

AI Assistants for Spaceflight Procedures: Combining Generative Pre-Trained Transformer and Retrieval-Augmented Generation on Knowledge Graphs With Augmented Reality Cues

Oliver Bensch^{*1,3}, Leonie Bensch^{1,2}, Tommy Nilsson², Florian Saling², Bernd Bewer², Sophie Jentzsch¹, Tobias Hecking¹, J. Nathan Kutz⁴

¹*German Aerospace Center, Institute for Software Technology*

²*European Space Agency, European Astronaut Center (ESA-EAC)*

³*AI Institute in Dynamic Systems, University of Washington, Seattle, WA USA*

⁴*Department of Applied Mathematics, University of Washington, Seattle, WA USA*

This paper describes the capabilities and potential of the intelligent personal assistant (IPA) CORE (Checklist Organizer for Research and Exploration), designed to support astronauts during procedures onboard the International Space Station (ISS), the Lunar Gateway station, and beyond. We reflect on the importance of a reliable and flexible assistant capable of offline operation and highlight the usefulness of audiovisual interaction using augmented reality elements to intuitively display checklist information. We argue that current approaches to the design of IPAs in space operations fall short of meeting these criteria. Therefore, we propose CORE as an assistant that combines Knowledge Graphs (KGs), Retrieval-Augmented Generation (RAG) for a Generative Pre-Trained Transformer (GPT), and Augmented Reality (AR) elements to ensure an intuitive understanding of procedure steps, reliability, offline availability, and flexibility in terms of response style and procedure updates.

1 Introduction

Astronauts embarking on space missions typically rely on standardized work procedures refined through decades of operational experience. These procedures guide a wide range of tasks, including maintenance, scientific experiments, and emergency response.

Historically, astronauts used relatively simple tools, such as paper checklists, to execute these procedures, often requiring oversight and assistance from mission control centres on Earth. While this approach was feasible for supporting activities in Low Earth Orbit, such as aboard the ISS, future missions will present new challenges. As the focus of human spaceflight gradually shifts to the Moon, with crewed missions to Mars envisioned to follow soon after, the number and complexity of relevant procedures will inevitably increase. Future astronauts will face unprecedented and often

unpredictable situations, all while frequent radio black-outs and communication delays will render direct support from Earth problematic. Consequently, there is a need for a new generation of tools that provide clear and contextually appropriate assistance, enabling astronauts to operate safely and efficiently on their own [1–5].

The rapidly evolving field of artificial intelligence has attracted significant attention in this context. Early experimental solutions have been centred around IPAs designed to facilitate natural, hands-free support and guidance for procedural tasks, thereby reducing the need for direct supervision from Earth.

In this regard, a key milestone was the 2018 ISS deployment of CIMON, a Watson-powered IPA [6]. Its successor, the CIMON-2, launched in 2019 and included enhanced features, such as visual recognition and emotion analysis, but required internet connection and lacked autonomous learning [7]. Similarly, in 2022, Amazon’s Alexa was integrated into the Artemis I lunar mission [8]. Yet, both IPAs are limited by their rule-based nature, meaning that they can only respond to a pre-defined set of questions regarding procedure steps. Meanwhile, the AI4U assistant, developed using a reinforcement learning approach, is being tested under simulated Mars conditions during a Mars analog study with the goal of eventually being deployed and further assessed aboard the ISS [9–11]. Unlike rule-based IPAs, like CIMON and Alexa, AI4U learns new skills during operation but requires extensive training sessions.

In summary, contemporary IPAs designed to support astronauts suffer from important limitations, including the lack of offline availability, limited flexibility regarding their range of answers, and delays in updating procedures (due to their rule-based nature or the requirement of extensive training). Moreover, these IPAs

^{*}Corresponding author.

provide relevant procedure information primarily in the form of text-based and/or spoken instructions. Although they sometimes make use of visual aids, including pictures and videos, they are currently incapable of delivering more contextually appropriate and spatially intuitive visual information, such as 3D representations of procedure steps when conducting complex operations.

