Toward an Earth-Independent System for EVA Mission Planning: Integrating Physical Models, Domain Knowledge, and Agentic RAG to Provide Explainable LLM-Based Decision Support

Kaisheng Li
Department of Mechanical and Aerospace Engineering, University of California, Davis, CA, USA

Richard S. Whittle[1]
Department of Mechanical and Aerospace Engineering, University of California, Davis, CA, USA

Abstract

We propose a unified framework for an Earth-independent AI system that provides explainable, context-aware decision support for EVA mission planning by integrating six core components: a fine-tuned EVA domain LLM, a retrieval-augmented knowledge base, a short-term memory store, physical simulation models, an agentic orchestration layer, and a multimodal user interface. To ground our design, we analyze the current roles and substitution potential of the Mission Control Center – identifying which procedural and analytical functions can be automated onboard while preserving human oversight for experiential and strategic tasks. Building on this framework, we introduce RASAGE (Retrieval & Simulation Augmented Guidance Agent for Exploration), a proof-of-concept toolset that combines Microsoft Phi-4-mini-instruct with a FAISS (Facebook AI Similarity Search)-powered EVA knowledge base and custom A* path planning and hypogravity metabolic models to generate grounded, traceable EVA plans. We outline a staged validation strategy to evaluate improvements in route efficiency, metabolic prediction accuracy, anomaly response effectiveness, and crew trust under realistic communication delays. Our findings demonstrate the feasibility of replicating key Mission Control functions onboard, enhancing crew autonomy, reducing cognitive load, and improving safety for deep-space exploration missions.

2012 ACM Subject Classification Human-centered computing → Interactive systems and tools; Human-centered computing → Interaction design process and methods

Keywords and phrases Human-AI Interaction for Space Exploration, Extravehicular Activities, Cognitive load and Human Performance Issues, Human Systems Exploration, Lunar Exploration, LLM

Digital Object Identifier 10.4230/OASIcs.SpaceCHI.2025.6

1 Introduction and Background

Current extravehicular activity (EVA) sorties are highly choreographed and rely on continuous support from ground control teams [5][6][33]. Flight controllers in Mission Control meticulously monitor the astronauts' suit telemetry and progress during EVA, managing safety and tasks in real time[33][39]. While this Earth-dependent model is feasible for current International Space Station (ISS) and lunar EVAs, future lunar habitats and surface infrastructure will necessitate multiple concurrent EVAs, and Mars missions will face significant communication latency and limited bandwidth[33][34][9]. Past missions have highlighted the challenges in on-the-fly EVA decision-making. For example, Apollo 14 astronauts attempted to reach Cone Crater without real-time navigation aid and had to abort when they became

[1] **Correspondence:** Department of Mechanical and Aerospace Engineering, UC Davis, One Shields Ave, Davis, CA 95616, USA

Advancing Human-Computer Interaction for Space Exploration (SpaceCHI 2025).
Editors: Leonie Bensch, Tommy Nilsson, Martin Nisser, Pat Pataranutaporn, Albrecht Schmidt, and Valentina Sumini; Article No. 6; pp. 6:1–6:17
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

disoriented and fatigued, despite coming within approximately 40 meters of the rim[32]. They carried only a paper map and made judgments on the ground while mission control tracked their vital signs[32]. This incident underscores the need for Earth-independent autonomous planning, monitoring, and decision-making support during EVAs.

NASA's Personalized EVA Informatics & Decision Support (PersEIDS) initiative represents an important step toward autonomy by leveraging the Crew State & Risk Model (CSRM) to project individualized crew health and performance parameters – such as metabolic rate, thermal load, fatigue, CO_2 dose, and decompression stress – over EVA timelines [35][1]. PersEIDS develops user interfaces, visualizations, and data-science methods that deliver actionable, data-driven options and recommendations to extravehicular and intravehicular crewmembers for safe, efficient planning and execution under high-latency conditions [35]. It also automatically tracks and probabilistically assesses suit consumables against flight rules to improve mission completion time and preserve consumables relative to unsupported operations [35]. However, while PersEIDS excels at biomedical decision support and consumable management, it does not provide end-to-end task planning, context-aware reasoning, or the fully autonomous "virtual flight controller" functionality needed when ground support is delayed or unavailable. Moreover, a fully autonomous decision-support system can directly interpret CSRM's predictive outputs, streamlining EVA planning without adding crew cognitive workload.

To fill this gap, an autonomous, Earth-independent artificial intelligence (AI) system – integrated within spacecraft, habitats, or embedded directly in astronaut suits – is envisioned by acting as a local "virtual flight controller" for the crew. Such a system could enhance astronaut safety and efficiency by dynamically replanning routes, monitoring astronaut health, providing task-related information, and advising the crew immediately, rather than waiting for delayed input from Earth. This can optimize EVA outcomes in real time, reducing reliance on ground support. Recent technological advancements underscore the feasibility of deploying sophisticated Large Language Models (LLMs) in space. In mid-2024, Booz Allen Hamilton, in collaboration with Hewlett Packard Enterprise (HPE), successfully deployed an LLM retrieval-augmented generation (RAG) application aboard the ISS using HPE's Spaceborne Computer-2 to help perform certain maintenance and repair procedures [17]. This milestone demonstrated that disconnected AI operations in austere environments are achievable.

In this paper, we introduce a framework for an Earth-independent AI system for EVA mission planning and decision support that integrates: (a) a small-parameter EVA domain LLM fine-tuned on astronaut-ground transcripts, NASA EVA handbooks, flight rules, and related literature; (b) a retrieval-augmented knowledge base of procedures, checklists, operational logs, and related documents; (c) a suite of physical simulation models (path planning, metabolic prediction, thermal load, radiation dose, life-support performance, communications quality, power consumption, and more); (d) a short-term memory store that persistently captures mission context (objectives, timeline), crew profiles (roles, expertise, health/status), prior user–agent interactions, and live sensor telemetry; (e) an agentic orchestration layer that sequences semantic retrievals, model invocations, ReAct-style iterative reasoning, and generation into fully traceable, explainable EVA recommendations; and (f) a multimodal user interface (text, voice and visual) that delivers concise guidance, annotated maps, task timelines, and underlying rationale with source citations. Building on this framework, we introduce RASAGE (Retrieval & Simulation Augmented Guidance Agent for Exploration), a proof-of-concept toolset that implements these components to deliver context-aware, quantitatively grounded decision support – emulating key Mission Control functions, enhancing crew autonomy, and safeguarding astronaut safety when Earth support is delayed or unavailable. This paper provides a high-level review of relevant literature, outlines our proposed architecture and experimental plan, and discusses associated risks and future steps toward realizing this vision.

