NASA/TM–20260000162

AMPLIFY: A Modular LLM Platform for Mission-Critical Retrieval and Reasoning

Jayden Ishihara, Sandeep Shetye, Moustafa Abdelbaky, Aiden Jones, and Stefan Schuet
National Aeronautics and Space Administration, NASA Ames Research Center, Moffett Field, California

Paul Kotchavong, Dan Liddell, and Kayshav Prakash
KBR, Inc., NASA Ames Research Center, Moffett Field, California

Sebastian Gutierrez-Nolasco, Besart Mujeci, and Olivia Alexander
Universities Space Research Association, NASA Ames Research Center, Moffett Field, California

January 2026

NASA STI Program. . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collection of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI Program, see the following:

- Access the NASA STI program home page at `http://www.sti.nasa.gov`
- E-mail your question to `help@sti.nasa.gov`
- Phone the NASA STI Information Desk at 757-864-9658
- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM–20260000162

AMPLIFY: A Modular LLM Platform for Mission-Critical Retrieval and Reasoning

Jayden Ishihara, Sandeep Shetye, Moustafa Abdelbaky, Aiden Jones, and Stefan Schuet
National Aeronautics and Space Administration, NASA Ames Research Center, Moffett Field, California

Paul Kotchavong, Dan Liddell, and Kayshav Prakash
KBR, Inc., NASA Ames Research Center, Moffett Field, California

Sebastian Gutierrez-Nolasco, Besart Mujeci, and Olivia Alexander
Universities Space Research Association, NASA Ames Research Center, Moffett Field, California

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

January 2026

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Abstract

This memorandum describes **AMPLIFY**, NASA's compound AI system, designed as a flexible, all-in-one platform for integrating, managing, and scaling large language models (LLMs). It supports composable workflows across NASA-hosted and third-party models, with optional use of external models for non-sensitive queries. The initial use case addresses the challenges that air traffic controllers face in retrieving information from large repositories of technical documentation in a short amount of time. The system applies a retrieval-augmented generation (RAG) architecture, provides observability through Langfuse tracing, and production-grade evaluation via VALOR. A novel element of AMPLIFY is its use of a Model Context Protocol combined with a Model Router, enabling seamless orchestration and dynamic selection of the most appropriate models for each query. Statistical metrics that measure the precision and reliability of our system show that AMPLIFY can deliver accurate and context-aware responses to user-entered queries. Future work will investigate a multi-agent, multi-step architecture to enhance modularity, interpretability, and governance.

1 Introduction and Background

Motivation

Mission-critical organizations rely on timely, accurate, and contextually relevant access to knowledge to ensure safety, compliance, and efficiency. In civil aviation, this need is particularly significant. Technical manuals, advisory circulars, and certification procedures often span thousands of pages of dense regulatory and engineering content. Inspectors and engineers are tasked with finding highly specific information under strict time pressures. Delays or inaccuracies in retrieving the right guidance can prolong inspections, delay certification, and, most importantly, introduce safety risks. As operations grow more complex and documentation expands, the demand for advanced systems capable of delivering precise, context-sensitive results continues to grow.

Limitations of Classical Search

Traditional keyword-based search systems, such as BM25 [1, 2], have been widely used because of their efficiency and simplicity. However, their limitations are evident in high-stakes technical environments. These methods operate primarily on token-level matches, meaning that if a query does not contain the exact keywords present in the documentation, the relevant passages may never be retrieved. For example, a pilot querying "stall speed" may fail to retrieve guidance described using terms such as "stalling airspeed," "minimum stalling speed," or the regulatory designation V_S, even though these refer to the same underlying aerodynamic limit. Beyond synonym mismatches, keyword-based systems tend to rank results using superficial features such as word frequency rather than contextual relevance. This surfaces long lists of partially related documents, forcing users to sift through noise and introducing inefficiency into time-critical workflows. In aviation contexts, where seconds can matter, such shortcomings translate directly into avoidable delays and increased risk of bottlenecks.

The AMPLIFY Approach

AMPLIFY was designed specifically to address these limitations by leveraging retrieval-augmented generation (RAG) [3–6]—a hybrid approach that integrates semantic search with large language model (LLM) synthesis. Rather than relying on purely lexical, keyword-based matching, AMPLIFY embeds documents and queries into high-dimensional vector spaces, enabling retrieval based on semantic similarity and allowing the system to recognize paraphrases, handle synonyms, and connect queries expressed using different vocabularies. The retrieved passages are then synthesized by a locally hosted LLM. A Model Context Protocol (MCP) [7] ensures that retrieved context, metadata, and supporting evidence are structured and passed to the model in a consistent and auditable manner, producing natural-language responses that are both contextually relevant and grounded in authoritative source documents.

AMPLIFY incorporates observability and evaluation as core components: retrieval scores, source citations, and model outputs are logged, auditable, and benchmarked against reference standards. Langfuse [8, 9] tracing underpins this observability layer, capturing end-to-end query flows and metrics to support debugging, moni-

toring, and long-term reliability. This transparency not only builds trust among users but also enables systematic monitoring, ensuring the system can evolve while maintaining safety in mission-critical settings.

