https://doi.org/10.31436/iiumej.v27i2.4061

Generative AI Models: A Comparison of Application Analysis on Web AI-Based Decision Support Systems for Satellite Anomaly Identification

ABDUL MUTHOLIB[1,2], NADIRAH ABDUL RAHIM[1*], TEDDY SURYA GUNAWAN[1], AHMAD SHAH HIZAM MD YASIR[3]

[1]*Department of Electrical and Computer Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, 53100, Gombak, Kuala Lumpur, Malaysia*
[2]*Department of Information Systems, Faculty of Science and Technology, UIN Syarif Hidayatullah Jakarta, Benten 15412, Indonesia*
[3]*Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates*

**Corresponding author: nadhirahabdulrahim@iium.edu.my*

(Received: 24 November 2025; Accepted: 19 January 2026; Published online: 10 May 2026)

ABSTRACT: The rapid innovation of Generative Artificial Intelligence (GenAI) has transformed Decision Support Systems (DSS) across various domains, including satellite operations. This paper presents a comparative analysis of four free Generative AI models, including Gemma 3 by Google, Llama 4 Maverick by Meta AI, Nemotron Nano 2 by NVIDIA, and Devstral Small by Mistral, in the context of integrating them into a Web AI-based Decision Support System (DSS) for satellite anomaly identification. Using a dataset of over 4,455 satellite anomaly records covering 1957 to 2024 provided by Seradata, with scalability and adaptability to diverse mission profiles. We evaluate these models by generating anomaly analyses for clarity, accuracy, completeness, and relevance across the Incident Overview, Reliability Trend, Insight, and Stakeholder Recommendation categories, using a 5-point Likert scale and Fleiss' Kappa for internal consistency. The comprehensive evaluation of GenAI models for Web AI-based DSS delineates a clear performance stratification, with scores of 4.44 and 4.39 for Nemotron Nano 2 and Llama 4 Maverick, respectively, confirming their positions as the leading systems based on overall Likert scores. However, the analysis further revealed a critical trade-off between absolute quality and internal consistency (Fleiss' Kappa). The superior models, Nemotron Nano 2 and Llama 4 Maverick, achieved high Likert scores by displaying pronounced performance peaks but suffered the lowest internal predictability, with $\kappa = 0.18$ on Llama 4 Maverick, indicating a highly volatile output structure in which strong clarity often masked critical incompleteness. Conversely, Devstrall Small, despite its suboptimal mean score of 3.44, demonstrated the highest internal consistency, with $\kappa = 0.66$. This robust predictability, even at a lower level, underscores a significant implication for DSS development. The model selection must prioritize the required balance between absolute performance ceiling and the predictability of the output structure. The findings highlight the potential of GenAI implementation in DSS to enhance the reliability of satellite operations, while exploring future directions for research and development in this area. This research contributes to the development of more resilient and intelligent satellite anomaly identification systems, with broader implications for space mission safety, resource optimization, cost reduction, and the future of AI-driven aerospace technologies.

ABSTRAK: Inovasi pesat Kecerdasan Buatan Generatif (GenAI) telah mengubah Sistem Sokongan Keputusan (DSS) merentasi pelbagai domain, termasuk operasi satelit. Kajian ini membentangkan analisis perbandingan empat model AI Generatif percuma termasuk: Gemma 3 oleh Google, Llama 4 Maverick oleh Meta AI, Nemotron Nano 2 oleh NVIDIA dan Devstral

Small oleh Mistral, dalam menyepadu Sistem Sokongan Keputusan (DSS) berasaskan web AI bagi mengenal pasti anomali satelit. Mengguna pakai lebih 4,455 rekod set data anomali satelit meliputi tahun 1957 hingga 2024 oleh Seradata, set data ini mempunyai ciri kebolehskalaan dan kebolehsuaian pada profil misi yang pelbagai. Kajian ini menilai kesemua model melalui penjanaan analisis anomali melalui kejelasan, ketepatan, kelengkapan dan kerelevanan merentasi kategori Gambaran Keseluruhan Insiden, Trend Kebolehpercayaan bagi, Wawasan dan Cadangan Pihak Berkepentingan, mengguna skala Likert 5-mata dan Kappa Fleiss kekonsistensi dalaman. Penilaian komprehensif model GenAI DSS berasaskan web AI menggariskan stratifikasi prestasi yang jelas pada skor 4.44 dan 4.39 untuk Nemotron Nano 2 and Llama 4 Maverick masing-masing, mengesahkan keupayaan sebagai sistem utama berdasarkan skor Likert keseluruhan. Walau bagaimanapun, analisis selanjutnya mendedahkan pertukaran kritikal antara kualiti mutlak dan kekonsistensi dalaman (Fleiss' Kappa). Model terbaik, Nemotron Nano 2 and Maverick Llama 4, mencapai skor Likert tertinggi dengan puncak prestasi yang tinggi tetapi keboleh ramalan dalaman terendah iaitu $\kappa \approx 0.18$ pada Maverick Llama 4, menandakan dapatan struktur yang sangat tidak menentu di mana kejelasan yang kuat sering menutupi ketidaklengkapan kritikal. Sebaliknya, Mistral Devstrall Small menunjukkan kekonsistensi dalaman tertinggi $\kappa \approx 0.66$ walaupun skor min suboptimumnya 3.44. Keboleh ramalan yang mantap ini, walaupun pada tahap lebih rendah, menggariskan implikasi ketara bagi pembangunan DSS. Pemilihan model mesti mengutamakan keseimbangan yang diperlukan antara siling prestasi mutlak dan keboleh ramalan dapatan struktur. Penemuan ini mengetengah potensi pelaksanaan GenAI DSS dalam meningkatkan kebolehpercayaan operasi satelit sambil meneroka hala tuju masa depan bagi tujuan penyelidikan dan pembangunan. Penyelidikan ini menyumbang kepada pembangunan sistem mengenal pasti anomali satelit berdaya tahan dan pintar, dengan implikasi yang lebih luas bagi keselamatan misi angkasa lepas, mengoptimum sumber, mengurangkan kos operasi dan masa depan teknologi aero angkasa yang dipacu AI.

KEYWORDS: *Generative AI, Gemma 3, Llama 4 Maverick, Nemotron Nano 2, Devstral Small, Satellite Anomaly Identification, Clarity, Accuracy, Completeness, Relevance, Decision Support Systems.*

1. INTRODUCTION

The escalating complexity of satellite operations demands robust, efficient anomaly-identification systems. Satellites, essential for communication, navigation, Earth observation, and scientific research, produce substantial quantities of telemetry data. Manually combing through this vast amount of data to identify anomalies from normal behavior is not only time-consuming but also prone to human mistakes. Conventional anomaly identification techniques often struggle with the complexity, high dimensionality, and class imbalance of the satellite dataset, which was labor-intensive and prone to human error [1]. This can lead to delayed identification, resulting in significant operational downtime, costly repairs, or even mission loss.

The growing field of Artificial Intelligence (AI), particularly Generative AI (GenAI) and Large Language Models (LLMs), offers a promising paradigm shift in addressing these challenges. These advanced models possess an unparalleled ability to learn intricate patterns, understand contextual nuances, and even generate human-readable explanations from complex datasets [1].

Web-based decision support systems (DSS) further enhance the advantage of generative models by enabling real-time collaboration, visualization, and automated analysis. However, the integration of lightweight, state-of-the-art generative models, such as Gemma 3, Llama 4 Maverick, Nemotron Nano 2, and Devstral Small, into web-based DSS for satellite anomaly

https://doi.org/10.31436/iiumej.v27i2.4061

identification remains largely unexplored. While these models are designed for efficiency and performance, their comparative effectiveness in high-stakes, real-time environments, particularly in terms of clarity, accuracy, completeness, and relevance, has not been systematically evaluated.