2 Current Roles and Substitution Potential of Mission Control Center

The Mission Control Center (MCC) serves as the operational hub for space missions, providing critical support across multiple domains [24]. For EVA operations specifically, MCC fulfills several essential functions that must be considered when designing an Earth-independent alternative. These functions include managing EVA timelines, tracking suit telemetry and consumables, providing contingency guidance during off-nominal situations, conducting real-time risk assessment and decision-making, optimizing resource usage, delivering specialized scientific guidance, coordinating communication between the crew and ground teams, monitoring astronaut physiological and behavior health, and strategically adjusting mission plans in response to discoveries or unexpected challenges [38][39][37][41][40][25][15]. Given these roles, what could be substituted with an onboard AI agent? The potential for substituting these functions with automated systems or alternative technologies is closely tied to the nature of the knowledge they require. The knowledge types employed by MCC can be categorized as:

- **Procedural Knowledge:** Formalized processes, checklists, and operational sequences documented in NASA handbooks and flight rules.
- **Experiential Knowledge:** Expertise developed through years of mission operations, including pattern recognition and analogical reasoning applied to novel situations.
- **Analytical Knowledge:** Quantitative assessment capabilities for system performance, environmental conditions, and mission constraints.
- **Domain-specific Knowledge:** Specialized expertise in areas such as geology, medicine, engineering, or other disciplines relevant to mission objectives.
- **Strategic Knowledge:** High-level mission planning capabilities that consider multiple interdependent factors and long-term consequences.

Tasks characterized by procedural and analytical knowledge – those governed by well-defined steps, rules, or data-driven logic – are generally more amenable to automation or delegation to technological systems. Conversely, tasks that depend heavily on experiential and domain-specific knowledge, such as those requiring nuanced judgment, intuition, or expertise developed through practical experience, pose greater challenges for automation [4]. Consequently, functions predominantly rooted in procedural or analytical knowledge exhibit a higher potential for substitution, while those reliant on experiential or strategic expertise demonstrate a medium substitution potential. Additionally, the psychological support function demonstrates limited substitution potential. Despite advancements in artificial intelligence, the nuanced interpretation of complex emotional signals, empathetic communication, and personalized adaptation to individual psychological requirements remain predominantly within the expertise of human professionals. Therefore, automated systems should be viewed primarily as supportive tools rather than comprehensive replacements in this deeply human-centered function [27]. This analysis highlights the potential to substitute many functions of the current Mission Control Center (MCC) with automated systems, particularly those rooted in procedural and analytical knowledge. However, it also reveals limitations in fully replacing tasks that demand advanced cognitive, contextual, and emotional capabilities. Table 1 provides a summary of the current functions of MCC and their substitution potential for EVA operation specifically.

Table 1 This table categorizes MCC responsibilities during EVAs and assesses their potential for AI substitution[38][39][37][41][40][25][15].

Function	Description	Knowledge Type	Substitution Potential
Procedure Management	Tracking EVA timeline, providing step-by-step guidance, managing deviations	Procedural	High
Systems Monitoring	Tracking suit telemetry, consumables, and environmental parameters	Analytical Procedural	High
Contingency Response	Providing troubleshooting guidance for off-nominal situations	Experiential Procedural	Medium
Risk Assessment	Real-time evaluation of mission risks and go/no-go decisions	Experiential Analytical	Medium
Resource Optimization	Managing consumables usage and recommending pace adjustments	Analytical	High
Scientific Guidance	Providing expertise on sample collection and experiment protocols	Domain-specific	Medium
Communication Coordination	Managing communication between EVA crew, vehicle crew, and ground	Procedural	High
Medical Monitoring	Assessing astronaut vital signs and health status	Analytical Domain-specific	Medium
Strategic Replanning	Substantial timeline modifications due to discoveries or failures	Experiential Strategic	Medium
Psychological Support	Monitoring astronaut psychological state, providing emotional support, and mitigating stress-related risks	Experiential Domain-specific	Low

3 Evolution from Text Generation to Agency

LLMs have evolved rapidly from early transformer-based models like BERT [13] to sophisticated systems like GPT-o1 [44], with capabilities extending far beyond text generation. This evolution has been marked by key innovations in both model architecture and application techniques. Recent advancements have transformed LLMs from passive text generators into systems with agent-like capabilities. These include self-reflection and output critique [51][31], effective use of external tools and APIs [50][45][56], complex task planning and decomposition [57][55], and multi-agent collaboration through structured dialogues [11][49]. Collectively, these developments enable LLMs to operate as increasingly autonomous, goal-directed systems.

Several key approaches have proven effective for domain-specific applications:

- **Domain Adaptation and RAG:** Fine-tuning on specialized corpora [18] and retrieval augmentation [30] enable LLMs to access and incorporate domain expertise.
- **Reasoning Enhancements:** Techniques like instruction tuning[60] and chain-of-thought reasoning [55] improve models' ability to follow constraints and demonstrate explicit reasoning.
- **Efficiency Optimizations:** Small-parameter models reduce computational requirements while maintaining performance on specialized tasks, making deployment possible in resource-constrained environments [47].

In space applications, these advances have enabled a spectrum of notable implementations. Mission support systems have seen early deployments through CIMON, an ISS-based AI assistant designed to reduce crew workload through natural dialogue [3], and NASA's Callisto experiment, which validated a voice assistant capable of handling procedural queries without Earth communication [16]. Autonomous decision support capabilities advanced significantly with the successful 2024 deployment of a generative AI system with RAG capabilities on ISS, validating the technical feasibility of operating advanced AI within space hardware constraints [17]. Meanwhile, an AI4U assistant – with a capability of learning new skills during operation as a result of its reinforcement learning approach – is being tested in the Mars Desert Research Station (MDRS) facilitates [42] [7]. More recently, CORE (Checklist Organizer for Research and Exploration) and METIS (Mars Exploration Telemetry-Driven Information System) have been proposed as offline-capable intelligent personal assistants that integrate knowledge graphs, retrieval-augmented generation, and augmented reality to deliver reliable, flexible, and intuitive procedural guidance for astronauts aboard the ISS, the Lunar Gateway, and future deep-space missions[7][8]. These developments collectively demonstrate the utility of AI and LLMs across multiple operational domains in space exploration.