Novel Contributions of AMPLIFY
Beyond these capabilities, AMPLIFY advances standard RAG implementations through two novel architectural mechanisms that are essential in safety-critical environments:

- *Model Context Protocol (MCP).* MCP provides a structured interoperability layer that enforces consistent context formatting, tool orchestration, and metadata handling across heterogeneous backends. This enables AMPLIFY to integrate multiple retrieval, evaluation, and synthesis components without introducing ambiguity or model-specific coupling.
- *Model Router.* The Model Router dynamically directs queries between secure internal models and external services based on compliance requirements, latency constraints, and expected performance. This routing logic embeds governance directly into the system architecture, ensuring that sensitive FAA or NASA data is handled appropriately while still allowing the use of high-performance external models when permitted.

Together, these innovations provide a modular, auditable, and policy-aware RAG framework, extending traditional retrieval–generation pipelines to meet the requirements of mission-critical domains.

2 System Design

2.1 Architecture Overview

The AMPLIFY architecture implements a retrieval-augmented generation (RAG) pipeline that integrates document ingestion, semantic retrieval, and controlled model routing within secure boundaries. Documents and associated metadata are normalized, chunked, and embedded into high-dimensional vectors, enabling semantic similarity search that matches queries to relevant content even when terminology differs. Figure 1 illustrates this end-to-end AMPLIFY pipeline, from document ingestion through retrieval, routing, and response generation.

At the core of this design are two mechanisms that extend beyond a conventional RAG pipeline. A dedicated Model Context Protocol (MCP) manages how embeddings, metadata, and contextual data are packaged and passed downstream, ensuring consistency across heterogeneous model backends and enabling structured tool integration for the LLM. This protocol not only enforces observability, making every step auditable, but also guarantees modularity, allowing retrievers and synthesis models to be swapped without breaking compatibility.

Complementing this, the Model Router dynamically directs queries to the most appropriate model backend. Sensitive queries are routed to self-hosted foundational

models within secure NASA infrastructure, while less sensitive queries may leverage managed external services. Routing decisions also account for latency and performance, allowing lightweight models to handle simple classification tasks while more complex synthesis is escalated to larger models. This adaptive routing enforces compliance, optimizes efficiency, and introduces governance capabilities not present in standard RAG implementations.

Together, these mechanisms—MCP and the Model Router—differentiate AMPLIFY from existing RAG architectures by embedding modularity, observability, and policy-driven adaptability as first-class design principles. By combining retrieval quality, transparency, and robust security controls, AMPLIFY delivers accurate, traceable, and compliant responses in high-stakes operational settings.

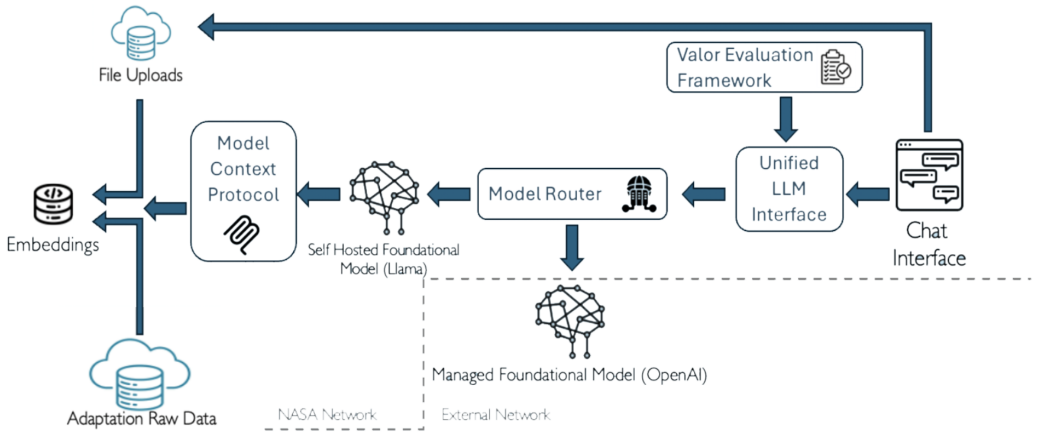

Figure 1. AMPLIFY system architecture showing the RAG pipeline from document ingestion to response generation.

2.2 FAA Use Case

Assisting Federal Aviation Administration (FAA) air traffic controllers (ATCs; see Fig. 2) with decision support and intelligent tools has been identified as a prevalent challenge, and many research efforts have explored AI and LLM-based approaches in this domain [10, 11].

