramount - it is a capability bottleneck.

Thus, the urgent need emerges: to evolve existing Defence training systems into intelligent, hierarchical, Agentic platforms, capable of transforming static workflows into adaptive, high-performance ecosystems. This transformation is not merely an enhancement - it is a necessity for the future of Defence capability development.

GENERAL AI TECHNICAL METHODOLOGY

The emergence of Gen AI presents a dichotomy of opportunities and challenges in Defence applications. On one hand, it offers the potential to significantly enhance capabilities across various use cases. On the other hand, ensuring the performance, availability, and reliability of AI-driven systems in mission-critical environments poses considerable difficulties. Furthermore, validating the outputs and safeguarding sensitive data and intellectual property (IP) are pressing concerns.

Through collaborative efforts with stakeholders from the Royal Air Force, we have endeavoured to comprehend these challenges and develop a viable approach to deploying generative AI within a secure framework. Specifically, we have integrated a secure Gen AI service into the Learning Lifecycle Management Software (ALaRMS) developed by Aquila Learning, leveraging API services to harness the capabilities of Gen AI tailored to the unique requirements of end-users - all while protecting the information to meet Defence requirements.

The implications of this integration are multifaceted. Notably, it enables users to access high-quality, AI-generated content at an expedited pace, without compromising on accuracy,reliability or security. By employing bespoke Retrieval Augmented Generation (RAG) techniques on reference documents, we can provide a transparent audit trail for Gen AI-generated content, thereby ensuring accountability and dependability. This, in turn, facilitates the creation of scalable, high-quality training solutions that meet stringent assurance standards..

To mitigate potential risks associated with the integration of LLMs, we have implemented a range of security measures, including advanced encryption, access controls, and continuous monitoring. By hosting LLMs within this secure environment, we can ensure the integrity of sensitive data and provide a secure framework for the Defence sector to harness the potential of Gen AI.

Retrieval Augmented Generation (RAG)

In order to generate relevant content utilising Gen AI, it is essential to employ appropriate reference documents to guide the LLM. This process is referred to as Retrieval Augmented Generation (RAG), a technique that enhances the capabilities of pre-existing, pre-trained LLMs with supplementary data sources. The symbiotic relationship between retrieval and generation is a cornerstone approach for systems like ALaRMS.

The fundamental principle of RAG is that, instead of relying solely on the pre-existing knowledge embedded within the parameters of the generative models, the system retrieves pertinent information from an additional knowledge source, thereby augmenting the models generation process. This enables the model to produce more accurate, relevant, and up-to-date responses. It is important to note that when the reference material is removed, the LLM reverts to its original state, as if the additional material was never introduced. This is a crucial step to ensure security and data protection, particularly when dealing with specialised or protected information, such as military content.

The RAG end-to-end workflow can be described as follows:

Question (query): This is provided to the system by the user.
Context: This refers to the processed additional documentation, often chunked for easier handling.
Vector Database: This is the result of the embedding process, where chunked documentation is stored as a vector database.
Nearest Neighbours: This step involves retrieving relevant documents from the additional source, which is the chunked documentation vector database.
Prompt: This is provided to the system by the user or in advance. The prompt must contain explicit instructions with three key elements: priming (e.g., assuming the role of an experienced Role Analyst), style and tone description (e.g., friendly, curious, professional, answering in British English), and instructions on how to handle errors and edge cases.
Enhanced Content: The retrieved documents are then utilised to augment the input (Prompt, Query, and Relevant Chunks) to the generative models.
Generation: The generative model uses the original query, prompt, and the retrieved context to generate a final response.

This process allows for a more accurate and secure generation of content, ensuring that the system's responses are up-to-date and relevant to the user's query. The RAG enhances the capabilities of LLMs in several key ways:

- Access to Supplementary Knowledge: Although LLMs are powerful tools pre-trained on extensive source material, their outputs are constrained by the limitations of their training data. They lack the ability to incorporate domain-specific or recently developed knowledge that was not part of their original training. RAG addresses this limitation by enabling the model to retrieve external, contextually relevant information during inference, thereby augmenting its knowledge base in real time.