Despite the rapid advancement of GenAI, its application, especially concerning DSS for satellite anomaly identification, faces several challenges:

- *Clarity*: Models must provide interpretable outputs to support human validation, yet few studies quantify interpretability across different architectures.
- *Accuracy*: High precision and recall are critical, but comparative analyses of these models' performance on satellite anomaly data are lacking.
- *Completeness*: Models must cover a wide range of anomaly types, but their ability to do so has not been systematically assessed.
- *Relevance*: Generated insights must be contextually appropriate for operational decision-making, a dimension rarely addressed in existing literature.

Prior research has primarily focused on larger models, such as GPT, Gemini, Grok, and DeepSeek, and not focused on real applications such as DSS, especially in the satellite anomaly area [2], leaving a gap in understanding how smaller, efficient generative models perform in constrained, real-time environments. The incorporation of GenAI into Decision Support Systems (DSS) is particularly promising for satellite anomaly identification, an essential task for assuring mission success and spacecraft longevity. Such a system could empower satellite stakeholders not only with anomaly identification but also with insightful, AI-generated diagnoses, significantly enhancing their decision-making capabilities.

This study addresses these gaps through a quantitative comparison of Gemma 3, Llama 4 Maverick, Nemotron Nano 2, and Devstral Small in a web-based DSS for satellite anomaly identification and evaluates the implementation of certain GenAIs within the DSS, identifying the recommended GenAI models for DSS for satellite anomaly identification.

The remainder of this paper is organized as follows: Section 2 describes the GenAI models, and Section 3 presents the research methodology for integrating these GenAI models into the Web AI-based DSS. Section 4 illustrates a comparative analysis of how GenAI models are implemented in our Web AI-based DSS. Finally, Section 5 concludes the findings, discusses the challenges and future GenAI models for DSS applications, and suggests directions for future research.

2. GENERATIVE AI IN DSS FOR SATELLITE ANOMALY IDENTIFICATION

2.1. Satellite Anomaly Identification

Satellite operations are inherently complex, involving numerous interconnected subsystems that must operate seamlessly in the hostile space environment [3, 4]. Satellite anomalies denote any deviation from the expected or nominal behavior of a satellite's components or overall system [3]. These can range from subtle sensor drifts to catastrophic failures in critical systems such as power generation, attitude control, or communication. Common categories found in the Seradata database related to satellite anomalies can be summarized as follows [3]:

- *Power System*: Fluctuations in solar array output, battery degradation, or power distribution issues.

https://doi.org/10.31436/iiumej.v27i2.4061

- *Attitude Control System*: Unintended rotations, thruster malfunctions, or gyroscope failures affecting satellite orientation.
- *Thermal Control System*: Overheating or undercooling of components.
- *Communication System*: Signal degradation, transponder failures, or antenna pointing errors.
- *On-board Data / Computer*: Software glitches, memory errors, or processor malfunctions.

Traditional methods for anomaly identification have many approaches, including statistical process control (SPC), trend analysis, and rule-based expert systems [5, 6]. While these methods are straightforward, they often suffer from high false-positive rates due to normal operational variations or miss true anomalies that manifest as subtle, non-linear patterns [5, 6]. They also require extensive domain expertise to define rules and thresholds, making them time-consuming and difficult to scale.

2.2. Generative AI (GenAI)

Satellite operations are inherently complex, involving numerous interconnected subsystems that GenAI refers to a class of artificial intelligence models capable of producing novel content, such as text, images, or data, that resembles the distribution of the data they were trained on. GenAI took on a specific form following the publication of Google's research in 2017 [7]. A team of scientists developed a neural network known as the Transformer that performed exceptionally in language translation, and carefully designed it to take advantage of the capabilities of parallel computing technology known as graphics processing units (GPUs) [7]. Large Language Models (LLMs) are a prominent subset of GenAI, characterized by their vast number of parameters and training on massive text datasets, enabling them to understand, generate, and manipulate human language with remarkable fluency and coherence. Beyond text generation, LLMs have demonstrated capabilities in complex pattern recognition, reasoning, summarization, and explanation generation, making them highly relevant for sophisticated data analysis tasks like anomaly identification. GenAI systems were initially single-modal, anticipating the input to possess a particular modality and generating output of a designated modality. In this case, the most popular GenAI systems were text-in, text-out single-modal systems and text-in, image-out single-modal systems [7].

GenAI comprises the hardware and software required to develop, train, and deploy models. Graphics Processing Unit, known as GPU, is a key part of the hardware supporting GenAI, with Nvidia as the principal entity [7]. The extensive matrix multiplication activities and the logic layers fundamental to neural networks are ideally suited to Nvidia's GPU architecture. In addition, the CUDA Software Development Kit (SDK) supports fast parallel processing by directly accessing sets of instructions in the GPU hardware [7].

This study focuses on a comparative analysis of four free GenAI models, including Gemma, Llama Maverick, Nemotron, and Devstral Small. The potential of these models for satellite anomaly identification lies not only in their ability to detect deviations but also in their capacity to interpret complex data patterns, synthesize information, and generate actionable explanations for operators. This moves beyond simple alerts to providing “why” an anomaly is occurring, which is crucial for effective decision-making.

2.2.1. Google Gemma 3

Gemma 3 is a family of advanced multimodal models developed by Google DeepMind, co-designed with the Gemini frontier model family [8]. Gemma 3 models are comparable in size to Gemma 2, with an additional 1B-parameter model and numerous new capabilities, including multimodality, long context, and multilingual [8]. In the context of Web AI-based

https://doi.org/10.31436/iiumej.v27i2.4061

DSS, multimodality enables simultaneous correlation of text data with a visual chart, while long context enables ingestion of the entire provided historical dataset. Gemma 3 comes in two size options: 7-billion- and 2-billion-parameter models. A 7 billion parameter model suitable for efficient development and deployment on GPU and TPU, while a 2 billion parameter model customized for CPU and on-device applications. Both parameter sizes are designed to adopt distinctive computational limitations, applications, and developer requirements [9].

The Gemma 3 models are constructed following the Transformer decoder architecture [10], as is illustrated in Fig. 1. Gemma 3 does not need the encoder part of the original Transformer architecture; it needs only the decoder, which is equivalent to an encoder, except for MASKING in the multi-head attention block, where the decoder only allows information to be collected from the previous words in the sentence [10]. Gemma 3 refined through structural enhancements and model optimizations to manage a 128K context window and training on up to 6T tokens [8]. To handle this scale, Gemma 3 employs Grouped-Query Attention (GQA), which decreases memory operating cost by sharing data across multiple processing heads, and Root Mean Square Layer Normalization (RMSNorm) in both pre-norm and post-norm configurations [8]. Pre-norm applies normalization before the attention or feed-forward layer, leading to significantly more stable training at enormous scales, while post-norm applies it afterward the remaining connection, which can improve the model's final representational power [8]. The models are powerful and adaptable, significantly outperforming their predecessors [8]. These architectural refinements are essential for Decision Support Systems (DSS) and anomaly identification, as they provide the structural stability and memory efficiency.

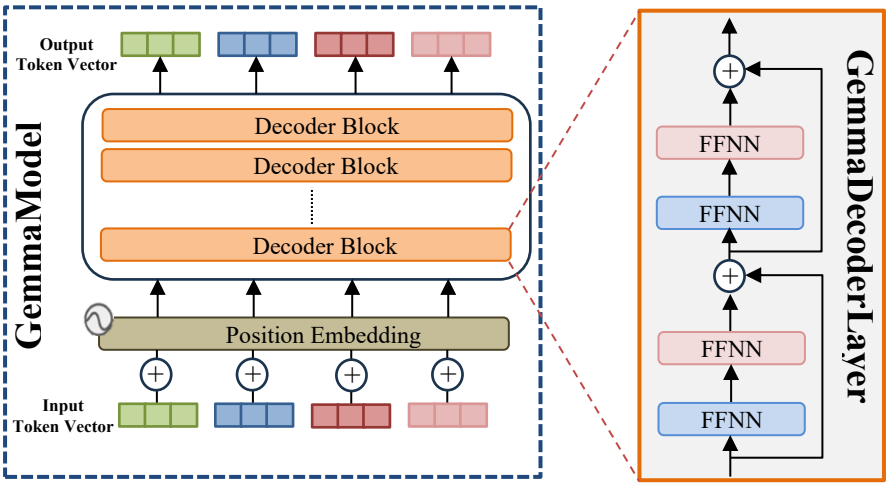

Figure 1. Gemma 3 model architecture.

