Transforming Technical Documentation into On-Demand Adaptive Training Content

Ernest V. Cross II, Jimena Guallar-Blasco, Matthew Miller, Leonard Eusebi
Charles River Analytics
Cambridge, MA
evcross@cra.com, jguallarblasco@cra.com, mmiller@email.com, leusebi@cra.com

ABSTRACT

In many contexts, trainees need individualized, on-demand content to refresh skills. To provide this capability, we designed an intelligent user interface (IUI) that uses a context-aware system for just-in-time training to convert dense, static technical manuals into dynamic, actionable training content. Traditional technical manuals and training materials are often lengthy and difficult to navigate, making it hard for users to quickly access the critical information they need, whether for routine maintenance or unexpected repairs. Furthermore, existing training materials have been bound by fixed formats and device dependencies, which limits accessibility and flexibility.

This research presents preliminary work on a system designed to overcome the usability constraints of current technical manuals, which impede efficient information retrieval during both standard operations and time-sensitive scenarios. Our system employs computational methods including deep learning, natural language processing techniques, and multimodal artificial intelligence to extract critical features of technical content—including procedural workflows, safety protocols, equipment specifications, and multimedia resources—from heterogeneous documentation repositories. This enables the creation of customized training modules through an integrated multimodel architecture that evaluates task requirements, training content, user expertise, environmental constraints, and device capabilities. This context-aware framework enables the system to deliver personalized training interventions—comprehensive guides for novices, concise refreshers for experienced personnel, and just-in-time support for immediate operational support—optimized for the specific operational setting and context (e.g., noise level, available interaction modalities, time constraints, and other factors).

We will describe our initial demonstration of a capability to systematically structure and remix training content from diverse sources into a standardized, machine-interpretable format suitable for dynamic distribution across multiple technological platforms, including mobile devices and tablets, desktop interfaces, and immersive reality environments. We will also discuss future applications of this work in both defense operations and commercial sectors, such as automotive repair, maintenance services, and safety procedures.

ABOUT THE AUTHORS

E. Vincent Cross II, PhD, is a Senior Research Scientist at Charles River Analytics. His work supports operational readiness across defense and commercial domains, with a focus on applying artificial intelligence–driven solutions to enhance performance, accelerate learning, and support decision-making in dynamic, high-stakes environments.

Jimena Guallar-Blasco is a Scientist II at Charles River Analytics. Her research focuses on applying natural language processing (NLP) and computational semantics to address complex tasks in multimodal real-time and resource-constrained contexts.

Matt Miller is a Senior Software Engineer at Charles River Analytics. He leads the development of the company's in-house component library and oversees full-stack deployment processes using Kubernetes, Helm, and modern DevOps practices. His work ensures that complex research systems are accessible and effective for end users across defense and commercial applications.

Leonard Eusebi is a Senior Scientist at Charles River Analytics. His research focuses on adaptive intelligent training systems, incentive engineering, and intelligent mission support systems.

Transforming Technical Documentation into On-Demand Adaptive Training Content

Ernest V. Cross II, Jimena Guallar-Blasco, Matthew Miller, Leonard Eusebi
Charles River Analytics
Cambridge, MA
evcross@cra.com, jguallarblasco@cra.com, mmiller@cra.com, leusebi@cra.com

INTRODUCTION

The need for fast, accurate, and accessible training content is increasingly critical in both military and nonmilitary operational environments. In modern military contexts, operations are becoming more distributed, time constrained, and technology intensive, driving a shift toward flexible, context-aware, and device-agnostic training solutions. This is especially evident in emerging doctrines such as Agile Combat Employment (ACE) and the rise of Multi-Capable Airmen (MCA), who must be rapidly trained and cross-functional across a range of domains, equipment, and procedures (USAF, 2022). Traditional training methods—relying on lengthy static manuals, scheduled classroom instruction, and prerecorded videos—struggle to meet the real-time demands and adaptability required in such environments (Lytell, 2025). These same challenges are mirrored in commercial sectors such as automotive repair, logistics, and industrial maintenance, where workforce shortages, increasing equipment complexity, and just-in-time service expectations demand faster onboarding skill development and reskilling (Kalejaiye, 2023). For example, within the oil and gas industry, technicians often work in remote or high-risk environments where they must follow intricate and verbose technical procedures under hazardous conditions, often with limited access to expert supervision. Equipment failure or procedural errors can result in significant safety risks, as well as costly downtime, and equipment damage (Sumbal et al., 2017). These parallels highlight a shared need across both commercial and military sectors for flexible training systems that can operate effectively at the point of need, especially in environments where traditional training pipelines are inaccessible or insufficient.

