Evaluation of Artificial Intelligence Implementation in Contaminant Impact Assessments

Jennifer G. Williams[1], Cedric L. Anderson,[2] Aman A. Butt[3], Kayla Carneal[4]

NASA Marshall Space Flight Center, Huntsville, Alabama 35812, USA

Every substance transported to the International Space Station (ISS) is rated for its potential to impact human toxicology, flammability, biosafety, and the environmental control life support systems (ECLSS). The results of these ratings are defined in a Hazard Mitigation Summary Table (HMST). The HMST ratings provide the basis for establishing safety guidance for designing for minimum risk during substance transportation, storage, and use. Rating the impact on human toxicology, flammability, biosafety, and ECLSS for every substance utilized on ISS is time-consuming and requires a large team of subject matter experts. This study evaluates using artificial intelligence (AI) to enhance ECLSS HMST assessment productivity. As a case study, the predicted impacts from particulate spills on ECLSS were evaluated. Our HMST application, Safety and EnvironmenTal Intelligence EvaLuator (SENTINEL), implements an AI-powered automated workflow that provides HMST ratings in seconds. SENTINEL employs Amazon Web Services (AWS) Textract and the Anthropic Claude model equipped with custom tool-calling capabilities to autonomously extract critical chemical properties. These properties feed into automated calculations that quantify the impact of particulate contaminants on ISS ECLSS filters. Specifically, the pressure drop across the filter and functional capacity consumption percentages are estimated, corresponding to HMST ratings. The workflow handles complex mixtures by performing proportional mass calculations for trace constituents based on concentration percentages, enabling comprehensive ECLSS impact assessments for multi-component substances. To support NASA engineering verification and eliminate AI misinformation risks, SENTINEL uses PyMuPDF to automatically highlight AI-extracted values directly on the original source portable document format (PDF) documents by employing Textract's spatial geometry data to precisely locate and mark each extracted value with its corresponding confidence score. This AI-driven approach significantly reduces the time required for ECLSS HMST evaluations while maintaining analytical rigor through validated engineering calculations and providing transparent validation mechanisms for subject matter expert review.

Acronyms and Nomenclature

AI	=	Artificial Intelligence
API	=	Application Programming Interface
AWS	=	Amazon Web Services
CAS	=	Chemical Abstract Service
CRF	=	Condition of Random Field
EAR	=	Export Administration Regulation
ECLSS	=	Environmental Control and Life Support Systems
FCC	=	Functional Capacity Consumption
GovCloud	=	Government Cloud
GROBID	=	GeneRation Of BIbliographic Data

[1] Technical Management (ECLSS Chemist), NASA, ECLSS Systems Team, Space Systems Dept., NASA MSFC, ES62
[2] Information Technology Specialist, NASA, Office of the Chief Information Officer, NASA MSFC, IS01
[3] Pathways Intern, NASA, Information Technology Specialist, Office of the Chief Information Officer, NASA MSFC, IS01
[4] Pathways Intern, NASA, ECLSS Systems Team, Space Systems Dept., NASA MSFC, ES62

HMST = Hazardous Material Summary Table
HVAC = Heating, Ventilation, and Air Conditioning
ISS = International Space Station
ITAR = International Traffic in Arms Regulations
JSON = JavaScript Object Notation
OCR = Optical Character Recognition
PDF = Portable Document Format
RAG = Retrieval Augmented Generation
SENTINEL = Safety and EnvironmenTal INtelligence EvaLuator
SDS = Safety Data Sheets
SME = Subject Matter Expert
SMA = Safety and Mission Assurance
YAML = YAML Ain't Markup Language

I. Introduction

A wide variety of substances travel to the International Space Station (ISS) to support missions. Every substance that travels to the ISS must be rated for its potential impact to human toxicology, biosafety, flammability, and the environmental control life support systems (ECLSS) if inadvertently released aboard the ISS. The ratings are performed by NASA Subject Matter Experts (SMEs) and the output of these ratings are defined in the Hazard Mitigation Summary Table (HMST). The severity of the HMST ratings informs the decision made by safety review panels regarding the levels of containment and relevant design and operational practices required for each substance to minimize risk during transport to and from the ISS as well as its use while on board the ISS. The definitions of the ratings are shown in Table 1. [1]

ECLSS Rating	**Functional Capacity Consumption (FCC) and System Maintenance Information**
E0	Functional capacity consumption is <2%. 100% of the functional margin is retained. No ECLS functional performance degradation. No change in scheduled maintenance.
E1	Functional capacity consumption is >2% and <10%. >10% of the functional margin is consumed. No ECLS functional performance degradation. No change in scheduled maintenance.
E2	Functional capacity consumption >10% and <25%. >25% of the functional margin is consumed. ECLS functional performance is degraded by <10%. Early replacement of consumable components may be necessary within nine months.
E3	Functional capacity consumption >25% and <50%. >50% of the functional margin is consumed. ECLS functional performance is degraded by >10% and <25%. Early replacement of consumable components may be necessary within six months.
E4	Functional capacity consumption >50% and <75%. >75% of the functional margin is consumed. ECLS functional performance is degraded by >25% and <50%. Early replacement of consumable components may be necessary within one month.
E5	Functional capacity consumption >75% and <90%. 100% of the functional margin is consumed. ECLS functional performance is degraded by >50% and <75%. System maintenance is required to restore functional performance within one week.
E6	Functional capacity consumption >90%. 100% of the functional margin is consumed. ECLS functional performance is degraded by >75%. System maintenance is required to restore functional performance within one day.
N/A	Does not impact the functional capacity consumption or scheduled maintenance.