2.2.2. Meta AI Llama 4 Maverick

Llama 4 Maverick is a high-capacity multimodal language model from Meta with 128K tokens and 17 billion active parameters [11]. It supports multilingual text and image input, and produces multilingual text and code output across 200 supported languages [11]. Building upon the success of previous Llama iterations, Llama 4 Maverick is expected to feature significant advancements in scale, efficiency, and reasoning capabilities.

Llama 4 Maverick is constructed using a mixture-of-experts (MoE) architecture as depicted in Fig. 2 [11]. In this architecture, a single token activates only a portion of the total

https://doi.org/10.31436/iiumej.v27i2.4061

parameters and exhibits superior computational efficiency for both training and inference, delivering higher quality compared to a dense model [11]. A Router analyses incoming data tokens and dynamically assigns them to specific Routed Experts or subnetworks that have been trained to specialize in niche areas such as mathematical reasoning or linguistic nuances [11]. Llama 4 Maverick employs a combination of dense and mixture-of-experts (MoE) layers to improve interpretability. The MoE comprises 128 routed experts in addition to a shared expert, with each token concurrently processed by the shared expert and one routed expert [11]. While the complete parameter set resides in memory, only a subset is activated during inference. As a result, this improves inference efficiency by reducing computational processing costs and latency [11]. This setup enables the model to sustain a large-scale knowledge base while dynamically engaging only the most appropriate experts for a given query. Such efficiency is particularly advantageous in real-time Decision Support Systems (DSS) and anomaly detection tasks, where rapid response and high precision are critical for identifying rare data anomalies.

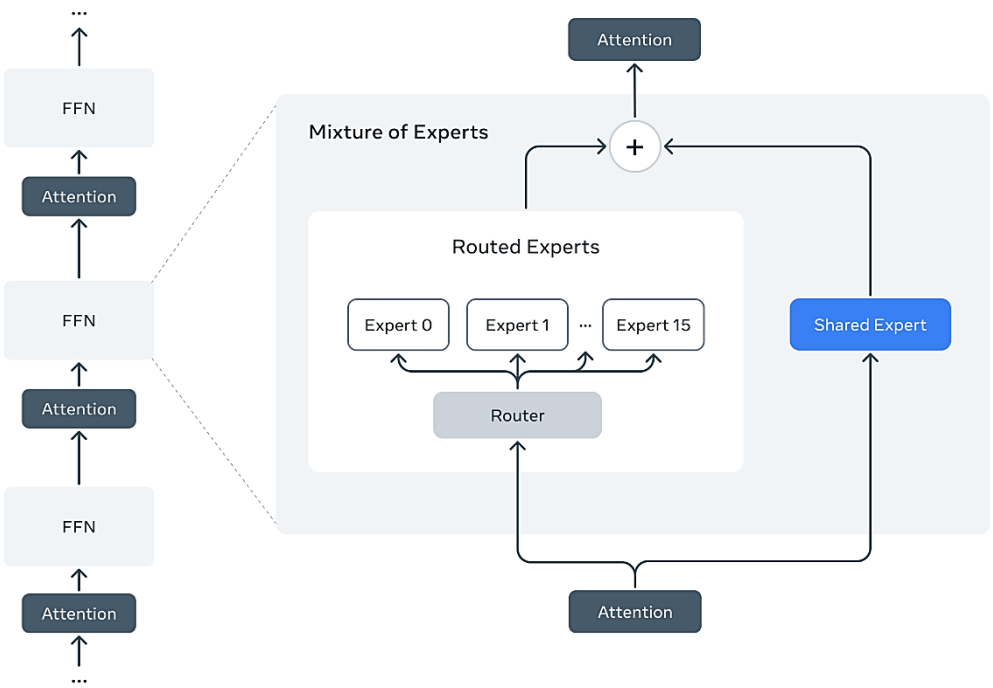

Figure 2. Llama 4 Maverick Mixture of Experts (MoE) architecture

2.2.3. NVIDIA Nemotron Nano 2

Nemotron Nano 2 is a diverse, open collection of thinking models created by NVIDIA, offering outstanding reasoning capabilities, efficient inference, and an open license for commercial applications [12]. Nemotron Nano 2 is a hybrid Mamba-Transformer reasoning model developed by NVIDIA to improve benchmark accuracy and handle heavy scenarios such as 1k input, 8k output, 8k input, and 16k output tokens [12]. It builds based on the architecture of Nemotron-H with new datasets and recipes for pre-training, aligning, pruning, and distillation [12].

The illustration of the Nemotron Nano 2 base-layer pattern, as shown in Fig. 3, comprises Mamba 2, self-attention, and FFN layers [12]. The architectural design comprises three distinct phases, with a total of 62 layers: 6 attention layers, 28 FFN layers, and 28 Mamba-2 layers [12]. The initial phase includes three repetitions of a Mamba-2 and feed-forward network (FFN) pair. This is followed by a central core, reiterated six times, which incorporates a single

https://doi.org/10.31436/iiumej.v27i2.4061

self-attention layer followed by a combination of Mamba-2 and FFN layers. The final phase concludes with a single Mamba-2 and FFN pair. Notably, the design applies only six attention layers, which is about 8% of the total depth, while relying on 28 Mamba-2 layers for linear-time processing. This selective distribution enables the model to deliver up to a sixfold higher implication throughput compared to standard dense architectures, while preserving competitive accuracy on complex reasoning tasks.

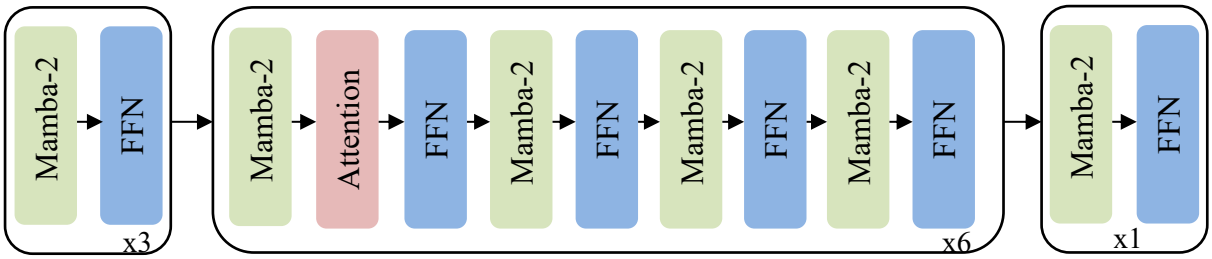

Figure 3. Nemotron Nano 2 Base Layer Pattern

Nemotron Nano 2 was pre-trained using DeepSeek's FP8 training recipe, which has precision over 20T tokens using a Warmup-Stable-Decay (WSD) learning rate schedule, and then undertook an endless pre-training long context addition phase to become 128K capable deprived without degrading other standard [12]. The post-training of Nemotron Nano 2 utilized a combination of Supervised Fine-Tuning (SFT), Group Relative Policy Optimization (GRPO), Direct Preference Optimization (DPO), and Reinforcement Learning from Human Feedback (RLHF), and was performed on roughly 90B tokens [12].

2.2.4. Mistral Devstral Small

Devstral is an agentic Large Language Model (LLM) for software engineering tasks, built in collaboration between Mistral AI and All Hands AI. Devstral excels at using tools to explore codebases, editing multiple files, and powering software engineering agents [13]. Devstral Small, as a more compact model, having only 24B parameters, could offer a compelling balance between performance and computational resource requirements [13]. Its efficiency could be particularly advantageous for deployment in web-based systems where resource optimization is a key consideration, while still retaining strong analytical capabilities for anomaly identification.

