A Pilot Study: Naive smart interfaces can cause accidents

Christian Arnold[1,*,†], Paul Robertson[2,†], Zoe Robertson[2], Robert Laddaga[2], Boris Katz[1] and Andrei Barbu[1]

[1]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[2]Dynamic Object Language Labs, Lexington, Massachusetts, USA

Abstract

We show that an agent with a simple interface, used sparingly, but proactively in alerting trained pilots of aircraft in simulated situations can dramatically improve aviation safety in recreated fatal accidents. However, naive implementation can increase the number of fatal accidents due to hallucinations from large language models. In our simulated evaluation of 23 pilots, we explore the utility and pitfalls of intelligent user interfaces in the cockpit to surprising results–ones that should serve as a cautionary tale to integrating seemingly intelligent systems in high-pressure situations. We then demonstrate that experienced pilots faced with simulated scenarios closely recreating real accidents benefit from such a system. The effects are remarkable: hardly any pilots would have survived in the control group, whereas almost all survived among those that used our proactive accident-aware agent. The control group was given access to a state-of-the-art audio LLM with an extensive aviation RAG, which even led directly to two crashes in our experiment, the first such result.

1. Introduction

Aviation depends on complex human-machine cooperation enabled by user interfaces that convey crucial information and manage cognitive load. Alone, these systems are clear, but in conjunction with multiple system failures, they can overwhelm even experienced pilots with cascading alerts, as observed in the Air France Flight 447 and Qantas Flight 32 accidents [1, 2]. These bototm-up systems focus on individual behaviors, often failing to highlight the root cause of problems, leading to cognitive overload in high-stress situations.

To address this, we introduce LISA (Lightweight Interaction and Storytelling Archive), a general-purpose, top-down alerting system that compares aircraft states to past accidents and provides concise, actionable alerts. We tested LISA against a control group using a general-purpose, voice-activated Large Language Model (LLM) with Retrieval-Augmented Generation (RAG) tuned for aviation knowledge. Pilots with LISA avoided most crashes, while those using the RAG LLM overwhelmingly failed. The study highlights the critical importance of interface design: tailored, succinct interventions significantly outperform general-purpose AI when

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
*Corresponding author: cmarnold@mit.edu.
†These authors contributed equally.
cmarnold@mit.edu (C. Arnold); paulr@dollabs.com (P. Robertson); zoe.robertson@dollabs.com (Z. Robertson); rladdaga@dollabs.com (R. Laddaga); boris@mit.edu (B. Katz); abarbu@mit.edu (A. Barbu)


responding to time-critical environments, such as aviation.

Additionally, we detail an experimental protocol for simulating aviation accidents, allowing future research on intelligent interface development in aviation. The findings caution against deploying generic AI systems without rigorous testing, as seemingly some AI behaviors, such as hallucinations, can have catastrophic consequences in critical domains like aviation. This work underscores the need for domain-specific solutions to improve safety and decision-making.

Our contributions are:

1. A new aviation interface with an intelligent agent trained on past accidents for the cockpit, LISA (the Lightweight Interaction and Storytelling Archive).
2. A demonstration of a naive, but plausible, implementation of a RAG LLM interface for aviation and its pitfalls.
3. A large-scale simulation framework for evaluating similar interfaces.
4. A demonstration that pilots remain susceptible to past accident conditions.
5. A description of components in interfaces that improve aviation safety and possibly other high-stress environments.
6. The first report that LLM hallucinations in high-stress environments can result in simulated fatal aviation accidents.

2. Related Work

HFACS. The Human Factors Analysis and Classification System (HFACS) is a framework used to analyze human error in aviation and other high-risk industries. Derived from James Reason's "Swiss Cheese Model"[3], HFACS categorizes errors into four levels: unsafe acts, preconditions for unsafe acts, unsafe supervision, and organizational influences. Developed in 2001 for the U.S. Navy and Marine Corps, HFACS has since been applied in civil aviation to investigate accidents[4]. Studies show that 60-80% of aviation accidents are caused by human error, often due to supervisory or organizational failures rather than individual mistakes[5]. With the rise of automation and AI in high-risk industries, HFACS is now being used to analyze human-machine interfaces and address errors associated with autonomous systems[6].

