



EXCERPT FROM THE
PROCEEDINGS
OF THE
TWENTY-THIRD ANNUAL
ACQUISITION RESEARCH SYMPOSIUM AND
INNOVATION SUMMIT

VOLUME III
“ACCELERATING WARFIGHTING CAPABILITIES”

HemingwAI: The Confidence Also Rises

Published: April 30, 2026

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

Disclaimer: The views expressed are those of the author(s) and do not reflect the official policy or position of the Naval Postgraduate School, US Navy, Department of Defense, or the US government.



The research presented in this report was supported by the Acquisition Research Program, Graduate School of Defense Management at the Naval Postgraduate School.

To request defense acquisition research, please contact:

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
E: arp@nps.edu
www.acquisitionresearch.net

Copies of Symposium Proceedings and Presentations; and Acquisition Sponsored Faculty and Student Research Reports and Posters may be printed from the **NPS Defense Acquisition & Innovation Repository** at <https://dair.nps.edu/>.



ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE, AND MANPOWER
NAVAL POSTGRADUATE SCHOOL

HemingwAI: The Confidence Also Rises

Alexandra Adams—holds a Master of Science in Data Science from the University of Virginia, with research experience in Natural Language Processing (NLP) of classical texts and medical imaging. Her published work addresses feature recognition in poorly focused MRI scans for use in Gamma Knife radiosurgery, addressing critical challenges in clinical image quality. She currently works as a Business Development Data Scientist at ASRC Federal, applying data science to key Operational Support initiatives and providing financial insights for executive leadership. Her research interests include NLP and AI-based image identification. [Aladams@asrcfederal.com]

Protima Banerjee—has held technical and leadership positions designing and developing mission critical systems for the DoW and commercial sectors for more than 30 years. Her contributions are part of combat systems delivered to the U.S. Navy, U.S. Coast Guard and associated foreign military sales. Protima is a Technology Fellow at ASRC Federal and an Adjunct Professor of Computer Science at Rowan University. Her research interests include conversational interfaces, semantic AI trust frameworks and real-time systems. She holds a BS in Electrical Engineering from Cornell, an MS in Computer Science from RPI and a doctorate in Computer Science from Drexel University. [PBanerjee@asrcfederal.com]

Theresa Cauble—brings deep expertise in uncovering and translating customer business needs into technical solutions that consistently exceed expectations. Through her leadership, ASRC Federal secured a significant NOAA contract expansion for an effort that began as a 2017 proof of concept and has since evolved into an enterprise-wide financial system. She leads ASRC Federal's applied research projects, including the HemingwAI Trust Framework highlighted in this paper and is committed to driving trusted AI practices across programs. Theresa holds an MS in Computer Science from Columbia University and has a proven track record for delivering innovations as reliable, production-scale solutions. [theresa.cauble@asrcfederal.com]

Noah Pape—is a Solutions Analyst who holds a BS in Information Technology from Oakland University. He currently supports ASRC Federal's Office of the Chief Technology Officer focusing on web applications development for various contracts from across the government space including real-time informational dashboards and decision-making tools. [NPape@asrcfederal.com]

Agam Singh—is completing his BS in Computer Science at the University of Colorado at Boulder. He began working with ASRC Federal in 2025 as a STRIDE Intern. With his passion for innovation, he hopes to continue researching and working with new forms and applications of AI such as HemingwAI. [AgSingh@asrcfederal.com]

Eric Toal—is a Data Science graduate student at Rowan University, where he also earned his BS in Computer Science with President's List honors. As a STRIDE intern at ASRC Federal since 2025, Eric continues to expand his analytical and technical expertise. He is eager to leverage his diverse strengths to support impactful, data-driven decision-making and deliver measurable results. [EToal@asrcfederal.com]

Abstract

As the Department of War (DoW) increasingly adopts Large Language Models (LLMs) to accelerate mission-critical functions, trust in model outputs becomes essential. While LLMs offer significant capability gains, their susceptibility to hallucinations presents an unacceptable risk in high-consequence environments. This paper introduces HemingwAI, an LLM-agnostic, modular trust framework designed to detect hallucinations and quantify hallucination risks. HemingwAI evaluates factual accuracy of LLM responses alongside response completeness, relevance, and subjectivity, producing actionable risk signals to support informed decision-making.

The HemingwAI framework is designed to integrate seamlessly into existing DoW workflows and to support deployment in secure and air-gapped environments. HemingwAI was evaluated using open-domain hallucination benchmarks (HaluBench); benchmark results show measurable improvement in hallucination detection rates above the baseline. More importantly, HemingwAI's operational relevance was evaluated through an internal Subject Matter Expert (SME) guided



pilot. ASRC Federal domain subject matter experts confirmed strong alignment between HemingwAI outputs and human evaluation, demonstrating that the tool can effectively reduce analyst review burden. The work presented in this paper positions HemingwAI as a foundational capability for trusted, mission-ready AI adoption across DoW programs.

Introduction

Large Language Models (LLMs) are transforming data analysis, decision support, and operational intelligence; their ease of use and flexibility makes them valuable for research, planning, and operational tasks. However, LLMs are susceptible to hallucinations, or outputs that are factually incorrect or fabricated. This significant drawback has the potential to fully or partially compromise LLM usability in high-stakes environments.