It's important to note that baseline LLMs have limitations that must be addressed for mission-critical use. They are prone to hallucination – confidently stating incorrect information – which is unacceptable when astronaut safety is on the line [21]. They also lack up-to-date awareness of the world beyond their training data cutoff. To mitigate these issues, the current state of the art uses Retrieval-Augmented Generation (RAG) and related techniques. RAG involves retrieving relevant documents or facts from an external knowledge source and feeding that into the LLM during query answering [58]. In our context, that means when the astronaut asks the AI a question (e.g., "What's the next step? My CO_2 level is high."), the system would fetch pertinent information (for example the CO_2 scrubber malfunction procedure from the EVA manual, or a rule about terminating EVA upon high CO_2), and the LLM would incorporate that into its response. This approach keeps answers grounded in authoritative sources and reduces the chance of the AI *"making stuff up"*. In NASA's own research, integrating RAG has been shown to improve the feasibility and correctness of LLM-generated solutions in a domain task [54].

The requirements for explainability, robustness to unexpected inputs, and operation within tight computational constraints still pose challenges in developing LLMs in safety-critical space applications. Astronauts must understand AI recommendations sufficiently to evaluate them in high-stakes situations, which require an appropriate level of trust and accuracy of information provided. Current LLM implementations, even those enhanced with basic RAG capabilities, fundamentally fail when confronting problems requiring precise analytical computation or physical simulation. These limitations are especially pronounced in EVA contexts where accurate metabolic modeling, dynamic path optimization, and physical environmental interactions demand capabilities beyond text generation and retrieval alone. Standard language models cannot effectively perform the mathematical modeling, spatial reasoning, or physical predictions needed for critical EVA parameters such as oxygen consumption rates, thermal regulation, or optimal traversal planning across variable terrain. To address these limitations, the agentic RAG paradigm extends the traditional framework by embedding an AI agent capable of orchestrating a sequence of actions. In this pipeline, the LLM is not merely a passive answer generator; it actively plans and executes multiple retrievals, engages external tools (such as a path planner, or other physical simulation APIs) and iteratively refines its reasoning [10]. For example, an EVA planner agent could break down a task, namely planning a route from point A to B, into a series of deliberate steps:

retrieving the mission map and constraints, invoking a path planning module to identify possible routes, consulting a metabolic cost prediction model to estimate oxygen usage, checking predefined procedures and rules, and finally comparing options against safety limits to produce a recommendation complete with explanations. Recent surveys highlight that this agentic RAG approach offers unparalleled flexibility and context-awareness, enabling dynamic retrieval strategies and multi-step reasoning that mirror the decision-making processes of a human flight controller [52]. We aim to leverage these benefits in the EVA domain, effectively creating a digital assistant that can think through an EVA plan similarly to how a human flight controller would – by consulting manuals, running calculations, and deliberating over options.

4 Proposed Framework and Toolset

The framework we propose consists of several integrated components to create a system that provides decision support comparable to Earth-based mission control while operating within the constraints of deep-space missions. An architectural schematic is presented in Figure 1. These components include:

- **Fine-Tuned EVA Domain LLM**: A language model (with relatively small parameters for on-board deployment) that has been fine-tuned on a corpus of EVA-related text. This includes transcripts of astronaut-ground conversations, EVA procedure documents, NASA operations handbooks, flight rules, and relevant academic literature. The fine-tuning process imbues the LLM with domain-specific vocabulary and an understanding of the structure of EVA activities. This LLM serves as the reasoning and language generation engine of the system – it will produce explanations, summaries, and recommendations in understandable terms. Crucially, it will not rely solely on its internal knowledge but will use the following modules to ground its outputs.
- **Retrieval-Augmented Knowledge Base (RAG Module)**: A database of EVA knowledge that the LLM agent can query as needed. This knowledge base could be implemented as a vector-indexed document store containing segmented texts from manuals, checklists, plans, and previous EVA logs. When the agent faces a question or task, it can perform a semantic search to fetch relevant snippets. Those snippets are then fed into the LLM's context window so that its decisions and responses cite actual references. The design ensures that even if the LLM's training data is outdated or if it "forgets" a detail, it can retrieve the latest ground-truth information. This is key for explainability – the AI can provide answers directly grounding its advice in source material.
- **Physical Simulation Models (Tool Use)**: A suite of domain-specific physical models that the LLM agent can invoke as needed via standardized APIs. Each module encapsulates a discrete aspect of EVA logistics – such as path planning, metabolic prediction, thermal load, radiation dose, suit life-support performance, communications link quality, and battery consumption – and returns quantitative metrics (e.g., distance, elevation gain, estimated O_2 usage, cumulative radiation exposure, power draw). When the agent evaluates a proposed EVA plan or contingency, it issues a structured query to the relevant model(s). The returned outputs are then incorporated into the LLM's context window so that recommendations are grounded in up-to-date, physical data. This modular approach ensures that the AI's reasoning is both objective – comparing candidate plans across multiple safety and performance dimensions – and explainable, since each decision point can be traced back to concrete model outputs rather than inferred solely from language patterns.

- **Agentic Orchestration Layer**: The central coordination engine that seamlessly integrates the components mentioned above into a unified decision-making workflow with four specialized agents – planning, retrieval, ReAct, and generation. The planning agent first defines task objectives and determines which information and simulations the LLM requires to address a given EVA scenario. The retrieval agent then performs semantic searches of the knowledge base for authoritative procedure snippets and invokes the appropriate physical models to generate quantitative metrics. The ReAct agent (Reason-and-Act) iteratively evaluates retrieved documents and model outputs against safety constraints and mission rules, refining queries or plan steps as needed to close any information gaps; and the generation agent synthesizes the validated context into a concise, actionable recommendation – complete with clear rationale and source citations. This layer ensures that each plan or contingency is grounded in up-to-date textual guidance, domain knowledge, and physical evidence, while logging each intermediate step for auditability and verification.
- **Short-Term Memory Store**: A rolling buffer that continuously captures and organizes critical contextual information – including overall mission objectives and timeline, individual crew profiles (roles, expertise, current health/status), the history of prior user–agent exchanges, and real-time sensory inputs (e.g., suit telemetry, environmental sensor data). This memory module supplies the Orchestration Layer with up-to-date situational awareness and conversational continuity, ensuring that every planning cycle, retrieval query, and recommendation remains consistent, personalized, and aligned with the evolving EVA context.
- **User Interface**: A multi-modal interaction layer that enables astronauts to seamlessly engage with the AI toolset via voice commands and/or a visual display (tablet or helmet HUD). Designed for hands-busy, high-noise EVA environments, the voice interface supports natural language queries and delivers spoken recommendations, while the visual display presents concise, contextually relevant information – such as annotated maps, task timelines, and highlighted procedure snippets. All outputs include clear rationale and source citations (e.g., flight rule references or model outputs) to ensure transparency and foster user trust. The interface dynamically adapts its presentation based on task urgency and crew workload, surfacing only the most critical information during high-stress scenarios and offering deeper explanatory detail when time permits.