Table 1: Definition of ECLSS ratings. [1]

For example, a substance that would severely degrade the ECLSS functional capacity would rate as an E6. The E6 rating informs the safety assessment that determines how that substance will be packaged and handled onboard ISS to achieve a design for minimum risk. Typically, an E6 rating will require that the item have two levels of containment in case there is a spill. [2] On the other hand, a substance that reduces the functional capacity by 3% would receive a E1 rating, which would help inform the packaging and usage requirements determined by the safety board. This careful preparation ahead of the launch promotes end-to-end mission safety.

Due to the number of substances transported to and used aboard the ISS, the time required by a team of subject matter experts to evaluate each one can be substantial. Artificial intelligence (AI) has been shown to have the potential to enhance productivity and was investigated for application to the HMST substance review process. As a proof-of-concept application, hypothetical solid chemical substances were evaluated by both AI and SMEs. For this proof-of-concept, the functional capacity consumption is defined as the predicted pressure drop across the ISS particulate HVAC filters caused by an accumulation of a solid substance under evaluation divided by the total expected pressure drop over the service life of the ISS HVAC filters. This proof-of-concept calculation mimics the actual calculations employed by SMEs to determine the ECLSS HMST rating for solid substances.

The artificial intelligence application Safety and EnvironmenTal INtelligence EvaLuator (SENTINEL) was developed for this proof-of-concept to provide a hazard analysis rating. Amazon Web Services (AWS) Bedrock was the artificial intelligence hosting platform. AWS was selected because it met the International Traffic in Arms Regulations (ITAR) and Export Administration Regulation (EAR) security requirements by being hosted in GovCloud. Initially, training material, including calculation spreadsheets and tutorials, were loaded into AWS; however, it was not able to learn from these materials to complete the HMST evaluations. Therefore, researchers selected another approach to define a specific pipeline that could be used as a proof-of-concept. The pipeline for SENTINEL is illustrated by Fig. 1. Anthropic's Large Language Model (LLM) Claude was used to calculate the functional capacity consumption and assign the ratings. Tools were written to compile information, perform calculations, and determine the ratings.

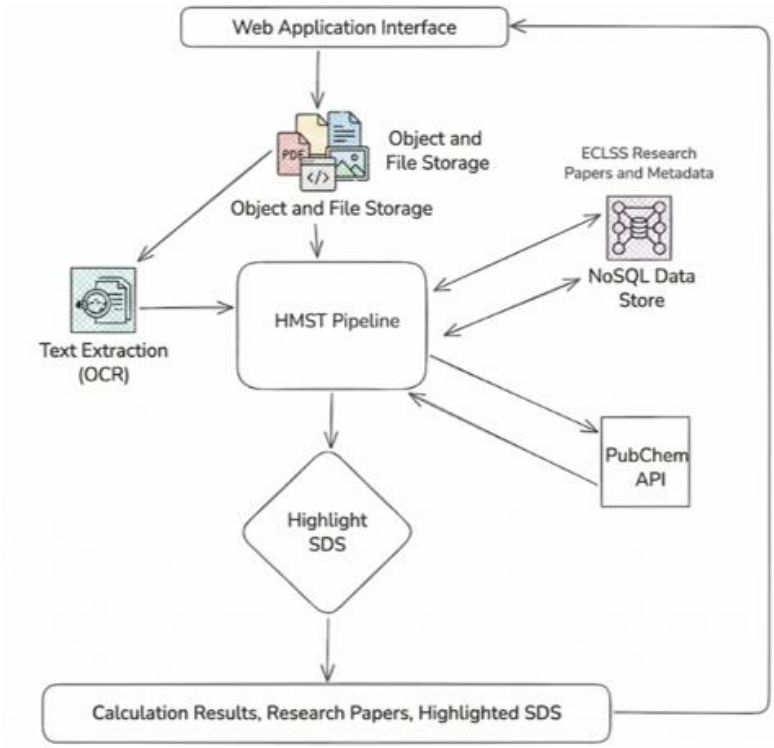


Figure 1. SENTINEL pipeline demonstrating the approach used for AI evaluations of HMST ratings.

In order for SENTINEL to perform the functional capacity consumption calculations, key chemical information such as density was required. This information is often contained within the chemical's safety data sheet (SDS) and must be extracted for use in the calculation. AWS Textract was the Optical Character Recognition (OCR) that

SENTINEL employed to extract key information from the SDS in portable document format (PDF). AWS Textract then formats the relevant content into a defined structure in a JavaScript Object Notation (JSON) file format which can be utilized for subsequent calculations. AWS Textract was selected for its demonstrated effectiveness with complex unstructured and structured data and capacity for high-throughput processing. [3] Once the necessary data were located in the SDS, SENTINEL employed PyMuPDF to highlight the extracted information to facilitate SME accuracy review. However, not all SDS documents contain the information needed to complete the ECLSS rating. When the required information was not available in the SDS, an AI orchestration service was developed in-house for SENTINEL to retrieve the information from specified reference sources. This service was limited to retrieving information from the chemical manufacturer listed on the SDS, the PubChem or Sigma-Aldrich websites, or published journal articles. Any information retrieved by SENTINEL was referenced and SMEs verified accuracy.