Accident Simulation. Simulated flight environments are vital for aviation safety investigations, enabling the analysis of pilot performance, decision-making, and accident hypotheses. They are often used in NTSB reports to determine the root cause of accidents[7]. Notable examples include US Airways Flight 1549, or the "Miracle on the Hudson"[8], and Air France Flight 447[1], where simulations revealed pilot confusion following autopilot disconnection. Likewise, analysis of Colgan Air Flight 3407 in 2009 led to improved stall recovery training[9]. These cases highlight the role of simulations in understanding human factors, such as workload, stress, and decision-making during emergencies, and their impact on aviation safety advancements.

3. Experimental Setup

All tests were performed in the same physical location. Pilots are first taken to the planning room where they are given a short description of the scenario (type of flight, aircraft, take-off and destination airport) and the weather conditions. Pilots have an unlimited amount of time to

plan their intended flight for the simulated scenario. They have access to ForeFlight, a popular flight planning tool, and are permitted to use a web browser to search for any information they deem useful. They are also given access to paper planning charts to plot their route if they wish to use a more traditional form of flight planning. ForeFlight comes pre-loaded with the Pilot Operating Handbook (POH) for the relevant aircraft, a Beech G36. The POH contains information regarding aircraft performance and limitations. Such planning before a flight is routine and each pilot has their own procedure for doing so. Our goal was to facilitate this process as a pilot would normally engage with it.

After the pilot finalizes the flight planning process, they proceed to the simulator room (shown in fig. 1) to commence the flight.

We simulated our flights with commodity hardware and Microsoft Flight Simulator 2020. This simulator is very popular with Flight Sim enthusiasts for its aircraft realism and expansive world, and is considered one of the best in-home flight simulators available.

Pilots interact with the simulator through the controls on the desk in front of them. A Honeycomb Alpha Yoke provides control input for pitch and roll axes of the aircraft, but also has added functions for aircraft pitch-trim, lights, and engine starter. To the right of the yoke is the Logitech Throttle Quadrant with three levers. From right to left they are the air-fuel mixture setting, the propeller RPM, and the engine throttle. Below the levers are buttons to raise and lower the aircraft landing gear. Below the yoke are Logitech Rudder Pedals, largely occluded by the chair in fig. 1, which control the yaw axis of the aircraft as well as the wheel-brakes.

Figure 1: The flight room is where participants are tested under each scenario in MSFS. Each primary control surface has its own dedicated external device that mimics the tactile feel of real-world flight controls.

On the left is a RealSimGear G1000 panel, that displays various instruments from the cockpit on a cohesive display. It simulates a real G1000 glass cockpit as found in many high-end general aviation and some commercial aircraft. Pilots were instructed not to use the autopilot during testing, as it was not relevant to either of the scenarios we tested. The pilots wore push-to-talk headphones to communicate with or receive aural alerts from the assistant.

A second identical station to the right is also shown, but all experiments took place with a single pilot at the controls, so this station is not used.

3.1. Baseline LLM with Aviation RAG

The control baseline assistant is a state-of-the-art LLM, Claude 3 Opus, extended with an aviation RAG that contained manuals of the relevant aircraft, videos about how to fly correctly collected from YouTube (which is a rich resource of information), and information about relevant runways and airspace classifications from published aviation documents. The RAG ensures that the same documents that were ingested into LISA were also available to the baseline. Claude was provided with a custom system prompt describing the aircraft being flown and the intended departure and arrival airports. Participants could interact with the baseline either via text in a