For the FAA, AMPLIFY provides a conversational interface that allows ATCs to query large repositories of documentation. Consider an ATC detecting an unidentified unmanned aerial vehicle (UAV) operating within controlled airspace without an active transponder or filed flight plan. Rather than manually consulting

Figure 2. Air traffic control environment showing the operational context for AMPLIFY deployment.

multiple FAA orders and interagency guidance
documents, the controller queries AMPLIFY: "What are the standard operating procedures for identifying, reporting, and managing a non-cooperative UAV in Class B airspace?" In practice, much of this work involves Standard Operating Procedures (SOPs), which form a primary target corpus for AMPLIFY in both testing and in deployment. AMPLIFY retrieves the relevant advisory circulars and SOP passages, highlights applicable sections, and synthesizes a grounded answer that includes citations. This workflow both accelerates decision-making and builds trust by surfacing the original source documents [12].

A central motivation for adopting AMPLIFY in this domain is security. Unlike external services such as OpenAI's hosted models, which require transmitting queries and sensitive data over public networks, AMPLIFY enables all retrieval, reasoning, and synthesis to occur within secure internal infrastructures. This ensures that proprietary technical data, regulatory materials, and operational details remain protected. In safety-critical contexts like aviation, where much of the documentation is non-public—including SOPs and other procedural manuals—the ability to deploy advanced retrieval-augmented generation entirely within organizational networks provides both a functional advantage and a compliance necessity [13].

2.3 Technology Stack

The AMPLIFY technology stack is engineered for modularity, scalability, and interoperability, ensuring that experimental research components integrate seamlessly into a production-grade environment.

- **Python with Asynchronous I/O** forms the backbone of the orchestration layer. By leveraging event-driven concurrency, AMPLIFY manages multiple retrieval and synthesis tasks in parallel, minimizing latency and maximizing throughput. This is critical in environments where queries must be resolved within seconds, such as air traffic management or emergency aviation operations.
- **FastAPI** provides the lightweight service framework. Its support for asynchronous request handling allows AMPLIFY to expose REST endpoints for ingestion, retrieval, and answer generation without bottlenecking under high query loads. FastAPI's automatic OpenAPI schema generation also facilitates interoperability with other FAA or NASA services.
- **LlamaIndex** orchestrates the retrieval-augmented generation (RAG) pipeline, integrating embedding generation, similarity search, and large language model (LLM) synthesis into a unified workflow. By abstracting low-level complexity, LlamaIndex enables rapid experimentation with different embedding models, retrievers, and synthesis strategies while preserving a consistent API.
- **Qdrant** is the vector database backend, optimized for high-dimensional similarity search across millions of document embeddings. Its distributed architecture and support for approximate nearest-neighbor search ensure AMPLIFY can scale from pilot deployments to enterprise-scale repositories without compromising retrieval quality or latency. In our deployment, collections are sharded across nodes, enabling queries to be processed in parallel, while replication provides

fault tolerance and high availability.

- **LibreChat** provides the user-facing interface, enabling inspectors, engineers, and controllers to interact with the system in a conversational manner. It also supports side-by-side comparison of responses from different models, directly aiding transparency, correctness assessment, and auditability.

Together, these components form a modular ecosystem. Each layer can be independently upgraded or replaced without disrupting the overall architecture.

3 Evaluation and Results

3.1 Observability

Observability refers to the ability to capture and analyze the internal operations of a system so that its behavior can be understood, audited, and improved [14]. Langfuse tracing provides full observability across the pipeline. Each query is logged with its input, retrieval results, system response, latency, and error traces. This granular tracking supports debugging, monitoring, and long-term performance analysis. In safety-critical environments like aviation, observability is not optional: it provides the evidence trail needed for auditing, certification, and regulatory review. Figure 3 shows the Langfuse dashboard used to surface these traces and performance metrics across the pipeline.

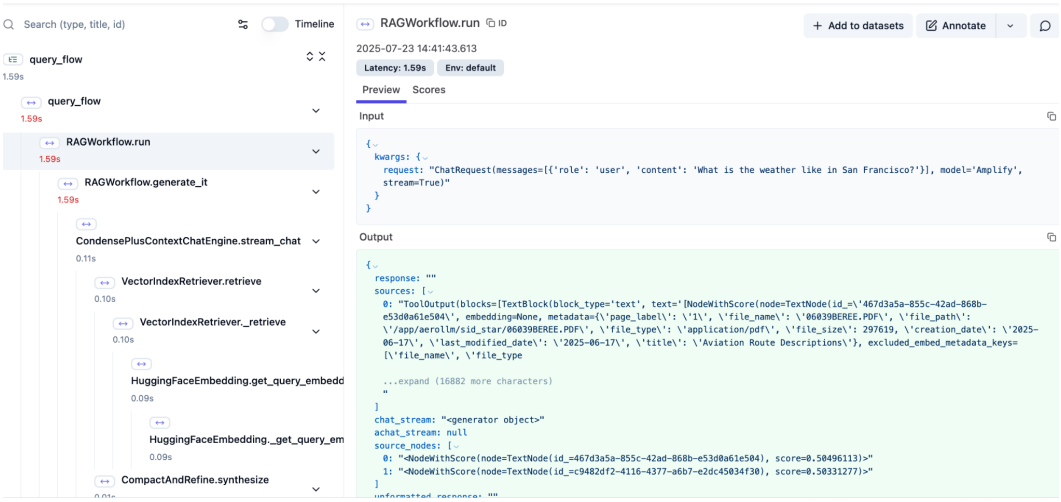

Figure 3. Langfuse dashboard showing query traces and performance metrics.