Once the chemical information was obtained, SENTINEL then called specific tools to perform the analysis. These tools defined the calculation steps necessary to estimate the service life reduction for a particular situation. To prevent misleading, inaccurate, or unpredictable results, SENTINEL was configured with highly restrictive parameters that suppressed creative reasoning and limited the LLM model to deterministic outputs. Complete process documentation was maintained for the sample set to ensure transparency, sources were cited, and accuracy was verified by a SME. [4] [5] [6] This allowed the creation of a proof-of-concept AI system that was transparent, reliable, validated, and trustworthy for this application.

II. Method

A. Evaluation Criteria

The dataset contained 30 sample HMST submissions that were evaluated by SMEs and the AI application SENTINEL. The accuracy and efficiency were measured as the evaluation criteria. Accuracy was defined as the total correct ratings divided by the total number in the sample set. Efficiency was defined by the time required to complete the task.

B. Software Architecture and Pipeline Implementation

The HMST AI pipeline SENTINEL was developed as a modular, extensible system implementing multiple software engineering design patterns to ensure maintainability and scalability. Engineering design patterns are reusable, proven solutions to commonly recurring software design problems that provide a template for structuring code to achieve specific architectural goals such as flexibility, maintainability, and extensibility. The architecture employs a strategy pattern for AI provider abstraction, enabling interchangeable LLM backends (Anthropic Claude and Google Gemini) without modifying core logic. A factory pattern facilitates runtime provider selection, while a dispatcher pattern routes function calls through a centralized class for tool handling. This design allows the system to seamlessly integrate multiple large language models while maintaining consistent query processing through a template method pattern implemented in the base AIProvider abstract class.

The pipeline for SENTINEL consists of four primary components: PDF document text extraction, AI provider integration, tool function execution, and asynchronous PDF document highlighting. Each component operates independently while communicating through well-defined interfaces, enabling parallel development and testing. The system was implemented in Python 3.11+ using modern asynchronous programming patterns. Key dependencies include PyMuPDF for PDF document manipulation, Boto3 for AWS service integration, Anthropic and google-generativeai for AI provider application programming interface (API), requests for PubChem queries, and PyYAML for configuration management. The codebase follows object-oriented design principles with clear separation of concerns across multiple modules handling document processing, provider implementations, tool schemas, execution logic, engineering calculations, AWS Textract operations, and PDF document annotation coordination.

C. Large Language Model Integration

Claude (Anthropic via Vertex AI, model: claude-sonnet-4@20250514) was the primary LLM utilized in this study, while Gemini (Google Generative Language API, model: gemini-2.5-flash-preview-05-20) was used for a comparison analysis for a subsequent integration. Both models are integrated through a unified abstract interface, with each provider implementing the identical four core methods that return the providers name, lists the tools available to the LLM, builds the request to send, and then makes the API call with that request. The response is then processed through

the method. The abstraction enables direct comparison of model performance while maintaining identical tool definitions and system prompts across providers.

The LLMs utilized by SENTINEL were equipped with the ability to utilize eight specified tools. Tools are software functions or programs that extend an LLM's capabilities beyond text generation. These tools enable the models to perform functions such as structured data extraction, external API calls, engineering calculations, and PDF document annotation requests. Tool definitions specify function names, parameter schemas with type validation, and detailed descriptions that guide the AI in appropriate tool selection and usage.

The SENTINEL pipeline implements an iterative conversation loop where the LLM's queries are processed through multiple tool-calling cycles. SENTINEL analyzes the query, determines required tools, executes tool calls, receives results, and repeats until sufficient information is gathered to provide a complete answer. This iterative approach enables complex multi-step workflows such as the full functional capacity consumption calculation sequence, which requires extracting density and particle size, calculating mixture-level filter impact, extracting composition information, searching PubChem for each component, and calculating component-level filter impacts.

D. Data Retrieval and Text Extraction

Information from SDS was needed to determine the ECLSS ratings. SDS were processed using the Python library PyMuPDF for the initial text extraction, preserving document structure and formatting. The extracted text serves as the primary context for AI analysis. For geometric data required in the highlighting workflow, AWS Textract performed asynchronous document analysis, generating detailed JSON output files containing word-level and line-level bounding boxes with associated confidence scores. The AWS Textract processing pipeline uploads local PDF document files to Amazon S3, initiates asynchronous analysis jobs, and polls for completion using exponential backoff to optimize resource utilization. AWS Textract JSON responses are cached locally to eliminate redundant API calls during iterative development and testing. AWS Textract accuracy was evaluated for its ability to extract relevant information from the SDS. To support subject matter expert verification and eliminate AI misinformation risks, a highlighting system was developed using AWS Textract geometric data and PyMuPDF rendering capabilities.

The SENTINEL workflow operates in three phases. First, AWS Textract processing uploads the input PDF to S3, initiates an asynchronous document analysis job, polls for completion, and downloads the resulting JSON file containing geometric data for every word and line in the document. Second, during the main query processing loop as SENTINEL extracts values from the SDS text, the orchestration layer calls the highlighting tool for each extracted data point, specifying the exact text string, category, reason, and match type. These requests accumulate in an array that tracks the highlight requests (highlight_requests). Third, array requests are converted to PDF coordinates using page dimensions from Textract metadata. The HighlightManager class matches each requested text string to AWS Textract blocks using exact matching (single words), contains partial matching, or phrase matching (multi-word sequences). After the blocks and values have been found, PyMuPDF applies yellow highlight annotations at the calculated positions. This approach provides transparent validation by visually marking every AI-extracted value in the source document, enabling rapid verification by NASA SMEs without manual text searching.