This limitation has already been noted to increase the workload for astronauts, as they had to switch their gaze between the checklist and the task at hand, while requiring complex mental rotations to interpret visual information from 2D pictures or videos, causing cognitive dissonance between procedure instructions and the actual task [12].

To address these shortcomings, we propose a novel IPA called *CORE*, specifically designed to support astronauts during upcoming human spaceflight missions. *CORE* is based on Generative Pre-Trained Transformers (GPT), Retrieval-Augmented Generation (RAG), Knowledge Graphs (KG), speech recognition, and speech synthesis. This unique combination enables *CORE* to effectively emulate human conversation, while the astronaut is visually supported using augmented reality (AR) elements displayed in a head-up display (HUD).

2 IPAs for Future Space Missions

Reliability: Reliability is a critical requirement for IPAs supporting astronauts during procedure completion, as it directly affects crew safety and the success of the mission. Firstly, any misinformation or errors in the system regarding procedure steps could potentially lead to fatal consequences. Secondly, unreliability in conversational agents often leads to user disengagement, distrust, and frustration [14]. Hence, it can be assumed that the slightest unreliability during procedure support would lead to the disengagement of the astronauts and a preference for the usage of traditional tools.

To significantly improve the reliability of astronautical procedure support, we propose the integration of GPTs with advanced data retrieval technologies.

GPT models, recognized for their ability to generate human-like responses without explicit instructions, play an important role in modern AI systems [15]. However, their reliance on training datasets poses challenges, as the information generated may be inaccurate or misleading, especially if it falls outside the scope of their training data [16]. While model fine-tuning can adapt GPTs to specific domains by incorporating domain-specific knowledge, this process is computationally intensive and lacks transparency, as it provides no sources for the generated responses [16].

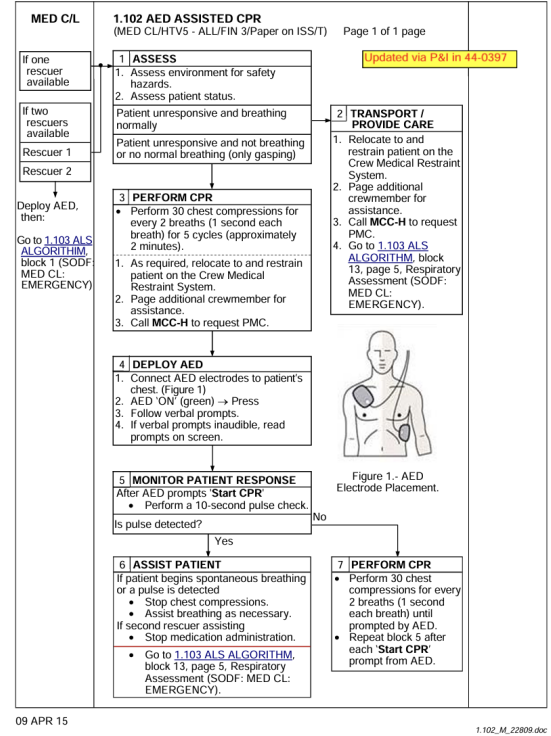


Figure 1: Example ISS procedure containing specific steps for a CPR, featuring pictures and location-dependent information about the placement of electrodes on the body of the patient [13].

To address these limitations and enhance reliability, we propose the integration of GPT models with RAG and KGs. This approach leverages the generative capabilities of GPTs alongside a robust retrieval system that utilizes a structured knowledge database [17]. By interlinking documents, metadata, and multimodal data within a KG, connected by references, topics, keywords, or related company or person information, such a system ensures that all information retrieved and utilized by the AI is accurate and up-to-date. User queries initiate a vector search within this database to identify relevant information, which is added to the GPT query [18]. This approach not only reduces the likelihood of erroneous outputs (hallucinations) but also improves the system's trustworthiness by allowing astronauts to verify the answer through the linked information used for its generation. Additionally, we suggest the enhancement of system transparency and user trust through the Graph of Thoughts (GoT) paradigm, which can be used to visualize the reasoning process of AI, enabling users to interact with and adjust the AI's thought process in real-time [19, 20].

