2025 10th International Conference on Machine Learning Technologies

Autonomous Multi-Agent AI Systems for Satellite Mission Design

Tomas Navarro
European Space Agency
ECSAT
Didcot, UK
tomas.navarro@esa.int

Ana Stroescu
European Space Agency
ECSAT
Didcot, UK
ana.stroescu@esa.int

Dario Izzo
European Space Agency
ESTEC
Nordwijk, Netherlands
dario.izzo@esa.int

Sergio Gálvez Rojas
University of Málaga
Language and Computer Science Department
Málaga, Spain
galvez@uma.es

Francisco López Valverde
University of Málaga
Language and Computer Science Department
Málaga, Spain
valverde@uma.es

2025 10th International Conference on Machine Learning Technologies (ICMLT) | 979-8-3315-3672-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMLT65785.2025.11192855

***Abstract*—The integration of Artificial Intelligence (AI) agents in supporting engineering design is rapidly gaining attention due to their potential to accelerate decision-making, optimise designs, and reduce costs. This paper presents a comprehensive evaluation of two different AI agentic systems, each system run by a different LLM (Large Language Model): DeepSeek-R1-70B and GPT-4o. The agents are evaluated in supporting satellite constellation design across key domains: market analysis, frequency filing, mission planning, payload feasibility, and cost analysis. Four distinct satellite designs were analysed per model, and expert evaluations were conducted to assess their effectiveness. This study highlights both the benefits and shortcomings of AI agents in satellite design, providing a comparative assessment and discussing implications for future AI-driven space mission planning.**

***Index Terms*—AI agents, LLM, Concurrent engineering, CDF, D2D, Autonomous design, SatCom.**

I. Introduction

Spacecraft design is a highly complex and interdisciplinary process that requires seamless integration of diverse expertise and significant time dedication. The design of a mission is a process that can take several months, with multiple iterations across different expert domains to refine technical, operational, and financial feasibility.

When this process is conducted through a concurrent engineering process or Concurrent Design Facility (CDF) session, dozens of experts from various disciplines work simultaneously in a collaborative, real-time environment. These experts interact dynamically under the guidance of a Study Manager to iteratively refine the mission architecture. This iterative cycle continues for several weeks, ensuring that the final design output aligns with mission objectives and customer requirements.

The process is, however, extremely time-consuming, requiring significant coordination efforts and extensive resources. Additionally, design iterations are often constrained by the following:

- Human cognitive limitations in handling large-scale trade-offs and complex multi-variable optimisations.
- Time availability of experts, leading to bottlenecks in decision-making and convergence of the design.
- Lack of automated exploration of a broader solution space, as many configurations are manually assessed.
- Subjective biases and knowledge gaps, where prior experience may lead to converging towards suboptimal, yet familiar solutions.

These limitations indicate a growing need for intelligent automation to support mission design, enabling faster exploration of optimal solutions, minimising expert workload, and enhancing decision-making efficiency. Recent advances in AI and multi-agent systems provide a promising opportunity to address these challenges.

This research explores the role of AI agents in satellite mission design, investigating their ability to replicate or enhance traditional concurrent engineering workflows, improve design efficiency, and reduce expert intervention, while adhering to mission objectives.

A. Objectives

The primary objective of this research is to quantify the potential time savings and assess the usefulness of AI-driven agentic workflows in the satellite mission design process. By leveraging state-of-the-art (SOTA) AI models such as DeepSeek-R1 [1] and GPT-4o [2], this study aims to evaluate how effectively these models, when integrated into a multi-agent system, can replicate, support, and potentially enhance the efficiency of traditional concurrent engineering design for space missions.

Specifically, this paper investigates the capabilities of advanced reasoning models in:

- Capturing built-in expert knowledge across mission analysis, payload design, cost estimation, market analysis and regulatory compliance, without Human-in-the-loop (HITL) support.

979-8-3315-3672-5/25/$31.00 ©2025 IEEE

- Automating design workflows to reduce manual workload, while maintaining high design fidelity.
- Generating optimised satellite configurations based on performance trade-offs, including link budget feasibility, coverage optimisation, and cost-effectiveness.

This study will compare AI-generated designs against human evaluations, measuring design accuracy, reproducibility, constraint adherence, hallucinations[1] rates, and expert intervention levels.