Beyond individual traces, Langfuse enables trend analysis across large volumes of queries, highlighting systemic issues such as degraded retriever recall or model drift over time. When such anomalies are detected, engineers can investigate specific executions to identify root causes. Figure 4 shows a representative detailed trace, exposing retrieval, prompt construction, and synthesis steps for a single query. These

Start Time ▼	Type	Name	Input	Output	Level	Latency	Tota
2025-08-07 09:14:29		OpenAI.stream_chat	{"messages":["ChatMessage(role=<MessageRole...	"assistant: I don't have any information about a \"...	DEFAULT	0.00s	$0.0(
2025-08-07 09:14:29		OpenAILike.stream_chat	{"messages":["ChatMessage(role=<MessageRole...		DEFAULT	0.00s	$0.0(
2025-08-07 09:14:29		LLM.stream	{"prompt":"ChatPromptTemplate(metadata={'pro...		DEFAULT	0.00s	$0.0(
2025-08-07 09:13:00		OpenAI.stream_chat	{"messages":["ChatMessage(role=<MessageRole...	"assistant: Based on the provided documents, I c...	DEFAULT	0.00s	$0.0(
2025-08-07 09:13:00		OpenAILike.stream_chat	{"messages":["ChatMessage(role=<MessageRole...		DEFAULT	0.00s	$0.0(
2025-08-07 09:13:00		LLM.stream	{"prompt":"ChatPromptTemplate(metadata={'pro...		DEFAULT	0.00s	$0.0(
2025-08-07 08:58:18		OpenAI.stream_chat	{"messages":["ChatMessage(role=<MessageRole...	"assistant: Based on the provided documents, he...	DEFAULT	0.00s	$0.0(
2025-08-07 08:58:18		OpenAILike.stream_chat	{"messages":["ChatMessage(role=<MessageRole...		DEFAULT	0.00s	$0.0(
2025-08-07 08:58:18		LLM.stream	{"prompt":"ChatPromptTemplate(metadata={'pro...		DEFAULT	0.00s	$0.0(
2025-08-07 08:57:01		OpenAI.stream_chat	{"messages":["ChatMessage(role=<MessageRole...	"assistant: I don't have any information on the \"U...	DEFAULT	0.00s	$0.0(
2025-08-07 08:57:01		OpenAILike.stream_chat	{"messages":["ChatMessage(role=<MessageRole...		DEFAULT	0.00s	$0.0(
2025-08-07 08:57:01		LLM.stream	{"prompt":"ChatPromptTemplate(metadata={'pro...		DEFAULT	0.00s	$0.0(
2025-08-07 08:57:00		OpenAI.complete	{"args":["\n Given the following conversation bet...	"What is the purpose and objectives of the UTMB...	DEFAULT	1.08s	$0.0(
2025-08-07 08:57:00		OpenAILike.complete	{"prompt":"\n Given the following conversation b...	{"text":"What is the purpose and objectives of th...	DEFAULT	1.08s	$0.0(

Figure 4. Detailed trace view in Langfuse showing retrieval and synthesis steps.

capabilities allow engineers to fine-tune components while providing auditors with a transparent, step-by-step record of system behavior. Observability therefore functions not only as a debugging tool but also as an accountability mechanism, ensuring that AMPLIFY can be trusted and maintained in operational environments.

3.2 Metrics

To evaluate AMPLIFY, we use metrics capturing retrieval effectiveness and output quality. Table 1 lists each metric with a brief description, moving from judgment-based measures (correctness, faithfulness, relevance) to classical IR metrics (recall, NDCG, MRR), and concluding with composite measures such as the retriever score.

Metric	**Description**
Correctness Score	Measures factual accuracy of system responses compared to authoritative sources. High correctness means fewer factual errors.
Recall	Fraction of relevant documents retrieved out of all possible relevant documents. High recall indicates coverage of relevant material.
Faithfulness Rate	Assesses whether generated responses remain grounded in retrieved passages, avoiding unsupported claims or hallucinations.

NDCG Score [15]	Normalized Discounted Cumulative Gain measures how well retrieved documents are ranked by relevance, giving more weight to higher-ranked items. Defined as: $$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \quad \text{DCG@}k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$ where rel_i is the graded relevance of the document at rank i, and Ideal Discounted Cumulative Gain (IDCG) is the Discounted Cumulative Gain (DCG) of the ideal ranking.
MRR (Mean Reciprocal Rank) [16]	Average reciprocal of the rank at which the first relevant document appears across queries: $$\text{MRR} = \frac{1}{\lvert Q \rvert} \sum_{j=1}^{\lvert Q \rvert} \frac{1}{\text{rank}_j}$$ where Q is the set of all queries, $\lvert Q \rvert$ is the total number of queries, and rank_j is the position of the first relevant result for query j.
Retriever Score	Lexical overlap measure between query and retrieved text, based on exact token intersections. Reflects surface-level matching without accounting for semantic similarity.
Relevance Score	LLM-judged measure of how well each retrieved passage semantically addresses the query.