Devstral Small is a compact yet powerful Transformer model with 24 billion parameters. Built based on Mistral Small 3, it features 40 layers and incorporates Grouped Query Attention (GQA) for efficient processing [13]. The model is trained on an extensive range of textual data, including both natural language and programming code. Its capabilities are further enhanced through a long context extension phase, enabling it to handle up to 128,000 tokens, which is ideal for tasks involving code agents and other applications demanding extended context windows [13] such as a Decision Support System (DSS).

The Devstral Small model is evaluated using the OpenHands scaffold, an open platform designed to develop AI agents that interact with the world similarly to human software developers [13]. It offers a comprehensive framework that allows agents to write code, interact with command lines, and browse the web within a secure, sandboxed environment.

2.3. Web AI-Based Decision Support Systems

A Decision Support System (DSS) is an information system that aids decision-making. It typically comprises three main components: a data management subsystem for data collection and organization, a model management subsystem for analytical tools and models, and a user interface subsystem for interaction [14]. Web-based DSS leverage internet technologies to provide accessible, scalable, and collaborative decision-making environments [15]. Their advantages include:

- *Accessibility*: Users can access the system from anywhere with an internet connection.
- *Scalability*: Easily accommodates a growing number of users and data.
- *Collaboration*: Facilitates shared access to information and collaborative decision-making.
- *Cost-effectiveness*: Reduces the need for specialized client-side software.

The role of AI in enhancing DSS capabilities is profound. AI can transform a traditional DSS into an intelligent DSS by automating data analysis, providing predictive insights, offering prescriptive recommendations, and generating explanations and descriptions [4, 16]. For satellite anomaly identification, an AI-based DSS can process vast streams of telemetry data, identify potential issues, and present them to operators with clear, concise explanations, thereby augmenting human cognitive abilities and accelerating response times. While specific public examples of Web AI-based DSS exclusively for satellite anomaly identification using advanced GenAI are limited, similar systems exist in other domains like industrial predictive maintenance, energy, healthcare, manufacturing, and satellite, demonstrating the feasibility and benefits of such an integration [17-20].

2.4. Systematic Literature Mapping

The uniqueness of this study, compared to previous studies presented in Table 1, lies in its focused comparative analysis of GenAI models (Gemma 3, Llama 4, Maverick, Nemotron Nano 2, and Devstral Small) integrated with a Web AI-based DSS designed for satellite anomaly identification. This research provides an empirical, multi-model evaluation tailored to the unique challenges and requirements of satellite operations, emphasizing not only identification accuracy but also the crucial aspect of explainability provided by GenAI models. This comprehensive approach aims to provide valuable insights into which GenAI models are best suited to enhance the resilience and efficiency of DSS for future satellite missions.

Table 1. Systematic Literature Mapping

Title	Author(s)	Contribution	Methods	Gap
Grok, Gemini, ChatGPT, and DeepSeek: Comparison and Applications in Conversational Artificial Intelligence	Murillo Edson de Carvalho Souza et al. [2]	Comparative analysis of LLMs (Grok, Gemini, ChatGPT, DeepSeek)	Transformer-based LLMs	Lack of standardized evaluation metrics across models; limited transparency in proprietary systems
Generative AI as a New Platform for Application Development	Michael A. Cusumano, Annabelle Gawer, David B. Yoffie [7]	GenAI as a foundational platform	Ecosystem mapping, governance	Insufficient frameworks for sustainable governance and open-source integration

Prototype Development of Web-AI-Based Decision Support System: Insights and Recommendations for Satellite Anomaly Identification	A. Mutholib, N. A. Rahim, and T. S. Gunawan [16]	Web-based DSS prototype for anomaly identification in satellites	Web-AI integration, anomaly identification models	Prototype stage, lacks large-scale validation
Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0	E. Mikołajewska, D. Mikołajewski, T. Mikołajczyk, and T. Paczkowski [17]	Digital twins for predictive maintenance	GANs, VAEs, Transformers	Limited real-world deployment evidence, need for federated and edge-based implementations
Using Machine Learning for Advanced Anomaly Detection and Classification	Lane et al. [19]	Anomaly detection in SSA	SVM, RF, k-NN, PCA	Lack of generative approaches for anomaly simulation; limited DSS integration for operational decision-making
Preliminary Design and Methodological Framework for an AI-Driven Decision Support System in Earth Observation Satellites	Abdallah Alabed, Tarık Özkul [18]	Framework for onboard DSS in EO satellites; drone-based simulation for cloud-aware imaging	YOLOv12 (cloud detection), SORT (tracking), adaptive mission planner	Limited to simulation; lacks orbital validation; planner rule-based (no reinforcement learning yet)

3. METHODOLOGY

This section discusses the methodical approach adopted to develop the Web AI-based Decision Support System (DSS) for satellite anomaly identification and to conduct the comparative analysis of the selected generative AI models.

3.1. System Architecture of the Web AI-Based DSS

The implementation of Web AI-based DSS uses a three-tier architecture, as illustrated in Fig. 4, including the frontend as the presentation layer, the backend as the logic tier, and the database as the data tier.

Figure 4. Web AI-based DSS Three-tier Architecture.

The frontend, known as the User Interface (UI), is built with modern web technologies, such as the Bootstrap framework, to provide an intuitive, interactive dashboard for satellite stakeholders. The UI displays anomaly data in both graph and table formats. It includes features for reviewing historical data, applying filters, and generating GenAI-generated analysis.

https://doi.org/10.31436/iiumej.v27i2.4061

The backend consists of an API, and the logic is implemented using a robust framework: Python with Django, to handle data ingestion, process frontend requests, orchestrate interactions with GenAI models, and manage the database. A PostgreSQL database is used to store satellite data, anomaly data, and GenAI-generated analysis.

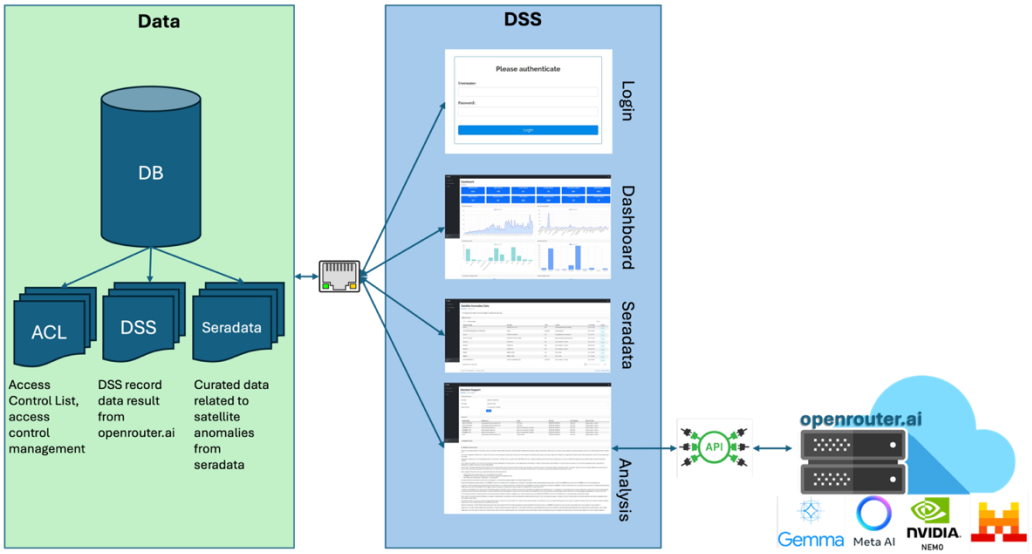

Figure 5. Web AI-based DSS System Architecture.

In addition, the system's high-level structure is defined by a blueprint that outlines how its components will interact and work together to meet the requirements. The details of the system architecture are illustrated in Fig. 5. The figure shows three sections: data, DSS, and Openrouter, which serve as GenAI models API providers.

The data section comprises a database that includes access control management, DSS analysis, and curated data from Seradata. This data is accessible via the local area network. It means that the data and the DSS application are in the same area network. This local access includes the authentication process, dashboard access, data presentation, and part of the data analysis. The other part of the data analysis requires an external connection or internet access.