To address these challenges, we are developing a personalized learning assistant that uses an intelligent user interface (IUI; Figure 1) to deliver context-aware JIT training (Dey, 2001; Gil et al., 2016; Hartmann, 2010; Miraoui, 2018). An IUI goes beyond traditional static interfaces by incorporating artificial intelligence (AI) to adapt its behavior, content, and interaction methods based on user needs and situational factors—enabling more intuitive and efficient human-computer interaction (Cao et al., 2023; Chromik & Butz, 2021; Miraz et al., 2021; Ntoa, 2025). Context-aware systems leverage real-time information about the user's environment, task, expertise level, and available resources to automatically customize system responses (Hong et al., 2009). Our IUI extends a context-aware approach to intelligently adapt and remix existing training materials to work across different hardware platforms (e.g., smartphones, tablets, AR headsets, or ruggedized laptops) and within environmental (e.g., high ambient noise), operator (e.g.), and operational constraints (e.g., limited connectivity, hands-free requirements, or compressed task timelines), providing tailored content selection, formatting, and delivery for operators in austere mission environments. As part of our solution, we use an in-house generative AI integration toolkit that utilizes open-source large language models (LLMs) in combination with other machine learning and rule-based systems to automatically deconstruct multimedia training materials—including text documents, videos, and PDFs—into modular data components (text, audio, video, images), enriching them with semantic labels and contextual metadata. This toolkit enables the creation of a complete, offline-capable system that supports context-aware retrieval, summarization, and adaptive content delivery. Using skill-based models and environmental factors, our IUI maps and delivers training content in the most appropriate format for each situation—whether interactive content via smartphone, presentation slides, local work cards, text summaries, audio instructions, or visual aids—ensuring effective training delivery across various interaction modalities available at austere locations. This approach offers several key benefits: it reduces cognitive load by filtering out irrelevant information, improves task efficiency through personalized content delivery, enhances safety by prioritizing critical procedures based on real-time conditions, and increases accessibility by tailoring content to the user's available devices and interaction modalities—ultimately improving both performance and user experience.

The primary limitation of forward operating base (FOB) training approaches is their inability to support “point of need” scenarios, where maintainers and operators require JIT access to knowledge in the field (Vanderpol, 2022). In dynamic expeditionary contexts, waiting for scheduled training or returning to FOB environments for refreshers is not feasible. Additionally, the rigid structure of conventional training fails to accommodate the variability in task complexity, environmental conditions, and hardware availability encountered during field operations.

This paper presents our preliminary work on developing an IUI that leverages context-aware computing to provide point-of-need training by transforming static technical documentation into dynamic, personalized training experiences that fit within operational constraints. Our system addresses three fundamental challenges in technical knowledge transfer: information accessibility, content personalization, and delivery optimization across diverse technological platforms and operational environments.

Figure 1: An intelligent user interface (IUI) uses a context-aware system to provide adaptive training content for multimission Just-in-Time training at the point of need

BACKGROUND

Historically, military technical training has been centralized at large FOBs or continental US-based training centers, where infrastructure, instructor support, and access to full-scale training equipment are readily available. These environments have enabled standardized, instructor-led training programs that follow a set curriculum. Training typically consists of classroom instruction, hands-on labs with operational systems, and simulation-based exercises when physical equipment is limited (Davis, 2021). These models are effective for building foundational and advanced skills across Air Force Specialty Codes (AFSCs), but they are heavily dependent on stable infrastructure, scheduled training rotations, and long lead times. This centralized approach assumes that personnel can be assigned dedicated time for training and that training hardware and subject matter experts (SMEs) are collocated with trainees. However, the evolving demands of modern warfare and contested logistic environments challenge these assumptions. Modern air operations increasingly require that Airmen and Joint personnel be prepared to execute tasks outside their core specialty in austere, distributed, or contested environments—without access to traditional training pipelines (Mills et al., 2020). In dynamic expeditionary contexts, waiting for scheduled training or returning to garrison environments for refreshers is not feasible. Additionally, the rigid structure of conventional training fails to accommodate the variability in task complexity, environmental conditions, and hardware availability encountered during field operations. These limitations have led to growing interest in decentralized, adaptive training systems that can be deployed across diverse platforms (e.g., tablets, AR headsets, or smartphones) and deliver learning tailored to a operator’s role, skill level, environment, and operational context. Our approach aims to fill this gap by providing modular, multimodal, context-aware training at the point of need.

RELATED WORKS

Recent advancements in intelligent training systems, digital job aids, and performance support platforms reflect a growing emphasis on delivering the right instructional content at the right time. In the military domain, several systems have established modular ecosystems for content authoring, credential management, and immersive learning delivery across a range of Air Force training pipelines. While these solutions enable integration with simulation and analytic tools, they focus primarily on predefined courseware and lack real-time, point-of-task content adaptation based on dynamic mission conditions. Similarly, the Army’s Synthetic Training Environment (STE) initiative has made significant strides in creating integrated training environments but remains constrained by the need for preauthored content and centralized infrastructure dependencies (Owens et al., n.d.). These limitations in military training platforms become particularly pronounced when considering the transition from centralized training models to distributed, expeditionary operations has created unprecedented demands for adaptive training systems.

While military platforms struggle with adaptation challenges, commercial solutions in digital work instruction and AR-guided procedure space have demonstrated the value of immersive and remote guidance tools for industrial maintenance. Platforms such as 360Learning, Immerse, and Scope AR have shown promising results in manufacturing and heavy industry contexts. However, these platforms often require content to be manually authored for each use case and generally lack automated mechanisms for repurposing and delivering content based on real-time context, such as device constraints, user skill level, or environmental conditions.