The HMST submission often is missing information necessary to complete the rating calculations. The process to retrieve information not contained in a SDS was developed in-house. In such cases, our tools to find external information were used. Given that accuracy is paramount, SENTINEL was limited to searching specified sources and cited all references. The LLM was instructed to first search for information on the chemical manufacturer's website listed in the SDS. If the information was still unavailable, SENTINEL would then search journal articles, the Sigma-Aldrich website, and the PubChem website. A SME reviewed and verified information gathered by SENTINEL for accuracy and relevant experimental conditions, such as the temperature at which the density values were measured.

E. Calculations and Assigning Hazard Scores

The functional capacity consumption analysis is nominally performed by a SME utilizing a Microsoft Excel spreadsheet with pre-loaded calculations. These calculations were converted into a tool and can be recalled and used by SENTINEL. No calculations outside of the defined calculations are allowed in order to eliminate unpredictability and inaccuracy.

For multi-component mixtures where the density is not defined in the SDS, the pipeline performs proportional mass calculations for each constituent based on concentration percentages extracted from Section 3 of the SDS. Component-specific densities are retrieved from various approved APIs, and individual calculations are executed for each constituent, enabling comprehensive impact assessment of complex chemical mixtures.

F. Configuration Management and Error Handling

System parameters are managed through a centralized YAML configuration file containing specified ISS environmental constants and AWS credentials (S3 bucket, region, Textract output prefix). This configuration-driven approach enables rapid adaptation to different spacecraft environments or mission parameters without code modification.

To prevent misleading, inaccurate, or unpredictable results, several validation mechanisms were implemented. AI creativity parameters were minimized to enforce deterministic behavior. Tool functions validate input parameters using type checking and range validation. External API calls include timeout handling and retry logic with exponential backoff. API searches return source attribution for all retrieved data. The highlighting system reports match confidence and unmatched requests to identify potential extraction errors. All calculations are logged with intermediate values to enable step-by-step verification by subject matter experts.

All extracted data are accumulated in a structured dictionary, deduplicated to prevent redundant information, and serialized to YAML format for output. The YAML structure includes metadata, the user query, a human-readable answer, extracted chemical properties, calculation results for mixture and components, and API search results with source attribution. SENTINEL's entry point accepts the PDF file path, natural language query, particle mass, AI provider selection, and optional flags for highlighting and output directory specification. Execution produces YAML output files, highlighted PDFs (when requested), and cached Textract JSON for reuse in subsequent runs.

G. Research Paper Integration and Ground Truth Validation

To support engineering validation and provide contextual grounding for HMST assessments, a research paper processing pipeline was integrated into the SENTINEL platform. This subsystem employs GeneRation Of BIbliographic Data (GROBID), an open-source machine learning library for extracting structured information from a database of peer-reviewed scientific literature. GROBID utilizes a cascade of Conditional Random Field (CRF) models, each specialized for different document zones including headers, body text, and references. [7] The system processes PDF documents through multiple stages: PDF layout analysis via pdfalto, document segmentation using a trained CRF model, zone-specific extraction using eleven specialized CRF models, and output generation in TEI-XML format conforming to the Text Encoding Initiative standard. [8]

In the implementation, GROBID is deployed as a containerized microservice (version 0.8.0), processing ECLSS technical papers retrieved from AWS S3 storage. The S3PaperProcessor class orchestrates the pipeline: PDF documents are downloaded via Boto3, submitted to GROBID's /api/processFulltextDocument endpoint, and the resulting TEI-XML is parsed to extract titles, authors, abstracts, and section content, as shown in Fig. 2. Extracted metadata is stored in DynamoDB with compound text indexes enabling full-text search across titles, abstracts, and section content. Processing achieves speeds of 2-5 seconds per page with over 90% extraction accuracy on well-formatted scholarly documents. The system implements idempotent processing with S3 key-based deduplication, enabling incremental corpus updates without reprocessing existing documents.

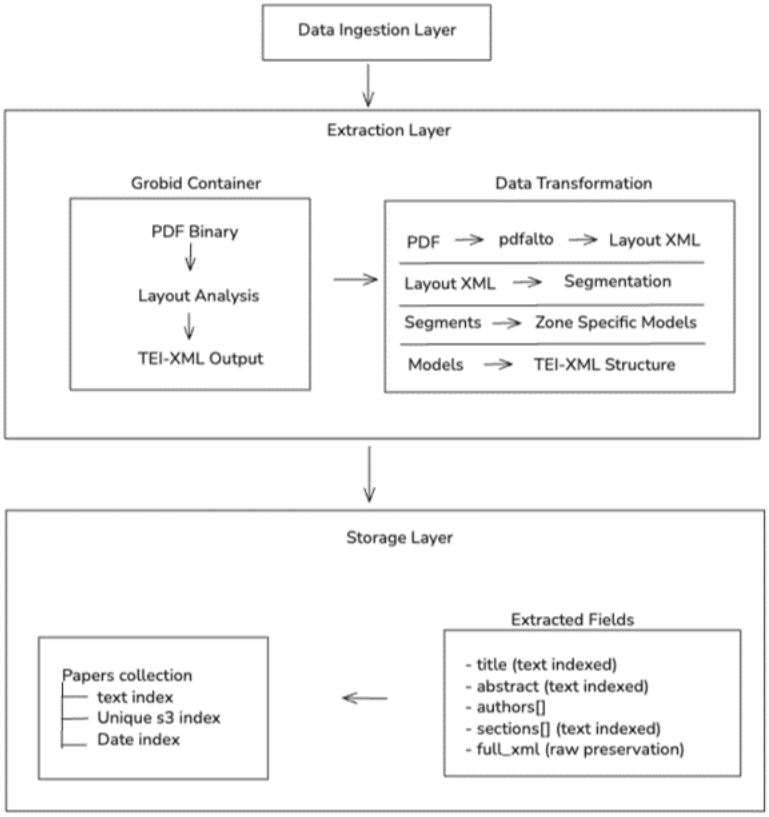


Figure 2. Pipeline for PDF document processing. [9]