This paper introduces HemingwAI, an LLM-agnostic, modular trust framework designed to detect and quantify hallucination risks. HemingwAI evaluates responses for factual accuracy, relevance, completeness, and subjectivity, providing indicators to enable confident, informed interpretation of LLM results. By combining advanced AI techniques with rigorous assessment metrics, HemingwAI transforms LLMs from a potential risk into a reliable and auditable tool. Its adoption strengthens confidence in AI-generated insights, enabling streamlined workflows while ensuring artificial intelligence (AI) outputs are trustworthy.

The remainder of this paper is organized as follows: the Problem Statement frames the technical and operational challenges we are addressing. Background and Related Work reviews prior work in hallucination detection and mitigation. Technical Approach describes HemingwAI's microservice architecture and technology stack. Evaluation and Benchmarking presents results from open-domain hallucination datasets. We conclude the paper with an in-depth description of our Subject Matter Expert (SME) guided experimentation pilot and the outcomes of operational validation.

Problem Statement

Despite extensive research on LLM capabilities, current methods for detecting and mitigating hallucinations are limited in scope, often model-specific, and rarely integrated into workflows that require auditable and scalable deployment. As a result, organizations face a tradeoff: leveraging the speed and flexibility of LLMs while accepting the risk of unreliable outputs, or, alternatively, limiting LLM use to low-risk tasks.

There is a clear need for a trust and validation framework that can reliably detect and quantify the risk of hallucinations across diverse models and domains. We characterize this demand signal in the research questions below:

1. **Trust:** What confidence metrics most effectively measure the accuracy of LLM outputs and the risk of hallucinations?
2. **Framework Design:** How can an LLM-agnostic framework be designed to reliably detect and quantify hallucinations across diverse models and domains?
3. **Operational Integration:** How can an LLM trust framework be designed so that it can be securely deployed and scaled in mission-sensitive environments?

In the remainder of this paper, we will show how HemingwAI provides a comprehensive solution to the above questions by detecting and quantifying hallucinations, measuring output trustworthiness, and supporting secure, scalable deployment in operational environments.



Background and Related Work

LLMs have demonstrated strong performance across a wide range of natural language tasks, but their outputs are susceptible to hallucinations, a well-documented limitation in current AI systems (Ji, 2023). In the context of LLMs, hallucinations refer to responses that appear fluent and plausible but are factually incorrect, misleading, or irrelevant to the given context. These errors often arise from incomplete or biased training data, limitations in world knowledge, or insufficient grounding in external sources (Ji, 2023; Maynez, 2020). As a result, LLMs may generate information that was not explicitly learned or supported, undermining trust in operational settings.

Closely related to hallucinations are the concepts of faithfulness and factuality, which are commonly used to evaluate LLM outputs. Faithfulness measures whether generated content is consistent with provided or cited sources, while factuality assesses whether the content is objectively correct or verifiable (Maynez, 2020). These concepts are related but not identical, as an output may faithfully reproduce an inaccurate source or be factually correct without being traceable to a specific reference. Distinguishing between these properties is critical for evaluating reliability in mission-sensitive applications.

Prior research categorizes hallucinations into intrinsic and extrinsic types (Ji, 2023). Intrinsic hallucinations occur when an LLM's output contradicts the input prompt or contextual information, including data provided through Retrieval Augmented Generation (RAG). Such contradictions may arise at the entity, relationship, or sentence level. Extrinsic hallucinations occur when outputs cannot be verified against known facts or trusted sources and often manifest as invented information, subjective claims presented as facts, or unverifiable statements. This taxonomy aligns with the framework proposed in Fine-grained Hallucination Detection and Editing for Language Models (Mishra, 2024).

A variety of techniques have been proposed to detect hallucinations. Traditional measures such as perplexity estimate how well a model's predicted probability distribution matches observed outputs, with higher perplexity sometimes correlated with hallucinated content (Holtzman, 2019). More recent approaches include SelfCheckGPT (Manakul, 2023), which introduces sentence-level metrics such as SelfCheck-BERTScore, SelfCheck-Ngram, and SelfCheck-MQAG to detect hallucinations without external references. Related techniques have been explored in systems such as ChatProtect (Mündler, 2023).

LLM-as-a-judge approaches have also gained traction. G-Eval evaluates responses by prompting an LLM to score outputs against predefined criteria (Liu, 2023), while GPTScore measures factual consistency by conditioning an LLM with task-specific instructions and evaluating perplexity (Fu, 2024). More recently, ChainPoll, developed by Galileo, estimates hallucination likelihood by repeatedly querying an LLM and aggregating its judgments into a probabilistic score, with variants that distinguish between correctness (intrinsic hallucinations) and adherence (extrinsic hallucinations; Friel, 2023).

While these methods provide valuable insights, most operate in isolation, are model-specific, or focus on narrow hallucination categories. They are rarely designed to work together within a unified, operational framework that supports secure deployment, auditability, and scalability. This gap motivates the need for an integrated approach that combines complementary metrics and evaluation techniques. HemingwAI builds on this body of work by unifying multiple hallucination detection methods within a single extensible framework designed for mission-critical use.