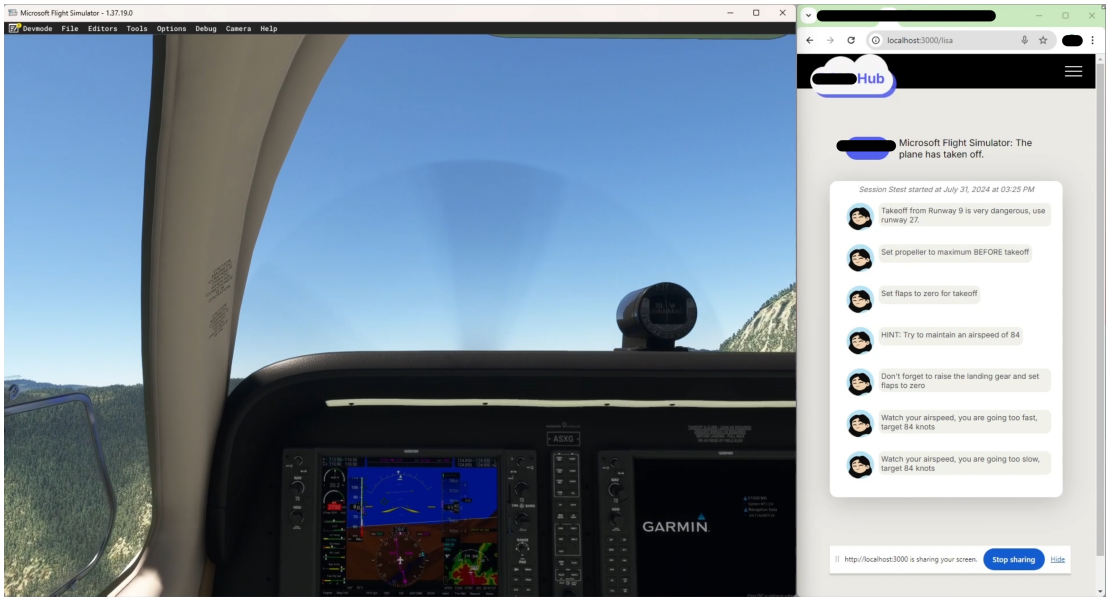

Figure 2: LISA outputs are presented on the right of the screen while the participant flies in the simulator. First LISA warns of the dangers of taking off from runway 9, then reminds the pilot to set the propeller and flaps before takeoff. Once off the ground, LISA states an ideal speed for climb, reminds the pilot to raise the landing gear, then provides time-critical alerts related to airspeed.

prompt window next to the flight simulator, or via a speech-to-text interface with Whisper [10]. Our instructions encouraged participants to use the system, but they were not required to use the system or listen to its recommendations.

We chose this baseline for two reasons. First, such systems are trivially buildable today and invariably, they will be proposed for cockpit operations. Knowing if they are effective as chat-based agents in avoiding accidents is critical. Second, without a baseline system, the experiment would have been unblinded because the participants would have been aware that they were testing an AI system for aviation.

3.2. LISA — Our Assistant

We developed LISA to be a proactive agent that interprets aircraft systems in a wholistic, top-down approach, from which it actively provides recommendations to the pilot for safe operation. LISA provides information that the pilot needs to know based on the state of the flight and the airplane rather than by responding to questions from the pilot. This is unlike the baseline and other chat agents which are reactive, and are driven by user interactions. LISA provides text and aural warnings to the pilot. To simplify the experimental variables, we did not provide a blend of chat-based reactive systems and a proactive LISA-based system, which actively provides suggestions to the pilot based on current conditions. Although, as will be seen later, this would not have made a difference: LISA alone avoided almost all accidents.

LISA is connected to the flight simulator through a custom plugin. It has access to the basic flight parameters (location, airspeed, attitude, bank angle, heading, fuel gauges, engine RPM, and oil pressure), just like the pilot does, with no privileged access to systems the pilot might not be able to know.

Internally, LISA reasons by reference to prior accident reports from the NTSB and near-

accident reports from ASRS, using a combination of rules-based logic and machine learning. The visual presentation of LISA and the baseline LLM are identical, and both use Whisper to speak to users as well as write responses. Figure 2 shows several alerts from LISA for one participant during scenario 2.