The analysis of data for decision-making uses the API provided by openrouter.ai to access GenAI models, including Gemma 3, Meta AI Llama 4 Maverick, Nvidia Nemotron Nano 2, and Mistral Devstral Small.

3.2. Dataset Selection

The effectiveness of any AI-driven anomaly identification system is heavily dependent on the quality and representativeness of its data. The data source used in this research is the Seradata SpaceTrak service [20]. Seradata is an open-source intelligence database that contains extensive information regarding international satellite and space launch vehicle activities, including failure reporting [3]. The selection of Seradata for this research is necessary due to its extensive depth, historical breadth, and thorough failure reporting, which are crucial for a meticulous analysis of satellite reliability [4]. With more than six decades of recorded data, Seradata provides a comprehensive perspective on satellite performance and failure characteristics, enabling researchers to understand the immediate and long-term effects of design decisions, operating conditions, and external environmental factors on satellite reliability [20].

https://doi.org/10.31436/iiumej.v27i2.4061

Spanning from 1957 to 2024, Seradata provides over 4,455 occurrences of spacecraft failures, rendering it an essential resource for examining trends and patterns in satellite anomaly identification [4]. The data used in this study have been preprocessed to minimize missing values, noise, and the possibility of outliers by using data cleansing. This preprocessing is very important to enhance raw data quality, consistency, and efficiency for further processing [3, 4]. The dataset classifies failures into nine main categories: attitude control, power, payload instrument, beam, control processor, telemetry, thermal, and transponder, providing a solid basis for analyzing the causes of satellite faults. Table 2 describes the satellite anomaly datasets used in this study.

Table 2. Description of satellite anomaly data

Attributes	Values
Number of failure categories	9
Data duration	1957 - 2024
Number of experimental datasets	4455

3.3. Generative AI Model Integration and Customization

Integrating various GenAI models into a DSS requires careful evaluation of their APIs and optimal prompting techniques. In this study, OpenRouter is utilized for model invocations, employing the following specific models as presented in Table 3, and all configured with default settings:

Table 3. GenAI models specification

Model Name	Params	Context length	Knowledge cutoff	Source
Gemma 3 [21]	27B	1M	December 2024	`google/gemma-3-27b-it:free`
Llama 4 Maverick [22]	17B	1M	August 2024	`meta-llama/llama-4-maverick-17b-128e-instruct:free`
Nemotron Nano 2 [23]	253B	131K	April 2025	`nvidia/nemotron-nano-9b-v2:free`
Devstral Small 2505 [24]	24B	33K	May 2025	`mistralai/mistral-small-3.2-24b-instruct:free`

Given the nature of LLMs, effective prompt engineering was essential to guide each model toward accurate anomaly identification and meaningful explanation generation. Prompts were carefully designed to encompass:

- *Contextual Information*: Current anomaly readings, historical trends, and relevant satellite operational modes.
- *Task Definition*: Clear instructions to identify anomalies and provide a concise explanation of the anomaly, its potential cause, its impact, and recommendations.

Each GenAI model was integrated into the Web AI-based DSS backend via its corresponding RESTful API. The responses from models' API, typically JSON objects containing the generated text, were subsequently parsed and processed by the Explanation Generation Module.

https://doi.org/10.31436/iiumej.v27i2.4061

3.4. Experimental Setup and Evaluation Metrics

The evaluation of each GenAI model was conducted in a controlled simulation environment to ensure reproducible, comparable results.

3.4.1. Simulation Environment

The Web AI-based DSS was deployed on both local and cloud platforms for the GenAI models API, with sufficient computational resources and GPUs on the local platform to accelerate inference and handle the workload of querying multiple LLMs simultaneously. The simulation involved feeding the preprocessed satellite anomaly data through the DSS. This setup allowed for accurate measurement of resource utilization under operational load.

Table 4. Hardware Specifications

Category	Specification
CPU	Intel(R) Core i9-14900K, 3.20 GHz
GPU	NVIDIA GeForce RTX 4090
RAM	64.0 GB DDR5
Storage (SSD)	500 GB
Storage (HDD)	6 TB

Table 4 presents the hardware configuration details, featuring cutting-edge components optimized for demanding computational operations. The Intel Core i9-14900K CPU, with a clock speed of 3.20 GHz, provides substantial processing power and efficiency for data preprocessing and model training. The NVIDIA GeForce RTX 4090 GPU enhances parallel processing, crucial for intricate ML algorithms and deep learning models. The machine, equipped with 64 GB of DDR5 RAM, accommodates large-scale memory-intensive tasks, such as processing the massive Seradata dataset. Storage comprises a 500 GB SSD for high-speed access to commonly utilized files and a 6 TB HDD for long-term storage, facilitating efficient data management.

The software environment presented in Table 5 was designed to meet the computational requirements of the Web AI-based DSS, ensuring speed, efficiency, and reproducibility. Python 3.11.9 as the foundation, Django 5.1.3 as the back-end web framework, PostgreSQL 14 as the database management system (DBMS), and Bootstrap 5 as the front-end toolkit are used to construct a Web AI-based DSS.

Table 5. Software Specifications

Category	Version
Windows Server Datacenter	2022
Python	3.11.9
Django	5.1.3
PostgreSQL	14
Bootstraps	5

3.4.2. Evaluation Metric

A comprehensive quantitative metric was used to evaluate each GenAI model's performance. A Likert scale and Fleiss' Kappa (κ) were employed in this study to provide a more comprehensive explanation of model performance. A Likert scale was utilized for scoring metrics, which provides five possible answers, ranging from 1 (strongly disagree) to 5 (strongly agree), to gauge the responses from the expert users who participate in the questionnaire to

indicate positive or negative strength of agreement of feeling regarding the question or statement [25].

Fleiss' Kappa is a statistical measure used to assess the level of agreement between three or more users who rate or respond when they classify or score items using categorical scales, such as a Likert scale with 1 to 5 categories [26]. To assess the robustness of the expert evaluation, Fleiss' Kappa was computed to quantify agreement among the expert evaluators. Fleiss' Kappa (κ) is suitable for assessing consistency when multiple raters assign categorical ratings. Fleiss' Kappa (κ) is formulated as in the following equation 1 [26].

$$k = \frac{\bar{P}_a - \bar{P}_c}{1 - \bar{P}_c} \quad (1)$$

where $\bar{P}_a$ is the mean agreement between raters and $\bar{P}_c$ is the mean probability of agreement on change. The metric quantifies agreement beyond chance, with commonly accepted thresholds of $\kappa \geq 0.41$ indicating moderate agreement, $\kappa \geq 0.61$ substantial agreement, and $\kappa \geq 0.81$ almost perfect agreement [26].

Both the Likert scale and Fleiss' Kappa were crafted to ensure comprehensive coverage of the dimensional metrics as presented in Table 6.

Table 6. Quantitative Metrics Description

Metrics	Description
Clarity	The explanation is easy to understand.
Accuracy	The explanation accurately describes the anomaly and its cause, or has signs of hallucination.
Completeness	The explanation provides sufficient detail for decision-making.
Relevance	The explanation is relevant to the observed anomaly data.

3.5. Questionnaire Responders

The expert users selected for this study typically represent an interdisciplinary cohort of spacecraft engineers, mission controllers, and space scientists with 5 to 10 years' experience and a proven ability to perform cross-domain diagnostics. Their nature is described by a deep experiential understanding of satellite health. The experts have backgrounds in satellite anomalies, particularly in spacecraft subsystems such as Attitude Control Systems (ACS), Electrical Power Systems (EPS), and communication systems, and have spent years studying signal anomalies that may indicate sensor noise, environmental interference such as solar flares, or genuine hardware degradation. Their primary role in the questionnaire process is to provide the qualitative weights and diagnostic logic that bridge the gap between raw statistical outliers and actionable operational decisions.