IUIs and context-aware learning systems have demonstrated the value of tailoring instructional content based on real-time situational awareness (Dey, 2001; Hong et al., 2009; Tintarev et al., 2014). These systems extend user modeling, environmental sensing, and adaptive algorithms to personalize learning experiences. However, prior academic and industry efforts have primarily explored mobile delivery, adaptive feedback, and role-based content customization within controlled environments, often relying on persistent connectivity and precurated training pipelines—assumptions that limit their effectiveness in dynamic or disconnected military operational environments. Given these identified limitations across military, commercial, and academic domains, our approach distinguishes itself by supporting the automated remixing of legacy training materials—including PDFs, videos, and slide decks—into semantically tagged, context-aware training content tailored to the user's situation and operational constraints. By embedding semantic reasoning and adaptive content delivery into an offline-capable architecture, our system addresses the unique requirements of military training at the point of need while building on established principles from IUI and context-aware computing. Our approach enables automatic assembly, formatting, and delivery of training content based on task demands, user expertise, environmental conditions, and mission requirements without the connectivity and infrastructure dependencies that limit existing solutions.

DESIGN AND DEVELOPMENT

Multitier Architecture

To meet the operational demands of Agile Combat Employment (ACE), we are extending a hierarchical training approach that dynamically adapts to operator familiarity, environmental conditions, operational constraints, and available equipment. Our system delivers on-demand, adaptive training across a multilevel structure, remixing existing training content such as manuals, videos, and instructional materials into semantically enriched modules tailored to the following context. *Level 3 (extensive pretask training)* provides foundational instruction for personnel preparing to execute unfamiliar tasks. This level includes procedural walkthroughs, safety protocols, and access to supplemental resources and is optimized for environments where time constraints are minimal and full learning resources are available. *Level 2 (contextual refreshers)* delivers targeted knowledge updates for personnel who possess baseline competence but require specific information updates or skill reinforcement. This tier focuses on procedural variations, safety reminders, and context-specific adaptations while assuming foundational knowledge. Content is structured for rapid consumption and immediate application. *Level 1 (interactive, real-time guidance)* provides step-by-step procedural support during task execution, functioning as an intelligent performance support interface. This level is designed for users who are actively performing a task and are using on-the-fly prompts, visual cues, or safety alerts to ensure accuracy while reducing the risk of error. Across all three levels, this adaptive training framework is powered by our IUI which uses semantic reasoning and contextual awareness to remix legacy training content to automatically determine the appropriate instructional level, format, and presentation modality.

To allow for this multilevel training approach within our IUI architecture, shown in Figure 2, we employ a distributed microservices architecture built around a RabbitMQ message bus and standardized common data model. The message bus facilitates asynchronous communication between all system components using publish-subscribe patterns, enabling services to operate independently while maintaining coordinated workflows. All interservice communication is standardized through a common data model that uses both JSON and Protocol Buffers serialization formats, ensuring compatibility across different service implementations while optimizing for both developer experience and production performance. The Persistence Service manages storage of message bus communications in PostgreSQL, providing audit trails and system state recovery capabilities. Data and LMS plug-ins provide standardized interfaces for connecting to external data sources such as data lakes and training platforms.

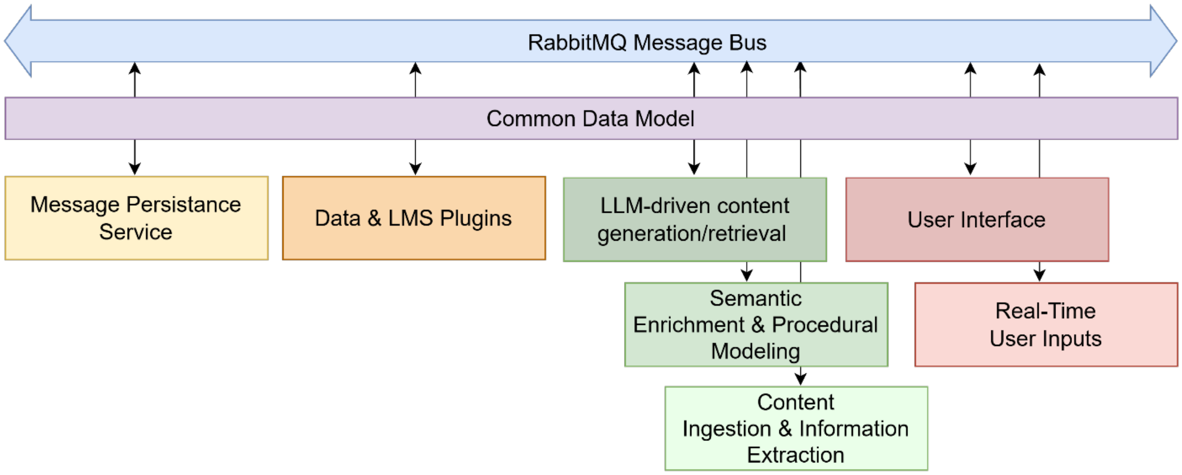

RabbitMQ Message Bus
Common Data Model
Message Persistance Service
Data & LMS Plugins
LLM-driven content generation/retrieval
User Interface
Semantic Enrichment & Procedural Modeling
Real-Time User Inputs
Content Ingestion & Information Extraction

Figure 2: On-demand adaptive training architecture

The core processing pipeline consists of specialized services that transform original training content into intelligent, adaptive JIT training experiences. This processing pipeline includes Content Ingestion, Information Extraction, and Semantic Enrichment services that work in concert to transform unstructured training materials into semantically rich, structured information while LLM-driven Content Generation and Retrieval leverages LLMs for dynamic content creation and intelligent content discovery. Technical implementation details for these content processing and AI components are covered extensively in the next section.