4. Flight Scenarios

We tested two scenarios mirroring two real accidents which took place under similar conditions in mountainous terrain [11, 12]. The conditions of our scenarios were designed specifically to mirror these two accidents, as well as other aviation accidents in mountainous conditions. For simplicity, both our scenarios take place at the Telluride Regional Airport (KTEX). To ensure that the scenarios are sufficiently difficult, we set the Outside Air Temperature (OAT) to 39°C. This corresponds to a density altitude of 14,065 feet. We loaded the aircraft to its maximum legal take-off weight, which places the aircraft at the limits of its performance envelope for take-off. We configured a 10 knot wind from the east to encourage pilots to take off from Runway 09, a very dangerous and fateful decision, but one that is easy to overlook. Runway 27 is the safer option, despite the unfavorable winds.

Our volunteers are asked to play the role of a commercial pilot. This provides a level of urgency to the flight. However, if a pilot was uncomfortable with any part of a flight, refusing to fly would be a natural and safe choice. In training, pilots are encouraged to do so when conditions exceed their personal comfort level. We only report the results of pilots who opted to fly the scenario despite the conditions.

4.1. Scenario 1

To pass Scenario 1, pilots encounter an engine failure on take-off. In the simulator, we engineered a partial failure of the engine at the beginning of the scenario. Given the environmental conditions, this means that a climb after take-off becomes nearly impossible even for the most skilled pilot. This failure is observable from the pilot's perspective, but only by observing the reduction in power by way of the tachometer. Under normal operations, the G36 indicates 2,650 to 2,700 RPMs on take-off; however, in this case, the engine only indicates 2,500 RPMs.

A participant is considered to have failed the scenario if the pilot attempts take-off.

4.2. Scenario 2

Immediately following Scenario 1, the simulator is reset. This time, the aircraft is mechanically sound throughout the scenario. The aircraft is placed on the departure end of Runway 09 at KTEX. Despite the fact that the winds favor taking off from Runway 09, this is far more dangerous than departing from the opposite direction, Runway 27, due to obstacles, a 0.4% upslope, and the presence of a box canyon immediately after departure from Runway 09.

Next, the pilots must overcome their unfamiliarity with the aircraft. Pilots must utilize the 419 page Pilot Operating Handbook (POH) in order to determine that the flaps should be set to UP for take-off in the simulated conditions. This aircraft procedure is counterintuitive to what most pilots learn in commonly available training aircraft.

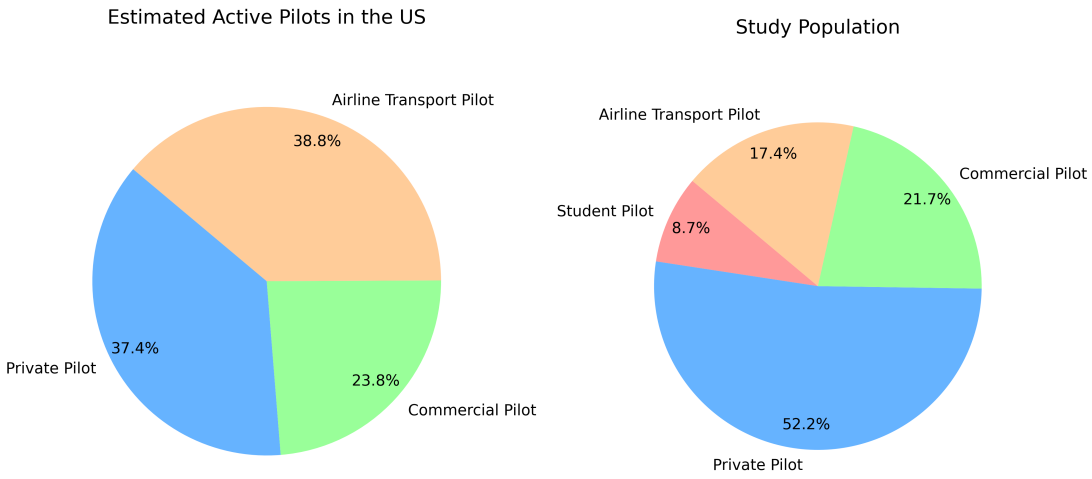

Figure 3: Distribution of pilot certificates by type for both the US population and for this paper's study. Private pilots are over-represented and Airline pilots are under-represented by the study volunteers. Student pilots are intentionally omitted from the first chart for greater ease in visual comparison.