Technical Approach

HemingwAI employs an evaluation framework to assess and enhance the reliability of LLM outputs. Rather than relying on a single validation method, HemingwAI applies multiple analytical techniques to detect hallucinations. This approach enables a broad, multi-faceted scrutiny of LLM-generated responses while remaining integrable with existing LLM architectures. The HemingwAI framework allows users to query their existing data using modern LLMs while gaining measurable confidence in the results. As illustrated in Figure 1, HemingwAI validates responses through structured fact-checking and verification processes, presenting scored and explainable results that enable users to quickly assess the trustworthiness of each output before acting upon it.

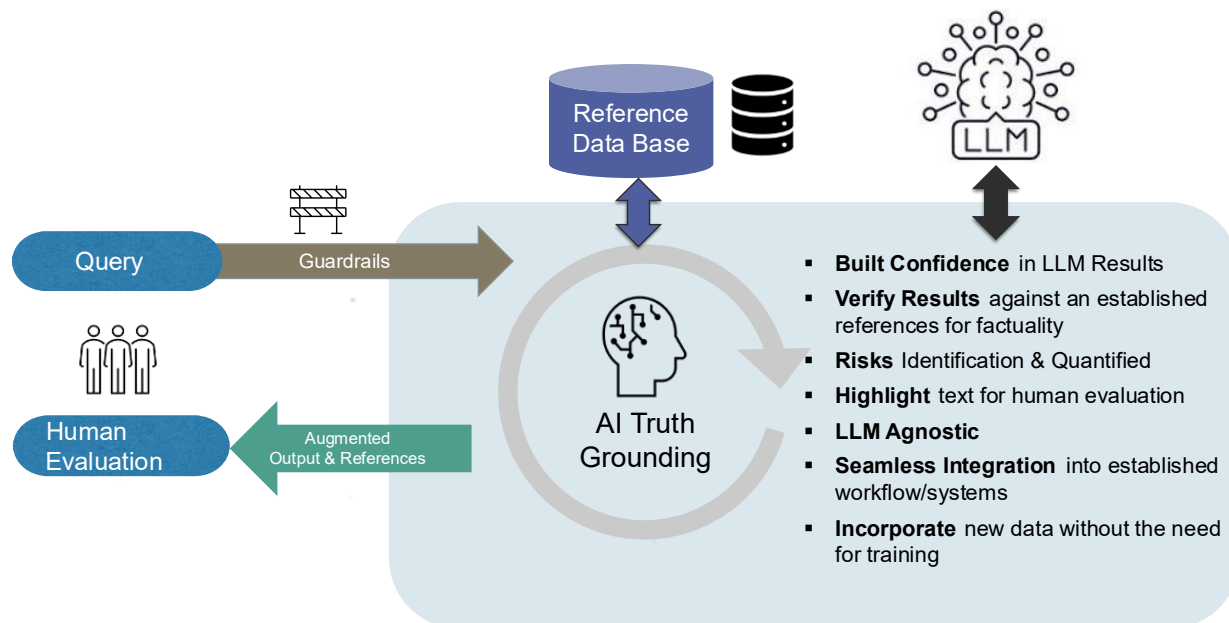


Figure 1. HemingwAI High Level Architecture

By integrating directly with a user's knowledge corpus and wrapping around any chosen LLM, HemingwAI ensures that responses are systematically fact-checked and analyzed for hallucinations. Using both industry-standard metrics and proprietary scoring techniques developed by ASRC Federal, the system applies analyses to deconstruct and validate LLM outputs.

HemingwAI is purpose-built to reduce hallucination risk by returning to the user a marked-up version of the LLM response in which unsupported facts and contradictory claims are explicitly flagged. From inception, the solution has been designed with security, scalability, and operational flexibility as foundational principles. Key discriminators of the HemingwAI solution include:

- **Quantifiable Confidence and Validation:** HemingwAI assigns explainable confidence scores by validating outputs against verified datasets using both industry-standard and customizable benchmarks. Users can tune system performance to prioritize speed, confidence, or an optimized balance of both.
- **Air-Gapped, Secure, and Scalable Deployment:** The platform operates in isolated environments using Department of War (DoW) Iron Bank-hardened containers to preserve data integrity and confidentiality. It scales efficiently within orchestration



platforms such as Kubernetes, supports high availability, and remains fully LLM-agnostic to prevent vendor lock-in.

- **Human-in-the-Loop Evaluation:** Incorporates expert review into the risk identification process, enabling humans to validate, refine, or override flagged unsubstantiated facts and contradictions. This ensures higher reliability, contextual accuracy, and continuous improvement of the system’s risk assessments.

Architecture and Design

HemingwAI’s architecture is composed of modular components that work together to detect hallucinations and perform confidence scoring. This modular design allows HemingwAI to wrap around any LLM, operate in secure or air-gapped environments, and scale to mission needs.

