ChatGPT and Gemini in the Tower: Generative AI's Role in Routine and Emergency ATC Communication

Billava Vasantha Monisha
Student ID: 758413
Masters in Automotive Software Engineering
billava-vasantha-monisha@s2022.tu-chemnitz.de

***Abstract*—The air traffic controllers (ATCOs) play a crucial role in the safe, orderly, and expeditious flow of air traffic in the global aviation system. This is indeed one of the most critical professions that requires precision, quick decision-making, and adaptability because even a small mistake can lead to disastrous situations. While Generative AI has been swiftly improving in many fields, someone might question how fully capable the generative AI models will ever be to take the place of human ATCOs in all eventualities-ranging from routine operations to emergency situations. Empirical research design is the approach followed in this paper, combining qualitative and quantitative features to explore the question above. To evaluate, two generative AI models have been selected as leading ones: ChatGPT and Gemini. Real-world air traffic control scenarios and conversations at routine/simple and emergency/complex levels were gathered and preprocessed. These scenarios were replayed with the AI models in order to analyze their performance with regard to giving responses on target and effectively. In this regard, the structured questionnaire targeted pilots and certified ATCOs in a methodical manner to validate the authenticity, accuracy, and adequacy of the AI responses. Findings will be analyzed in search of proof that Generative AI is performant at ATC tasks and what possible limitations exist when working in safety-critical contexts. This study will thus provide insight into how far AI can go in integrating with air traffic control systems and where human expertise remains indispensable.**

***Index Terms*—Generative AI, Air Traffic Controllers, ChatGPT, Gemini, Routine Scenarios , Emergency Scenarios**

I. Introduction

Aviation is a very complex and safety-critical environment where clear communication and rapid decision-making are required. It is the job of air traffic controllers to manage this complexity for safe and efficient aircraft movements. However, growing air traffic volumes, human workload, and the possibility of communication errors have developed an urgent need for innovative technological solutions that would enhance operational efficiency and safety. Generative AI systems are the latest promising frontier of AI, whereby human-like communications can be emulated-for example, ChatGPT and Gemini. These developments spur research into possible applications within air traffic control (ATC) scenarios.

Applications based on Generative AI have already shown promising usage in customer support and real-time communication, among other industries [1]. The same is bound to help ATCOs do their routine and emergency communication jobs much more swiftly in aviation and reduce the workload on humans. International Air Transport Association, or IATA in short, sees this upcoming technology as having immense potential for making operations better in all aspects of flying an aircraft, addressing concerns of inclusivity and technological accessibility [1].

Despite these advances, there are still significant concerns about the deployment of generative AI in safety-critical systems because of latent errors, a lack of contextual awareness, and unintended behaviors that may pose risks in situations involving human lives [2]. Ebeni's work emphasizes how much more stringently safety needs to be implemented and how designs should be improved to meet the challenges in aviation, among other critical systems [2]. Furthermore, research into generative AI chatbots has shown their prowess in answering routine questions and increasing efficiency, but they also have their shortcomings when it comes to complex or ambiguous situations [3].

This research will seek to empirically test the capabilities of generative AI models, ChatGPT and Gemini, in mimicking human responses in ATC scenarios. Focusing on both simple and critical scenarios, the research will try to establish:

- Participants' ability to distinguish between AI-generated and human-generated ATC responses.
- The performance of ChatGPT and Gemini based on clarity, accuracy, and effectiveness metrics.

The outcomes of this research will contribute to the broader understanding of generative AI's role in safety-critical communication systems. By identifying strengths, limitations, and potential risks, this study aims to provide actionable insights for the aviation industry as it explores the integration of AI into its operations.

II. Study Design

A. Research Goals

The goal of the research are to investigate the capabilities of generative AI models, such as ChatGPT and Gemini, to emulate human air traffic controllers in safety-critical communication tasks. The main objectives are:

- Performance Evaluation of Generative AI Models: Investigate realistic/authentic, accurate, and safety-compliant ATC conversations generated by AI models in routine/simple and complex/emergency situations.
- Contrast Communication by AI and Humans: Assess whether AI-driven conversations differ from human-generated conversations based on the perceived sense of authenticity and accuracy among participants.
- Feasibility Testing of Partial Automation: Assess the suitability of Generative AI models for partial automation in handling the air traffic control task, particularly for routine situations, by identifying their limitations in complex situations.