Real-Time Model Inputs handle dynamic data streams including device capabilities, environmental context, network conditions, and UI patterns, enabling the system to adapt training content based on current operational conditions. The UI provides the presentation layer and builds on a responsive template approach for dynamic delivery of training material. This modular, containerized architecture supports deployment flexibility from full enterprise Kubernetes clusters to lightweight development environments, with services that can be developed, tested, and deployed independently while maintaining overall system functionality.

Content Ingestion and Information Extraction

To provide training content across a range of users, environments, tasks, and operational constraints, we first need to ingest and decompose multimedia training materials (e.g., documents, slides, videos, audio recordings, diagrams, and images). These materials are typically stored in formats such as PDFs, PowerPoint files, Word documents, MP4 videos, and MP3 audio, and they often represent decades of accumulated instructional knowledge that remains locked in static, difficult-to-adapt formats. Our ingestion process transforms this heterogeneous input into a structured, machine-readable format that serves as the foundation for delivering adaptive training across all three of the aforementioned instructional levels. To support this transformation, we developed a modular ingestion pipeline capable of extracting three standardized modalities of multimedia content: text, images, and audio. For example, from PDF and Word documents, the system extracts paragraphs, lists, titles, and embedded graphics via a text document processing system uses a layout-aware parsing model to analyze each page's visual structure rather than simply reading raw text line by line. A transformer-based layout-detection engine separates elements like headers, footers, paragraphs, tables, and embedded graphics, enabling precise tagging and semantic enrichment downstream. With this approach the content is parsed while preserving the hierarchical structure of the document (e.g., maintaining the relationship between sections, subsections, and body text). The system then uses transformer-based models to enrich this structure-aware text with additional semantic information, enabling intelligent chunking that aligns with the document's logic and instructional flow rather than arbitrary character counts or page breaks.

PowerPoint presentations are parsed to retrieve slide text, speaker notes, visual elements, and animation references, which are combined and chunked using our text processing module. Instructional videos are processed using lightweight video parsing techniques to extract representative keyframes, while narration and dialogue are transcribed into text using integrated automatic speech recognition models. Audio is similarly transcribed and time-stamped for temporal alignment with visual materials or procedural steps. This multimodal breakdown is essential because different users, environments, and operational context demand different types of content presentation; for example, a novice maintainer may need a visual walk-through, while an experienced user in a noisy environment may benefit from step-by-step text with embedded warnings. As before, the text extracted from video and audio is analyzed via a text processing module, which intelligently determines sections and text hierarchy for best chunking for downstream semantic tagging. By extracting and indexing each modality independently, we ensure that instructional components can be remixed (i.e., filtered, reordered, or emphasized) based on several factors such as the user's profile, the environment and operational constraints.

Each extracted element is converted into a normalized internal format with consistent metadata annotations capturing information such as modality, position in the original source, and instructional intent. Text elements are tagged with headings, step identifiers, or warning classifications; images and figures are annotated with captions and contextual usage; and audio content is aligned with its corresponding visual or procedural reference. This multimodal decomposition allows the system to treat previously siloed training artifacts as interoperable components, enabling flexible reassembly into adaptive training experiences. It allows the system to treat instructional content not as fixed documents or videos but as searchable, combinable building blocks that can be tailored to the needs of individual users, tasks, and operational contexts. For example, a 10-minute narrated maintenance video is broken down into a visual sequence of keyframes, a time-stamped text transcript of each instruction, and a set of modular audio prompts aligned to procedural segments. These pieces can be assembled in different combinations depending on a range of operational and user constraints such as noise level, time pressure, available devices, hands-free requirements, interface limitations, experience level, and/or skill gaps. By transforming legacy multimedia materials into this structured and accessible form, we can expand the utility of existing training resources while ensuring compatibility with low-bandwidth and device-constrained environments.

Semantic Enrichment and Procedural Structuring

After the multimedia training content is ingested and extracted into its constituent elements, we next perform semantic enrichment and procedural modeling to transform the raw instructional material into structured, machine-interpretable training data. The objective of this phase is to extract the meaning, intent, and instructional function of each content element and organize them into coherent task representations that support downstream personalization, adaptation, and delivery. This step bridges the gap between static information and actionable knowledge, enabling the ability to match the training to user needs, task context, and device capabilities. Semantic enrichment involves analyzing textual and audio-transcribed content to identify and classify distinct instructional elements. These include procedural actions (e.g., "connect the hose," "power off the unit"), safety warnings, prerequisite conditions, equipment or tool references, and goal-oriented task statements. Where possible, structural indicators from the original documents—such as bullet points, numbered steps, bold or italicized text, and callout boxes—are also used to guide segmentation and labeling. Each identified element is annotated with semantic tags that describe its instructional function, such as whether it constitutes a standalone step, a conditional instruction, or an auxiliary note. These tags are then linked to metadata about the content modality (text, image, audio), original position in the source material, and any inferred dependencies or coreferences with other elements. For example, a visual showing a tool may be linked to a specific procedural step and annotated with a label indicating its instructional value (e.g., visual cue, demonstration, or reference). This creates a network of interrelated training elements—each aware of its position, purpose, and instructional relationship to other elements.