- **Containerization, Scalability, and Security:** HemingwAI is built as a collection of containerized microservices on a hardened Iron Bank container base image, ensuring secure, compliant, and resilient deployment.
- **Guardrails:** HemingwAI employs domain specific textual instructions that enforce security best practices and ensure domain relevance of the response.
- **Reference Database:** HemingwAI incorporates a database layer to support RAG and to manage storage of reference documents to validate LLM outputs. These reference documents serve as a source of truth against which LLM responses can be verified.
- **LLM:** HemingwAI is LLM-agnostic and has been tested with multiple LLMs including OpenAI, Gemini, and multiple open source LLMs running in the Ollama framework. Its default configuration is a locally hosted Llama3.2 model.
- **AI Truth Grounding:** HemingwAI’s fact checking process verifies LLM accuracy by extracting knowledge claim triplets from each response, retrieving references for each claim, and then evaluating the claim triplet against the retrieved references. Each claim is then determined to be supported by the references, contradicted by the references, or unverifiable. The LLM response is then further examined by a Small Language Model (SLM) Judge for additional validation checks including relevance, subjectivity and completeness.
- **Human in the Loop:** HemingwAI returns a report to the user with the details of its truth grounding process presented in a explainable format. This final step enables the user to make more informed decisions on the risk profile of the LLM response. The Human-Machine Interaction (HMI) design of the HemingwAI tool provides an easy-to-use front-end interface. There is no need for AI/machine learning (ML) or software expertise and minimal training required for users.

To support operational flexibility, HemingwAI exposes an API that enables integration into pre-existing enterprise workflows. This allows truth-grounding results and risk assessments to be embedded within other user interfaces or mission systems as appropriate, ensuring that HemingwAI can augment—not disrupt—established processes.

HemingwAI is also designed to incorporate a supervised learning feedback loop based on user evaluations of individual claims (e.g., thumbs up/down assessments). While this represents a planned improvement, it reflects the system’s architecture for continuous improvement. User feedback will be captured and leveraged to refine model performance, strengthen truth-grounding accuracy, and adapt system behavior over time—ensuring that human judgment remains central to model evolution.



Technology Stack

HemingwAI’s deployment architecture consists of three services that collectively host its modular framework. Each service encapsulates a distinct system responsibility within the overall deployment:

- **Knowledge Store**—Provides access to curated domain knowledge to support RAG and enable source verification.
- **HemingwAI Core Service**—Processes user requests, performs AI Truth Grounding workflows, integrates with Knowledge Store to produce LLM output validation results.
- **HemingwAI User Interface**—Enables user engagement with the system, including query submission, visualization of confidence scores, and summary and details of AI Truth Grounding process. Enables retrieval of past queries and their associated AI Truth Grounding results.

Our implementation technology stack is shown in Figure 2.

	Functionality	Technology
RAG Toolset	Embedding Model/Similarity Scoring	Sentence Transformers model all-mpnet-base-v2
	Document Ingestion	LLM Sherpa
	RAG Database	Neo4J Community Edition: DoD Iron Bank Base Image
LLM Toolset	ML Infrastructure Libraries	PyTorch, CUDA
	Inference Engine	Ollama, running locally on AWS compute
	LLMs	Llama3.2 and Phi3.5
REST / Web Interface	Web UI	Python Django
	REST API	Python Flask
	Primary Development Language	Python 3.11: DoD Iron Bank Base Image
Infrastructure	Containerization	Docker
	Hardware	Amazon Web Services (AWS) GovCloud (1 NVIDIA GPUs and 16 vCPUs)

Figure 2. HemingwAI Technology Stack

Evaluation and Benchmarking

We performed benchmarking using the Halubench dataset (Ravi, 2024) which contains approximately 15,000 annotated question–answer pairs. Each data point includes a question, a generated answer, a reference passage representing ground-truth information, and a binary PASS/FAIL label indicating whether the answer contains a hallucination. The PASS/FAIL annotation serves as the ground-truth signal against which all hallucination detection methods are evaluated. To establish a baseline, we first evaluate a standalone LLM, initially Llama 3.2, as a hallucination detector. For each question–answer pair, the model is prompted to determine whether the answer contains hallucinated content. The model’s binary classification output is then compared against the HaluBench ground-truth label to compute a baseline hallucination



detection accuracy score. This framework is model agnostic; Llama 3.2 can be replaced with other LLMs.

We then evaluated HemingwAI’s trust framework approach by ingesting the reference passages from HaluBench into the HemingwAI retrieval system. For each question–answer pair, the system retrieves relevant reference material and evaluates whether the answer is entailed by, contradicted by, or unsupported by the retrieved evidence. These entailment-based judgments are mapped to a binary hallucination classification and compared against the ground-truth PASS/FAIL labels to compute a HemingwAI hallucination detection accuracy score. The primary outcome measure is the comparative performance between the baseline LLM hallucination detection accuracy and the HemingwAI truthfulness score. This comparison quantifies the HemingwAI’s hallucination detection performance over and above a baseline LLM. Although HaluBench contains approximately 15,000 data points, we conducted our experiments on randomized runs of 100 samples to mimic real-world LLM usage scenarios.