B. Prior Work

The integration of AI agents in engineering design has gained significant attention in recent years, with researchers exploring various applications, including decision support tools, evolutionary design systems, and interactive frameworks that facilitate human-AI collaboration [3]. Agent-based approaches have been successfully implemented in evolutionary design systems, where autonomous agents leverage search, interface, and information processing capabilities to enhance design exploration and optimisation [4].

In the field of space mission design, several studies have explored AI-based Design Engineering Assistants (DEAs) to enhance knowledge management during concurrent engineering sessions [5], [6]. These systems integrate natural language processing (NLP), machine learning (ML), and knowledge representation techniques to extract information from large historical mission datasets [7]. By utilising ontology learning methods, DEAs structure mission data for efficient retrieval, helping experts navigate complex design processes. In addition, AI-driven data mining techniques have been proposed for space system modelling, helping to identify interactions between complex subsystems and improving feasibility assessments [8].

While these prior efforts have focused on knowledge retrieval and structuring, they do not engage in autonomous generative design. The approach of this study extends beyond these methods by introducing a multi-agent AI framework that actively designs and optimises satellite missions. Key differences include:

- Multi-agent debate and self-correction, where AI agents iteratively refine designs through self-improving loops.
- Real-time AI-driven trade-off analysis, dynamically adjusting constellation sizing, link budgets, and cost estimates.
- AI-driven generative design, moving beyond data retrieval to propose new, optimised mission configurations.

Furthermore, challenges remain in ensuring that AI-generated designs align with real-world engineering constraints [3]. While AI has proven effective in assisting decision-making, its application in fully autonomous spacecraft design is still evolving. This study demonstrates how multi-agent collaboration can enhance AI's role in early-stage mission design, enabling faster conceptualisation, improved explainability, and reduced human workload.

II. Methodology

This research implements an AI-driven agentic workflow with a hierarchical structure, where an AI Study Manager oversees the satellite mission design process and dynamically assigns tasks to specialised expert AI agents. These agents collaborate iteratively to refine mission parameters, payload configurations, and financial feasibility assessments.

Given the computational cost and processing time constraints, the number of AI experts was limited to the most critical roles in this study. However, a more exhaustive setup could include additional domain-specific experts such as structural engineers, propulsion specialists, ground segment experts, user terminal specialists, power engineers, and deployment mechanism experts. Incorporating these additional agents would lead to a more comprehensive mission design, but at the cost of increased inference time and computational resource requirements.

The LLM models used in this research are GPT-4o and DeepSeek-R1-70B, deployed locally on a server with 64GB RAM memory and with a dedicated NVIDIA A40 GPU. These models were chosen for their advanced analytical tasks and technical problem-solving capabilities.

For the AI agent system, CrewAI [9] has been used as a proven framework to facilitate the design of AI-driven autonomous workflows [10]. As shown in Figure 1, the AI Study Manager acts as a central coordinator, receiving human input only once as key mission requirements. It dynamically assigns tasks to specific expert agents and can request iterative refinements if a generated output is unsatisfactory.

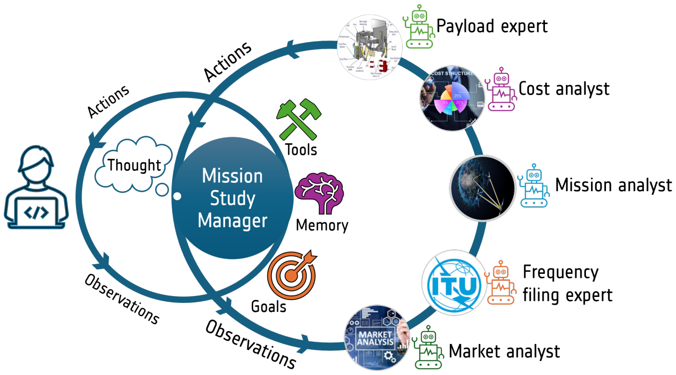

Fig. 1. Agentic AI architecture for satellite mission design.

The crew of AI agents involved in this study is composed of the following agents:

1) Market Analyst Agent: Identifies key markets, estimates potential user base and geographic demand and analyses economic feasibility.
2) Frequency Filing Expert Agent: Determines the most suitable frequency bands and ensures compliance with International Telecommunications Union (ITU) regulations.

[1] In the context of AI, *hallucinations* refer to inaccurate or fabricated outputs generated by models, especially when such outputs are not supported by the training data or real-world facts.