- Enhanced Accuracy: RAG architectures improve the accuracy of responses by integrating document retrieval with the generation process. This ensures that generated outputs are grounded in verifiable sources. Such grounding is particularly vital for high-stakes applications like fact-checking, where verifiability and precision are critical. Moreover, because the source documents used in generation are traceable, the transparency and reliability of model outputs are significantly increased.

- Improved Topical Coverage: Pre-trained LLMs may lack information on highly specialised or niche domains - such as those related to national security - especially if such knowledge was absent from their training corpus. RAG mitigates this issue by retrieving relevant documents that contain the required information, thereby extending the model’s coverage. Importantly, this process does not alter or re-train the underlying model, preserving its original architecture while expanding its applicability.

- Reduction in Hallucinations: A persistent challenge in generative modeling is hallucination, wherein the model produces information that appears credible but is factually incorrect. By grounding generation in retrieved, authoritative documents, RAG significantly mitigates this issue. In some cases, hallucination can be further minimised or even eliminated through targeted fine-tuning and optimisation of the retrieval and generation components.

The Critical Role of Chunking in Generative AI-Based RAG Applications

Chunking represents a pivotal factor influencing the quality and reliability of outputs generated by RAG applications built on generative AI systems. While many users may be familiar with uploading documents to tools such as ChatGPT for general analysis, this typically involves default or generic chunking strategies designed to serve broad, non-specialised use cases. Therefore, when it comes to using LLMs that have already built in RAG applications, where the user is not in control, there are potential issues that need to be considered, particularly in a military context. Incorrect chunking can lead to:

- Accuracy issues: If the input to LLM is not chunked correctly, it may not provide accurate or relevant responses. This could lead to misunderstandings or incorrect decisions being made.
- Hallucinations: LLM may extract incorrect chunks to generate responses. This could be particularly problematic in a military context, where accurate information is critical for decision-making.

This means, such an approach - though adequate for general-purpose applications - lacks the specificity and precision required for secure, high-stakes, and domain-specific implementations. Keeping this in mind, highly customised chunking methodologies are essential to achieve accurate, contextually relevant results.

Therefore the precision of chunking strategies are critical to the effectiveness of generative AI-based RAG systems. Tailored chunking not only improves retrieval efficiency but also ensures that outputs are accurate, role-specific, and semantically coherent.

Embedding

In RAG systems, embeddings play a central role in enhancing the accuracy, efficiency, and contextual relevance of generated outputs. Embeddings serve as high-dimensional vector representations of text, capturing the semantic meaning of document segments ("chunks") and user queries. These dense numerical vectors enable the system to identify relationships between linguistic elements that go beyond surface-level word matching, thereby supporting semantic search and improving the quality of retrieval.

In a standard RAG pipeline, embeddings are generated for both the query and the segmented input document. These embeddings are stored in a vector database and compared using similarity measures such as cosine similarity or Euclidean distance to identify the document chunks most semantically aligned with the query. Once relevant chunks are retrieved, they are passed to the generative model as contextual input, thereby grounding its responses in relatively accurate, task-specific information. This architecture significantly reduces the risk of hallucination and increases the factual reliability of generated content.

Key Contributions of Embeddings to the RAG Process

- Semantic Enrichment of Retrieval: embeddings allow for deep semantic understanding by capturing latent textual relationships, enabling the system to retrieve relevant documents even when the query employs synonyms or domain-specific terminology. This is particularly advantageous in technical fields such as military operations, where fine-tuning embedding models on domain-specific corpora further improves the system's ability to interpret specialised jargon and phraseology.
- Contextual Precision in Generation: because the generative model receives semantically relevant information, the quality and coherence of its output are markedly improved. The use of embedding-based retrieval ensures that contextual cues align with the intent of the query, facilitating more accurate and situation-specific responses.
- Scalability and Computational Efficiency:embeddings facilitate scalable document storage and retrieval by reducing complex search tasks to efficient vector operations. This enables rapid querying over large datasets without sacrificing performance or accuracy.