If the pilot correctly chooses to take off on Runway 27 and set the flaps to UP, they still need to contend with the reduced performance due to the very high density altitude. A theoretical 500 foot per minute climb is possible, but with the addition of turbulence and imperfect control inputs from pilots, it is very easy for a pilot to find themselves momentarily unable to climb. Close to the ground, especially right after take-off, this usually triggers an instinctive, but often fatal, reaction: the pilot pulls back on the control yoke in an effort to climb. This reduces airspeed, further reducing the rate of climb. Unchecked, this intuitive, but wrong, reaction leads to a feedback loop and an eventual stall followed by impact with terrain.

Once the pilot leaves the runway environment and has climbed their first 1,000 feet of altitude, flying becomes psychologically easier as the ground is further away. However, the participant has one final challenge to overcome: climbing over mountainous terrain that rises more than 14,000 feet to the east. Due to the extreme temperatures combined with the aircraft capabilities, it is only capable of maintaining a mediocre 300-400 foot per minute climb.

For the purposes of the experiment, the scenario ends when the pilot reaches 14,500 feet, which is high enough to clear the surrounding terrain.

5. Experiments

The experiment itself is a single-blind study. Upon arrival at the testing facility, participants are informed that two systems have been developed and they have been randomly assigned to one of those systems. They are told that both systems use some form of automated assistance and that they will be evaluating how well that system helped them in their flight.

As part of the recruitment process, all participants were informed that they would participate in simulated flights with an AI assistant in scenarios based on real aviation accidents. However, the solution to those scenarios is trivial if presented with the right information.

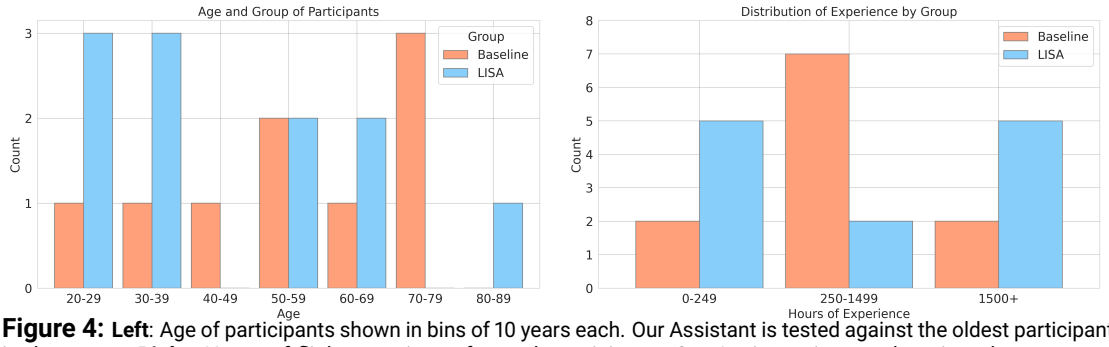

Figure 4: **Left**: Age of participants shown in bins of 10 years each. Our Assistant is tested against the oldest participant in the group. **Right**: Hours of flight experience for each participant. Our Assistant is tested against the two most inexperienced pilots in the group.

5.1. Participants

Participants were recruited from the local aviation community, with many pilots eager to contribute to aviation safety research. The ease of recruitment suggests potential for future studies. Volunteers were randomly assigned to either the control baseline or LISA groups, with assignments finalized on the evaluation day to accommodate scheduling flexibility.

The U.S. has over 800,000 active pilots, categorized into Student, Private, Commercial, and Airline Transportation Pilots, with 39% being student pilots[13]. Our recruitment focused on pilots with at least a Private Pilot certificate, though students were not excluded.

From an estimated 7,000 pilots in the study's metropolitan area, 23 volunteered. The study's demographic and the broader pilot population are detailed in fig. 3. Women, who comprise less than 7% of non-student pilots, were overrepresented with three female volunteers participating in our study. The average age of U.S. pilots is 42.8, while our volunteers averaged 49.85 years old (fig. 4). There is no age limit for pilots in the US.