Ultimately, semantic enrichment allows us to transform loosely organized, static instructional content from the ingestion process into a dynamic knowledge base that is modular, context aware, and ready for adaptive delivery. This phase is essential for enabling the system to deliver tailored guidance that is aligned with the user's role, task demands, environmental constraints, and available interaction modalities. Without this semantic foundation, adaptive training would not be possible at the speed, specificity, and scale that modern operational environments demand.

LLM-Driven Content Generation and Retrieval

Traditional technical manuals present significant barriers to efficient information retrieval during both routine operations and high-pressure scenarios. Static documentation formats cannot adapt to user expertise levels, fail to provide conversational interaction capabilities, and offer no mechanism for real-time content summarization or contextual question answering. These limitations become critical in operational environments where users operating outside of their AFSC need immediate access to procedural explanations, task-specific clarifications, or condensed briefings tailored to their current situation.

The deployment environments we are focused on (e.g., Forward Arming and Refueling Points (FARPs) present unique computational and connectivity constraints that fundamentally shaped our approach to LLM integration. FARPs represent agile, rapidly deployable support nodes providing essential services including recovery, refueling, rearming, launching, and defense operations for aircraft in close proximity to frontlines (Greer, 2022); (Mills et al., 2020). These environments frequently lack access to cloud infrastructure, rely on low-power computing devices, and operate under time constraints where access to the right information can directly impact safety, performance, or mission success. Given these operational realities, we established four critical requirements for our LLM implementation: local deployment capability without Internet dependency, efficient performance on consumer-grade hardware without high-end GPUs, real-time responsiveness to support JIT information needs, and robust integration support with retrieval-augmented generation workflows. Our LLM evaluation focused on open-weight models capable of operating reliably within our target constraints. We established five primary evaluation criteria:

1. Hardware efficiency, which is the ability to run on consumer-grade devices with limited GPU/CPU resources, prioritizing smaller parameter models (≤ 8B parameters) that maintain acceptable performance
2. Inference speed, which involves low-latency response generation suitable for real-time operational environments
3. Task performance, which features strong capabilities across summarization, question answering, and retrieval-augmented generation workflows
4. Integration ecosystem, ensuring compatibility with established frameworks such as LangChain, model-serving tools like Ollama, and retrieval-augmented generation (RAG) architectures to support modular, scalable deployment
5. Licensing, which is open-source availability for both commercial and research applications

Given our emphasis on local deployment, it was critical that the LLMs we evaluated could run efficiently on consumer-grade hardware with limited GPU or CPU resources. This constraint guided our selection toward models that offered smaller parameter sizes (e.g., 8B) without significantly compromising performance. Fast inference speed was also a key factor, as the system is designed to operate in real-time environments where users rely on quick access to critical information. Any noticeable delay in response could be disruptive during critical moments or reduce the system's effectiveness in decision-making contexts. Therefore, we prioritized models and tooling that could deliver low-latency performance even on modest hardware. In addition, we prioritized models with strong community support and robust ecosystems—those offering tools, wrappers, and integration capabilities that reduce development overhead. Compatibility with RAG workflows, especially integration with frameworks like LangChain (Topsakal & Akinci, 2023) or Ollama (Liu et al., 2024), was essential. Ollama is an open-source tool designed to configure, run, and deploy LLMs on local hardware. Finally, an open-source license was a requirement, ensuring the models could be freely used and adapted for both commercial and research applications. This ruled out several proprietary and cloud-dependent solutions and led us to focus on open-source model families such as LLaMA (Touvron et al., 2023, p. 2), Mistral (Jiang et al., 2023) (Jiang et al., 2023), Falcon (Penedo et al., 2023), and Gemma (Team et al., 2024). These models are well regarded for their performance on standard natural language processing (NLP) tasks, and many offer multiple model size variants to balance performance versus compute requirements. We were specifically looking for models that could support a range of NLP tasks including summarization, question answering (QA), and RAG. We evaluated the following open-weight model families against these criteria:

We chose LLaMA 3 as our primary model family for summarization and interactive QA due to its combination of performance, flexibility, and ability to run locally via Ollama. Using Ollama drastically simplified running and managing the LLaMA models locally. Llama 3 models demonstrate state-of-the-art accuracy among open-source models, with LLaMA 3 8B achieving high accuracy on several widely adopted benchmarks, including Massive Multitask Language Understanding (MMLU) and General Purpose Question Answering (GPQA) on which the Llama models rival or exceed other open-source models such as Mistral and Qwen and approach non-open-source models such as Claude and Gemini 1.5 (Touvron et al., 2023). In terms of local generation speed, LLaMa 3 8B via Ollama, which is optimized for local inference, achieves 30–55 tokens/second on consumer GPUs and 4–15 token/seconds on CPUs, depending on the hardware and quantization level. This makes the LLaMa models practical for lightweight local inference even without GPU acceleration. Additionally, Ollama supports a wide range of LLaMA variants, including the latest LLaMA 3, and provides built-in support for tasks like summarization and question answering. Ollama also integrates with RAG (Lewis et al., 2021, p.) workflows, making it ideal for building retrieval-based applications. Furthermore, Ollama is highly portable, which allows for easy packaging of the selected models and runtime dependencies, enabling us to easily deploy across different machines and environments with minimal setup (Grattafiori et al., 2024). This, combined with LLaMA's strong performance and range of model sizes, made it the best fit for addressing the aforementioned challenges. The result is a system that retains the interpretive flexibility of advanced language models while preserving the security, portability, and resilience needed for use in disconnected and mission-critical settings.