BESPOKE AI METHODOLOGY IMPLEMENTED FOR LLMS APPLICATION (ALaRMS)

The technology stack we are using is all Commercial Off The Shelf. It includes Snowflake for the data management pipeline, data warehouses and serverless GPUs (for flexible and elastic compute for ML and AI), Lambda and AWS Gateway APIs. Over the last 18 months, we have architected, built and deployed a Data Intelligence platform, based on Secure by Design principles, which is accredited to NIST Cybersecurity Framework 2.0 for the British Army. Currently built on Amazon Web Services, but it is cloud agnostic. This means that every data manipulation, and processing required for an AI application is all achieved inside the secure platform, using technology that was not available until we started. This further enhances security.

Hierarchical Chunking

In the present case study, source documentation - in the form of an Aircraft Operational Data Manual - was retrieved via API from the ALaRMS system and securely stored within a dedicated account on our Secure Data Platform. This document, comprising 2,274 pages and incorporating extensive diagrams, images, and tables, was used to conduct a Role Analysis for a Helicopter Pilot position, culminating in the generation of a Role Performance Statement.

To ensure both the fidelity of output and the efficiency of computational resources, a comprehensive preprocessing phase was undertaken. This involved segmenting the document into smaller, semantically coherent chunks, typically at the paragraph, page, or section level. Chunking is particularly critical when processing large-scale documents exceeding 2,000 pages, as it allows for efficient information retrieval without overburdening the system.

As part of the preprocessing pipeline, non-textual elements - such as images and tables - were initially extracted, given that the LLMs employed in this study were optimised for text-based input. Text was then extracted from

images using optical character recognition (OCR) techniques. Although tabular data was not utilised in this specific application, it will be addressed for training design content generation tasks.

Three distinct chunking strategies were employed:

- Section-Level Chunking for Duty Generation: documents were initially divided into large chunks based on their sectional structure. Where necessary, these sections were further segmented into smaller units while preserving semantic integrity. These refined segments were then raised using an LLM, and the resulting summaries formed the basis for generating discrete duties.

- Dynamic Medium-Sized Chunking for Task Generation: the section-level chunks were further divided into medium-sized chunks of variable length, designed to maintain topic coherence. This variability allowed the system to retain contextual continuity while enhancing the granularity required for Task identification.

- Paragraph-Level Chunking for Sub-Tasks, Performance, Conditions and Standards: For the generation of Sub-tasks, Performance, Conditions, and Standards, the documentation was segmented at the paragraph level. This finer level of granularity facilitated a detailed and accurate mapping of responsibilities and performance

Optimising Information Retrieval in RAG Architectures through Embeddings, Fine-Tuning, and Snowflake's Hybrid Search Framework

Therefore, to further optimise the RAG pipeline, we integrated Snowflake's Cortex Search, which introduces a hybrid retrieval architecture combining both vector-based semantic search and keyword-based lexical search. This dual approach enhances retrieval robustness by accounting for both semantic and literal textual similarity. Unlike traditional vector search implementations that require explicit similarity calculations, Cortex Search abstracts this step by internally ranking the relevance of returned results. This significantly streamlines the retrieval process, reducing system complexity and potential points of failure.

Furthermore, we implemented fine-tuning strategies to align Cortex Search with domain-specific vocabularies i.e. military specific keyword jargon, particularly in Defence-sector applications where precise understanding of operational language is essential. This calibration ensures that both the vector and keyword components of the hybrid search mechanism are proficient in domain-relevant terminology, thereby increasing retrieval precision and avoiding unnecessary misalignments.

The integration of embedding techniques, domain-specific fine-tuning, and Snowflake's Cortex Search has yielded substantial performance gains in our bespoke RAG systems. These enhancements have proven especially beneficial in DSAT processes, where the accuracy, traceability, and contextual accuracy of generated content are critical. The resulting system is capable of delivering fact-checked, role-specific outputs with high efficiency and minimal manual intervention.

For the hierarchical structure of the DSAT process we had to construct a hierarchical structure within the outputs of LLMs, we employed a multi-stage architecture in which the output of one model served as the input for subsequent models. The generated outputs of the first model and input to the sequential model were parametritisided, which made it possible to call them individually in the API deployment. This approach enabled the generation of increasingly granular information through a structured, sequential reasoning process. Specifically, we implemented a Pipeline Processing framework, wherein multiple LLMs were executed in sequence to progressively refine and contextualise the outputs.