The present research points to the strengths and weaknesses that generative AI exhibits when integrated into safety-critical systems and forms a blueprint for future directions toward the use of AI in air traffic control processes.

Research Question: Can generative AI models (e.g.: ChatGPT and Gemini) replace safety critical systems like the conversation between the Air traffic Controllers and the pilots?

This research focuses on the below variables:

Independent Variables: Source of conversations- Generative AI (Gemini and ChatGPT) vs Humans.

Dependent Variables:

- Authenticity: How realistic is the conversation.
- Accuracy in routine and complex situations: Weather the information provided aligns with aviation safety standard.
- Distinguishability: The ability to distinguish between the AI generated and the human conversations

B. Hypotheses

In the research of Generative AI replacing the ATCOs, the hypotheses and the null hypotheses along with their alternatives would be:

H_0: Generative AI models like ChatGPT and Gemini can produce realistic air traffic control conversations in routine/simple scenarios, but they struggle in complex/emergency scenarios, this limits in replacing the human air traffic controllers.

H_1: Generative AI models struggle in complex air traffic control scenarios, they are considerably better than human controllers in routine situations, indicating partial automation of ATCOs tasks.

N_0: Generative AI models like ChatGPT and Gemini do not outperform human air traffic controllers in either routine/simple scenarios or complex/emergency scenarios, indicating that they cannot replace human air traffic controllers.

N_1: Generative AI models do not perform better than human controllers in routine situations, nor do they show significant differences in handling complex air traffic control scenarios, indicating no potential for partial automation of ATCOs tasks.

C. Experiment Planning

In this section, how the experiment is planned to conclude on the hypothesis will be discussed. Both qualitative and quantitative approach will be carried out. The dataset would be a collection of conversations between ATCOs and pilots from various platforms. Data pre-processing involves classification of the data into two kinds of scenarios: simple scenarios, such as standard and routine communications in landing or takeoff clearances; and complex scenarios, representing any kind of emergency or safety-critical situations where engine failure, runway incursion, or bad weather conditions may occur. Specific plots and situations are extracted from these datasets to serve as inputs for the evaluation process.
These extracted scenarios are then used to present the Generative AI models, Gemini and ChatGPT, with carefully crafted prompts to generate the matching ATC conversations. Their AI-generated conversations, along with human-generated ones from the datasets, are included for blind assessment. In this testing, targeted audience participants are given conversations, in a randomized and blinded method, such that the participant does not know whether a conversation is AI or human-generated. Participants are tasked with rating the conversations on key criteria. They are also asked to state the source of each conversation: AI or human.
From these results, the hypothesis of the research will be reviewed for validity, and conclusions on the future of generative AI replacing or supporting human air traffic controllers in safety-critical tasks will be outlined. The figure below shows the steps involved in this experiment.

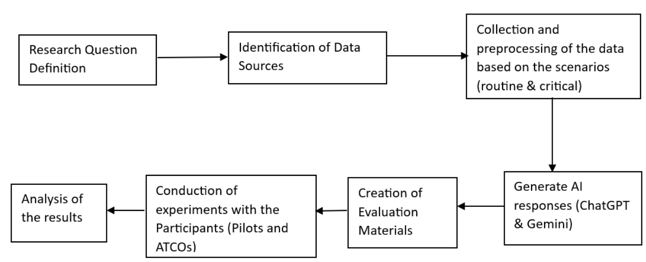

Fig. 1. Expected steps of the Empirical research. Source: By Author, produced by Microsoft word

D. Participants

The evaluation is conducted with the participation of:

- **Certified Air Traffic Controllers:** Professionals with hands-on experience in managing air traffic control systems and pilot communications.
- **Pilots:** Experienced aviators accustomed to ATC interactions during routine and emergency scenarios.