Table 1: Metrics used to evaluate AMPLIFY's retrieval and generation performance.

Taken together, these metrics provide a comprehensive evaluation framework for AMPLIFY. Correctness, faithfulness, and relevance directly capture the quality of generated answers, while recall, NDCG, and MRR quantify how effectively the retriever surfaces supporting evidence. By combining system-level measures (e.g., Retriever Score) with task-specific judgments, evaluators can not only assess overall performance but also identify where errors originate in the pipeline. This multi-dimensional view makes it possible to target improvements in a principled way, such as tuning retrieval strategies when recall is low, or refining synthesis when faithfulness lags. This ultimately ensures that AMPLIFY remains both reliable and trustworthy in mission-critical contexts.

3.3 Methodology

Evaluation in AMPLIFY is conducted using Validation for Aerospace LLM Output and Reasoning (VALOR) [17], a rubric-based framework developed to provide sys-

tematic, repeatable scoring of LLM outputs. VALOR submits predefined queries to the system and compares generated answers against vetted reference responses. For this study, evaluation was conducted on a curated subset of FAA technical documentation, including Standard Operating Procedures and advisory circulars. The corpus comprised several dozen documents totaling a few hundred pages, paired with a test set of representative queries. Queries were designed to reflect realistic inspector and engineer information needs, ensuring that scoring captured performance in operationally relevant contexts. VALOR decomposes performance into accuracy, retrieval quality, robustness, and uncertainty, offering a multidimensional view of system quality. Figure 5 shows the VALOR evaluation interface used to compute and visualize these metrics across experimental runs.

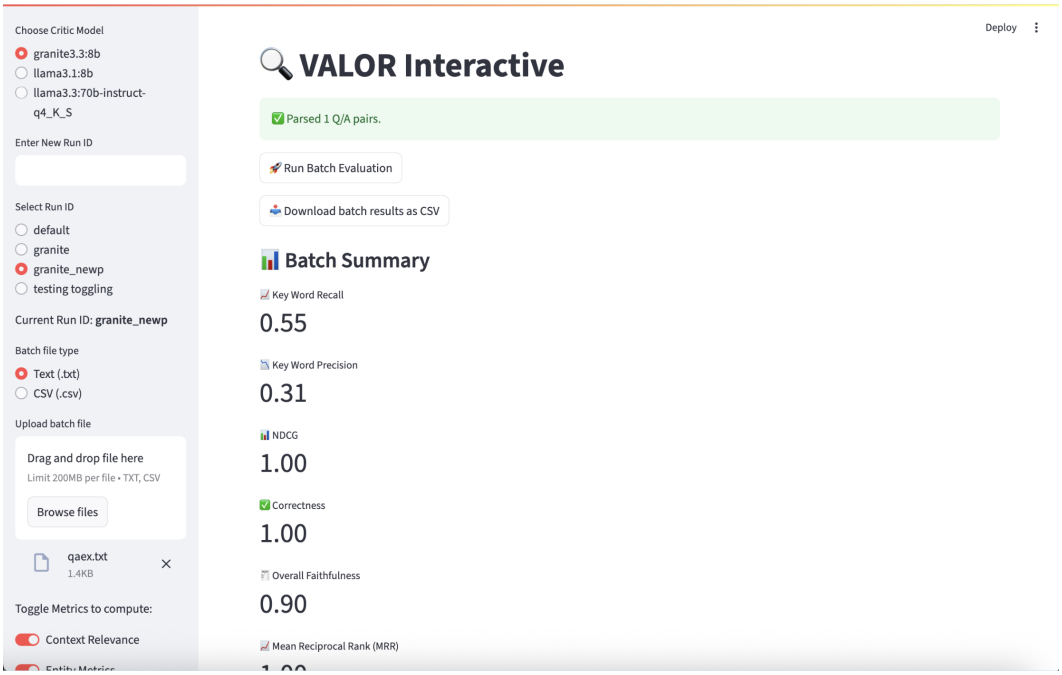

Figure 5. VALOR evaluation interface showing systematic scoring of LLM responses across multiple quality dimensions.