Requirements for the three major licenses vary based on demonstrated flying aptitude, knowledge, and raw flight time. A pilot must obtain a minimum amount of flight time to test for the next license; 40 hours for Private, 250 hours for Commercial, and 1,500 hours for Airline. Thus, license is held as a common proxy for pilot experience and is shown in fig. 4.

Although the baseline and LISA groups differed slightly in age and experience, these differences were not statistically significant in affecting the study results.

5.2. Participant Instructions

Twenty-four hours prior to the evaluation, participants are informed that they will be flying a Beech G36 in a simulator. At the testing facility, participants are granted the opportunity to initially fly the simulator for ten minutes in a Cessna 172 departing from a local major airport that is near sea level. This time allows pilots to familiarize themselves with the physical controls, locations of the buttons, and how the simulator responds to inputs. Participants are allowed to do anything they wish with this time, with no stated goals or objectives provided for them.

After practice flights, each participant goes to the planning room where they are given a simple instruction: *You are a commercial pilot for Mountain Air Cargo flying a Beech G36 at max gross weight from Telluride Airport (KTEX) to Mineral County Airport (C24). Your goal is to complete the flight safely.* This instruction is accompanied by the following Meteorological Aerodrome Report (METAR), a cryptic text containing the weather report that would only be interpretable by a trained pilot: `METAR KTEX [Current Date and Time]Z AUTO`

`09010KT 10SM CLR A2992 RMK AO2`. As described above, participants are allowed to use any resource that they wish for flight planning. This also includes an off-the-shelf unmodified Claude. Participants are allotted an unlimited amount of time for flight planning.

5.3. Simulated Flights

Once flight planning is complete, participants return to the flight simulator for Scenario 1. They may only use one attempt to pass Scenario 1, and the only way to pass Scenario 1 successfully is to reject take-off due to an engine malfunction.

Immediately following the conclusion of Scenario 1, the proctors inform the participant that Scenario 1 involves an engine failure, and the correct answer was to reject take-off.

Then participants start Scenario 2. They are permitted up to two attempts to pass Scenario 2. This is done to allow for additional familiarization with the flight simulator and eliminate that as a source of failure for pilots. Flying in real life relies heavily on haptic feedback from the aircraft, commonly referred to as "Flying by the seat of your pants." Because there is very little sensation given to the pilot coupled with the loss of peripheral vision, additional experience with the simulator is necessary.

Once the participant either crashes twice or reaches 14,500 feet in altitude, Scenario 2 concludes, and the participant takes a short post-flight survey.

The scenarios are completed in this order because scenario 2 naturally builds on scenario 1. Although it was not considered, it would have been possible to perform scenario 2 prior to scenario 1, but we do not believe this order would have impacted the results of the experiment.

5.4. Survey and Feedback

The survey asks participants to rate the assistant's helpfulness using the Likert Scale, with 1 being "Not Very Helpful" and 5 being "Very Helpful". After this single question, they may write additional freeform comments. Participants are then unblinded and informed of the system they evaluated only after completing the survey. Some pilots deeply affected by their accidents participated in a detailed exit interview. Weeks later, they reflected on their experience by responding to all questions from the System Usability Scale (SUS)[14].

6. Results

The topline summary of each accident scenario is summarized in fig. 5. No participants using the baseline succeeded in passing Scenario 1. In effect, all fell victim to some of the same blind spots that caused the original real-world accident. Every pilot took off, experienced a loss of control, and entered a deadly stall-spin, which resulted in impacting the ground at over 100 mph. This is despite having access to the baseline LLM. With our accident-aware system, LISA, almost all pilots (80%) avoided the pitfalls of the scenario and would not have crashed.

In Scenario 2, only 36% of participants using the baseline reached 14,500 feet and cleared the mountainous terrain, while the remaining 64% lost control and crashed shortly after takeoff. In contrast, all participants using LISA safely climbed above 14,500 feet. Moreover, they achieved this consistently safer and more efficiently, using 10.8% less time on average to reach the required height, demonstrating enhanced efficiency due to LISA.