Alongside standard generation via prompting, we have implemented a RAG architecture that combines fast, local semantic search with on-device language model inference to support fast, localized content retrieval and real-time, context-aware generation of instructional response. All ingested technical manuals or training materials, after being decomposed and semantically enriched, are processed using a section-aware chunking strategy, which breaks down documents into meaningful units based on content structure and instructional relevance. These chunks are embedded using lightweight sentence transformers and indexed using a Facebook AI Similarity Search (FAISS) vector store (Johnson et al., 2017), enabling low-latency similarity-based retrieval. Index creation is efficient, typically taking less than one second for every 10,000 tokens processed. When a user submits a query, whether through voice, text, or a contextual action, our system performs a semantic search over the indexed content to retrieve the most relevant segments. These are then processed by the locally hosted LLM, which generates a grounded, contextually aware response using the retrieved content as its basis. This approach allows users to request general document summaries, ask natural language questions about specific technical content, or receive output that conforms to predefined operational schema. These schemas can be defined and uploaded by training administrators and used to enforce consistent formatting across responses. In this way, we can deliver accurate, context-aware content while ensuring standardized formatting aligned with operational doctrine—all within a rapid retrieval and generation time of 5–15 seconds, depending on the volume of requested text. If conforming to a predefined schema, the system takes 15–25 seconds depending on the length or complexity of the text or the schema, Additionally, FAISS indexes are reusable across machines, allowing deployments to avoid recomputation when distributing or updating content libraries, which significantly enhances deployment speed and efficiency in distributed field operations.

CONCLUSIONS

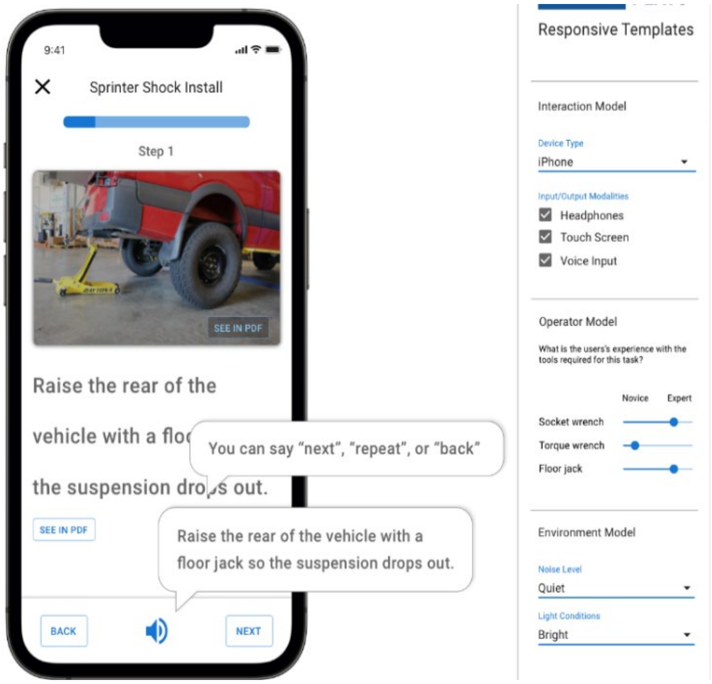
This work addresses a fundamental gap in modern training and operational readiness: the need for fast, personalized, and context-sensitive access to procedural knowledge in environments where conventional classroom training and static manuals fall short. In increasingly distributed, time-constrained, and resource-limited operational settings—particularly those aligned with ACE doctrine—warfighters and maintainers must execute complex tasks without direct access to instructors or full-scale equipment. We are developing an IUI in response to these challenges, offering an intelligent, device-agnostic training solution that adapts instructional content to the user, the task, and the mission context.

Our technical approach integrates content decomposition, semantic enrichment, and local LLM-powered reasoning to transform dense, static technical materials into dynamic training support. Using a hierarchal skill tree framework, we organize procedural, content, user, and environmental models into a unified reasoning engine that drives real-time instructional decision-making. Our adaptive training solution leverages a RAG architecture—based on FAISS indexing and local language models hosted via Ollama—to enable fast, grounded QA and summarization from decomposed training material. This approach ensures that users can access relevant content even in disconnected or austere environments, with delivery tailored to their skill level, situational constraints, and available devices. This approach demonstrates that intelligent content delivery can be achieved without continuous connectivity or specialized hardware. The system delivers meaningful, JIT training and decision support through lightweight, modular components deployable at the tactical edge. By modeling procedures, skills, user states, and device constraints in an integrated framework, the system can present content that is operationally relevant and appropriately scoped—reducing cognitive load, supporting performance, and improving safety.