The participants will be briefed about the main objectives of the experiment and asked to complete a survey questionnaire. The experiment will involve a blind test, where participants will review conversations and attempt to identify the source (human or AI) based solely on the prompts. Participants will also evaluate each conversation on key metrics—authenticity (How realistic is the communication?), accuracy (Did the response follow ATC protocols?) , and Distinguishability (Can participants identify whether the source is AI or human?) —using Likert scales as part of the quantitative approach. These ratings will provide measurable and comparable data to assess the performance of AI models (ChatGPT and Gemini) against human-generated conversations.

As part of the qualitative approach, participants will be asked to provide open-ended feedback explaining why they rated the conversations a certain way. This feedback will offer deeper insights into the strengths and weaknesses of the AI models, particularly in handling complex scenarios.

E. Material Creation

For this research, the following materials and data sources were utilized:

Data Collection:

- Transcripts of live air traffic control (ATC) conversations were collected from publicly available YouTube channels such as ATC AUDIO [6] and You Can See ATC [7].

Official ATC communications were sourced from credible organizations, including:

- The National Transportation Safety Board (NTSB) [8],
- The Air Traffic Control Complete Corpus (LDC94S14A) [9],
- The Air Safety Institute [10].

Scenario Categorization:

The collected dataset was analyzed to identify and categorize conversations into two groups: Simple Scenarios: Routine operations such as landings, takeoffs, and weather updates. Critical Scenarios: High-pressure situations such as emergencies, priority landings, and communication failures. Each category was curated to include a diverse set of conversations to evaluate AI performance under different levels of complexity.

Generative AI Training:

Selected conversations from the categorized dataset were used to pre-train two generative AI systems: ChatGPT (Version GPT-4) and Gemini (Version 2.0) The training involved providing contextual and scenario-specific details to ensure the AI models could generate realistic and contextually appropriate responses. After training, both models were used to generate responses for the test scenarios. These responses were evaluated by participants during the experiment.

Evaluation Tools:

The evaluation of the collected data will be conducted using Microsoft Excel. This tool was utilized for:

- Ensuring completeness and accuracy by removing invalid or incomplete responses from participants.
- Structuring responses into categories based on the scenarios, ratings, and participant feedback for easier analysis.
- Generating statistical insights, such as average ratings for clarity, accuracy, and effectiveness. Creating charts and graphs to visualize participant responses and identify trends.

Microsoft Excel is chosen for its versatility in handling quantitative data and its ability to generate clear visual representations of the results.

F. Questionnaire

The questionnaire is designed to assess the performance of AI-generated air traffic control communications compared to human responses. It has the following parts:

- **Introduction and Study Objective:**An introduction page indicating the purpose of the study and objectives of the same, along with clear instructions on how to fill up the Google Form.
- **Participant Background:**Roles (for example, pilot, ATC professional) and years of experience in the aviation field are collected to place the results in context.
- **Scenario Identification:**The participants will go through two simple and two critical scenarios, each with one original and one AI-generated response. They will identify which response is AI-generated and give a reason for their decision. Ratings are not required in this section.
- **AI Performance Evaluation:** For the same scenarios, responses from two AI systems (for example, ChatGPT and Gemini) will be provided. The participants will rate each response on clarity, accuracy, and effectiveness using a scale.

After each case, participants will explain their reasoning for rating the answers in such a manner. This method ensures that thorough and detailed analysis is undertaken for participants' ability to identify human and AI answers as well as their appraisal of AI-generated communication concerning performance metrics.

G. Task

In this qualitative research, comprehension tasks were designed and presented using Google Forms, accessible at Survey Link. The survey is divided into two distinct sections to evaluate participants' ability to differentiate between human and AI-generated air traffic control (ATC) responses and to assess the performance of generative AI systems.

Scenario Identification (Scenarios 1–4)

Participants were provided with responses for four different ATC scenarios and were tasked to carefully read the responses and identify whether they were generated by AI or a human controller. This section was designed to test the participants' ability to distinguish between human and AI responses and to gather qualitative feedback explaining their choices.