3) Payload Expert Agent: Calculates the power requirements, antenna sizing, and overall link budget feasibility.
4) Mission Analyst Agent: Determines the optimal satellite constellation configuration.
5) Cost Analyst Agent: Evaluates the financial feasibility of the AI-generated mission design.

III. Selected Mission Design

The satellite mission type selected for this study is Direct-to-Device (D2D) communications, which presents a unique set of complex design challenges that must be carefully addressed to ensure technical feasibility, regulatory compliance, and economic viability. These challenges include:

- Frequency allocation and regulatory compliance: Selecting the appropriate frequency band while ensuring ITU compliance and mitigating interference with terrestrial mobile networks.
- Link budget and power constraints: Closing the link with low-power mobile devices, requiring high satellite transmission power, large deployable antennas, and efficient power budgeting.
- Satellite Constellation and orbital trade-Offs: Balancing coverage, signal strength, and latency by optimising satellite constellations sizing and orbital altitude.
- Antenna Sizing and deployment feasibility: Designing large, lightweight, and reliable deployable antennas that meet performance and integration constraints.
- Network load and multi-beam management: Handling thousands of simultaneous mobile connections with dynamic beam allocation, frequency reuse, and traffic balancing strategies.
- Economic feasibility: Ensuring cost-effectiveness by optimising satellite manufacturing, launch, and operational expenses while maintaining competitive service pricing.

Fig. 2. Advanced concept of a D2D satellite system proposed by ESA.

IV. Results

The evaluation of AI-generated satellite mission designs incorporates both quantitative and qualitative scoring methodologies to ensure a comprehensive assessment. For qualitative scoring, a 1–5 scale has been used, integrating both measurable metrics where numerical validation is feasible, and expert judgment where qualitative assessments are necessary. Four different subject-matter experts evaluated the designs based on accuracy, feasibility, and alignment with real-world constraints. The quantitative assessment focuses primarily on link budget evaluation, where AI-generated results are compared against manually computed values using fundamental link budget equations. Key performance parameters, such as free-space path loss (FSPL), carrier-to-noise ratio (C/No) and received signal power are calculated to determine the AI's accuracy.

A. Comparison of AI models for satellite mission design across expert domains

A summary of the comparative analysis of the two agentic systems, which evaluates their performance across expert domains, is presented below. Table I summarises their respective capabilities and limitations in satellite mission design. The scoring represents an average across four different iteration designs, factoring in the inputs of at least two experts for each expert domain.

1) DeepSeek-R1-70B:

- Strengths: Demonstrated strong mission analysis capabilities, producing reasonable satellite constellation designs and altitude selections. It selected appropriate frequency bands, avoiding unrealistic high-frequency bands in frequency filing and it provided reasonable capital expenditures (CapEx) and operating expenses (OpEx) estimates, making it more reliable in cost analysis compared to GPT-4o.
- Weaknesses: Struggled with market analysis, failing to provide competitive analysis, affordability models, and realistic demand figures. It performed poorly in link budget calculations and while cost estimates were more structured than GPT-4o's, they lacked detailed return on investment (ROI) and breakeven calculations. It also failed to assess regulatory compliance comprehensively.

2) GPT-4o:

- Strengths: Improved accuracy in constellation sizing and link budget calculations, and it provided more realistic antenna size estimates across different designs. It performed better than DeepSeek-R1 in market analysis, correctly identifying underserved markets and structuring demand assessment.
- Weaknesses: GPT-4o lacked financial feasibility analysis, failing to estimate ROI and market competitiveness. It had gaps in frequency filing, with incorrect assumptions regarding bandwidth allocation and interference considerations. It overestimated satellite constellation sizing, making mission analysis inconsistent across different iterations. It also misinterpreted operational costs and failed to provide comprehensive economic benchmarking.

GPT-4o outperformed DeepSeek-R1-70B in constellation sizing accuracy, link budget calculations, and market analysis by providing more structured demand assessments and realistic