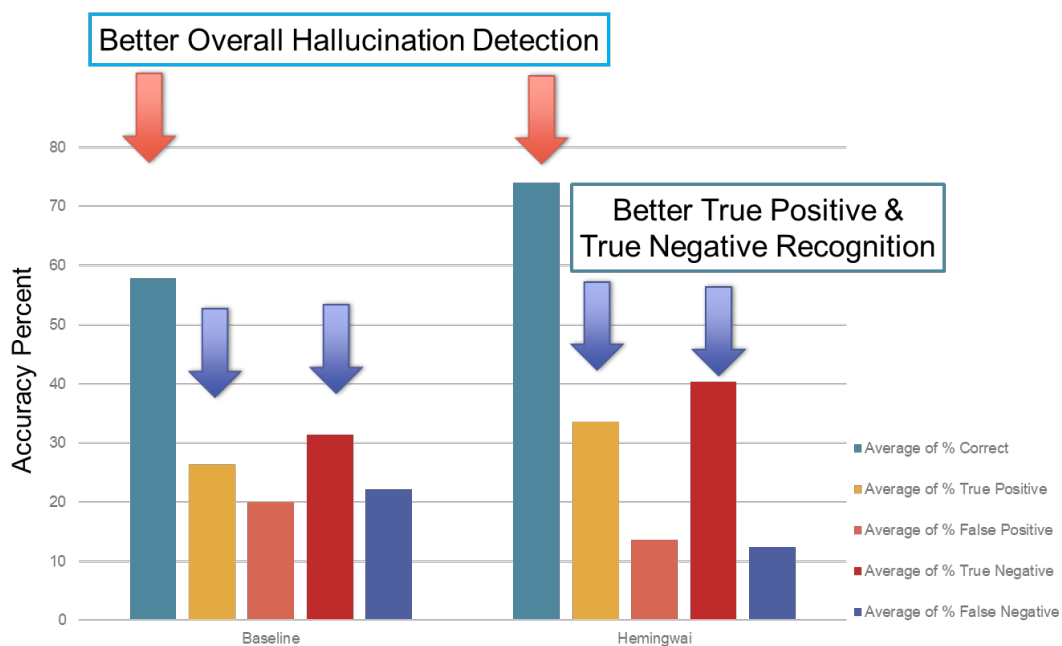


Figure 3. Hallucination Detection Benchmark (HaluBench) Results Against Llama3.2 Baseline

As shown in Figure 3, HemingwAI consistently outperformed the baseline in hallucination detection accuracy. HemingwAI demonstrated an 11%–21% improvement in accurately identifying factually correct (non-hallucinated) and factually incorrect (hallucinated) statements relative to the baseline. This performance gain was stable across randomized evaluation subsets, indicating robustness rather than isolated variance effects. From these results, we extrapolate that HemingwAI can produce a measurably more reliable signal for downstream tasks than a baseline model, thereby enhancing the trustworthiness of outputs used in subsequent decision-making or analytical pipelines compared to LLMs alone.

SME Guided Experimentation

The HemingwAI team engaged with ASRC Federal domain experts early in the development cycle and maintained regular meetings throughout the effort. Domain SME partnership, collaboration, and engagement were foundational to the success of HemingwAI. Our approach emphasized open communication and transparency in sharing findings between



domain SMEs and the development team. Through continuous dialogue, we established a shared understanding; SMEs gained insights into LLM capabilities and limitations, while the development team deepened their knowledge of the domain problem space. This mutual learning fostered alignment on objectives, constraints, terminology, and a common definition of what constitutes a successful outcome.

“When AI talent and military expertise work shoulder-to-shoulder, innovation stops being theoretical and becomes a measurable advantage for the Warfighter. By fusing deep mission understanding with advanced data science, this partnership delivers the accuracy, trust, and impact needed to give our forces the edge in today’s operational environment.” –Rich Smith, ASRC Federal Program Manager

Real-World Data Constraints

The domain use case was built upon a curated corpus of real-world Army doctrines. The documents we used were sourced from openly available materials from the Army Publications directorate (Department of the Army [DA], 2013, 2019a, 2019b, 2019d, 2020, 2021a, 2021b, 2023a, 2025a). Each of these documents is several hundred pages in length and includes dense narrative text, structured tables, diagrams, figures, and embedded imagery. Collectively, this corpus is representative of the structural and semantic complexity of military domain documentation.

While the document corpus provided an authentic foundation, ingesting them into the HemingwAI framework surfaced several practical challenges. The following sections describe the primary challenges encountered, along with the considerations that shaped the solutions.

Challenge 1: Scale and Structural Complexity of Source Documents: A primary real-world challenge involved managing the structure and complexity of the Army Publications documents. Information on a single topic is often distributed across chapters, appendices, annexes, and tables, with key concepts defined in one document and operationalized in another. Supporting authentic user queries therefore requires retrieval workflows capable of consolidating information across documents while preserving context and user intent. Iterative refinement conducted in collaboration with domain SMEs was essential in ensuring that retrieval outputs were not only technically accurate, but also operationally meaningful and aligned across all referenced sources.

Challenge 2: Operational Traceability: A second real-world challenge centered on source document traceability. Initial system outputs correctly identified relevant doctrinal sources but required users to manually navigate lengthy publications to locate supporting language. Feedback from SMEs emphasized that operational users require immediate, precise page-level traceability to efficiently validate responses. To address this, the system was enhanced to return page-specific citations that linked users directly to the exact location within the source document.

Challenge 3: Data Security Considerations: A third challenge arose from the security characteristics of the curated corpus. Portions of the reference documents were partially or fully encrypted, preventing ingestion through the standard document processing pipeline. Excluding these documents would have reduced the operational fidelity of the dataset. Instead, controlled decryption workflows leveraging trusted document management tools were implemented and integrated into the ingestion pipeline in a manner consistent with security requirements. This approach ensured comprehensive corpus coverage while maintaining compliance, reflecting the practical constraints inherent in working with authentic domain documentation.