Furthermore, to further improve the astronauts understanding of the systems capabilities, we propose displaying confidence scores next to retrieved data

during RAG, indicating the reliability of the information retrieved [21, 22]. This method, complemented by using KGs linked with established ontologies like schema.org [23], ensures astronauts have access to rich, well-organized, and easily navigable information, crucial for decision-making in mission-critical scenarios [24–26].

Flexibility: Importantly, the assistant’s adaptability to new situations and requirements, ranging from routine procedures to unexpected emergencies, is a critical factor. Furthermore, the assistant should be flexible enough to accommodate the varied personal styles of astronauts during procedure completion. For instance, one astronaut might prefer interacting with the assistant in a voice-based manner, while another may opt for displaying written procedure steps. Additionally, preferences in interaction styles may differ among astronauts, as one may prefer an assistant that explains just the necessary steps, while another might favor an assistant with its "own character" that can engage in casual conversations, similar to a colleague. Lastly, one crucial aspect for such a system is that procedures can be flexibly adapted due to updates in the procedures themselves.

Regarding the need for flexible response styles, such as shifting between casual and formal tones, GPT-based assistants exhibit a level of versatility, unlike classic IPAs limited to predefined responses. GPT models can be steered continuously depending on requirements and preferences using steering vectors [27]. The latest GPT models are capable of document writing and editing and are evaluated to modify, add, or remove components of Knowledge Graphs or document databases through natural language conversations [28]. Furthermore, RAG enables such a system to access the latest updated information directly after indexing into a vector database, without the need for retraining of the GPT model [18]. Linked information in a mission KG can be added or edited in a natural conversation without changing the approved procedure’s content.

Offline Availability: Given the risk of radio blackouts and communication delays inherent in human spaceflight (and lunar south pole operations in particular), one additional crucial aspect is the need for offline availability when developing future IPA concepts to support astronauts during procedure completion.

While KG databases are widely available for offline deployment, leading GPT models like Gemini 1.5 or GPT-4 and leading speech synthesis modules are only available online [29–31] due to their extensive computational requirements and intellectual property protection. However, published open-source GPT models like Meta’s Llama 3 or Mistral’s Mixtral 8x7b [32] are narrowing the performance gap while also reducing model size [33] and, consequently, hardware require-

ments. These models can benefit from further optimization methods such as low-rank adaptation (LoRa) [34] and low-bit quantization [35], which enable small models like Microsoft’s Phi-3 to run on mobile devices without the requirement of a dedicated GPU [34, 36, 37]. This approach allows for a streamlined and efficient AI assistant setup in astronaut suits for deep space missions, where critical functionalities like speech recognition are embedded in the suit, while more computationally intensive tasks are handled by the larger GPT model and database located at a nearby base, connected through reliable and high-speed radio communication links.

Overall, we propose that an offline IPA based on a GPT model with an RAG architecture querying a KG for astronauts with low hardware requirements is already feasible. Although the quality of answers generated by open-source GPT models outperforms that of GPT-3.5, they still lag behind GPT-4 [33]. Furthermore, the quality of voices generated by open-source Text-To-Speech models has yet to match that of online available models like OpenAI’s [38].

Combining Voice, Text, and Visual 3D Information:

Text and picture-based procedures pose the risk of significantly increasing the workload of astronauts due to 1) the need to switch attention back and forth between the visual spaces of the task and the procedure instructions and 2) the mental effort of mapping instruction steps to the real equipment [12, 39].