Fine-tuning strategy and data availability

Within the proposed framework, the domain LLM will be adapted through a two-step, resource-aware pipeline: (i) task-adaptive pre-training (TAPT) on the full raw EVA corpus, approximately 18 M token outlined in Table 2, to familiarise the backbone with mission-specific vocabulary and discourse [18], and (ii) parameter-efficient QLoRA fine-tuning on a carefully curated dataset of several thousand question–answer pairs, inserting low-rank adapters while the 4-bit base weights remain frozen [12]. Source PDFs are processed with olmOCR, which converts each page image into clean, ordered text, removes noise, and drops duplicates before embedding [48]. The cleaned corpus will be also segmented, embedded, and stored in the Retrieval-Augmented Knowledge Base to ensure that training and real-time grounding reference the same documents. LoRA/QLoRA studies demonstrate that models in this parameter range can achieve substantial gains from just a few thousand high-quality instruction–response examples [22][59]; therefore, our curated Q&A dataset fully supports effective fine-tuning. All sources are public-domain U.S. Government works. This workflow

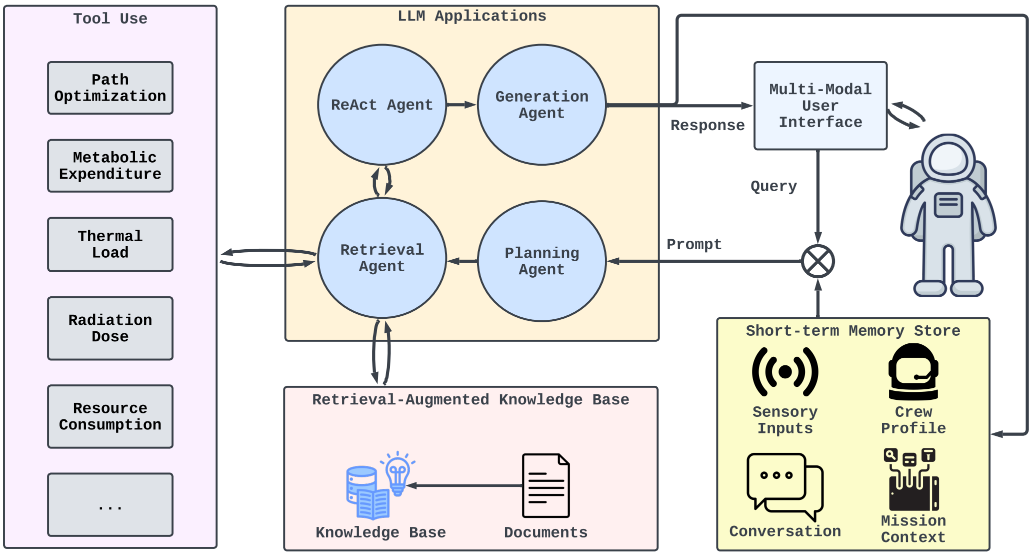


Figure 1 Conceptual Architecture of the AI EVA Planner.

minimizes on-orbit compute demands and allows future document updates to be re-embedded and adapter-patched without modifying the frozen backbone, preserving consistency across the integrated framework.

Physical-Simulation Suite

Each physics model is encapsulated as a uniform JSON-RPC endpoint that the LLM can invoke on demand [26]. The exemplar modules listed in Table 3 – the path planner, the hypogravity metabolic-cost model, the thermal-load and radiation-dose estimators, and the suit-consumables tracker – are drawn from literature with extensive prior studies demonstrating their suitability for seamless integration. For every call, the agent submits a compact JSON request and receives a structured reply that returns hard numbers (e.g., way-points, ascent, ETA, $\dot{V}O_2$, coolant margin, O_2). Each request–response pair is logged, routed through downstream checks, and ultimately surfaced to the crew with full parameter context, yielding an auditable *retrieve → compute → evaluate* loop. Because the framework is inherently versatile, additional or higher-fidelity simulators can be added at any time by exposing the same endpoint schema, allowing the toolset to grow without altering the surrounding architecture.

Concept of Operations

In operation, the proposed AI toolset would support EVA mission planning at every stage – from high-level strategy during pre-mission planning to dynamic decision support in real time. During mission planning, astronauts and flight planners can simulate multiple EVA scenarios, asking "what if" questions such as "What if we add a geology stop here – can we still return to the lander in time given our oxygen reserves and estimated metabolic expenditure?" to refine task sequences and resource allocations. Once on the surface, the system helps brief and verify the day's EVA plan, enabling the crew to request a final check – "Are all tasks feasible within a six-hour EVA given current consumables and predicted metabolic

Table 2 Summary of Extravehicular Activity Document Corpus (token counts estimated using the conservative heuristic of 250 words per page and 1.3 tokens per word[53][43]).

Document Set	**Pages (approx.)**	**Words ($\times 10^3$)**	**Tokens ($\times 10^3$)**	**Source**
NASA Mercury–Apollo–Gemini Communications Transcripts	46,000	11,500	15,000	NASA JSC History Collection
EVA Console Handbook & Flight-Control Operations Manual	1,800	450	585	JSC-26843 JSC-29229 JSC-20597
Apollo & Space Shuttle & ISS Flight Rules	5,000	1,250	1,625	NSTS-12820 Apollo 8 - 17 Final Flight Mission Rules
Exploration EVA System Concept of Operations & Technical Standards	200	50	65	EVA-EXP-0042 EVA-EXP-0034
NASA Technical Standards (Human Factors and Health)	900	225	290	NASA-STD-3000 NASA-STD-3001
EVA Tools and Equipment Reference Book	750	190	250	JSC-20466
xEMU Data Book	600	150	195	JSC-E-DAA-TN55224
Selected EVA technical library documents (roadmaps, analog debriefs, etc.)	1,600	400	520	NASA EVA Technical Library
Total (approx.)	**57,000**	**14,200**	**18,500**	

load?" – with immediate confirmation or warnings if constraints would be violated. During the EVA, the tool continuously monitors progress: if an astronaut falls behind schedule or deviates from the planned route, the astronauts can ask the AI to recompute the timeline and recalculate metabolic costs for alternative routes to ensure safe return margins. In an anomaly – such as a malfunctioning tool – the AI retrieves step-by-step troubleshooting procedures while simultaneously assessing whether continued work would exceed safe workload or consumable limits, leveraging its embedded metabolic prediction model to advise whether to proceed, adjust tasks, or terminate the EVA. Crucially, all functionality operates offline, providing immediate, explainable guidance even when communications with Earth are delayed or unavailable, while still complementing Mission Control for higher-level decisions when connectivity permits.