At the initial stage, section-level chunks of the source document were processed by a summarisation model to reduce complexity and extract key thematic content. The resulting summarisations served as inputs for the subsequent model, which generated high-level role descriptors referred to as duties. This modular structure facilitated a clear mapping between large-scale textual content and operational role descriptions.

Following the generation of duties, we applied a Hierarchical Reasoning methodology, wherein the outputs from one model at a higher level of abstraction were systematically utilised by another model to generate more specialised and detailed outputs. This methodology - also referred to as multi-layered reasoning - mirrors hierarchical decision-making processes and enables LLMs to simulate structured cognitive workflows.

In our implementation algorithm, we employed duty-level chunks as the foundation for generating duties through pipeline processing. These duty-level chunks were then integrated with medium-dynamic, chunk-level inputs in a subsequent hierarchical stage to produce Tasks. These tasks enhanced with paragraph-level chunks were subsequently transmitted to a third-level LLM, which in turn generated sub-tasks, performance , conditions, and standards. This hierarchical structure ensured a logically consistent and contextually coherent output across all layers.

Given the complexity and volume of model-to-model interactions necessitated the development of an automated orchestration framework to govern the hierarchical generation process. This framework leveraged nested loop constructs within the codebase, thereby facilitating the dynamic capture and redirection of parametrised outputs from LLMs as inputs for subsequent processing stages. A crucial aspect of this approach was the rigorous control maintained over output formatting at each level, ensuring compatibility and seamless transitions between successive LLMs. By doing so, the framework significantly enhanced the efficiency, scalability, and reproducibility of hierarchical output generation, while minimising the requirement for human intervention in complex multi-step reasoning tasks.

Validation & Implementation

By imposing this structured governance, the framework substantially enhances the efficiency, scalability, and reproducibility of hierarchical generation workflows, while simultaneously minimising the need for manual intervention in complex reasoning tasks. Within the context of our bespoke RAG systems for Gen AI applications, a key advantage lies in the end-to-end control over each stage of the generation process. This includes fine-tuning parameters such as chunk granularity, embedding i.e. retrieval mechanics, prompt composition, output tone, and response formatting.

Critically, the system maintains a detailed trace of which content chunks were retrieved and utilised by the generative model to produce a given output. This traceability facilitates a robust mechanism for post-generation interpretability and validation. Moreover, the implemented system for AI allows for control over the degree of invention in the generation process, ensuring that outputs can be tuned for creativity or factual precision depending on the use case. These factors empower subject matter experts (SMEs) in the TNA process to perform informed evaluations of model outputs by referencing the original source materials. This capability not only fosters transparency but also enables systematic ‘peer review’ of generative responses.

The validation process is further supported by a structured human-in-the-loop feedback mechanism. When outputs are deemed accurate and contextually appropriate, Training Analysts may approve them for deployment. Conversely, outputs may be revised or rejected to ensure compliance with domain-specific quality standards. All decisions and annotations from this validation step are programmatically relayed back to the data platform through a dedicated API, thereby contributing to a continuous learning feedback loop. Over time, this iterative refinement mechanism yields compounding improvements in system performance, reliability, and domain alignment.

As we forced the generated outputs to be in the format that training management software can seamlessly integrate, the final deployment is relatively easy. We created a parametric API for each level of the generated outputs. This allows ALaRMS software to pull the results into their system and populate within software instantaneously, in a format and presentational style familiar to the users.

AI PRINCIPLES

Whilst implementing our methodology, we developed 6 x AI principles (see Figure 1). These principles are central to ALaRMS’s AI enhancements and directly tackle known concerns with the use of AI to support the Analysis and Design of Defence training activity.