To address these issues, past research has demonstrated that HUD AR technology can ease the work of astronauts during procedural tasks onboard the ISS by superimposing computer-generated instructions directly onto real equipment [12, 40, 41]. Additionally, the combination of AR and voice commands offers hands-free interaction without requiring astronauts to divert their gaze to another display [42, 43]. Moreover, recent research has shown that combining LLMs and AR HUDs is beneficial for the completion of procedural maintenance tasks [44]. Therefore, we suggest combining the advantages of AR technology and LLMs to allow for intuitive interaction between astronauts and procedure elements. We furthermore propose that graphically displaying relational information from the KG could further aid astronauts in understanding linked concepts, components, documents, or images during critical situations, enhancing the assistant’s explainability [45, 46].

3 Results - CORE Prototype

In our initial experiments, we indexed publicly available procedures [13] into a vector database and combined the documents with metadata such as the edit date

and related images in the *ArangoDB* graph database. Subsequently, we implemented a RAG system based on GPT-4 as a GPT model and indexed textual information into the Elasticsearch vector database. A speech user query is first converted into text using *Whisper* [31]. Relevant procedure information is retrieved from the vector database and added to the user query, along with linked information in the graph, using the following prompt template:

You will be presented with a matching procedure enclosed by three quotation marks ("). If a question is asked, respond with either the next step or only the first step. Specify the relevant step and repeat the text of the procedure step verbatim. If the information does not correspond to the question or if information is missing, state that this is the case.
 ""[Procedure] [Graph information] ""

Furthermore, using the following system-prompt, we directed the GPT-model to respond exclusively to procedure-related questions and return the current step as well as information about linked images in a separate format to be displayed to the user via the interface if necessary:

You are a helpful assistant for astronauts, answering questions about provided procedures. If asked a question, respond with either the next step or the first step only. Name the corresponding step as «STEP [NUMBER]» and repeat the text of the procedure step word for word. If a figure or other data is referenced, include «SHOW FIGURE [NUMBER]» in your answer.

Our first tests have shown that depending on the system-prompt, the GPT-4 model is able to interpret the given procedures and answer related user questions correctly. For example procedure steps were returned verbatim for the user question:

Hi, I have a person that is not breathing. I have already requested PMC. What was the fourth step of the ISS CPR procedure?

The system returns the corresponding step word by word and adds additional information to correctly display the step and the linked image.

«STEP 4» - DEPLOY AED: Connect AED electrodes to patient's chest. (See Figure 1) AED ON (green) → Press Follow verbal prompts. If verbal prompts inaudible, read prompts on screen. Continue with "Step 5" «SHOW FIGURE 1»

When the system is asked for graph data that is not added as a full sentence, such as "Last update: 09 April 2015" the system formulates a full sentence. For instance, if the user asks:

When was the procedure last updated?

the system replies:

The CPR procedure on the ISS was last updated on 09 April 2015.

Moreover, unrelated questions regarding other topics than procedure-related information are not answered by CORE.

4 Discussion

Our work highlights the significance of IPA technologies for the future of human space exploration. In response to unprecedented challenges, we propose a solution based on state-of-the-art natural language processing methods. Our primary focus is to develop a versatile IPA that seamlessly adapts to a range of tasks and interaction styles through the use of GPT models. These models surpass current IPAs, being capable of handling unforeseen situations with enhanced reasoning, creativity, and task-solving abilities. Reliability is another crucial factor, as it directly impacts astronauts' trust and engagement with the system. We recommend the integration of RAG and KG technology with GPT models, ensuring accurate information retrieval and enhancing trustworthiness, which is vital for high-stakes space operations. Additionally, we emphasize the need for offline availability of these systems, considering the communication limitations in space. Recent technological advancements suggest the feasibility of efficient, offline IPAs that require minimal hardware resources. Finally, we propose a combination of voice-based and visual information, utilizing AR cues to enhance the intuitive understanding of complex spatial information such as checklists, navigation, or KG relations.

In the second version of CORE, we plan to integrate AR HUD features into our prototype and test it in a user study at the European Astronaut Center (ESA EAC) to assess user experience and performance metrics.