Below is an example of the questionnaire presented under simple and complex scenarios, where the participants had to choose which response was AI-generated.

Simple Scenario:: **Response 1:**

Pilot: Lancaster Ground, Cherokee 8121K, west ramp, VFR, 4,500 to Frederick with Sierra.

ATCO: Cherokee 8121K, Lancaster Ground, taxi to Runway 26.

Pilot: Taxi to Runway 26, Cherokee 8121K.

Pilot: Lancaster Tower, Cherokee 8121K, Runway 26, ready for takeoff.
ATCO: Cherokee 8121K, Runway 26, cleared for takeoff.
Pilot: Cleared for takeoff Runway 26, Cherokee 8121K.

Response 2:
Pilot: Lancaster Ground, Cherokee 8121K, west ramp, VFR, 4,500 to Frederick with Sierra.
ATCO: Cherokee 8121K, Lancaster Ground, taxi to Runway 26 via Taxiway Alpha, the runway is clear.
Pilot: Taxi to Runway 26, Cherokee 8121K.

Pilot: Lancaster Tower, Cherokee 8121K, Runway 26, ready for takeoff.
ATCO: Cherokee 8121K, Runway 26, cleared for takeoff.
Pilot: Cleared for takeoff Runway 26, Cherokee 8121K.

Complex Scenario:: **Response 1:**
Pilot: MAYDAY MAYDAY MAYDAY, Tallinn Tower, Regional Jet 347, Embraer 195 engine failure, continuing SID, passing TN 850, 2000 ft, climbing to altitude 4000 ft, endurance 2 ½ hrs, 112 POB.
ATCO: (MAYDAY) Regional Jet 347, Tallinn Tower, roger MAYDAY.

Pilot: MAYDAY MAYDAY MAYDAY, Kastrup Tower, Big-prop 1233, ATR72, right-hand engine fire, continuing approach Runway 22L, after landing stopping on the runway, 5 miles final, altitude 1600 ft descending, 64 POB.
ATCO: (MAYDAY) Big-prop 1233, Kastrup Tower, roger MAYDAY, runway 22L cleared to land, wind 260 degrees 10 knots, emergency services alerted.

Response 2:
Pilot: MAYDAY MAYDAY MAYDAY, Tallinn Tower, Regional Jet 347, Embraer 195 engine failure, continuing SID, passing TN 850, 2000 ft, climbing to altitude 4000 ft, endurance 2 ½ hrs, 112 POB.
ATCO: (MAYDAY) Regional Jet 347, Tallinn Tower, roger MAYDAY. Maintain 4000 ft. Turn left heading 310 to ensure safe routing. Advise intentions.

Pilot: MAYDAY MAYDAY MAYDAY, Kastrup Tower, Big-prop 1233, ATR72, right-hand engine fire, continuing approach Runway 22L, after landing stopping on the runway, 5 miles final, altitude 1600 ft descending, 64 POB.
ATCO: (MAYDAY) Big-prop 1233, Kastrup Tower, roger MAYDAY. Runway 22L cleared to land. Winds 260 degrees at 10 knots. Emergency services alerted and standing by. Report when on short final.

AI Performance Evaluation (Scenarios 5–7)

For the same scenarios, responses from two AI systems (ChatGPT and Gemini) were provided. Participants were instructed to evaluate and compare the responses based on three metrics: clarity, accuracy, and effectiveness, using a provided rating scale.

Below is an example of the questionnaire used for evaluation:

Scenario: Bird Strike and Go-Around: **Response from ChatGPT:**
Pilot: MAYDAY, MAYDAY, MAYDAY. [Station Name], [Flight Number]. Bird strike, bird strike, go-around.
ATCO: [Flight Number], roger MAYDAY. Confirm intentions.
Pilot: Requesting permission to land on Runway 19 from the opposite direction. [Flight Number].
ATCO: [Flight Number], roger. Stand by for clearance.
ATCO: [Flight Number], cleared to approach Runway 19. Winds 190 degrees at 8 knots. Emergency services will be on standby. Report established on final.
Pilot: Cleared for approach Runway 19. Will report when established on final. [Flight Number].