Testing and Evaluation

Testing and evaluation was led by ASRC Federal domain SMEs, who were granted direct secure user access to the HemingwAI system. These SMEs submitted representative queries derived from their domain expertise and conducted additional ad hoc testing to assess operational utility. This approach ensured that evaluation reflected authentic analytical workflows rather than artificial benchmark scenarios.

Answer

✓ This is a reliable answer with high confidence.

FM 3-0 (MAR25) builds upon the foundation established in FM 3-0 (OCT22), integrating intelligence preparation of the operational environment (IPOE) into multidomain operations. The process of IPOE remains unchanged, with its four steps mirroring those outlined in IPB, as described in ATP 2-01.3. This continuity ensures that the systematic analysis of mission variables, including enemy, terrain, weather, and civil considerations, continues to be a critical component of operational planning.

The integration of IPOE into multidomain operations is facilitated by FM 3-0's emphasis on understanding the complex interplay between domains, including military, economic, social, cultural, and technological factors. This holistic approach recognizes that the operational environment is not solely the domain of the military, but rather a dynamic system influenced by various actors and factors.

FM 3-0 (MAR25) also acknowledges the importance of integrating IPOE into the operational planning process, ensuring that commanders and staff consider the potential effects of their actions on the operational environment. This requires a deep understanding of the complex relationships between mission variables and the ability to anticipate and adapt to changing circumstances.

In contrast to FM 3-0 (OCT22), which may have provided more explicit guidance on integrating IPOE into multidomain operations, FM 3-0 (MAR25) takes a more implicit approach. Instead, it relies on the reader's understanding of the operational environment and its complexities, as well as their ability to apply the principles of IPOE in a multidomain context.

Overall, FM 3-0 (MAR25) represents a significant evolution in the integration of IPOE into multidomain operations, building upon the foundations established in previous editions while acknowledging the complexity and nuance of modern operational environments.

Response Metrics

Confidence Score 100%

Answered in 38.59 seconds

Figure 4. HemingwAI Response to Doctrine Analysis Request

One representative query, shown in Figure 4, asked: “Analyze how FM 3-0 (MAR25) integrates intelligence preparation of the operational environment (IPOE) into multidomain operations compared to FM 3-0 (OCT22).” The resulting output, shown in the diagram, includes fact verification that was confirmed to be correct by the SME team. The HemingwAI system verified that each claim in the LLM’s response could be traced to the source documents; these are highlighted in green. Clicking on a green highlight directs the user to the reference material substantiating the claim, showing how HemingwAI links its verification to authoritative sources. Since all claims in the LLM response could be verified, HemingwAI assigned a confidence score of 100% to the LLM response.

To illustrate the role of guardrails within HemingwAI, Figure 5 depicts how the system appropriately handles queries that fall outside its validated knowledge scope. When a user submits a prompt that exceeds the system’s policy constraints, the guardrail layer intervenes and returns a structured notification indicating insufficient knowledge, lack of verified sources, or domain misalignment. This controlled behavior ensures that the system does not fabricate information or respond beyond its validated expertise, reinforcing trustworthiness and operational safety while maintaining transparency with the user



The screenshot displays a user interface for HemingAI Guardrails. At the top, a purple banner reads "No fact checking was done on this query, proceed with extreme caution." Below this, the AI response states: "I'm not able to provide an answer to this question as it falls outside of the domain of Army Doctrine. The query pertains to baking a cake, which is a culinary activity unrelated to military operations or doctrine. I can only provide information and guidance within the scope of Army Doctrine and related publications. If you have a question about a specific aspect of Army Doctrine or operations, I would be happy to try and assist you." Underneath the response is a "Response Metrics" section showing a "Confidence Score" of 0%. At the bottom of the response area, it says "Answered in 46.47 seconds." and includes icons for refresh, close, share, and copy. Below the response is a "Query" section with a text input field containing "How do you bake a cake?" and a green "Send" button. The bottom left corner features the "ASRC FEDERAL" logo and the text "Press Ctrl+Enter to send".

Figure 5. Example of HemingAI Guardrails Intercepting an Out-of-Scope Query



The screenshot displays an AI-generated answer with a yellow warning banner: "This answer may not be fully reliable. Please verify important information." The answer text states that according to Army Doctrine, a Productive Intelligence Requirement (PIR) is not always directly connected to a commander's decision point. A red arrow points from the circled term "Productive Intelligence Requirement (PIR)" in the answer to a reference box. The reference box, titled "Reference for: 'Productive Intelligence Requirement (PIR)'" and marked as a "Contradiction", provides document details (JP 3-0, Page 82, Confidence 93.3%), a complete claim, and a reference text. The reference text defines PIRs as tools for intelligence collection and analysis. A second red arrow points from the reference text to a specific paragraph in a source document titled "3. CCIRs", which defines PIRs as elements of information critical to decision-making and lists categories like Priority Intelligence Requirements (PIRs) and Friendly Force Information Requirements (FFIRs).