References

1. *Apollo 11 Preliminary Science Report* <https://history.nasa.gov/alsj/a11/a11psr.html>. 1969.
2. Mindell, D. A. *Digital Apollo: Human and machine in spaceflight* (2011).

3. Hersch, M. H. Checklist: The secret life of Apollo's 'fourth crewmember'. *The Sociological Review* **57**, 6–24 (2009).
4. Marshburn, T. *et al.* An independent human factors analysis and evaluation of the emergency medical protocol checklist for the International Space Station tech. rep. (2003).
5. Kintz, N. M. & Palinkas, L. A. Communication delays impact behavior and performance aboard the international space station. *Aerospace Medicine and Human Performance* **87**, 940–946 (2016).
6. Deutsches Zentrum für Luft- und Raumfahrt (DLR). *CI-MON gelingt Weltpremiere mit Alexander Gerst auf der ISS* https://www.dlr.de/de/aktuelles/nachrichten/2018/4/20181116_cimon-weltpremiere-alexander-gerst-iss. Nov. 2018.
7. Schmitz, H.-C. *et al.* Towards robust speech interfaces for the ISS in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (2020), 110–111.
8. Amazon. *Amazon and Lockheed Martin are sending Alexa to space as part of Callisto, a technology demonstration on NASA's upcoming Artemis I mission.* <https://www.aboutamazon.com/news/devices/alex-take-me-to-the-moon>. 2022.
9. Shashkovaa, E., Navarrob, G., Royc, R. N., Pailletd, A. & Truntzlere, L. Study and development of an AI assistant for future Moon and Mars stations. (2022).
10. Navarro, G., Desjean, M.-C. & Paillet, A. *ECLSS Technology Roadmap at Spaceship FR* in (2023).
11. SPooN. *AI-4U: A personal assistant for astronauts* <https://www.spoon.ai/cnes-ai-4u>. 2024.
12. Braly, A. M., Nuernberger, B. & Kim, S. Y. Augmented reality improves procedural work on an international space station science instrument. *Human factors* **61**, 866–878 (2019).
13. National Aeronautics and Space Administration. *Emergency Medical Procedures Manual for the International Space Station (ISS)* (National Aeronautics and Space Administration, Washington, DC, 2016).
14. Tulshan, A. S. & Dhage, S. N. *Survey on virtual assistant: Google assistant, siri, cortana, alexa in Advances in Signal Processing and Intelligent Recognition Systems: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers 4* (2019), 190–201.
15. Bubeck, S. *et al.* *Sparks of Artificial General Intelligence: Early experiments with GPT-4* 2023. arXiv: 2303.12712 [cs.CL].
16. Zheng, S., Huang, J. & Chang, K. C.-C. Why Does ChatGPT Fall Short in Providing Truthful Answers? *ArXiv preprint, abs/2304.10513* (2023).
17. He, X. *et al.* *G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering* 2024. arXiv: 2402.07630 [cs.LG].
18. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020).
19. Besta, M. *et al.* *Graph of Thoughts: Solving Elaborate Problems with Large Language Models* 2023. arXiv: 2308.09687 [cs.CL].
20. Wu, T., Terry, M. & Cai, C. J. *AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts* in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, , New Orleans, LA, USA, 2022). ISBN: 9781450391573. <https://doi.org/10.1145/3491102.3517582>.
21. Rechkemmer, A. & Yin, M. *When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models* in *Proceedings of the 2022 chi conference on human factors in computing systems* (2022), 1–14.
22. Zhang, Y., Liao, Q. V. & Bellamy, R. K. *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making* in *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), 295–305.
23. Guha, R. V., Brickley, D. & Macbeth, S. Schema.org: evolution of structured data on the web. *Communications of the ACM* **59**, 44–51 (2016).
24. Auer, S. *et al.* *DBpedia: a nucleus for a web of open data* in (Springer-Verlag, Busan, Korea, 2007), 722–735. ISBN: 3540762973.
25. Vrandečić, D. *Wikidata: a new platform for collaborative data collection* in *Proceedings of the 21st International Conference on World Wide Web* (Association for Computing Machinery, Lyon, France, 2012), 1063–1064. ISBN: 9781450312301. <https://doi.org/10.1145/2187980.2188242>.
26. Zhu, X. *et al.* Multi-Modal Knowledge Graph Construction and Application: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 1–20. ISSN: 2326-3865. <http://dx.doi.org/10.1109/TKDE.2022.3224228> (2022).
27. Konen, K. *et al.* Style Vectors for Steering Generative Large Language Model. *arXiv preprint arXiv:2402.01618* (2024).
28. Pan, S. *et al.* Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 1–20 (2024).
29. Team, G. *et al.* *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* 2024. arXiv: 2403.05530 [cs.CL].
30. OpenAI *et al.* *GPT-4 Technical Report* 2024. arXiv: 2303.08774 [cs.CL].
31. Radford, A. *et al.* *Robust speech recognition via large-scale weak supervision* in *Proceedings of the 40th International Conference on Machine Learning* (JMLR.org, Honolulu, Hawaii, USA, 2023).
32. Touvron, H. *et al.* *Llama 2: Open Foundation and Fine-Tuned Chat Models* 2023. arXiv: 2307.09288 [cs.CL].
33. Jiang, A. Q. *et al.* *Mixtral of Experts* 2024. arXiv: 2401.04088 [cs.LG].
34. Hu, E. J. *et al.* *LoRA: Low-Rank Adaptation of Large Language Models* 2021. arXiv: 2106.09685 [cs.CL].
35. Zhao, Y. *et al.* *Atom: Low-bit Quantization for Efficient and Accurate LLM Serving* 2023. arXiv: 2310.19102 [cs.LG].
36. Carreira, S., Marques, T., Ribeiro, J. & Grilo, C. *Revolutionizing Mobile Interaction: Enabling a 3 Billion Parameter GPT LLM on Mobile* 2023. arXiv: 2310.01434 [cs.CL].