5 The RASAGE Project

Currently, we are developing a proof-of-concept toolset named Retrieval & Simulation Augmented Guidance Agent for Exploration (RASAGE). RASAGE is designed as an integrated Earth-independent decision support system that combines a compact on-board LLM with a structured retrieval framework and physical simulation in an agentic RAG pipeline. Its core language component, Microsoft-Phi4-mini-instruct, is an open-source 3.8 billion-parameter model quantized to 4 bits and optimized for small-scale GPUs [2]. We will fine-tune Phi-4-mini on a curated EVA corpus – NASA EVA handbooks, flight rules, astronaut-ground transcripts,

Table 3 Example Physics-Based Simulation Tools.

Module	**Inputs**	**Outputs (per call)**
Path planner	Start/goal, DEM tile, slope limits, oxygen constraint	Optimal path, distance, ascent, travel-time, predicted O_2 usage
Hypogravity metabolic-cost model	Speed, grade, suit mass, g	$\dot{V}O_2$, kcal h^{-1}, O_2 usage
Thermal load estimator	Solar angle, wind, suit RL parameters	Skin/core T traces, coolant margin
Radiation dose calculator	SPENVIS pre-computed look-up & shielding depth	mSv per EVA
Suit-consumables tracker	Task timeline, suit power profile	Remaining Wh, O_2, CO_2 absorbent

anomaly reports, and analog mission debriefs. Complementing the LLM is a vector-indexed knowledge base composed of metadata-annotated, semantically segmented document chunks, built using FAISS (Facebook AI Similarity Search) – an open-source library designed for efficient similarity search and clustering of dense vectors[23][14].

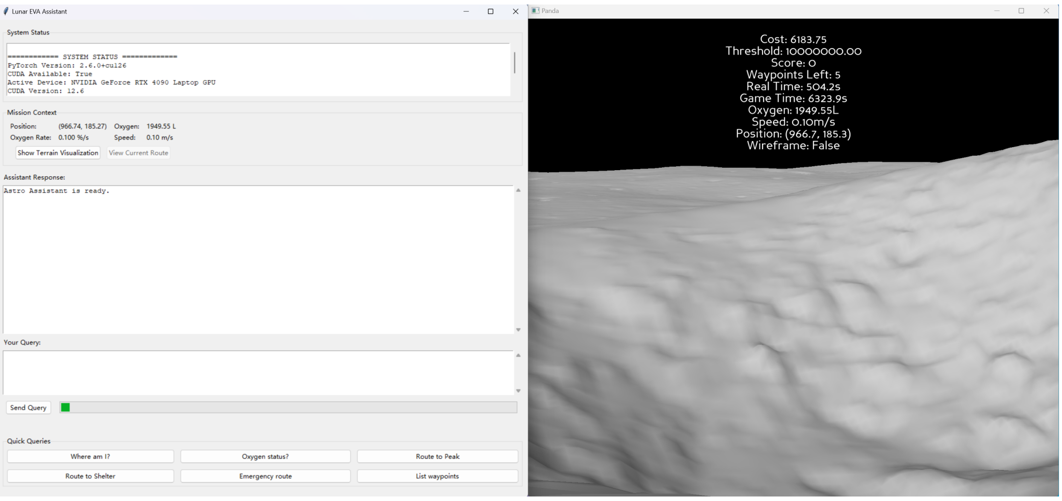
In our system, raw EVA documents are converted to plain text and segmented into coherent chunks using a recursive splitting algorithm that preserves procedural structure and context. Each chunk is enriched with detailed metadata (source, section, page, timestamp) for precise traceability, and is transformed into a dense vector representation via Phi-4-mini's embedding head. These embeddings, along with their metadata, are then indexed in a FAISS vector store, enabling rapid semantic searches. During operations, RASAGE performs semantic retrievals of authoritative procedure snippets, grounding every recommendation in up-to-date source material and ensuring full traceability and explainability.

RASAGE's agentic coordination engine orchestrates a seamless retrieval-augmented workflow that integrates both textual guidance and quantitative simulation outputs via standardized function calls. Domain-specific physical modules, including an A*-based path planner [19] over high-resolution lunar digital elevation models (DEMs) and a hypogravity ambulation metabolic prediction model [28], are indexed and fused into the LLM's context window. The resulting recommendations undergo iterative refinement against safety constraints and mission rules, with every decision logged for auditability. A multi-modal interface (text, voice and tablet display) delivers concise, contextually relevant information – complete with clear rationale and source citations – dynamically adapting output detail based on task urgency and crew workload to maximize situational awareness and trust during EVA.

5.1 Experimental Design

To evaluate the feasibility and effectiveness of the proposed AI system, we will first conduct a series of controlled digital simulation experiments in an EVA environment modeling lunar terrain, a physical astronaut model, and a library of representative tasks. The prototype of the digital simulation interface using Apollo 14 EVA scenarios is shown in figure 2. First, we will evaluate route-planning performance by tasking the AI to generate navigation plans under predefined time and resource constraints and comparing its chosen paths, EVA duration, and energy use against baseline strategies (shortest-distance routing and expert-derived plans). Next, we will validate metabolic model accuracy by benchmarking the AI's consumable-usage predictions (O_2 and battery) against published NASA EVA metabolic data and historical EVA logs [46], quantifying prediction error and iteratively adjusting model parameters until errors consistently fall within predefined safety thresholds.