TABLE I
COMPARISON OF AI MODELS FOR SATELLITE MISSION DESIGN ACROSS EXPERT DOMAINS

Aspect evaluated	DeepSeek-R1-70B expert rating [1-5]	DeepSeek-R1-70B comment	GPT-4o expert rating [1-5]	GPT-4o comment
Market analysis	2	Identified underserved regions and some market segments, but lacked depth in economic feasibility and competitive analysis.	3	Identified market segments more accurately than DeepSeek-R1 but did not provide market figures or affordability models.
Frequency filing	2.5	Correctly selected appropriate frequency bands (e.g. L-band, MSS frequencies), but failed to evaluate ITU regulatory compliance thoroughly.	2.5	Provided better-structured frequency band selection, but lacked bandwidth assumptions, provided wrong details about commercial D2D frequency bands, did not address ITU compliance properly.
Payload analysis	2	Considered appropriate satellite antenna sizes, but performed poorly in overall link budget calculations and lacked beamforming analysis and power requirements.	3	Had realistic antenna size estimates and valid link budget margins, but lacked beamforming assessment, power system assessment, and spectral efficiency assumptions.
Mission analysis	3	Had a reasonable satellite constellation sizing and credible mission design, but lacked quantitative justification, and made inconsistent trade-offs between payload and coverage.	3	Attempted more detailed calculations but misinterpreted antenna beamwidth and satellite footprint, leading to incorrect satellite count estimates despite structured methodologies.
Cost analysis	3	Provided more structured CaPex and OpEx estimates, but occasionally contained errors in calculations.	2	Struggled with cost estimations, often misinterpreting mass, launch costs, and economic feasibility, making its financial projections incomplete or incorrect.
Final score	2.5		2.7	

antenna size estimates. However, it lacked financial feasibility analysis, made incorrect assumptions about bandwidth allocation and interference, and overestimated the required number of satellites.

DeepSeek-R1-70B was slightly better in mission analysis, offering reasonable satellite constellation designs and more structured CaPex and OpEx estimates, making it more reliable in cost analysis. However, it struggled with market analysis, failed to provide affordability models, and lacked comprehensive regulatory compliance assessments.

Overall, both models performed similarly, with GPT-4o performing better in design structuring and DeepSeek-R1-70B being more reliable in cost assessments, but neither model was fully consistent or reliable for autonomous mission planning. Both models require significant improvements in market feasibility analysis, competitor comparisons, regulatory compliance, and redundancy strategies to be fully reliable for mission planning.

Future work should integrate Human-in-the-loop (HITL) and Retrieval-Augmented Generation (RAG) techniques to improve knowledge retrieval, numerical consistency, and ITU compliance validation, to enhance the overall effectiveness of multi AI-Agents for satellite design.

B. Evaluation of Study Manager reasoning capabilities in task delegation/reasoning capabilities

The AI agent acting as the Study Manager was evaluated based on its interactions with a crew of expert agents, accross four iterations of the satellite mission design described in Section III. This evaluation followed CrewAI's hierarchical process, which structures task management by simulating traditional organisational hierarchies. This approach ensures efficient task delegation and execution.

The following key considerations have been taken into account when evaluating the agent reasoning capabilities. The AI study manager:

1) Should effectively delegate tasks to expert agents.
2) Should ensure logical iteration of design improvements.
3) Must verify whether agents respond to requested modifications correctly.
4) Should demonstrate emergent reasoning and self-correction.

The results of the evaluation of reasoning capabilities for both AI systems can be found in Table II and Table III.

The study concludes that GPT-4o outperforms DeepSeek-R1-70B in overall study manager reasoning. GPT-4o's Study Manager was more structured in delegating tasks and ensuring expert feedback integration. DeepSeek-R1-70B's Study Manager had major gaps in oversight and iterative improvement, leading to incomplete designs.

C. Quantitative assessment of link budget calculations

Accurate link budget calculations are crucial for assessing the feasibility and performance of satellite communication systems. This part of the study evaluates the capability of the two different LLM models (DeepSeek-R1-70B and GPT-4o) in predicting key link budget parameters compared to expert-validated results. Specifically, it examines the models' accuracy in calculating FSPL and effective throughput, identifying potential errors and limitations in their estimations.

The basic link budget equation (dB scale) is as follows:

$$C/N_0 = P_t + G_t + G_r - L_{\text{fs}} - L_{\text{other}} - N_0, \qquad (1)$$

where:

- C/N_0 = carrier-to-noise density [dB-Hz]