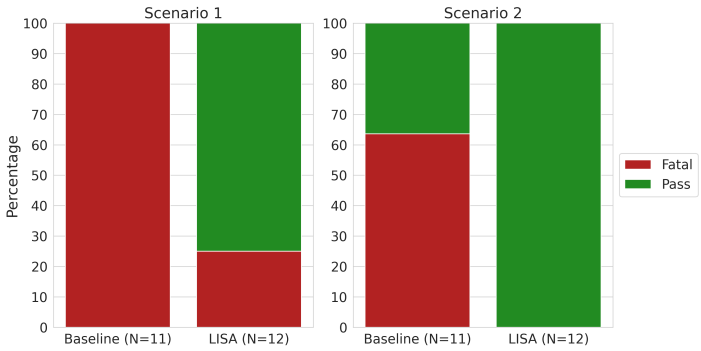

Figure 5: Results from each scenario divided by group. For each bar chart, Baseline participants are on the left and LISA participants on the right. All participants in the Baseline failed to identify the engine failure in Scenario 1, and only some were able to successfully fly the aircraft to 14,500 feet in Scenario 2.

Interestingly, nine out of the eleven baseline participants chose not to use the baseline LLM at all during scenario 2, which is an indication of their frustration and loss of trust in the system after using it in scenario 1. The two participants that used the baseline for scenario 2, received invalid information that reduced the performance of the aircraft and contributed to the pilot's loss of control and subsequent deadly impact with terrain. The LLM hallucinated the response to a question regarding the least-optimal flap setting for take-off. Because the vast majority of trainer aircraft exhibit improved take-off performance with the flaps slightly deployed, the LLM assumed that this applied to the Beech G36. This confirmed the incorrect assumption that most pilots had, which contributed to their accidents. In effect, the LLM made an already bad situation much worse: a pilot who had doubts had them put to rest by a hallucinating LLM which then resulted in a crash. This is the worst-case outcome for deploying such agents in cockpits, one that had been feared before by aviation regulators.

We found no statistically significant factor (age, experience, recency of experience) other than system used that explain the results of the experiments.

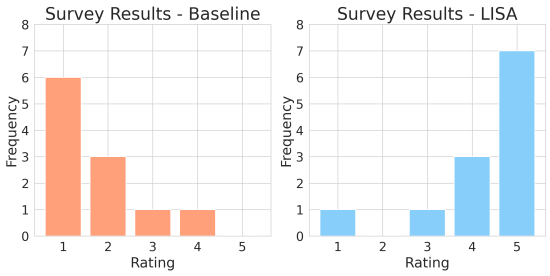

Figure 6: Participants were asked to answer the question "Were the responses or suggestions provided by the AI Assistant helpful?" using a Likert Scale of 1 being "Not Very Helpful" and 5 being "Very Helpful".

Pilots scored each system on a scale of 1 to 5 after each scenario. As expected, pilot opinions of the system were correlated with their own performance. LISA received much more favorable scores compared to the baseline; see fig. 6. Curiously, in the exit interview, outliers that rated the baseline either a 3 or 4 did so because they felt they didn't engage with it and thus could not accurately give it a score and because they did not want to offend the evaluation team. These effects explain why the experiment needs to be blinded. One subject rated LISA as 1, "Not Very Helpful". That subject, a new pilot, intentionally ignored advice from the agent and then in their comments blamed the agent for not preventing the accidents. Otherwise, the freeform feedback for LISA was overwhelmingly positive.

System Usability Scale (SUS) [14] questions were asked after the experiment was unblinded, but their scores aligned well with the previous single Likert scale helpfulness question. This is a well-known effect where even binary net promoter scores, NPS, correlate highly with SUS. We found a strong positive correlation, 0.81, between the single helpfulness question and the ten-question SUS survey.