Figure 6. Reference Linking Directly to Individual Pages within a Source

As noted in the Real-World Data Constraints, it is essential not only to cite a source document but to link directly to the specific page or paragraph that substantiates a claim. Page-level traceability ensures transparency, enables rapid verification, and reduces ambiguity, especially when a single term might have multiple interpretations. As a case in point, the example shown in Figure 6 was drawn from a test conducted during a live system demonstration. During the session, the LLM response incorrectly expanded the acronym PIR as *Productive Intelligence Requirement*. HemingAI identified this as a contradiction (highlighted in yellow) and directed the user to the exact paragraph in the source document containing the correct definition, *Priority Intelligence Requirements*. By linking directly to the precise location of



the authoritative definition, the system demonstrated its ability to detect errors in generated content and immediately ground corrections in verified sources.

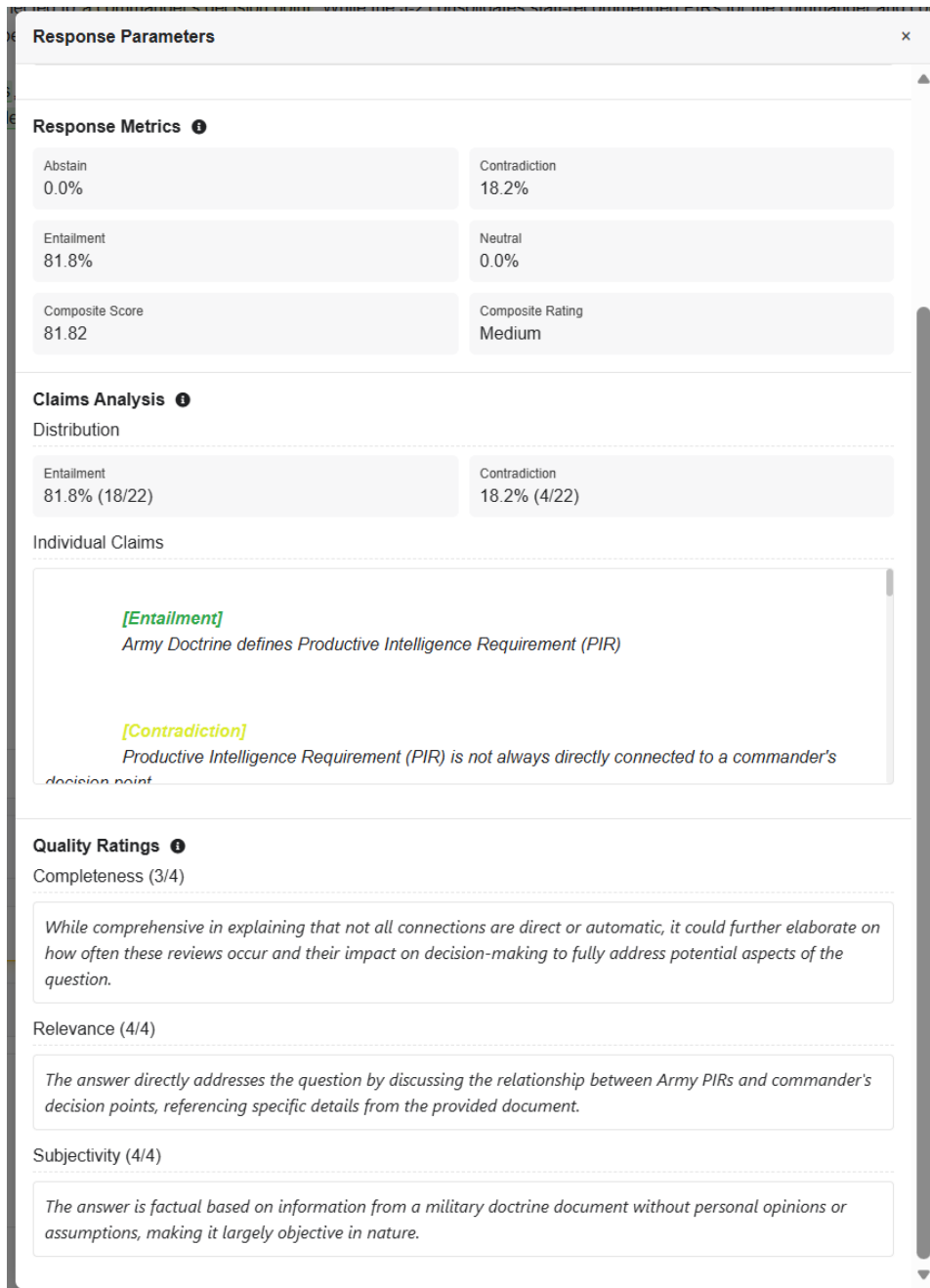


Figure 7. Advanced Trust Framework Metrics with Claim Level Verification Scores and Response Quality Metrics

For more advanced users, Figure 7 presents the detailed LLM response assessment generated by HemingwAI. The interface decomposes the response into discrete claims and evaluates each against the source corpus, categorizing them as entailed, contradicted, or unsubstantiated (neutral), with aggregate percentages summarizing overall factual alignment. In

the response above, since only 18 of the 22 total claims in the response could be verified, HemingwAI indicates a Confidence Score of 18/22 or 81.8%.