The platform implements a Retrieval-Augmented Generation (RAG) architecture that combines GROBID-extracted content with large language models (AWS Bedrock Claude) for intelligent research assistance. User queries trigger DynamoDB text searches that retrieve relevant papers based on GROBID-extracted content. The system assembles context from paper titles, abstracts, and key sections, then submits this context along with the user query to the LLM for response generation. This approach grounds SENTINEL AI responses in actual ECLSS research literature while providing automatic citation of source papers, thus reducing misinformation risks through retrieval constraints and enabling domain-specific knowledge augmentation.

The GROBID-processed corpus integrates directly with SENTINEL to provide research context for engineering calculations. When SENTINEL performs functional capacity consumption calculations, it automatically generates search terms from the calculation context and retrieves relevant ECLSS papers to support and validate results. Papers are presented alongside calculations for engineering review, enabling SMEs to verify that SENTINEL AI-generated assessments align with established research findings. This integration ensures that HMST ratings are grounded in peer-reviewed literature and provides transparent validation mechanisms that enhance confidence in SENTINEL AI assisted safety assessments for critical aerospace applications.

H. Deterministic System Validation

To demonstrate the importance of structured AI tooling for mission critical engineering calculations, a consistency study was conducted comparing the structured SENTINEL HMST AI pipeline against raw LLM prompting approaches. This validation addresses fundamental concerns about AI reliability in safety critical aerospace applications where accuracy and repeatability are paramount. Five representative test cases were randomly selected from the validation dataset, covering a range of particle masses (1 g to 1200 g), particle densities (1.165 g/cm^3 to 11.36 g/cm^3), particle diameters (0.149 mm to 1.5 mm), and expected filter life reduction values (0.12% to 38.8%). Test materials included iron(III) oxide, lead, silica gel, ion exchange resin, and JSC-1A lunar regolith simulant. Each test case was evaluated using four methods: SME ground truth (expert-verified calculations

serving as baseline), the structured SENTINEL HMST-AI pipeline (tool-calling architecture with programmatic functional capacity consumption equation calculation), raw Claude prompting (direct API calls with formula and inputs), and raw Gemini prompting (direct API calls with formula and inputs).

Raw LLM methods received detailed prompts containing all input parameters, functional capacity consumption calculation instructions, and explicit requests to return only numerical results. Each raw LLM method was called five times per test case with identical prompts to measure accuracy (proximity to SME ground truth), consistency (variance across multiple runs with identical inputs), and reliability (whether the model produces valid numerical output).

III. Results

This initial proof-of-concept AI-HMST pipeline, SENTINEL, evaluated a dataset of 30 example compounds and matched the SME generated results with 90% accuracy for that dataset, as shown in Table 2. Three correctable differences were noted between the SME rating and SENTINEL's rating. In Test 18 (activated charcoal), the bulk density was listed in the SDS and was used for the calculation by SENTINEL, yielding a very conservative rating of E6 compared to the SME rating of E2. While bulk density is useful in calculating the pressure drop across a filter, our example calculation is programmed to utilize intrinsic density. Using the bulk density in place of intrinsic density yielded an error we anticipated when selecting this SDS for the example dataset. SENTINEL will be updated to include an option for bulk density calculations, enabling the pipeline to distinguish between bulk density and intrinsic density to prevent future errors. If the E6 rating had been used without SME verification, it would have likely resulted in the safety panel recommending more conservative packaging and use aboard ISS.

The other two examples that did not match SME ratings were Test 20 (mercury in iodized carbon) and Test 27 (silica inorganic sorbent). In these examples, the density was not listed in the SDS and had to be obtained by SENTINEL. In the majority of examples, SENTINEL obtained an accurate density. However, in these particular examples the density differed from what the SME used. For mercury in iodized carbon, SENTINEL estimated the intrinsic density as 1.5 g/cm^3 while the SME estimated the density to be 2.1 g/cm^3. This yielded a more conservative rating of E3 by SENTINEL whereas the SME rated it as an E2. In this example, an E2 or E3 rating would likely yield one level of containment required by the safety panel. However, if there was slightly more of the substance being flown to ISS and SENTINEL returned a rating of E4 while the SME rated it as an E3, this could be the difference between one and two levels of containment as directed by the safety panel. In Example 27 (silica inorganic sorbent), the SDS reported the composition was 90-100% silicon dioxide, which should have yielded a density around 2.2 g/cm^3. However, SENTINEL used a density value of 4.6 g/cm^3, yielding a less conservative rating of E2 instead of the SME rating of E3. Currently, SENTINEL is instructed to find five sources of density if density is not listed in the SDS and use the highest value. However, this does not yield a conservative result, so SENTINEL will be updated in future work to select the lowest value from the five intrinsic densities and flag any discrepancies for SME review, thus yielding a more conservative answer. Ideally, there is a balance between conservative ratings and levels of containment for safety that potentially restrict experimental processes. If too many levels of containment are required, it may hinder experimental progress, but if too few levels of containment are designated, it could pose a risk to the crew in the event of a spill. Based on these two examples, future work will explore refining the density estimations to further increase accuracy while maintaining the appropriate level of containment for a safe mission.