4. Results and Discussions

This section presents the empirical findings from the comparative analysis of Gemma 3, Llama 4 Maverick, Nemotron Nano 2, and Devstral Small within the developed Web AI-based Decision Support System (DSS) for satellite anomaly identification. We discuss each model's individual performance, conduct a direct comparison across key metrics, provide insights from the DSS implementation, and acknowledge the study's inherent limitations. The evaluation was conducted using a questionnaire and the Web AI-based DSS by a human expert in the satellite field.

https://doi.org/10.31436/iiumej.v27i2.4061

4.1. Performance of Generative AI Models

Assessing the performance of GenAI models is more complex than that of conventional classification models. In a generative model, there is no single "correct" answer; instead, there is a continuum of acceptable outputs. However, the core concepts of precision, recall, and F1 score can be adapted to evaluate different aspects of a generative model's performance [27]. The evaluation disclosed distinct performance profiles for each generative AI model across various anomaly types and operational scenarios within the simulated satellite environment.

In this study, several scenarios were applied to prompts and GenAI models that were implemented in the Web AI-based DSS. Table 7 presents the scenarios applied.

Table 7. Prompts Scenarios

Scenarios	Prompts
Incident overview	As an expert in the satellite area, write the incident overview related to the mission's primary, orbit category, mass at launch, age, and design life, especially for ages less than design life, based on the data provided.
Reliability Trends & Insights	As an expert in the satellite area, write the Reliability Trends & Insights related to mission primary, orbit category, mass at launch, age, and design life, especially for ages less than design life, based on the data provided.
Stakeholder Recommendations	As an expert in the satellite area, write the Stakeholder Recommendations related to mission primary, orbit category, mass at launch, age, and design life, especially for ages less than design life, based on the data provided.

4.1.1. Likert Scale Analysis

Fig. 5 shows the overall model performance across all categories (Incident Overview, Reliability Trend, Insights, and Stakeholder Recommendations) and metrics (Clarity, Accuracy, Completeness, and Relevance).

From Fig. 6, Nemotron Nano 2 outperforms other models, scoring 4.4 out of 5, followed closely by Llama 4 Maverick, with an overall score of 4.39 out of 5. It shows that both models consistently scored between "Agree" and "Strongly Agree". On the other hand, Gemma 3 shows moderate performance, while Devstral Small achieved the lowest score among the models, averaging closer to "Neutral" and "Agree."

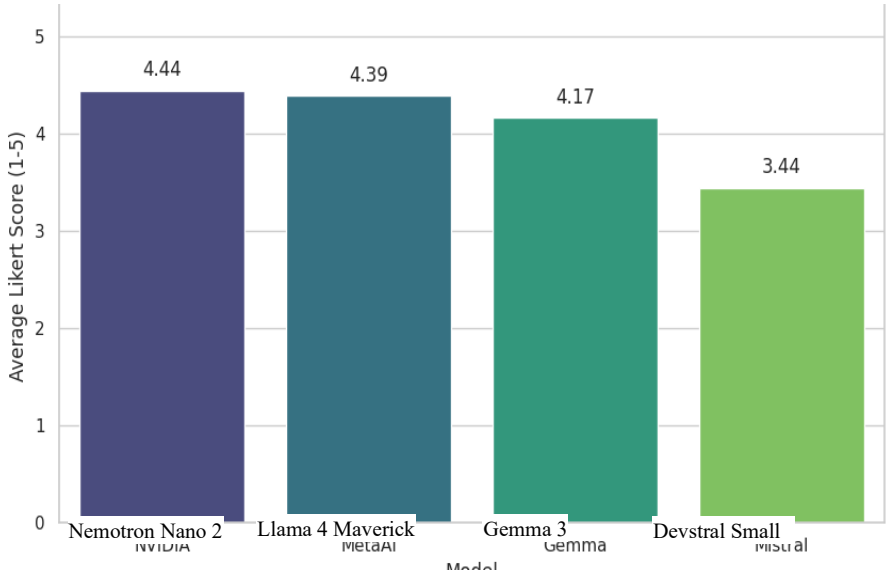

Figure 6. Overall Average Assessment Score by Model

https://doi.org/10.31436/iiumej.v27i2.4061

The categorical assessment, as shown in Fig.7, reveals a distinct hierarchy in model efficacy, with Nemotron Nano 2 and Llama 4 Maverick consistently outperforming their counterparts across the evaluated domains of Incident, Reliability, and Recommendation.

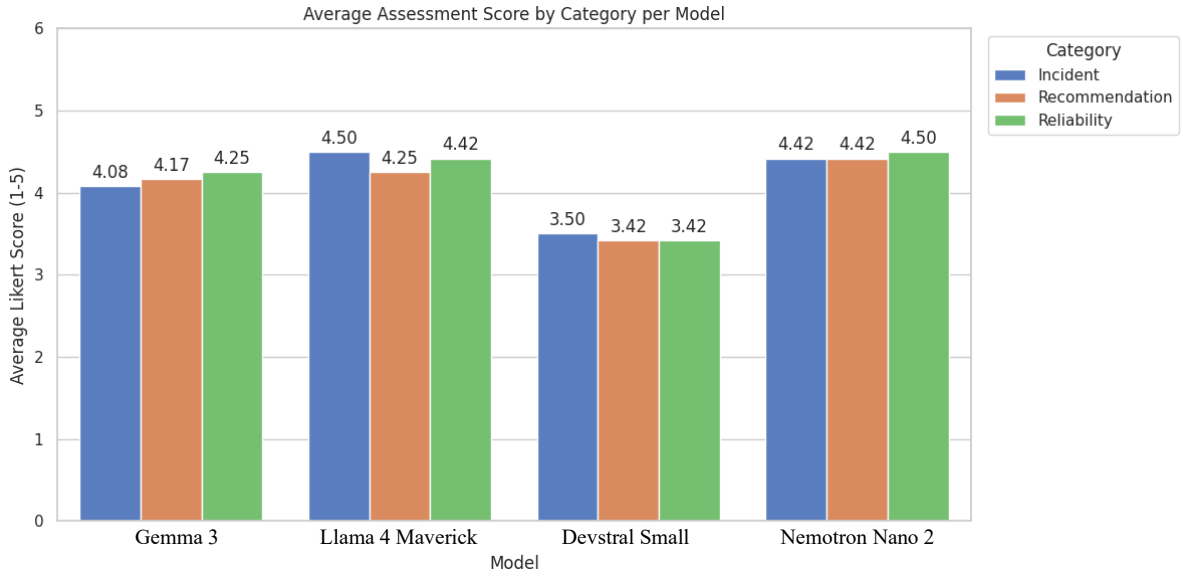

Figure 7. Average Assessment Score by Category per Model

Nemotron Nano 2 established pre-eminence in the Reliability and Recommendation categories, achieving mean Likert scores of 4.50 and 4.42, respectively, indicating a robust capacity for consistent and actionable output. Conversely, Llama 4 Maverick demonstrated a slight advantage in the Incident category, achieving 4.50, suggesting specialized proficiency in handling event-specific queries. While Gemma 3 maintained competitive performance thresholds, particularly in Reliability, Devstrall Small exhibited a marked divergence from the leading models, recording the lowest aggregate scores across all categories, notably falling to 3.42 in both Reliability and Recommendation assessments.

The heatmap for the detailed average assessment score by metric and category is presented in Fig. 8. It can be summarized that the heatmap presents a granular comparative analysis of model performance across the distinct dimensions of Incident, Reliability, and Recommendation, further ranked by the qualitative metrics of Accuracy, Clarity, Completeness, and Relevance. Visualization uses a continuous color gradient to represent the mean Likert-scale evaluations, with darker hues corresponding to higher performance ratings. This matrix shows that Nemotron Nano 2 and Llama 4 Maverick consistently achieve high scores, especially in the domains of Reliability and Incident overview, as evidenced by a predominance of values above 4.33. In contrast, the lighter color associated with Devstral Small across multiple vectors is most pronounced in Recommendation Accuracy and Completeness, revealing a significant performance disparity and underscoring its suboptimal assessment relative to the benchmark established by the leading models.

https://doi.org/10.31436/iiumej.v27i2.4061

Figure 8. Detailed Average Assessment Score by Metric and Category

4.1.2. Inter-Rater Reliability Results

The inter-rater reliability analysis yielded a Fleiss' Kappa for Gemma 3 as presented in Table 8, showing that Gemma 3 demonstrates substantial internal consistency for the Reliability (κ=0.692) and Recommendation (κ=0.636) dimensions. This means that when the model performs well, it tends to be correspondingly strong across Clarity, Accuracy, Completeness, and Relevance. However, the notably lower score for the Incident dimension (κ=0.404) highlights a specific performance imbalance during complex tasks. The data shows that Gemma 3 often maintains high scores for Clarity and Relevance, declaring confidence and on topic, even when its Accuracy and Completeness drop slightly. This suggests that while Gemma 3 is a communicative and reliable tool overall, its hallucination risk is highest during specific incidents where it prioritizes sounding clear over being perfectly accurate.