Ultimately, our solution represents a shift toward modular, AI-enhanced training ecosystems—where instructional content is no longer constrained by static formats, but dynamically structured around mission requirements, user readiness, and operational context. Rather than requiring trainers or instructional designers to manually reformat or rebuild existing content for each scenario, our adaptive training approach enables organizations to leverage their existing static training materials and automatically adapt them for a wide range of use cases. Through intelligent content decomposition, semantic structuring, and real-time reasoning, we are able to transform legacy documents, slide decks, and multimedia into dynamic, situationally aware training experiences. This not only reduces the time, effort, and cost typically associated with content redevelopment, but also significantly increases the accessibility and operational relevance of training resources. By fusing intelligent reasoning with practical delivery constraints, we can empower warfighters, technicians, and training developers alike, ensuring that knowledge is delivered when, where, and how it is most needed.

FUTURE WORK

A major focus of the next phase of development is enhancing its adaptive content delivery capabilities through a flexible, responsive template-based formatting system. Responsive templates are a well-established concept in modern digital interfaces, as is analogous responsive design, where websites adapt their layout and content when accessed on various devices (Voutilainen et al., 2015). This approach allows training content to be dynamically structured and rendered in a platform-agnostic manner, meaning that the same core instructional material can be automatically formatted for smartphones, tablets, desktops, or immersive devices depending on the operational context. Each template specifies not only how content should appear on different devices, but also how it should be sequenced and labeled.



Raise the rear of the vehicle with a floo

the suspension dro s out.

You can say “next”, “repeat”, or “back”

Raise the rear of the vehicle with a floor jack so the suspension drops out.

Responsive Templates

Interaction Model

Device Type: iPhone

Input/Output Modalities
☑ Headphones
☑ Touch Screen
☑ Voice Input

Operator Model

What is the users's experience with the tools required for this task?

	Novice	Expert
Socket wrench		●
Torque wrench	●	
Floor jack		●

Environment Model

Noise Level: Quiet

Light Conditions: Bright

Figure 3: Web-based interface that allows trainers to visualize *remixed* training content prior to deployment

To support trainers and training content creators in leveraging this system without requiring deep technical expertise, we are developing a web-based authoring interface that enables dynamic configuration of training delivery parameters. Figure 3. Through this tool trainers and training content creators can preview the training content for different operational contexts and end-user profiles. For example, SMEs can select a specific device type (e.g., iPhone, tablet, AR headset), enable or disable supported input/output modalities such as voice commands, touch interaction, or headphone audio, and adjust environmental variables such as noise level and lighting conditions. The interface also includes adjustable sliders for tool-specific user expertise, allowing trainers to simulate how content will be presented to a novice or experienced user. These responsive templates let SMEs test multiple delivery configurations without writing code, ensuring that instructional content is not only accurate but also accessible and effective across a wide range of real-world deployment scenarios. Once finalized, the content can be published in a form that adapts automatically to warfighter needs during execution, especially in resource-constrained environments like FARPs.

While the current authoring interface allows trainers and content developers to manually configure operational constraints and preview how training material will appear to different types of users across various devices, these previews are inherently hypothetical. In austere, forward-deployed environments, such as FARPs, trainers rarely have visibility into the warfighter’s actual conditions at the time of training execution. Factors like noise levels, device availability, task urgency, and user expertise may differ significantly from what was anticipated during content creation. As such, to move beyond manually configured previews and enable real-time adaptation for the warfighter, during the next phase we will integrate hierarchal skill models that will serve as the semantic and computational reasoning core of the system. Hierarchal skill models are a structured modeling methodology that represents domain tasks, associated skills, user knowledge, and contextual constraints in a machine-interpretable form. It provides a foundation for reasoning about the instructional requirements of a given task and the conditions under which that task must be performed—whether those conditions involve time pressure, limited visibility, hands-free requirements, or constrained hardware capabilities. These models include procedural relationships, preconditions, performance metrics, and skill linkages, enabling the system to align instructional content with both learner needs and mission realities (Bauchwitz et al., 2018). Integrating hierarchal skill trees will allow our system to dynamically evaluate mission context, user skill level, device constraints, and environmental factors in real time—automatically selecting not just what content to deliver, but how to deliver it, based on encoded task models, content metadata, and user profiles.

Finally, we plan to demonstrate this capability at the 2025 Air Force Northern Strike exercise. This operational demonstration will focus specifically on KC-135 hot pit refueling procedures. The event will provide a valuable opportunity to assess the system's ability to deliver personalized training across varying levels of operator proficiency, environmental constraints, and device configurations in a realistic expeditionary setting. We aim to validate how the system supports all three levels of adaptive training, from extensive pre-task preparation to real-time procedural guidance, while collecting feedback from Airmen and maintainers performing live operations. Lessons learned will guide further refinement of our delivery logic, user interface, and model integration, informing continued development and operational integration across broader Air Force readiness initiatives.

ACKNOWLEDGMENTS

This material is based upon work supported by the Small Business Innovation Research (SBIR) Program through the US Air Force under Contract No. FA238424CB020 with the Air Force Research Laboratory, 711th Human Performance Wing, Wright-Patterson Air Force Base. The authors thank Mr. Stephen McGee for his support and direction on this project. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Air Force.