Additionally, in Test 17 (molecular sieve), SENTINEL increased rating accuracy by locating a more accurate density reference than what was originally estimated by the SME. SENTINEL located the exact composition of the molecular sieve in the SDS and then located a reference density for that material that had not been located by the SME. The SME updated their calculations to reflect the density provided by SENTINEL.

Of note, SENTINEL did an excellent job of interpreting the definition of compounds that should rate as "N/A" and applying it to the four example cases – Tests 5, 21, 22, and 25. In these instances, SENTINEL had to interpret the definition of substances that rate N/A and correctly apply that definition to substances it had not evaluated previously. This pleasantly surprised the team as it required applying abstract definitions accurately to concrete examples.

Table 2. SENTINEL demonstrated a 90% accuracy when compared to SME ratings. *[a] More conservative rating due to use of bulk density from the SDS. [b] Difference in rating due to density used by the pipeline.*

Dataset Results						
Test ID	**Chemical**	**SME FCC (%)**	**AI FCC (%)**	**Delta**	**SME Rating**	**AI Rating**
Test 1	iron (III) oxide pellets	8.2	8.2	0.0	E1	E1
Test 2	lead	0.1	0.1	0.0	E0	E0
Test 3	iron (III) oxide	0.0	0.0	0.0	E0	E0
Test 4	silica gel	12.2	12.2	0.0	E2	E2
Test 5	bovine serum albumin	-	-	-	N/A	N/A
Test 6	Dithiothreitol	0.1	0.0	0.1	E0	E0
Test 7	gold powder	0.9	1.6	0.7	E0	E0
Test 8	ion exchange resin	38.8	38.8	0.0	E3	E3
Test 9	iron metal filings	0.2	0.2	0.0	E0	E0
Test 10	JSC-1AF regolith	39.8	39.8	0.0	E3	E3
Test 11	JSC-1A regolith	0.1	0.1	0.0	E0	E0
Test 12	JSC -1AC regolith	0.0	0.0	0.0	E0	E0
Test 13	Nickel	0.1	0.1	0.0	E0	E0
Test 14	Zirconium	0.1	0.2	0.1	E0	E0
Test 15	polyfill_30	0.0	0.0	0.0	E0	E0
Test 16	Desiccant	15.6	15.3	0.3	E2	E2
Test 17	molecular sieve	7.4	7.4	0.0	E1	E1
Test 18	activated charcoal (Brand 1)	15.9	113.3	+97.4	E2	E6[a]
Test 19	activated charcoal (Brand 2)	27.5	27.5	-0.0	E3	E3
Test 20	mercury in activated carbon	24.2	33.9	+9.7	E2	E3[b]
Test 21	Interleukin	-	-	-	N/A	N/A
Test 22	RNA Later	-	-	-	N/A	N/A
Test 23	aluminum oxide pellets	0.1	0.1	0.0	E0	E0
Test 24	glass beads	0.0	0.0	0.0	E0	E0
Test 25	Antibody	-	-	-	N/A	N/A
Test 26	paraffin wax pellets	0.3	0.3	0.0	E0	E0
Test 27	silica inorganic sorbent	43.1	22.1	21.0	E3	E2[b]
Test 28	silica gel spherical pellets	93.8	93.8	0.0	E6	E6
Test 29	nylon pellets	22.1	22.1	0.0	E2	E2
Test 30	$CaCl_2$ desiccant	15.6	15.6	0.0	E2	E2

In addition to accuracy, efficiency was also measured. The average analysis time for each example calculation was 5.2 minutes by a SME versus 1.2 minutes by SENTINEL, indicating a 4.3 times (~77%) reduction in analysis time in this initial proof-of-concept, as shown in Fig. 3. While 5 minutes may seem minimal, that is the time to evaluate one chemical in this data set and sometimes there can be hundreds of chemicals submitted for HMST evaluation. Future work will reduce SENTINEL analysis time to approximately 12 seconds per compound by changing the code to only search for external resources if they are not available in the SDS and running certain portions of code simultaneously instead of sequentially.