7. Conclusion

We find that the thoughtful implementation of intelligent user interfaces is incredibly important to their effectiveness and impacts on safety in high-stress environments like those found in aviation. Based on the results of our study, and from the solicited feedback of participants, we find that good interface for these environments is one that is succinct, accurate, and timely. Many users of the baseline simply stopped using the LLM after a few interactions, and many became frustrated with its verbosity, which is a common feature of modern LLMs today.

We also find that alerting systems that aggregate and present information in a top-down systems-focus approach are superior when timely decisions are the difference between life and death. Current bottom-up, systems-focused approaches that rely on users to interpret the root cause of multiple, sometimes seemingly disjointed system failures can cause cognitive overload. Once cognitive overload is encountered, humans fall back on instinctive responses that are often catastrophically wrong for environments for which humans are not adapted.

Finally, we provide a word of caution: just because a system is knowledgeable, appears to be helpful, and is even often helpful, does not mean that it should be deployed in high-stress situations. Our results show that providing the wrong answer to a question or offering the wrong information can lead to fatal outcomes. Systems like LLMs are inherently susceptible to hallucinations, and no current approach can prevent this. Their deployment is not a neutral decision with only upside; it is a deliberate choice to trust these systems that can have disastrous downsides. We hope that this can inform policymakers as the integration between humans and intelligent agents progresses.

8. Acknowledgements

This work was supported by the Center for Brains, Minds, and Machines, NSF STC award CCF-1231216, the NSF award 2124052, the MIT CSAIL Machine Learning Applications Initiative, the MIT-IBM Watson AI Lab, the CBMM-Siemens Graduate Fellowship, the DARPA Mathematics for the DIscovery of ALgorithms and Architectures (DIAL) program, the DARPA Knowledge Management at Scale and Speed (KMASS) program, the DARPA Machine Common Sense (MCS) program, the Air Force Office of Scientific Research (AFOSR) under award number FA9550-21-1-0399, the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, and the Developing the Airmen We Need–Education (DAWN-ED) PhD fellowship.

9. Disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force, the United States Space Force, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Christian Arnold is an active duty officer of the United States Space Force on a Department of the Air Force funded fellowship.

References

[1] Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile, Final Report On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris, Technical Report, Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile, 2009.

[2] Australian Transport Safety Bureau, In-flight uncontained engine failure Airbus A380-842, VH-OQA, overhead Batam Island, Indonesia, on 4 November 2010, Technical Report, Australian Transport Safety Bureau, 2013.

[3] J. Reason, Human error, Cambridge university press, 1990.

[4] D. A. Wiegmann, S. A. Shappell, A human error approach to aviation accident analysis: The human factors analysis and classification system, Routledge, 2017.

[5] S. Shappell, C. Detwiler, K. Holcomb, C. Hackworth, A. Boquet, D. Wiegmann, M. Cilnic, Human error and commercial aviation accidents: A comprehensive, Fine-Grained Analysis Using HFACS (2006).

[6] M. Chignell, L. Wang, A. Zare, J. Li, The evolution of hci and human factors: Integrating human and artificial intelligence, ACM Transactions on Computer-Human Interaction 30 (2023) 1–30.

[7] B. Elias, Flight simulation as an investigative tool for understanding human factors in aviation accidents, in: 2005 International Symposium on Aviation Psychology, 2005.

[8] National Transportation Safety Board, Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River US Airways Flight 1549, Technical Report, National Transportation Safety Board, 2009.

[9] National Transportation Safety Board, Loss of Control on Approach Colgan Air, Inc. Operating as Continental Connection Flight 3407, Technical Report, National Transportation Safety Board, 2009.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.

[11] National Transportation Safety Board, Aircraft Accident Report: Beech S35, N4444K, Telluride, CO, Technical Report, National Transportation Safety Board, 2024.

[12] National Transportation Safety Board, Aircraft Accident Report: Beech A24, N8020R, Keene, NH, Technical Report, National Transportation Safety Board, 2024.

[13] Federal Aviation Administration, U.s. civil airmen statistics, 2024.

[14] J. Brooke, SUS – a quick and dirty usability scale, Taylor and Francis, 1996.