REFERENCES

Bauchwitz, B. R., Niehaus, J. M., & Weyhrauch, P. W. (2018). Modeling and Comparing Nurse and Physician Trauma Assessment Skills. *Military Medicine*, *183*(suppl_1), 47–54. https://doi.org/10.1093/milmed/usx139

Cao, J., Lam, K.-Y., Lee, L.-H., Liu, X., Hui, P., & Su, X. (2023). Mobile Augmented Reality: User Interfaces, Frameworks, and Intelligence. *ACM Comput. Surv.*, *55*(9), 189:1-189:36. https://doi.org/10.1145/3557999

Chromik, M., & Butz, A. (2021). Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (pp. 619–640). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_36

Davis, J. R. (2021). The Air Force's True Expeditionary Roots: Historical Context and Lessons for the Agile Combat Employment (ACE) Concept. *US Army School for Advanced Military Studies*.

Dey, A. K. (2001). Understanding and Using Context. *Personal and Ubiquitous Computing*, *5*(1), 4–7.

Gil, D., Ferrández, A., Mora-Mora, H., & Peral, J. (2016). Internet of things: A review of surveys based on context aware intelligent services. *Sensors*, *16*(7), 1069.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models* (arXiv:2407.21783). arXiv. https://doi.org/10.48550/arXiv.2407.21783

Greer. (2022, June 16). *Agile Combat Employment: UOTT Takes Part in Lightning carrier concept aboard USS Tripoli*. Air Force Operational Test & Evaluation Center.

Hartmann, M. (2010). *Context-aware intelligent user interfaces for supporting system use* [Doctoral Dissertation]. Technische Universität.

Hong, J., Suh, E., & Kim, S.-J. (2009). Context-aware systems: A literature review and classification. *Expert Systems with Applications*, *36*(4), 8509–8522. https://doi.org/10.1016/j.eswa.2008.10.071

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (arXiv:2310.06825). arXiv. https://doi.org/10.48550/arXiv.2310.06825

Johnson, J., Douze, M., & Jégou, H. (2017). *Billion-scale similarity search with GPUs* (arXiv:1702.08734). arXiv. https://doi.org/10.48550/arXiv.1702.08734

Kalejaiye, P. O. (2023). Addressing shortage of skilled technical workers in the USA: A glimpse for training service providers. *Future Business Journal*, *9*(1), 50. https://doi.org/10.1186/s43093-023-00228-x

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv. https://doi.org/10.48550/arXiv.2005.11401

Liu, F., Kang, Z., & Han, X. (2024). *Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models* (arXiv:2408.05933). arXiv. https://doi.org/10.48550/arXiv.2408.05933

Lytell, M. C. (2025). *Developing combat support mission ready airmen for agile combat employment*. RAND.

Mills, P., Leftwich, J. A., Drew, J. G., Felten, D. P., Girardini, J., Godges, J. P., & Wirth, A. J. (2020). Building Agile Combat Support Competencies to Enable Evolving Adaptive Basing Concepts. *Rand Arroyo Center Santa Monica*.

Miraoui, M. (2018). A context-aware smart classroom for enhanced learning environment. *International Journal on Smart Sensing and Intelligent Systems*, *11*(1), 1–8.

Miraz, M. H., Ali, M., & Excell, P. S. (2021). Adaptive user interfaces and universal usability through plasticity of user interface design. *Computer Science Review*, *40*, 100363. https://doi.org/10.1016/j.cosrev.2021.100363

Ntoa, S. (2025). Usability and User Experience Evaluation in Intelligent Environments: A Review and Reappraisal. *International Journal of Human–Computer Interaction*, *41*(5), 2829–2858. https://doi.org/10.1080/10447318.2024.2394724

Owens, K., Goldberg, B., Robson, R., Hoffman, M., Ray, F., Colburn, A., Hernandez, M., Divovic, M., Blake-Plock, S., & Casey, C. (n.d.). *Synthetic Training Environment Experiential Learning for Readiness (STEEL-R) Demonstration*.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., & Launay, J. (2023). *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only* (arXiv:2306.01116). arXiv. https://doi.org/10.48550/arXiv.2306.01116

Sumbal, M. S., Tsui, E., See-to, E., & Barendrecht, A. (2017). Knowledge retention and aging workforce in the oil and gas industry: A multi perspective study. *Journal of Knowledge Management*, *21*(4), 907–924. https://doi.org/10.1108/JKM-07-2016-0281

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., … Kenealy, K. (2024). *Gemma: Open Models Based on Gemini Research and Technology* (arXiv:2403.08295). arXiv. https://doi.org/10.48550/arXiv.2403.08295

Tintarev, N., O'Donovan, J. T., Brusilovsky, P., & Felfernig, A. (2014). *Proceedings of the Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*.

Topsakal, O., & Akinci, T. C. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences*, *1*(1), 1050–1056. https://doi.org/10.59287/icaens.1127

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

USAF. (2022). *Agille combat employment* (Air Force Doctrine Note 1-21; p. 13). USAF. https://www.doctrine.af.mil/Portals/61/documents/AFDN_1-21/AFDN%201-21%20ACE.pdf

Vanderpol, C. A. (2022). *Prioritizing mobility Air Force airborne information exchange requirements* (Graduate Research Paper AFIT-ENS-MS-22-J-057; p. 41). USAF Air University. https://apps.dtic.mil/sti/trecms/pdf/AD1183387.pdf

Voutilainen, J.-P., Salonen, J., & Mikkonen, T. (2015). On the Design of a Responsive User Interface for a Multi-device Web Service. *2015 2nd ACM International Conference on Mobile Software Engineering and Systems*, 60–63. https://doi.org/10.1109/MobileSoft.2015.16