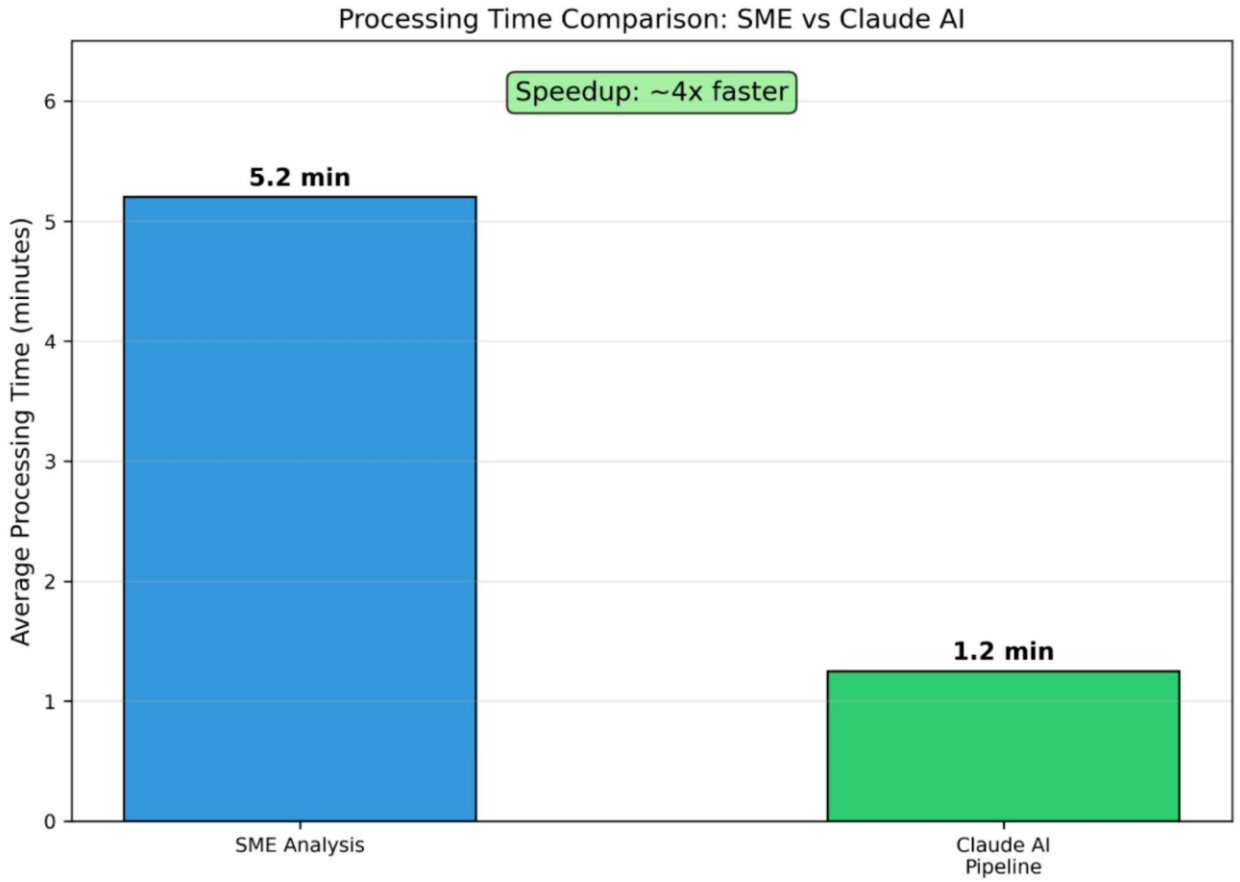


Figure 3. SENTINEL improved efficiency by reducing analysis time by 4.3 in this initial proof-of-concept.

SENTINEL results were compared to results from commercial-off-the-shelf versions of LLMs, Claude and Gemini. For the five randomly selected representative deterministic validation cases, SENTINEL demonstrated measurable accuracy, matching SME ground truth within 0.01% for all test cases with zero variance across runs due to deterministic programmatic calculations. In contrast, raw Claude results consistently calculated values 5-10 times lower than correct values, though it maintained a measured variance of (0.05-0.17 standard deviation) indicating consistent systematic error. Raw Gemini results exhibited catastrophic failures, frequently returning input values instead of calculations (e.g. mass = 1000 g returned as 1000% life reduction) with extreme variance across runs (0.4-575 standard deviation), rendering it unreliable for engineering calculations with only approximately 60% of runs producing valid numerical output. Results are depicted in Fig. 4.

These results demonstrate that while LLMs can follow complex instructions, they are fundamentally unreliable for precise engineering calculations and require a structured pipeline architecture, such as was done for SENTINEL, for success. The structured SENTINEL pipeline architecture, which leverages LLMs for information extraction but delegates calculations to deterministic code, provides the accuracy, consistency, and reliability required for safety critical aerospace applications. This validation study confirms that the hybrid approach, using AI for natural language understanding and information extraction while also employing traditional programming for mathematical precision is essential for gaining trust in AI systems at the mission level where full transparency and repeatability are not suggestions, but requirements.