TABLE II
STUDY MANAGER REASONING EVALUATION - DEEPSEEK-R1-70B

Iteration #	Task delegation efficiency	Iteration quality	Expert responses	Overall decision-making	Score [1-5]
1	Inconsistent delegation to agents.	Minimal re-evaluations, lacks depth.	Some expert inputs ignored by the manager.	Lacks structured decision flow.	2
2	Assigns tasks but lacks oversight.	Few iterations, leading to incomplete design.	Provides partial validation, but does not refine errors.	Decisions feel arbitrary, lacks oversight.	2
3	Well-structured delegation with multiple iterations.	Revisits prior assumptions and refines designs.	Experts comply with tasks, feedback loops work well.	Best decision-making structure with clear task refinement.	5
4	Delegates tasks but lacks thorough validation.	Some iteration requests but lacks rigorous validation.	Misses the opportunity for full validation of design choices.	Some reasoning errors persist, lacks accountability.	3
Final score					3

TABLE III
STUDY MANAGER REASONING EVALUATION - GPT-4O

Iteration #	Task delegation efficiency	Iteration quality	Expert responses	Overall decision-making	Score [1-5]
1	Better delegation, but sometimes redundant tasks.	Some iteration, but limited re-analysis of past errors.	Experts provide useful inputs, but manager does not always follow up.	More structured than DeepSeek-R1-70B, but decisions sometimes inconsistent.	3
2	Assigns tasks correctly, but does not always ensure execution.	More iteration cycles than DeepSeek-R1-70B, but still limited.	Partial validation, sometimes revisits decisions, but inconsistently.	Logical, but lacks deeper validation.	3
3	Highly structured delegation, logical breakdown of tasks.	Best iterative process, refining errors and rechecking assumptions.	Best feedback loop, ensuring expert responses are iterated upon.	Best structured decision-making process with accurate refinements.	5
4	Mostly effective, but lacks validation steps in some cases.	Lacks a robust second-pass correction method.	Misses some expert refinements in the final version.	Some decisions lack clear justification.	4
Final score					3.75

- P_t = transmit power [dBm]
- G_t = transmit antenna gain [dBi]
- G_r = receive antenna gain [dBi]
- L_{fs} = free-space path loss [dB]
- L_{other} = additional system losses [dB], (e.g., atmospheric losses)
- N_0 = noise power spectral density [dBm/Hz]

FSPL is calculated using the following equation:

$$L_{\text{fs}} = 20\log_{10}\left(\frac{4\pi df}{c}\right), \tag{2}$$

where:

- d = distance [m], calculated from altitude and Earth's radius.
- f = frequency [Hz].
- c = speed of light [m/s].

The Carrier-to-Noise Ratio (C/N) [dB] represents the signal strength relative to noise and is computed using the Carrier-to-Noise Density Ratio (C/N_0) as follows:

$$C/N = C/N_0 - B, \tag{3}$$

where B is the system bandwidth [dB-Hz], calculated as:

$$B = 10\log_{10}. \tag{4}$$

This equation accounts for the signal power received over the available bandwidth, determining how well the system can mitigate noise effects.

The Shannon-Hartley Theorem defines the maximum achievable channel capacity [bps] (theoretical Shannon capcity), given by:

$$C_{\text{Shannon}} = B\log_2(1 + SNR), \tag{5}$$

where SNR is the Signal-to-Noise Ratio, given by:

$$SNR = 10^{(C/N)/10}. \tag{6}$$

In real-world systems, due to modulation, coding schemes, and system overhead, the Effective Throughput is typically lower than Shannon capacity. The effective (actual) throughput [bps] can be expressed as:

$$C_{\text{Eff}} = \eta C_{\text{Shannon}}, \tag{7}$$

where η is the efficiency factor, ($0 < \eta < 1$), depending on coding, modulation, and protocol overhead.

To evaluate the accuracy of the AI-generated link budget parameters, two key metrics are considered: Free-Space Path Loss (FSPL) and Effective Throughput.

1) FSPL Error Calculation: FSPL is a logarithmic value, so it must be converted to its linear scale before computing the percentage error. The error is calculated as:

$$E_{\text{FSPL}} = \frac{1}{N} \sum_{j=1}^{N} \left| \frac{10^{X_{\text{AI, FSPL},j}/10} - 10^{X_{\text{Human, FSPL},j}/10}}{10^{X_{\text{Human, FSPL},j}/10}} \right| \times 100, \tag{8}$$

where:

- E_{FSPL} represents the average percentage error for FSPL across N link budget scenarios,
- $X_{\text{AI, FSPL},j}$ and $X_{\text{Human, FSPL},j}$ are the AI-generated and human-calculated values of FSPL, measured in [dB], for the j-th link budget.