Response from Gemini:
Pilot: Mayday Mayday Mayday! [Flight Number], bird strike, bird strike, go-around!
ATC: [Flight Number], [Station Name], say again, nature of emergency?
Pilot: Bird strike, severe engine damage, requesting immediate go-around!
ATC: [Flight Number], [Station Name], go-around approved. Squawk 7700. Maintain visual separation from other traffic. Vectors for Runway 19 approach will follow.
Pilot: [Flight Number], go-around initiated, requesting immediate vectors for Runway 19 approach.
ATC: [Flight Number], [Station Name], vectors for Runway 19 approach. Turn left heading 270 degrees. Maintain 1500 feet. Expect further vectors.

Participants rated each response on the following questions using a scale of 1 (Not Good) to 5 (Excellent): - How clear is the response? - How accurate is the response? - How effective is the response?

H. Study Design

In this empirical study, a qualitative method was employed to explore the ability of participants to identify and evaluate AI-generated responses in air traffic control (ATC) scenarios. The study was carefully designed to collect meaningful insights through a structured approach, as described below:

1) Data Collection: The study began with a participant background questionnaire to gain insights into their expertise and experience in their respective domains, particularly in aviation-related fields such as piloting and air traffic control. This helped establish a foundation for understanding the participants' familiarity with the tasks and scenarios presented.

Following the background survey, participants were provided with a series of comprehension tasks.

2) Data Cleaning: The survey collected 27 responses from participants with varying levels of expertise and experience. To ensure the integrity of the analysis, the data was carefully reviewed. No responses were excluded, as each response was considered valuable for understanding the performance of the AI systems and the participants' perspectives.

Although some participants provided partial responses, these were included to provide additional context to the analysis. This approach ensured a holistic representation of the collected data while acknowledging variations in participation levels.

3) Data Selection: Only complete responses from participants who fully engaged with the survey were considered for the final analysis. This included evaluating their ability to identify AI-generated responses and their ratings of AI systems' performance across clarity, accuracy, and effectiveness metrics.

The responses were categorized by the complexity of the scenarios (simple and critical) to ensure a comprehensive evaluation of the AI systems under varying task demands. This structured approach allowed us to derive meaningful insights into the capabilities and limitations of generative AI in replicating human communication in ATC scenarios.

4) Study Objectives: The primary objective of this study was to assess the effectiveness of generative AI systems, specifically ChatGPT and Gemini, in replicating human-like responses in ATC communication. Additionally, the study aimed to:

- Evaluate the participants' ability to distinguish between human and AI-generated responses.
- Analyze participant ratings of the AI systems in terms of clarity, accuracy, and effectiveness.
- Gain insights into the potential applications and limitations of generative AI in safety-critical domains like air traffic control.

By systematically collecting, cleaning, and analyzing the data, this study provides valuable insights into the capabilities of generative AI systems and their role in augmenting or potentially replacing human operators in complex and safety-critical environments.

III. Results

A. Participants and Expertise

The participants involved in this study came from diverse professional backgrounds in aviation-related fields. Their expertise primarily fell into two categories: pilots and ATC (Air Traffic Control) professionals. To better understand the participants' ability to assess air traffic control scenarios, it was essential to categorize them by their duration of experience in their respective fields.

Figure 2 shows the breakdown of the participants' expertise and experience. Among the respondents, 37% identified as pilots, while 63% were ATC professionals. Furthermore, the participants' years of experience in their fields varied significantly. A majority, 55.6%, had less than one year of experience, while 25.9% had 1–3 years of experience. A smaller proportion reported 3–5 years (11.1%) and over 5 years of experience (7.4%).