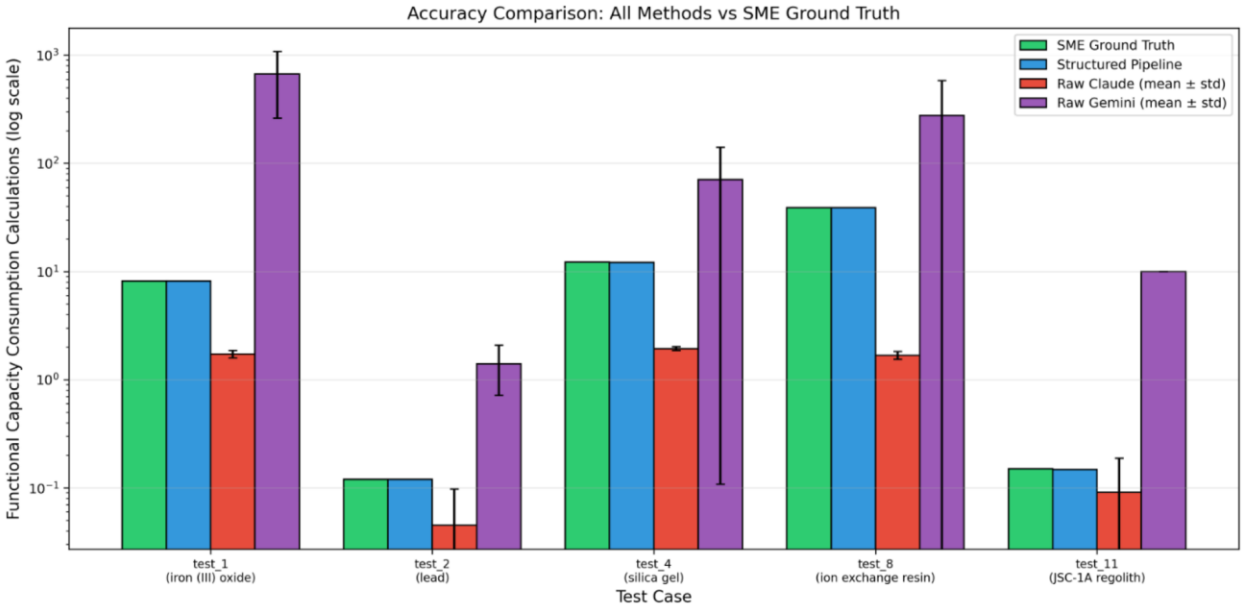


Figure 4. Deterministic validation test cases demonstrating the need for the SENTINEL pipeline compared to raw Claude and Gemini LLM results. *SENTINEL pipeline results in blue match the SME generated results shown in green. Large variations in results are seen in the raw Claude results in red and raw Gemini results in purple. This demonstrates the need for the structured SENTINEL pipeline architecture.*

IV. Conclusion

The AI HMST application, SENTINEL, rapidly retrieved physical property data from reliable sources, performed calculations, and made rating recommendations. This initial proof-of-concept demonstrates that SENTINEL reduced the time of hazard analysis by 4.3 times (~77%) with an accuracy of 90% while maintaining analytical rigor for accuracy and reliability as well as documentation for transparency. The root cause of the three ratings that did not match SME ratings were identified and will be corrected. While AI cannot yet replace SME evaluations, we demonstrated that it can greatly improve the efficiency at which the evaluations are performed and can be used as a tool by the SME, as was demonstrated when SENTINEL located more accurate density references or rapidly retrieved physical properties needed for the calculations. To prevent SENTINEL from being a separate tool SMEs must learn, we envision a future state where this application would be integrated into the current HMST submission platform, pending program adoption. Within the existing HMST submission platform, we envision the missing physical property information in HMST submissions would be input by SENTINEL with references, user submitted physical properties would be verified for accuracy, and suggested calculations and ratings would be displayed for SME verification. This seamless integration into the existing platform would greatly enhance the SME rating experience, ease adoption, and reduce SME evaluation time. Additionally, integration into the existing platform allows for documentation of calculations and reference physical parameters that the SME may have not located alone. We also envision providing users with an HMST screening tool that offers insight into ratings prior to submission. Many HMSTs are updated and resubmitted after the user receives E6 ratings. A screening tool may eliminate some of the re-work and give users insights on their expected HMST ratings earlier in the process allowing them more time to make updates prior to launching their experiment.

Near-term future work will include increasing the efficiency another order of magnitude by streamlining SENTINEL, improving accuracy by refining API parameters, and evaluating a larger sample set to improve statistical significance. Future work may also include easy to implement analysis like flammability ratings alongside more complex analysis for rating liquid and gas spills aboard ISS. The ISS is set to sunset around 2030, and this architecture was developed to be easily adapted to support future crewed vehicles, space stations, and space exploration missions. LLMs will continue to advance, and newer models are expected to have improved performance. The rigorous SENTINEL pipeline is designed with the LLM extracted to easily leverage the state-of-the-art LLM as the technology advances.

Acknowledgments

The authors gratefully acknowledge Jay Perry, Matthew Kayatin, and Robby Newton for mentorship and discussions on HMST; Keyton Fike for discussions on Safety and Mission Assurance; Scott Tashakkor for conversations regarding data quality; and Anthony Maenza for reviewing our pipeline architecture design.

References

[1] J. Perry, "The Role of System Compatibility and Cabin Environmental Impact Assessment in Environmental Control and Life Support System Design and Flight Operations," in *48th International Conference on Environmental Systems*, Albuquerque, New Mexico, 2018.

[2] *NASA SSP 51721 ISS Safety Requirements Document.*

[3] D. Nancy, S. Alwyn, A. Vinora, M. Soundarya, S. A. G. Mohammed and A. S. Mohammed, "Efficient Information Retrieval: AWS Textract in Action.," in *Proceedings of International Conference on Communication and Computational Technologies*, Japur, 2023.

[4] N. Maleki, B. Padmanabhan and K. Dutta, "AI Hallucinations: A Misnomer Worth Clarifying," in *2024 IEEE Conference on Artificial Intelligence (CAI)*, Singapore, 2024.

[5] D. S. Z. Z. Y. W. Yujie Sun, "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content," *Humanities and Social Sciences Communications ,* vol. 11, p. 1278, 2024.

[6] S. M. W. a. V. Prybutok, "The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation," *Information,* vol. 15, no. 6, p. 200, 2024.

[7] P. Lopez, "Automatic Extraction and Resolution of Bibliographical References in Patent Documents," in *Proceedings of the 3rd International Workshop on Patent Information Retrieval (PaIR'10)*, Toronto, 2024.

[8] P. Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," in *Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, Corfu, 2009.

[9] L. Romary and P. Lopez, "GROBID - Information Extraction from Scientific Publications," *ERCIM News,* pp. 31-32, 2015.