3.4 Analysis of Results

Preliminary results demonstrate promising performance in the FAA domain. On a curated test set of FAA-relevant queries, AMPLIFY achieved strong retrieval quality and synthesis scores (evaluated on a real data subset of the corpus):

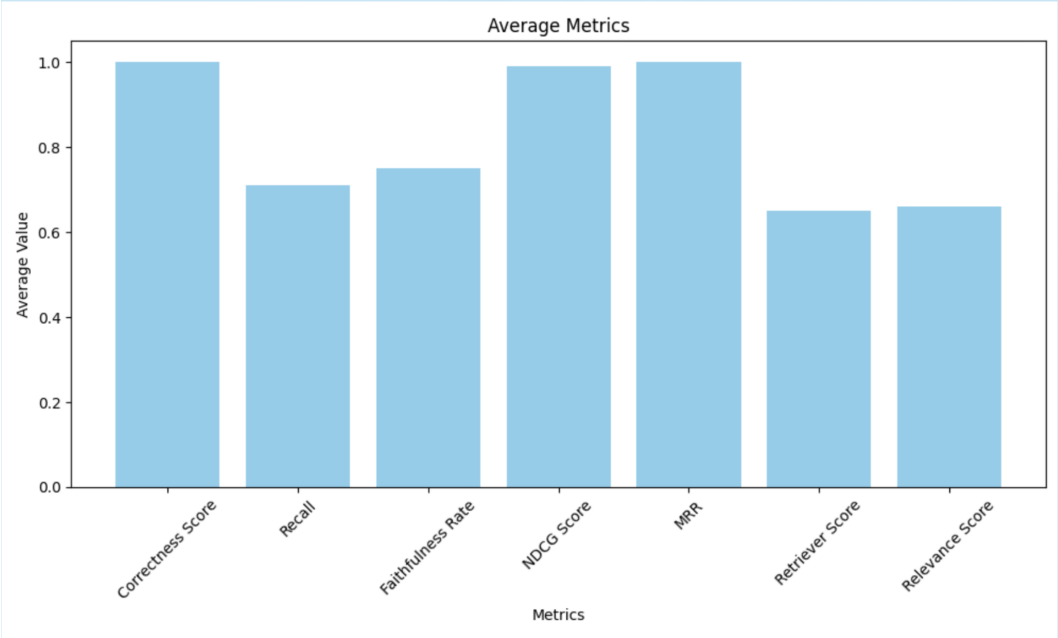

Figure 6. VALOR metrics visualization showing performance across different evaluation dimensions.

Each metric captures a distinct retrieval or generation property; interpretations below describe observed behavior under this evaluation setting.

On this curated dataset shown in Figure 6, we observe a **Correctness Score of 1.0**, indicating consistent factual accuracy, an essential requirement in safety-critical contexts. The **Faithfulness Rate of 0.75** indicates that most responses were grounded in retrieved passages, with the remaining cases involving paraphrasing or synthesis beyond the exact text.

Perfect recall (1.0) is rarely attainable in semantic retrieval, where models retrieve passages based on meaning rather than exact term matching. Thus, AMPLIFY's **Recall of 0.70** is consistent with reported performance ranges for untuned embedding-based semantic retrievers, which commonly trade exact coverage for conceptual matching. This value therefore reflects a known design trade-off rather than an optimized operating point. The near-perfect **NDCG of 0.99** and **MRR of 1.0** further show that the most useful supporting material consistently appeared at the top of the ranked list. Under this evaluation setup, high NDCG and MRR values reduce the likelihood that inspectors must scan lower-ranked results to locate supporting material.

The **Retriever Score (0.65)** and **Relevance Score (0.65)** reflect trade-offs in semantic retrieval. This behavior is consistent with embedding-based semantic retrieval, which prioritizes conceptual similarity over exact token overlap. Because semantic retrievers prioritize conceptual similarity over lexical overlap, maximizing the Retriever Score toward 1.0 would bias retrieval toward exact term matches, po-

tentially reducing coverage of semantically relevant but differently phrased passages. The observed score of 0.65 reflects this trade-off rather than an optimization target. Likewise, perfect semantic relevance (1.0) is uncommon, as LLM-based relevance scoring is conservative and technical domains often contain partially aligned but not strictly direct passages. Within the constraints of the curated FAA subset and limited query set, these scores indicate functional semantic matching, though their generality cannot be assumed without broader evaluation.

Together, these results indicate that LLM system implementations built using the AMPLIFY framework can deliver reliable and trustworthy answers under the evaluated conditions. The framework's transparent metrics provide clear guidance for targeted refinements in future iterations.

Nonetheless, several limitations warrant discussion. First, the evaluation was conducted on a curated subset of FAA technical documents with a relatively small number of test queries. While the observed Correctness, NDCG, and MRR scores are strong, they may not fully capture variability across the broader operational corpus or across different implementations that adopt the AMPLIFY architecture. Second, the Faithfulness Rate of 0.75 indicates that approximately one-quarter of responses incorporated paraphrasing or synthesis beyond retrieved passages. In safety-critical settings, such deviations may introduce risk, underscoring the need for deeper analysis of paraphrasing patterns and more constrained generation strategies at the level of the deployed LLM system rather than the framework itself. Third, the moderate Retriever and Relevance Scores highlight the need for systematic retriever optimization, including evaluation of domain-adapted embedding models, hybrid retrieval architectures, terminology normalization, and improved indexing strategies. These improvements pertain to specific system configurations enabled by AMPLIFY, rather than inherent limitations of the framework. More extensive testing on larger and more diverse datasets will be essential to demonstrate robustness and ensure that LLM systems implemented using AMPLIFY maintain reliability under real-world operational demands.