2) Effective Throughput Error Calculation: Since Effective Throughput is measured in Mbps (linear scale), the percentage error is computed directly as:

$$E_{\text{Eff\_Throughput}} = \frac{1}{N} \sum_{j=1}^{N} \left| \frac{X_{\text{AI, Eff},j} - X_{\text{Human, Eff},j}}{X_{\text{Human, Eff},j}} \right| \times 100, \tag{9}$$

where:

- $E_{\text{Eff\_Throughput}}$ represents the average percentage error for Effective Throughput across N link budget scenarios,
- $X_{\text{AI, Eff},j}$ and $X_{\text{Human, Eff},j}$ are the AI-generated and human-calculated values of Effective Throughput, measured in [Mbps], for the j-th link budget.

The accuracy of AI models DeepSeek-R1-70B and GPT-4o was assessed across downlink and uplink link budgets. The results, averaged over multiple link budget scenarios, are presented in Table IV.

TABLE IV
ACCURACY OF AI MODELS ACROSS DOWNLINK AND UPLINK LINK BUDGETS

Metric	DeepSeek-R1-70B	GPT-4o
Downlink Accuracy [%]		
FSPL (%)	62.81	62.95
Eff Throughput (%)	57.84	94.41
Uplink Accuracy [%]		
FSPL (%)	56.76	56.89
Eff Throughput (%)	51.42	14.82

In summary, both models required multiple attempts to refine the prompts to generate meaningful outputs. In the initial attempts, both models produced completely different and inaccurate C/N values, leading to untrustworthy results. However, once the prompts were correctly formulated, both models demonstrated a solid understanding of link budgets and satellite communications principles.

For FSPL, both models made similar assumptions about propagation losses and correctly computed the slant range. However, they struggled with realistic assumptions about additional path losses, which led to significant errors. Regarding effective throughput, both models exhibited systematic errors due to incorrect estimations of system losses and unrealistic efficiency assumptions. These inconsistencies highlight their limitations in non-deterministic aspects of link budget calculations. Overall, both models demonstrated similar knowledge of link budgets, but their reliability is limited when dealing with assumptions beyond standard calculations. While AI models can be valuable tools for supporting such analyses, human oversight is essential to verify assumptions and ensure accurate results. Given the high sensitivity of link budget calculations to even minor errors, AI-generated outputs cannot be fully trusted without expert validation.

Finally, it is important to emphasise that these results reflect the built-in capabilities of the AI models without any fine-tuning or knowledge enhancement through retrieval-augmented generation (RAG). The models were not explicitly trained or guided to perform link budget calculations according to specific standards. Future work will focus on enhancing the models' expertise by fine-tuning them and incorporating RAG-based approaches to improve their accuracy and alignment with industry-standard link budget methodologies.

D. *Evaluation of key performance metrics*

The conclusions that can be drawn from this evaluation, which can be found in Table V, are:

- GPT-4o demonstrated superior accuracy, greater consistency, and fewer hallucinations, making it a more reliable AI model for supporting satellite mission design.
- Both models struggled with complex engineering trade-offs, particularly regulatory compliance, cost estimation, and link budget considerations.
- Neither model is currently reliable enough for fully autonomous mission planning, highlighting the need for expert oversight in critical design phases.

V. CONCLUSIONS AND FUTURE WORK

This study provides a preliminary evaluation of how current state-of-the-art AI models, including DeepSeek-R1 and GPT-4o, perform in reasoning and technical knowledge for satellite mission design. While AI agents demonstrate significant potential in accelerating feasibility studies and trade-space exploration, their reliability is not yet sufficient for autonomous mission planning requiring human in the loop in key critical design milestones. The key takeaways are:

- Multi-AI agents can rapidly generate structured satellite mission designs, but struggles with complex physics-based calculations and cost estimations.
- GPT-4o outperforms DeepSeek-R1 in accuracy and it exhibited fewer hallucinations. Constraint adherence and reproducibility were considered poor for both models.
- Mission Analyst and Payload Expert Agents provided the most useful outputs, while the Market Analyst, Frequency Filing Expert and Cost Analyst Agents showed frequent errors.
- AI models hallucinate critical engineering parameters, requiring expert validation to avoid design flaws.

To improve the technical accuracy and domain knowledge of AI-driven mission design agents, future research will explore the use of Retrieval-Augmented Generation (RAG) techniques. RAG will enable AI agents to access validated engineering knowledge from technical databases and technical documents,

TABLE V
COMPARISON OF AI MODELS FOR SATELLITE MISSION DESIGN ACROSS EXPERT DOMAINS.