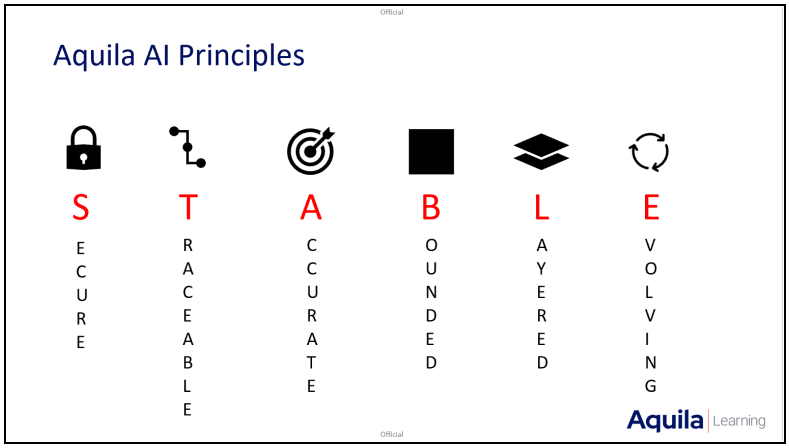

Figure 1 - Aquila AI Principles

Secure. Given the Defence customer, security is at the core of the process and is needed to protect data and functionality. Sharing any training related data in an uncontrolled way on the world wide web is not an option. Aquila Learning is secure by design and ALaRMS includes built-in security, whilst the Data Environment provided by Mission Decisions is also developed along Secure by Design Principles with its data and AI Models hosted within a secure environment.

Traceable. The importance of maintaining an audit trail to verify outputs is a central tenet of the ALaRMS software and therefore any AI outputs must also adhere to this, to ensure trust in data generation. This is achieved via automated references and links to source documents to enable the user to quickly and easily navigate to the relevant section of the document to check the source. These automated checks and balances minimise the verification burden on the user and help build their confidence in the accuracy and validity of the LLM outputs.

Accurate/Authorised. By directing the LLM at specified, secure data sets, we maximise the likelihood that its outputs will be valid and accurate. If we just point the LLM at internet data that can be produced by anyone, there is more likelihood that invalid, erroneous information will be generated. Outputs are also appropriately structured, (configured into the format required by Defence training policy in JSP 822) and are integrated directly into ALaRMS, meaning they are immediately ready for onward use, removing the need for any labour intensive data export or import. The outputs are then subject to native ALaRMS workflows and authorisations which assure and control onward processing of data throughout the rest of the DSAT process.

Bounded. Users are able to define their own parameters for control and specificity, for example the user can define a minimum and maximum number of data elements produced at each level, and whether certain data elements must start with a verb, or whether conditions and standards should apply to an entire section of the analysis or just an individual node. Users can also determine whether they receive all data in one go or whether the output builds iteratively and how similar the degree of invention is to the reference docs.

Layered. A step-by-step approach is adopted enabling, refinement and progressive development of LLM outputs, rather than a 1-hit-wonder. Outputs are developed iteratively (if required), meaning they are more manageable to review, digest and assure. It is possible for the user to gradually build the data picture and refine it incrementally, progressing onto the next step when they are happy, rather than receiving a data-vomit which can be overwhelming and is much harder to assure. The user has the ability to Accept/Modify/Reject all LLM generated outputs, ensuring there is always a human in the loop. Significantly, if a user wishes to modify or reject an LLM generated output, it will be fed back (with reasons) to enable the LLM to learn and improve performance.

Evolving. ALaRMS technology, LLM technology and the customer requirement for AI are all evolving. The core ALaRMS application is the start point but this is constantly evolving, with the Product Roadmap covering both AI and non-AI enhancements. We work with a wide range of current customers, industry experts and wider stakeholders to identify areas where AI will add most benefit. We are not just using open source commercial Gen AI

(like ChatGPT, Gemini or CoPilot). We draw on multiple LLMs appropriate to the reference material requiring analysis and use new LLMs as they emerge. Finally, our process learns over time and gets better the more it is used.

RESULTS

The implementation of our RAG and Hierarchical Reasoning framework has demonstrated substantial productivity gains and operational improvements in the design of training content. Across trials, initial content accuracy - measured against human analyst standards - ranged between 80–90%. Following iterative fine-tuning and feedback-informed retraining, this accuracy increased to 96–99%, markedly reducing the time required for role analysis and Training Needs Analysis (TNA). Analysts now benefit from automated outputs that are not only accurate but also immediately actionable, streamlining quality assurance processes and accelerating insight generation.