We will then conduct human-in-the-loop analog trials that simulate off-nominal EVA events (e.g., rising suit CO_2 levels) under realistic communication delays. In these scenarios, we will measure the AI's response latency, correctness of guidance against NASA flight rules, and success in resolving anomalies before Earth-based support could intervene. By comparing paired scenarios with and without the AI assistant, we will quantify impacts on crew task efficiency, error rates, and cognitive workload – demonstrating whether RASAGE maintains or improves EVA performance and safety under deep-space conditions. Additionally, we will log and categorize every participant query to the system, using these data to identify usability gaps and inform iterative interface and capability improvements.



```
============ SYSTEM STATUS =============
PyTorch Version: 2.6.0+cu126
CUDA Available: True
Active Device: NVIDIA GeForce RTX 4090 Laptop GPU
CUDA Version: 12.6
```

Position: (966.74, 185.27) Oxygen: 1949.55 L
Oxygen Rate: 0.100 %/s Speed: 0.10 m/s

```
Astro Assistant is ready.
```

```
Cost: 6183.75
Threshold: 10000000.00
Score: 0
Waypoints Left: 5
Real Time: 504.2s
Game Time: 6323.9s
Oxygen: 1949.55L
Speed: 0.10m/s
Position: (966.7, 185.3)
Wireframe: False
```

Figure 2 Prototype digital simulation interface in Apollo 14 EVA scenarios, showing the Lunar EVA Assistant's user interface (left) and a 3D lunar terrain visualization (right).

After software validation, our next step is a preliminary field evaluation in a nearby outdoor environment that mimics EVA conditions (e.g., uneven terrain, limited visibility, and communications delay). In this lab-led field test, participants wearing mock-up suit gear will use RASAGE to plan routes, respond to simulated anomalies, and complete task checklists while we measure task completion rates, time to resolution, error incidence, and log every user query for iterative design feedback. Building on these results, we will then propose deploying the system in established analog facilities (including NASA facilities such as Desert RATS, NEEMO, or HERA, and other facilities such as the Mars Desert Research Station Utah, MDRS) during designated "crew-autonomous" EVA days under a 20-minute round-trip communication delay. These higher-fidelity trials will collect both quantitative mission-outcome metrics (objective achievement, anomaly resolution speed) and qualitative insights (usability, clarity, trust) to refine RASAGE's interface, guidance phrasing, and operational integration.

5.2 Evaluation Metrics Summary

Across these experiments, success will be measured by both quantitative metrics (e.g., reduction in excess distance traveled, percentage of scenarios where the AI correctly averts a problem, accuracy of model predictions) and qualitative assessments (crew confidence in the system, perceived workload reduction, trust in recommendations). We expect the AI to demonstrate measurable improvements in EVA planning efficiency and contingency response. For instance, we aim for a >20% improvement in energy efficiency of routes (compared to baselines) and near-100% compliance with safety constraints (the AI should never violate a known flight rule). Additionally, we'll evaluate explainability: we will survey users on whether they felt the AI's explanations were sufficient and clear. The ideal outcome is that users not only accept the AI's advice but can also articulate why the plan is what it is, indicating successful knowledge transfer from the AI.

5.3 Risk Analysis

Developing and deploying an AI-driven EVA planning system carries several technical and operational risks that must be carefully managed:

- **Reliability & Safety:** LLM hallucinations or incorrect suggestions could jeopardize crew safety. Mitigations include retrieval-grounded responses, strict constraint enforcement, mandatory source citation for safety-critical advice, and a formal verification & validation process analogous to flight-software certification.
- **Scope & Competence:** The AI's knowledge is inherently limited to its training corpus. RASAGE will self-assess uncertainty ("I don't know") and defer novel or ambiguous scenarios to human judgment, with clear documentation of system boundaries.
- **Computational Constraints:** Space hardware has limited processing capacity. We mitigate this by using a quantized Phi-4-mini model optimized for onboard accelerators, designing a modular architecture that isolates faults, and ensuring graceful degradation if simulation modules fail.
- **Human–System Integration:** Trust and usability are essential in EVA operations. The user interface will surface only mission-critical information under high workload, provide transparent rationale for recommendations, and be refined through analog mission testing to prevent information overload.
- **Accountability & Bias:** All agent reasoning steps, retrieved sources, and simulation outputs are logged for post-event auditing. Crew training exercises – both with and without AI assistance – will calibrate trust and mitigate automation bias.

By embedding these safeguards and following established NASA and industry guidelines for trustworthy AI [36][29][20], RASAGE is designed as a support tool that errs on the side of caution and enhances – not replaces – human decision-making.

6 Future Work

Building on the envisioned proof-of-concept system introduced in Section 4, multiple avenues remain for advancing the technical robustness, operational readiness, and domain applicability of RASAGE and similar AI-driven EVA support systems.

- **Extending Simulation Fidelity:** While our initial focus is on path planning and metabolic modeling, the environment can be expanded to include detailed rover integration, robotic assistance, and realistic environmental hazards like regolith dust accumulation or dynamic lighting conditions in crater shadows. Incorporating these additional parameters would refine route optimization and contingency responses to better mirror actual planetary surface operations.

- **High-Fidelity Analog Testing and Crew Training:** Following controlled digital simulations, the next step involves extended field trials in environments such as the NASA NEEMO underwater habitat, Mars Desert Research Station (MDRS), the Desert RATS analog, or the HERA (Human Exploration Research Analog) facility. By subjecting the system to mission-length EVA simulations and real-time anomalies under operational constraints, we can collect user performance metrics and refine interface design, agent orchestration algorithms, and knowledge-base coverage.
- **Coping with Partial or Degraded Data:** EVA telemetry may be noisy or incomplete, particularly in harsh environments. Future versions of the system can incorporate robust sensor-fusion and error correction mechanisms, using Bayesian inference or state-estimation techniques to ensure reliable feed-forward to the LLM and simulation modules. This will help maintain safe operation even with intermittent sensor failures or degraded communications.
- **Adaptive Autonomy Levels:** The system currently provides planning, retrieval, and reasoning capabilities. Ongoing research could explore dynamic adjustments to the "level of autonomy" based on crew workload, mission criticality, and time constraints. This feature would allow RASAGE to operate in a more advisory capacity under normal conditions, yet assume higher authority for rapid decision-making in critical or time-sensitive scenarios when Earth-based assistance is unavailable.
- **Cross-Mission and Cross-Domain Integration:** Many of the functionalities outlined – such as retrieval-augmented guidance, physical modeling, and multi-modal user interfaces – are relevant beyond EVA, potentially assisting with in-vehicle maintenance, habitat operations, or scientific payload management. Expanding the knowledge base to additional mission domains would increase the utility and cost-effectiveness of onboard AI systems.