Metric	**Definition**	**DeepSeek-R1-70B rating [1-5]**	**Justification**	**GPT-4o rating [1-5]**	**Justification**
Design accuracy	Alignment of AI-generated designs with validated models.	2.5	Had reasonable satellite constellation sizing and credible design but, lacked detailed cost and power assessments.	3.5	Demonstrated improved accuracy, especially in constellation sizing and link budgets.
Constraint adherence	Compliance with physical and regulatory constraints.	2.5	Handled power and satellite constellation sizing constraints and properly identified frequency bands, but lacked interference assessment and regulatory compliance review.	2.5	Handled power and satellite constellation sizing constraints better and provided more realistic market segmentation, however it lacked power system assessment and spectral efficiency assumptions.
Reproducibility	Consistency of results across different runs.	1.5	Was fairly inconsistent across multiple iterations. Different user needs, different payload and constellation sizing.	2	The outputs were more stable across reruns when compared to DeepSeek-R1-70B, and provided clearer explanations of design choices but they were in general inconsistent from one run to another.
Hallucination rate	Incorrect or unverifiable AI-generated outputs.	2.5	Had a higher rate of hallucinated outputs, especially in link budgets and power requirements.	3.5	Reduced hallucinated outputs but still had some inconsistencies in both payload and mission analysis.
Final score	Average of all ratings	2.2		2.9	

and ITU regulations, reducing hallucinations. RAG will also improve explainability and numerical consistency by backing AI decisions with retrieved, verifiable data.

Thanks to RAG, trusted engineering datasets and computational models (e.g., Python-based physics models, Excel datasheets, and link budget simulators) can be integrated, allowing AI agents to cross-validate their outputs with existing industry-standard calculations. This will significantly increase the reliability of AI-generated satellite architectures, reducing the need for manual correction by human experts.

Beyond RAG, future work will also focus on incorporating Human-in-the-Loop (HITL) feedback to enhance the credibility of AI-driven mission design. By integrating expert oversight at critical decision points, AI agents can refine their reasoning, correct biases, and improve overall feasibility assessments. However, this study deliberately focused on a fully human-off-the-loop approach to evaluate the independent reasoning capabilities of AI agents. This preliminary evaluation provides a baseline understanding of how autonomous AI systems perform in mission design without human intervention, serving as a foundation for future HITL-enhanced frameworks.

ACKNOWLEDGMENT

Special thanks to Daniel Dinis, a satellite communications expert, for supporting the quantitative assessment of link budgets. We also extend our gratitude to Barnaby Osborne, an expert in small satellite technology with a background in orbital mechanics and mission analysis, for contributing to the mission analysis and cost evaluation. Finally, we acknowledge Clement Cesbron, a business analyst and expert in frequency filing regulations, for assisting with the market analysis and regulatory compliance assessment. Their valuable insights and expertise, as colleagues at European the Space Agency, were instrumental in supporting this study.

REFERENCES

[1] DeepSeek-AI et al.,”DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.”, arXiv, 2501.12948, 2025.

[2] Aaron Hurst et al., ”GPT-4o System Card”, OpenAI, arXiv, 2410.21276, 2024.

[3] M. Shergadwala and M. Seif El-Nasr, ”Agent-based support within an interactive evolutionary design system. AI EDAM.”, 2021

[4] D. Cvetkovic and I. C. Parmee, ”Agent-based support within an interactive evolutionary design system.”, AI EDAM, 2002.

[5] M Murdaca, et al., ”Artificial intelligence for early design of space missions.”, SECESA, 2018.

[6] A. Berquand, et al., ”Towards an artificial intelligence-based design engineering assistant.”, 2018.

[7] A. Berquand, et al., ”AI-based design engineering assistant for space mission design.”, 2019.

[8] C. Guariniello, et al., ”AI agents for space systems modeling.”, IEEE Aerospace Conference, 2020.

[9] crewAI, 2018, Source: https://www.crewai.com/ (Cited on 26th February 2025).

[10] Tomas Navarro, Ana Stroescu, Dario Izzo, Sergio Gálvez Rojas, Francisco López Valverde, ”Decision Making for Planetary Landing Applications using AI Agents and Reinforcement Learning”, ESA Spaice Conference, September 2024. Source: https://zenodo.org/records/13889941 (Cited on 26th February 2025).