4 Conclusions

Ultimately, AMPLIFY provides a streamlined approach for operationalizing retrieval-augmented generation for mission-critical knowledge access. By combining semantic retrieval, LLM synthesis, and robust observability, AMPLIFY provides users with fast, accurate, and traceable answers. Its successful application to FAA documentation highlights its potential for broader high-stakes domains such as healthcare, engineering, and emergency management.

Looking ahead, a future version of AMPLIFY may evolve into a multi-agent system, extending its current modular design and observability. Figure 7 illustrates a representative multi-agent retrieval and reasoning workflow, demonstrating how planning, sub-query decomposition, retrieval, extraction, and synthesis can be coordinated across specialized agents. Such an architecture could incorporate specialized roles, including: (1) **Routing Agents** [18] that dynamically select among retriev-

ers or models based on sensitivity, latency, or accuracy requirements; (2) **Query Planning Agents** [18] that decompose complex information needs into structured sub-queries, improving retrieval coverage and precision; (3) **ReAct Agents** [18] that integrate reasoning with action, deciding when to retrieve additional evidence versus when to synthesize; and (4) **Plan-and-Execute Agents** [18] that coordinate multi-step workflows, sequencing retrieval, validation, and synthesis into transparent, auditable pipelines.

By embedding these agentic roles within the existing Model Context Protocol and Model Router, AMPLIFY could move beyond a linear RAG pipeline toward a coordinated framework of interacting agents. This evolution would enhance interpretability, reduce error propagation, and enable more adaptive responses in mission-critical environments. In parallel, ongoing work will incorporate adaptive retrieval strategies and strengthened safety controls, ensuring that AMPLIFY not only delivers practical value today but also scales into a trustworthy, future-proof platform for critical knowledge management.

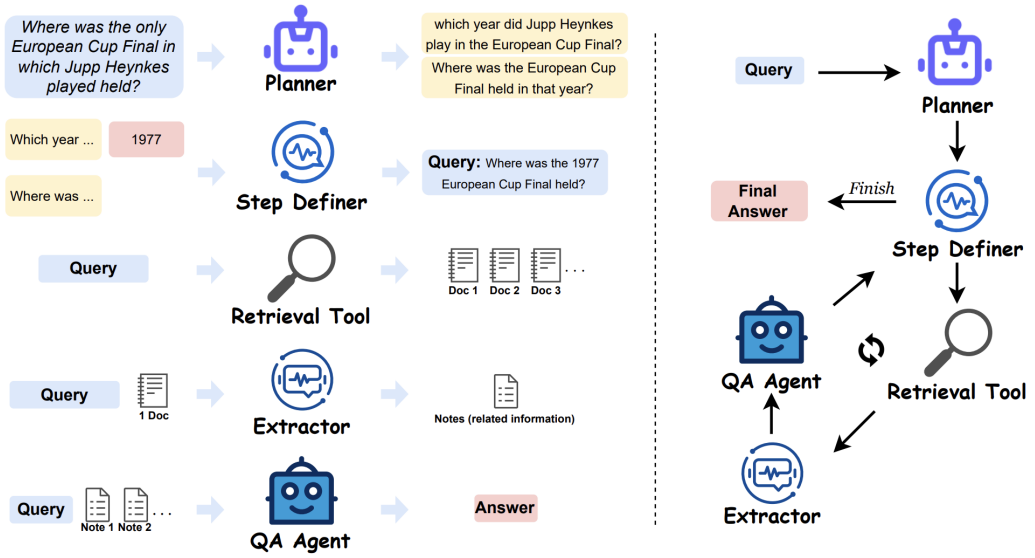

Figure 7. Example multi-agent retrieval and reasoning workflow [19]. This conceptual architecture illustrates how planning, sub-query decomposition, retrieval, extraction, and answer synthesis can be coordinated across specialized agents.

5 Appendices

5.1 Acknowledgments

The authors gratefully acknowledge the guidance and expertise of Richard Papasin and David Alfano. The authors also thank NASA Ames Research Center and the Intelligent Systems Division for providing the opportunity and environment to conduct this work.

5.2 Acronyms

- **AI** – Artificial Intelligence
- **API** – Application Programming Interface
- **ATC** – Air Traffic Controller
- **BM25** – Best Matching 25
- **DCG/NDCG** – Discounted Cumulative Gain / Normalized Discounted Cumulative Gain
- **FAA** – Federal Aviation Administration
- **IDCG** – Ideal Discounted Cumulative Gain
- **I/O** – Input/Output
- **LLM** – Large Language Model
- **MCP** – Model Context Protocol
- **MRR** – Mean Reciprocal Rank
- **NASA** – National Aeronautics and Space Administration
- **RAG** – Retrieval-Augmented Generation
- **REST** – Representational State Transfer
- **SOP** – Standard Operating Procedure
- **TM** – Technical Memorandum
- **VALOR** – Validation for Aerospace LLM Output and Reasoning

References

1. Stephen E Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. pages 109–126, 1995.

2. Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389, 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL `http://dx.doi.org/10.1561/1500000019`.

3. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL `https://arxiv.org/abs/2005.11401`.

4. Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, 2020. URL `https://arxiv.org/abs/2002.08909`.

5. Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. URL `https://arxiv.org/abs/2007.01282`.

6. Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3789–3803, 2021. doi: 10.18653/v1/2021.emnlp-main.547.

7. Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions, 2025. URL `https://arxiv.org/abs/2503.23278`.

8. Langfuse. Langfuse observability and tracing overview. Online documentation, 2025. URL `https://langfuse.com/docs/observability/overview`. Accessed: 2025-09-21.

9. Langfuse Documentation. Observability data model. Online; Langfuse Docs, 2025. URL `https://langfuse.com/docs/observability/data-model`. Accessed: 2025-09-21.

10. K. Kalyanam. Application of ai/ml tools for air traffic management. NASA Technical Reports Server (NTRS), 2024. URL `https://ntrs.nasa.gov/api/citations/20240012467/downloads/DASC%202024%20Tutorial%20Slides.pdf`. Tutorial slides, NASA/FAA collaborative effort.

11. Justas Andriuškevičius and Junzi Sun. Automatic control with human-like reasoning: Exploring language model embodied air traffic agents, 2024. URL https://arxiv.org/abs/2409.09717.

12. Federal Aviation Administration. Air traffic control decision support tool handbook. Technical report, Federal Aviation Administration, 2019. URL https://hf.tc.faa.gov/publications/2019-atc-decision-support-tool/. Accessed: 2025-09-20.

13. Federal Aviation Administration. Roadmap for artificial intelligence safety assurance. Technical report, Federal Aviation Administration, 2024. URL https://www.faa.gov/media/82891. Accessed: 2025-09-20.

14. Deepa Sanjay Singh. Observability for AI systems: Tracing, drift, and SLAs. *International Journal of Research and Applied Innovations*, 8(2):11952–11955, March 2025. doi: 10.15662/IJRAI.2025.0802002. URL https://ijrai.org/index.php/ijrai/article/view/93.

15. Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

16. Nick Craswell. Mean reciprocal rank (mrr). In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 1703–1703. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_488. URL https://doi.org/10.1007/978-0-387-39940-9_488.

17. K. S. Prakash, S. Schuet, K. Wheeler, K. Krishnakumar, S. Gutierrez-Nolasco, and S. Shetye. VALOR: Validation for aerospace llm output and reasoning. Technical Memorandum NASA/TM–2025–20260000076, National Aeronautics and Space Administration, 2025.

18. Ivan Belcic and Cole Stryker. What is agentic rag? https://www.ibm.com/think/topics/agentic-rag, 2025. Accessed: 2025-09-29.

19. Thang Nguyen, Peter Chin, and Yu-Wing Tai. Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2505.20096.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-01-2026	Technical Memorandum	

4. TITLE AND SUBTITLE
AMPLIFY: A Modular LLM Platform for Mission-Critical Retrieval and Reasoning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jayden Ishihara, Sandeep Shetye, Moustafa Abdelbaky, Paul Kotchavong, Sebastian Gutierrez-Nolasco, Kayshav Prakash, Besart Mujeci, Olivia Alexander, Dan Liddell, Aidan Jones, Stefan Schuet

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Ames Research Center
Moffett Field, California 94035

8. PERFORMING ORGANIZATION REPORT NUMBER
L–

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TM–20260000162

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified-Unlimited
Subject Category 64
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES
An electronic version can be found at http://ntrs.nasa.gov.
This work was performed during a summer internship with Code TI.

14. ABSTRACT

This memorandum describes **AMPLIFY**, NASA's compound AI system, designed as a flexible, all-in-one platform for integrating, managing, and scaling large language models (LLMs). It supports composable workflows across NASA-hosted and third-party models, with optional use of external models for non-sensitive queries. The initial use case addresses the challenges that air traffic controllers face in retrieving information from large repositories of technical documentation in a short amount of time. The system applies a retrieval-augmented generation (RAG) architecture, provides observability through Langfuse tracing, and production-grade evaluation via VALOR. A novel element of AMPLIFY is its use of a Model Context Protocol combined with a Model Router, enabling seamless orchestration and dynamic selection of the most appropriate models for each query. Statistical metrics that measure the precision and reliability of our system show that AMPLIFY can deliver accurate and context-aware responses to user-entered queries. Future work will investigate a multi-agent, multi-step architecture to enhance modularity, interpretability, and governance.

15. SUBJECT TERMS
Retrieval-Augmented Generation, Large Language Models, Evaluation, FAA, Observability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Information Desk (help@sti.nasa.gov)
U	U	U	UU	22	19b. TELEPHONE NUMBER *(Include area code)* (757) 864-9658