7 Conclusion

This paper outlines a vision – and an early technical framework – for an Earth-independent, onboard AI system capable of providing real-time decision support during extravehicular activities (EVAs). By combining a large language model fine-tuned on EVA procedures, retrieval-augmented knowledge bases, physical simulation modules, and a lightweight agentic orchestrator, the proposed system seeks to emulate key aspects of ground-based flight-controller expertise directly within the spacesuit. Planned experiments in route optimization, metabolic load modeling, and anomaly response are designed to quantify improvements in crew efficiency, resource management, and operational safety. Initial results from the RASAGE prototype suggest that a self-contained, offline-capable AI tool can successfully manage complex EVA tasks while respecting established flight rules and safety margins.

Transitioning this concept into a flight-worthy capability entails addressing several persistent challenges. Chief among these are ensuring robust performance under uncertain or incomplete data, generalizing across diverse mission architectures, and meeting the rigorous safety, verification, and transparency requirements of human-rated flight software. Despite these hurdles, recent advances in large language models, retrieval augmentation, and physical simulation point toward a future in which onboard autonomy becomes indispensable for deep-space exploration. As crewed missions extend to the lunar surface, Mars, and beyond, Earth-independent AI decision support promises to enhance astronaut safety, maximize scientific return, and reduce dependence on ground control – thereby laying a critical foundation for sustainable human presence beyond low Earth orbit.

References

1 Andrew F.J. Abercromby, Grace L. Douglas, Kent L. Kalogera, Jeffrey T. Somers, Rahul Suresh, Moriah S. Thompson, Scott J. Wood, Emma Y. Hwang, B. Kyle Parton, and James L. Broyan. NASA Crew Health & Performance Capability Development for Exploration: 2021 to 2022 Overview. In *Proceedings of the 51st International Conference on Environmental Systems (ICES-2022-299)*, St. Paul, Minnesota, July 2022.

2 Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs, 2025. doi:10.48550/arXiv.2503.01743.

3 Airbus. Hello, I am CIMON. Press release, Airbus, February 2018. URL: https://www.airbus.com/en/newsroom/press-releases/2018-02-hello-i-am-cimon.

4 David Autor. Polanyi's Paradox and the Shape of Employment Growth. Working Paper 20485, National Bureau of Economic Research, Cambridge, MA, September 2014. doi:10.3386/w20485.

5 Ernest Bell and David Coan. A Review of the Approach to ISS Increment Crew EVA Training. In *HRSE-16: ExtraVehicular Activity (EVA) Exploration: Stepping Forward into the Universe, AIAA SPACE 2007 Conference & Exposition, Long Beach, California*, 2007. doi:10.2514/6.2007-6236.

6 Ernest Bell, David Coan, and David Oswald. A Discussion on the Making of an EVA: What it Really Takes to Walk in Space. In *SpaceOps 2006 Conference, Rome, Italy, Session: MM-1: Mission Management I*, 2006. doi:10.2514/6.2006-5570.

7 Oliver Bensch, Leonie Bensch, Tommy Nilsson, Florian Saling, Bernd Bewer, Sophie Jentzsch, Tobias Hecking, and J. Nathan Kutz. AI Assistants for Spaceflight Procedures: Combining Generative Pre-Trained Transformer and Retrieval-Augmented Generation on Knowledge Graphs With Augmented Reality Cues, 2024. doi:10.48550/arXiv.2409.14206.

8 Oliver Bensch, Leonie Bensch, Tommy Nilsson, Florian Saling, Wafa M. Sadri, Carsten Hartmann, Tobias Hecking, and J. Nathan Kutz. Towards a Reliable Offline Personal AI Assistant for Long Duration Spaceflight, 2024. doi:10.48550/arXiv.2410.16397.

9 David Bodkin, Paul Escalera, and Kenneth Bocam. A Human Lunar Surface Base and Infrastructure Solution. In *SAS-10: Space Architecture Symposium: Lunar and Planetary Surface Systems and Construction I, Space 2006*, 2006. doi:10.2514/6.2006-7336.

10 Erika Cardenas and Leonie Monigatti. What is Agentic RAG, 2024. Published November 5, 2024; Accessed: March 20, 2025. URL: https://weaviate.io/blog/what-is-agentic-rag.

11 Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors, 2023. arXiv:2308.10848.

12 Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 10088–10115, 2023.

13 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

14 Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library, 2025. arXiv:2401.08281.

15 Charles Dukes. NASA's behavioural health and performance services for long duration space missions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 94(12):e2, 2023. doi:10.1136/JNNP-2023-BNPA.15.

16 Erika Peters. Callisto Technology Demonstration to Fly Aboard Orion for Artemis I. Press release, NASA Johnson Space Center, January 2022. URL: https://www.nasa.gov/missions/callisto-technology-demonstration-to-fly-aboard-orion-for-artemis-i/.

17 Sandra Erwin. Booz allen deploys advanced language model in space. *SpaceNews*, 2023. Accessed: Mar 17, 2025. URL: https://spacenews.com/booz-allen-deploys-advanced-language-model-in-space/.

18 Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020. doi:10.18653/v1/2020.acl-main.740.

19 Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Trans. Syst. Sci. Cybern.*, 4(2):100–107, 1968. doi:10.1109/TSSC.1968.300136.

20 High-Level Expert Group on AI, European Commission, Directorate-General for Communications Networks, Content and Technology. Ethics guidelines for trustworthy AI, 2019. doi:10.2759/177365.

21 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2), 2025. doi:10.1145/3703155.

22 Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. LIMIT: Less Is More for Instruction Tuning Across Evaluation Paradigms, 2023. doi:10.48550/arXiv.2311.13133.

23 Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi:10.1109/TBDATA.2019.2921572.

24 Michael Peter Johnson. *Mission Control: Inventing the Groundwork of Spaceflight*. University Press of Florida, 2015. Ebook; Published October 18, 2015; Language: English.

25 Ernest R. Bell Jr., Victor Badillo, David Coan, Kieth Johnson, Zane Ney, Megan Rosenbaum, Tifanie Smart, Jeffry Stone, Ronald Stueber, Daren Welsh, Peggy Guirgis, Chris Looper, and Randall McDaniel. Mission control team structure and operational lessons learned from the 2009 and 2010 NASA desert RATS simulated lunar exploration field tests. *Acta Astronautica*, 90(2):215–223, 2013. doi:10.1016/j.actaastro.2012.11.020.