In addition to claim-level verification, the display includes Quality Ratings (judgement scores) assessing completeness, relevance, and subjectivity. Positioned alongside the factuality metrics, these evaluations offer a multidimensional assessment of overall response quality.

Overall, the outcome of this effort demonstrates the value of aligning system development with real-world constraints, including document scale, varied internal document content types, and security requirements. Feedback from domain experts highlighted improvements meaningful to their use case, particularly in the area of enhanced reference traceability. These insights and lessons learned emphasize the importance of early SME engagement, iterative refinement, and designing pipelines capable of operating effectively within the realities of DoW environments.

Summary

HemingwAI represents a decisive step forward in making LLMs mission-ready for DoW applications. Rather than asking operators and analysts to simply “trust” AI outputs, HemingwAI provides measurable, explainable, and operationally grounded verification that supports faster, more confident analyst workflows. Its LLM-agnostic, modular architecture allows it to integrate seamlessly into existing workflows while remaining adaptable to evolving model ecosystems. By evaluating factual accuracy alongside completeness, relevance, and subjectivity, HemingwAI transforms raw model output into structured, actionable risk signals enabling informed human judgment.

What differentiates HemingwAI is its ability to operate as a verification and accountability layer over any LLM. Through page-level traceability to authoritative source documents, contradiction and entailment detection, and transparent confidence scoring, the framework ensures that every substantive claim can be audited against trusted references. This capability directly addresses one of the most significant barriers to AI adoption in high-consequence environments: the inability to verify outputs quickly and confidently. In secure and air-gapped settings, where external validation is not an option, HemingwAI provides a self-contained trust infrastructure that strengthens analyst confidence while reducing review burden.

Acknowledgements

The authors would like to acknowledge the outstanding support provided by the ASRC Federal domain experts. In particular, we extend our sincere appreciation to Richard Smith and Kirk McIntosh for their leadership, guidance, and sustained engagement throughout the effort. We also gratefully recognize Andrew Jackson and Kevin Browning for their hands-on support during testing activities, as well as Richard Smith’s direct involvement in those efforts. Richard Smith’s deep subject matter expertise and provision of critical domain data were instrumental to the success of this work. The collaboration and commitment of these domain SMEs was invaluable in achieving both the technical rigor and operational relevance of the HemingwAI solution.

The authors also wish to thank and recognize Dr. Eric Adams for championing the foundational concept behind HemingwAI. His early vision and leadership were instrumental in shaping what has become the HemingwAI framework described here.

References

Department of the Army. (2013). *Information Collection* (FM 3-55). Army Publications.
https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN35577-FM_3-55-000-WEB-0.pdf



- Department of the Army. (2019a). *Intelligence* (ADP 2-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN18009-ADP_2-0-000-WEB-2.pdf
- Department of the Army. (2019b). *Intelligence Preparation of the Battlefield* (ATP 2-01.3). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN43736-ATP_2-01.3-003-WEB-7.pdf
- Department of the Army. (2019c). *Joint Fires Support* (JP 3-09). Army Publications.
- Department of the Army. (2019d). *Operations Process* (ADP 5-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN18126-ADP_5-0-000-WEB-3.pdf
- Department of the Army. (2020). *Intelligence Analysis* (ATP 2-33.4). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN43899-ATP_2-33.4-001-WEB-3.pdf
- Department of the Army. (2021a). *Armies, Corps, and Division Operations* (FM 3-94). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN34770-FM_3-94-000-WEB-1.pdf
- Department of the Army. (2021b). *Brigade Combat Team* (FM 3-96). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN31505-FM_3-96-000-WEB-1.pdf
- Department of the Army. (2022a). *Commander and Staff Organization and Operations* (FM 6-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN35404-FM_6-0-000-WEB-1.pdf
- Department of the Army. (2022b). *Joint Campaigns and Operations* (JP 3-0). Army Publications.
- Department of the Army. (2022c). *Joint Intelligence* (JP 2-0). Army Publications.
- Department of the Army. (2023a). *Intelligence* (FM 2-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN39259-FM_2-0-000-WEB-2.pdf
- Department of the Army. (2023b). *Targeting* (FM 3-60). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN39048-FM_3-60-000-WEB-1.pdf
- Department of the Army. (2024). *Planning and Operations* (FM 5-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN42404-FM_5-0-000-WEB-1.pdf
- Department of the Army. (2025a). *Operations* (ADP 3-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN43323-ADP_3-0-000-WEB-1.pdf
- Department of the Army. (2025b). *Operations* (FM 3-0). Army Publications. https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN43326-FM_3-0-000-WEB-1.pdf
- Friel, R. a. (2023). Chainpoll: A high efficacy method for LLM hallucination detection. *arXiv preprint arXiv:2310.18344*.
- Fu, J. N. (2024). Gptscore: Evaluate as you desire. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576.
- Holtzman, A. B. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Ji, Z. L. (2023). Survey of hallucination in natural language generation. *ACM computing surveys* 55, 1–38.
- Liu, Y. I. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Manakul, P. L. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the 2023 conference on empirical methods in natural language processing*, 9004–9017.
- Maynez, J. N. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Mishra, A. A. (2024). Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.
- Mündler, N. H. (2023). Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.
- Ravi, S. M. (2024). Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*.





ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE, AND MANPOWER
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

WWW.ACQUISITIONRESEARCH.NET