A significant outcome of this methodology is the efficiency of resource usage: the automation framework delivers estimated time and labor savings. Integration with Snowflake's Cortex hybrid search, embedding techniques, and military-specific fine-tuning enables robust semantic and lexical retrieval performance. This ensures outputs remain contextually precise even within complex domains such as DSAT. The chunking strategies - sectional, paragraph-level, and dynamic - proved crucial in balancing model load with granularity, facilitating efficient processing of large-scale documents like the 2,274-page helicopter pilot source material.

Whilst out of scope of this purely technical presentation, the human factors considerations for this process should not be underestimated. We worked closely with the training specialists at Aquila Learning to develop the appropriate Human-AI workflow. The Training Analyst is at risk of being overwhelmed, unengaged and lost in the training analysis process if they are fire-hosed with near instant generation of the entire analysis outputs. Therefore, some steps should be necessarily manual, and a continuous gap analysis of 'what is missing' is fundamental to ensure confidence in the end result.

CONCLUSION

This work directly addresses the significant challenges faced in the analysis phase of the DSAT process - namely, its manual and time-consuming nature, dependence on variable human expertise, complexity of high-volume data processing, and the slow responsiveness to evolving operational needs.

This real use case presents a comprehensive enhancement of the RAG pipeline through the integration of hybrid search technologies, hierarchical reasoning structures, and automated operation frameworks. By embedding Snowflake's Cortex Search into the architecture, we achieved a robust hybrid retrieval mechanism that significantly improves both semantic and lexical precision. The inclusion of domain-specific fine-tuning, particularly tailored to Defence-sector jargon, ensures that the system accurately captures and reproduces contextually sensitive information.

The implementation of a multi-stage, hierarchical generation framework further advances the system's ability to model structured reasoning processes. This modular pipeline - progressing from document summarisation to the generation of duties, tasks, and sub-tasks - mirrors real-world decision-making hierarchies, thereby aligning closely with operational and training structures i.e. the DSAT process. By applying this multi-layered interaction through an automated, parametrised framework, we reduced system complexity, increased scalability, and ensured consistent formatting across all stages of generation.

Essentially, this framework also emphasises interpretability and human oversight. Through rigorous traceability of content origins and a structured validation loop involving subject matter experts, the system maintains a high standard of transparency and accountability. This validation feedback is not only instrumental for quality assurance but also sets the foundation for a self-improving system through continuous learning.

Finally, the integration of a parametric API ensures that the outputs are readily digestible by training management software such as ALaRMS, currently in use by RAF enabling immediate operational deployment. While the feedback loop for fine-tuning based on user interaction has yet to be activated, the architecture is fully prepared to support this next phase. This positions the system for ongoing evolution, guided by real-world use and expert feedback, ensuring sustained relevance, precision, and utility in domain-specific applications.

To quantify the benefits of this approach, it is estimated by a current UK Defence training practitioner that it took approximately 6 months for 2 analysts to complete a TNA and turn a 14000 page technical document for a new military vehicle into 3 roles with associated duties, tasks, sub-tasks, conditions and standards. Whilst on average it takes between 6-12 months to complete a non-equipment TNA without AI - applying the estimated savings could see this reduced to 2 weeks. This would provide significant cost savings, free up internal resources to focus on other high value activity, such as training evaluation, and minimise the delay between defining a training need and delivering the training, ultimately expediting force readiness.

This work contributes a scalable, interpretable, and operationally aligned architecture for Gen AI in structured domains, offering a replicable model for future advancements in intelligent training and decision support systems.

REFERENCES

Aquila ALaRMS, Retrieved June 27, 2025, https://aquilalearning.com/product/

Defence Systems Approach to Training, Defence direction and guidance for training and education (JSP 822), Retrieved June 27, 2025, https://www.gov.uk/government/publications/jsp-822-governance-and-management-of-defence-individual-training-education-and-skills

Snowflake Cortex Search, Retrieved June 27, 2025, from https://docs.snowflake.com/en/user-guide/snowflake-cortex/cortex-search/cortex-search-overview