26 JSON-RPC Working Group. JSON-RPC 2.0 Specification, 2013. URL: https://www.jsonrpc.org/specification.

27 Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review, 2024. doi:10.48550/arXiv.2401.01519.

28 Logan Kluis. *A Novel Metabolic Model for Ambulation in Hypogravity and Applications to Planetary Extravehicular Activity.* Doctoral dissertation, Texas A&M University, College Station, TX, August 2024.

29 Bhavya Lal and Kate Calvin. NASA's Responsible AI Plan. Technical Report NASA-20220013471, NASA Office of the Chief Scientist (OCS) and Office of Technology, Policy, and Strategy (OTPS), September 2022.

30 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. arXiv:2005.11401.

31 Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback, 2023. arXiv:2303.17651.

32 Jessica J. Marquez, M. L. Cummings, N. Roy, M. Kunda, and D. J. Newman. Collaborative human-computer decision support for planetary surface traversal. In *Infotech@Aerospace*, 2005. doi:10.2514/6.2005-6993.

33 Jessica J. Marquez, Matthew J. Miller, Tamar Cohen, Ivonne Deliz, David S. Lees, Jimin Zheng, Yeon J. Lee, Bob Kanefsky, Johannes Norheim, Matthew Deans, and Steven Hillenius. Future Needs for Science-Driven Geospatial and Temporal Extravehicular Activity Planning and Execution. *Astrobiology*, 19(3):440–461, 2019. doi:10.1089/ast.2018.1838.

34 Natalie A. Mary. Surface EVA Architectural Drivers. Technical report, National Aeronautics and Space Administration, 2023. White paper from the 2023 Moon to Mars Architecture Concept Review.

35 Tim McGrath, Jason Norcross, Jon Morris, Federico Piatti, Fernando Figueroa, Brianna Sparks, and Jeffrey Somers. A decision support system for extravehicular operations under significant communication latency. In *Proceedings of the 52nd International Conference on Environmental Systems (ICES-2023-327)*, Calgary, Canada, July 2023.

36 Edward McLarney, Yuri Gawdiak, Nikunj Oza, Chris Mattman, Martin Garcia, Manil Maskey, Scott Tashakkor, David Meza, John Sprague, Phyllis Hestnes, Pamela Wolfe, James Illingworth, Vikram Shyam, Paul Rydeen, Lorraine Prokop, Latonya Powell, Terry Brown, Warnecke Miller, and Claire Little. NASA Framework for the Ethical Use of Artificial Intelligence (AI). Technical Report NASA/TM-20210012886, NASA, April 2021.

37 Matthew J. Miller and Karen M. Feigh. Assessment of Decision Support Systems for Envisioned Human Extravehicular Activity Operations: From Requirements to Validation and Verification. *Journal of Cognitive Engineering and Decision Making*, 14(1):54–74, 2020. doi:10.1177/1555343419871825.

38 Matthew J. Miller, Kerry M. McGuire, and Karen M. Feigh. Information flow model of human extravehicular activity operations. In *Proceedings of the 2015 IEEE Aerospace Conference*, Big Sky, MT, USA, March 2015. IEEE. doi:10.1109/AERO.2015.7118942.

39 Matthew J. Miller, Kerry M. McGuire, and Karen M. Feigh. Decision Support System Requirements Definition for Human Extravehicular Activity Based on Cognitive Work Analysis. *Journal of Cognitive Engineering and Decision Making*, 11(2):136–165, 2016. doi:10.1177/1555343416672112.

40 NASA. Flight Control Operations Handbook (FCOH) Station Operations. Technical Report JSC-29229, Johnson Space Center, Mission Operations Directorate, Flight Directors Office, 2009. Latest Release.

41 NASA. Flight Control Operations Handbook (FCOH) Shuttle Operations. Technical Report JSC-26843, Johnson Space Center, Mission Operations Directorate, Flight Directors Office, 2010. Final Release.

42 Gregory Navarro, Marie-Christine Desjean, and Alexis Paillet. ECLSS Technology Roadmap at Spaceship FR. In *Proceedings of 52nd International Conference on Environmental Systems (ICES-2023-147)*, Calgary, Canada, July 2023.

43 OpenAI . What are tokens and how to count them?, January 2025. URL: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them.

44 OpenAI. OpenAI o1 System Card. Technical report, OpenAI, September 2024. System Card. URL: https://cdn.openai.com/o1-system-card.pdf.

45 Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large Language Model Connected with Massive APIs, 2023. doi:10.48550/arXiv.2305.15334.

46 Heather L. Paul. Energy Expenditure During Extravehicular Activity Through Apollo. In *Proceedings of the 42nd International Conference on Environmental Systems*, San Diego, CA, USA, July 2012. doi:10.2514/6.2012-3504.

47 Ruslan O. Popov, Nadiia V. Karpenko, and Volodymyr V. Gerasimov. Overview of Small Language Models in Practice. In *Proceedings of the 7th Workshop for Young Scientists in Computer Science & Software Engineering (CS&SE@SW 2024), Virtual Event, Kryvyi Rih, Ukraine, December 27, 2024*, volume 3917 of *CEUR Workshop Proceedings*, pages 164–182. CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3917/paper28.pdf.

48 Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models, 2025. doi:10.48550/arXiv.2502.18443.

49 Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative Agents for Software Development, 2024. arXiv:2307.07924.

50 Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models, 2024. arXiv:2304.08354.

51 Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. arXiv:2303.11366.

52 Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG, 2025. doi:10.48550/arXiv.2501.09136.

53 Tameri Guide for Writers. Word Counts of Common Forms. URL: https://www.tameri.com/format/word-counts/.

54 Christopher Toukmaji and Allison Tee. Retrieval-augmented generation and llm agents for biomimicry design solutions. In *Proceedings of the AAAI Spring Symposium Series*, Stanford, CA, USA, March 2024. doi:10.1609/aaaiss.v3i1.31210.

55 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023. arXiv:2201.11903.

56 Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action, 2023. doi:10.48550/arXiv.2303.11381.

57 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, 2023. arXiv:2210.03629.

58 H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu. Evaluation of Retrieval-Augmented Generation: A Survey. In *Proceedings of 12th CCF Conference, BigData 2024*, pages 102–120. Springer Nature, Singapore, 2025. doi:10.1007/978-981-96-1024-2_8.

59 Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method, 2024. doi:10.48550/arXiv.2402.17193.

60 Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction Tuning for Large Language Models: A Survey, 2024. arXiv:2308.10792.