Table 8. Fleiss' Kappa Analysis for Gemma 3

Dimension	Fleiss' Kappa (κ)	Agreement Level	Interpretation
Incident	**0.404**	Moderate	Significant consistency break.
Reliability	**0.692**	Substantial	Highly consistent performance.
Recommendation	**0.636**	Substantial	Strong consistency.

Based on Table 9, Llama 4 Maverick exhibits low internal consistency on complex tasks, as evidenced by the slight-to-fair agreement scores for Incident (κ=0.180) and Reliability (κ=0.219). While the Recommendation score (κ=0.692) remains strong, the detailed metrics

https://doi.org/10.31436/iiumej.v27i2.4061

expose a critical gap in challenging scenarios; the model maintains perfect scores for Clarity and Relevance but suffers a significant drop in Completeness. This statistical divergence points to a superficial fluency issue: Llama 4 Maverick generates highly sophisticated, on-topic responses that sound convincing but fail to capture all necessary information, resulting in answers that are clear but materially incomplete.

Table 9. Fleiss' Kappa Analysis for Llama 4 Maverick

Dimension	Fleiss' Kappa (κ)	Agreement Level	Interpretation
Incident	**0.180**	Slight	Extreme divergence between metrics.
Reliability	**0.219**	Fair	Very low consistency.
Recommendation	**0.692**	Substantial	Strong consistency.

In Table 10, Nemotron Nano 2 demonstrates Fair internal consistency (κ≈ = 0.32–0.38), which is noticeably lower than that of the previous models. The Fair agreement here is not due to random noise; it is due to a consistent performance gap in complex tasks. This score is driven by a systematic Form over Substance split in challenging scenarios, where the model maintains good scores for Clarity and Relevance, but drops to average scores for Accuracy and Completeness. Unlike Gemma 3, which had high highs or Llama 4 Maverick, which had polished omissions, Nemotron Nano 2's data suggests a general degradation of performance in complex tasks, where both the delivery Clarity and the content Accuracy suffer, but the content suffers more.

Table 10. Fleiss' Kappa Analysis for Nemotron Nano 2

Dimension	Fleiss' Kappa (κ)	Agreement Level	Interpretation
Incident	**0.317**	Fair	Moderate split between form and substance.
Reliability	**0.385**	Fair	Weak consistency, but systematic.
Recommendation	**0.317**	Fair	Consistency mirrors Incident performance.

Devstral Small exhibits the highest consistency, reaching Substantial agreement for Incident and Recommendation as shown in Table 11. This is because the scores across all four criteria are very tightly grouped. This high Kappa score indicates the model is highly predictable; if it is accurate, it is also clear, complete, and relevant. The caveat is that its absolute performance ceiling is lower than the other models, but its performance is balanced across all four metrics.

Table 11. Fleiss' Kappa Analysis for Devstral Small

Dimension	Fleiss' Kappa (κ)	Agreement Level	Interpretation
Incident	**0.657**	Substantial	Highest consistency for Incident metric.
Reliability	**0.556**	Moderate	Strong overall consistency.
Recommendation	**0.657**	Substantial	Highly consistent.

To sum up, the comprehensive evaluation of GenAI models for Web AI-based DSS reveals a critical trade-off between absolute quality and internal consistency. Nemotron Nano 2 and Llama 4 Maverick emerge as the superior systems in terms of absolute quality, achieving overall Likert scores of 4.44 and 4.39, respectively. However, this superior performance comes at the cost of output predictability, as evidenced by their low Fleiss' Kappa scores (κ≈ 0.32 for Nemotron Nano 2 and κ≈ 0.18 for Llama 4 Maverick). This low consistency suggests that, while these models achieve high scores, the reliability of their output structure is volatile, with

significant internal discrepancies, such as the gap between Llama 4 Maverick's Clarity (Strongly Agree) and Completeness (Neutral). In contrast, Devstral Small, which recorded a suboptimal Likert score of 3.44, indicating a need for substantial improvement in absolute quality, demonstrated the highest internal consistency ($\kappa = 0.66$). This suggests that while Devstral Small's output quality is lower, it is highly predictable, with all criteria rising and falling in a uniform manner. Gemma 3 performed as a competitive mid-tier option, scoring 4.17 and exhibiting moderate consistency ($\kappa = 0.40$), balancing strong performance with moderate internal predictability. Ultimately, model selection must therefore prioritize the required balance between absolute performance ceiling (Likert Score) and output predictability (Fleiss' Kappa) for the specific DSS application.

4.2. Insights from Web AI-Based DSS Implementation

The development and deployment of the Web AI-based DSS delivered valuable insights for integrating GenAI models into operational decision-making workflows. One important challenge was optimizing prompt engineering to consistently obtain accurate anomaly identifications and relevant explanations from each LLM. Different models responded best to varying prompt structures and levels of detail, demanding iterative refinement. Another challenge involved managing API request limits and ensuring robust error handling for continuous data streams. Real-time data processing needs careful orchestration to minimize latency between telemetry ingestion and anomaly presentation to the user.

Despite these challenges, the implementation was a significant success in representing the feasibility and benefits of such a system. The web-based interface proved highly accessible and intuitive for simulated stakeholders, including owner, manufacturer, and operator. The integration of GenAI explanations alongside anomaly data visualizations significantly improved stakeholder awareness and reduced the time required to understand anomaly nature.

The feedback from the experts was overwhelmingly positive. Users highlighted the clarity and helpfulness of the GenAI explanations, the AI-based DSS significantly reduced diagnostic time, and provided a starting point for investigation that was not available before. The AI-based DSS effectively transformed raw anomaly data into actionable intelligence, demonstrating its potential to enhance operational efficiency and resilience.

4.3. Limitations of the Study

While this study provides valuable insights, it is important to acknowledge certain limitations:

- *Simulated Data*: The study primarily utilized a simulated satellite anomaly dataset. While designed to mimic real-world conditions and anomaly types, it may not fully capture the full complexity, noise, and unforeseen events present in live satellite data. Future work will aim to validate these findings in real-world settings, where permissible.
- *Model Versions*: The performance of GenAI models is continuously evolving. The results presented are based on the specific versions of Gemma 3, Llama 4 Maverick, Nemotron Nano 2, and Devstral Small available at the time of this research. Newer iterations may exhibit different performance characteristics.
- *Scope of Anomaly Types*: While a diverse set of anomaly types was injected, the study did not cover every conceivable satellite malfunction. The models' performance on highly novel or infrequent anomalies might vary.

- *Prompt Engineering Dependency*: The performance of LLMs is highly sensitive to prompt engineering. Although significant effort was invested in optimizing prompts, different prompting strategies may yield slightly different results.
- The Fleiss' Kappa score indicates only fair agreement among evaluators. This may be due to the subjective nature of assessing generative text explanations. Future work will expand the evaluator pool and refine scoring guidelines to reduce variability in interpretation.
- *Computational Resources*: The resource utilization metrics are specific to the cloud environment and configurations used in this study. Performance may vary in different hardware or infrastructure setups.