37. Abdin, M. *et al.* *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone* 2024. arXiv: 2404.14219 [cs.CL].
38. OpenAI. *TTS (text-to-speech) model* <https://platform.openai.com/docs/models/tts>.
39. Yeh, M., Wickens, C. D. & Seagull, F. *Effects of frame of reference and viewing condition on attentional issues with helmet mounted displays* MA thesis (University of Illinois at Urbana-Champaign, 1998).
40. Markov-Vetter, D. & Staadt, O. *A pilot study for Augmented Reality supported procedure guidance to operate payload racks on-board the International Space Station in 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2013), 1–6.
41. Helin, K., Kuula, T., Vizzi, C., Karjalainen, J. & Vovk, A. User experience of augmented reality system for astronaut's manual work support. *Frontiers in Robotics and AI* **5**, 106 (2018).
42. Cardenas, I. S., Powlison, K. & Kim, J.-H. *Reducing cognitive workload in telepresence lunar-martian environments through audiovisual feedback in augmented reality in Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (2021), 463–466.
43. Rometsch, F. A. *et al.* Design and evaluation of an Augmented Reality tool for future human space exploration aided by an Internet of Things architecture. *Advances in Space Research* **70**, 2145–2166 (2022).
44. Xu, F., Nguyen, T. & Du, J. Augmented Reality for Maintenance Tasks with ChatGPT for Automated Text-to-Action. *Journal of Construction Engineering and Management* **150**, 04024015 (2024).
45. Nararatwong, R., Kertkeidkachorn, N. & Ichise, R. *Knowledge graph visualization: challenges, framework, and implementation in 2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (2020), 174–178.
46. Serrano, F., Nunes, S. & Pesquita, C. *VOWLExplain: Knowledge graph visualization for explainable artificial intelligence in Proceedings of the Seventh International Workshop on the Visualization and Interaction for Ontologies and Linked Data* **3253** (2022), 29–40.