These limitations highlight areas for future research and development to further enhance GenAI's robustness and applicability for satellite anomaly identification.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a comprehensive comparative analysis of four state-of-the-art GenAI models, Gemma 3, Llama 4 Maverick, Nemotron Nano 2, and Devstral Small, within a Web-based AI-based Decision Support System (DSS) for satellite anomaly identification. The study successfully demonstrated the feasibility and significant benefits of integrating advanced GenAI into this critical domain, moving beyond simple identification to providing actionable, AI-generated explanations.

The empirical findings reveal a clear performance hierarchy in model efficacy. Nemotron Nano 2 and Llama 4 Maverick emerged as the superior architectures, demonstrating high-fidelity performance across all evaluated metrics, with Nemotron Nano 2 establishing a benchmark in the Reliability Trend, Insight, and Stakeholder Recommendation tasks (avg. > 4.40). The data suggests that while current top-tier models are highly capable of handling complex incident overview and reliability queries, significant performance disparities remain among standard open-weight models. Notably, Devstral Small's consistent underperformance (avg. 3.44), particularly in the accuracy and completeness of recommendations, underscores the critical need for rigorous pre-deployment validation.

Crucially, Fleiss' Kappa analysis of internal consistency presents a necessary counterpoint to the absolute performance scores. The highest-scoring models, Llama 4 Maverick $\kappa \approx 0.18$, Slight Agreement, and Nemotron Nano 2 $\kappa \approx 0.32$, Fair Agreement, exhibit the lowest output predictability. This indicates that their superior Likert scores are often achieved at the cost of internal structural reliability, marked by significant discrepancies between criteria, such as high Clarity masking low Completeness. In direct contrast, Devstral Small, despite its consistent underperformance (avg. 3.44), demonstrated the highest output consistency ($\kappa = 0.66$), Substantial Agreement. This suggests Devstral Small's, though suboptimal in quality, is the most predictable model, with all output metrics aligning tightly.

The primary implication of this research is that selecting a GenAI model for a Web AI-based DSS in satellite operations is not a one-size-fits-all decision. The optimal model depends directly on the mission's priorities: for critical, high-value assets where detailed diagnostics are essential. This study fills a significant research gap by providing the first empirical evidence of these performance trade-offs, thereby guiding future development and deployment strategies for intelligent satellite anomaly-identification systems.

This research lays a foundation for further advancements in AI-driven satellite operations. Future work should focus on several key areas, including prioritizing the expansion of the

evaluation corpus to include a wider diversity of incident scenarios to verify whether Nemotron Nano 2's reliability advantage holds across edge cases. Additionally, investigating the specific failure modes of the Devstal Small model through qualitative error analysis could provide insights into architectural limitations. Further research could also explore fine-tuning strategies or Retrieval-Augmented Generation (RAG) implementations to potentially bridge the performance gap observed in mid-tier models like Gemma 3 and Devstal Small.

ACKNOWLEDGEMENT

This research is fully supported by the Asian Office of Aerospace Research and Development (AOARD) under the grant scheme numbers FA2386-23-1-4073 and SPI23-179-0179.

REFERENCES

[1] N. Kazanskiy, R. Khabibullin, A. Nikonorov, and S. Khonina, "A Comprehensive Review of Remote Sensing and Artificial Intelligence Integration: Advances, Applications, and Challenges," Sensors, vol. 25, no. 19, p. 5965, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/19/5965.

[2] M. Edson de Carvalho Souza and L. Weigang, "Grok, Gemini, ChatGPT and DeepSeek: Comparison and Applications in Conversational Artificial Intelligence," vol. 1, 02/18 2025, doi: 10.5281/zenodo.14885243.

[3] A. Mutholib, N. A. Rahim, T. S. Gunawan, and A. A. Ahmarofi, "Performance Comparison of Data Preprocessing Methods for Trade-Space Exploration with AI Model: Case Study of Satellite Anomalies Detection," in 2024 IEEE 10th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), 2024: IEEE, pp. 271-275.

[4] A. Mutholib, N. A. Rahim, T. S. Gunawan, and M. Kartiwi, "Trade-Space Exploration with Data Preprocessing and Machine Learning for Satellite Anomalies Reliability Classification," IEEE Access, 2025.

[5] P. H. Tran, A. Ahmadi Nadi, T. H. Nguyen, K. D. Tran, and K. P. Tran, "Application of machine learning in statistical process control charts: A survey and perspective," in Control charts and machine learning for anomaly detection in manufacturing: Springer, 2022, pp. 7-42.

[6] H. Akbarian, "Deep Learning Based Anomaly Detection in Space Systems and Operations," Florida Atlantic University, 2024.

[7] Cusumano, M. A., V. F. Farias, and R. Ramakrishnan, "Generative AI as a New Platform for Applications Development," An MIT Exploration of Generative AI no. September, 2024, doi: https://doi.org/10.21428/e4baedd9.f189351f.

[8] G. Team et al., "Gemma 3 technical report," arXiv preprint arXiv:2503.19786, 2025.

[9] G. Team et al., "Gemma: Open models based on gemini research and technology," arXiv preprint arXiv:2403.08295, 2024.

[10] T. Kao. "Why Gemma adopts the Decoder-Only Transformer architecture?" https://makerpro.cc/2024/04/why-gemma-adopts-decoder-only-transformer-architecture/ (accessed 20 October, 2025).

[11] "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation." https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (accessed 11/07/2025.

[12] A. Basant et al., "Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model," arXiv preprint arXiv:2508.14444, 2025.

[13] A. Rastogi et al., "Devstral: Fine-tuning Language Models for Coding Agent Applications," arXiv preprint arXiv:2509.25193, 2025.

[14] E. Turban, J. E. Aronson, and T. P. Liang, Decision Support Systems and Intelligent Systems, 7th Edition ed. New Delhi: Prentice Hall of India, 2007.

[15] R. Islam et al., "The future of cloud computing: benefits and challenges," International Journal

[16] A. Mutholib, N. A. Rahim, and T. S. Gunawan, "Prototype Development of Web AI-Based Decision Support System: Insights and Recommendations for Satellite Anomaly Identification," in Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS), Maui, Hawaii, September 16-19 2025.

[17] E. Mikołajewska, D. Mikołajewski, T. Mikołajczyk, and T. Paczkowski, "Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0," Applied Sciences, vol. 15, no. 6, p. 3166, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/6/3166.

[18] A. Alabed and T. Özkul, "Preliminary Design and Methodological Framework for an AI-Driven Decision Support System in Earth Observation Satellites," in 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA),

[19] B. Lane, M. Poole, M. Camp, and J. Murray-Krezan, "Using machine learning for advanced anomaly detection and classification," in Advanced Maui Optical and Space Surveillance Tech. Conf.(AMOS), 2016.

[20] T. M. Grile and R. A. Bettinger, "Statistical reliability estimation for satellites operating from 1991-2020 with payload reliability focus," in 2022 6th International Conference on System Reliability and Safety (ICSRS), 2022: IEEE, pp. 378-386.

[21] "Google: Gemma 3 27B (free)." https://openrouter.ai/google/gemma-3-27b-it:free (accessed 16 October, 2025).

[22] "Llama 4 Maverick 17B 128E Instruct." https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct (accessed 11/07/2025.

[23] N. Corporation. "NVIDIA-Nemotron-Nano-9B-v2." https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-9B-v2 (accessed 12 October, 2025).

[24] J. Denize. "Devstral Small" https://huggingface.co/mistralai/Devstral-Small-2505 (accessed 11/07/2025.

[25] A. Marengo, F. G. Karaoglan-Yilmaz, R. Yılmaz, and M. Ceylan, "Development and validation of generative artificial intelligence attitude scale for students," Frontiers in Computer Science, vol. 7, p. 1528455, 2025.

[26] R. Cole, "Inter-Rater Reliability Methods in Qualitative Case Study Research," Sociological Methods & Research, vol. 53, no. 4, pp. 1944-1975, 2024, doi: 10.1177/00491241231156971.

[27] A. Bonnet. "AI Metrics that Matter: A Guide to Assessing Generative AI Quality." https://encord.com/blog/generative-ai-metrics/ (accessed 15/09, 2025).