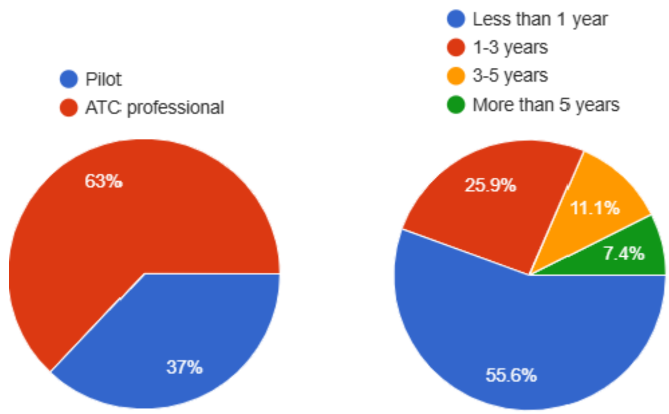

Fig. 2. Participants' professional expertise and experience in their respective industries.

B. Responses and Interpretation for Scenarios 1–4

In the first set of tasks, participants were presented with four scenarios, two classified as "simple" and two as "critical." Each scenario included one human-generated response and one AI-generated response. Participants were required to identify which response was human-generated.

The results, shown in Figure 3, reveal the following trends:

- **Scenario 1 (Simple):**
  - Response 1 (Human) was identified incorrectly by 44.4% of participants as AI generated.
  - Response 2 (AI) was correctly identified as AI by 48.1%.
  - 7.4% of participants were uncertain.
- **Scenario 2 (Critical):**
  - Response 1 (Human) was identified incorrectly by 51.9% of participants.
  - Response 2 (AI) was correctly identified as AI by 33.3%.
  - 14.8% of participants were uncertain.
- **Scenario 3 (Simple):**
  - Response 1 (AI) and Response 2 (Human) each received 40.7% of responses.
  - 18.5% of participants were uncertain.
- **Scenario 4 (Critical):**
  - Response 1 (Human) was identified incorrectly by 63% of participants.
  - Response 2 (AI) was correctly identified as AI by 14.8%.
  - 22.2% of participants were uncertain.

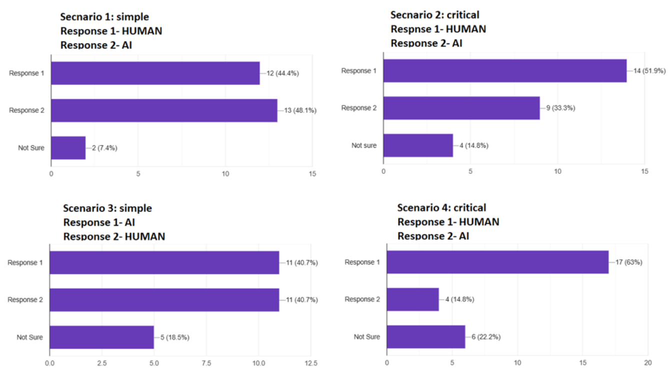

Fig. 3. Participant responses for Scenarios 1–4, indicating their ability to identify human-generated responses.

To further analyze the results, Figure 4 provides a cumulative view of participants' ability to identify human and AI-generated responses across all four scenarios. This visualization highlights three key observations:

1) **Simple Scenarios:** Participants were more successful in recognizing AI-generated responses in simple routine scenarios. These results suggest that AI performs well in structured, routine communications where responses follow a predictable pattern. Participants were able to identify AI with higher confidence, indicating that AI-generated responses may already be reliable for simple ATC exchanges, such as standard taxiing instructions, clearances, or routine pilot-controller exchanges
2) **Critical Scenario:** In more complex, high-stakes ATC scenarios, participants found it harder to differentiate between AI and human responses. This suggests that AI-generated responses are more difficult to distinguish from human responses in high-stakes scenarios, such as emergency situations, rapid decision-making, or unexpected ATC changes. It also indicates that participants were more skeptical about critical communications, possibly due to the expectation that AI might struggle in such situations.
3) **Uncertainty:** The percentage of "Not Sure" responses was higher in critical scenarios (e.g., 22.2% in Scenario 4), indicating that participants struggled to make definitive choices in high-stakes situations. This reflects the inherent difficulty in distinguishing nuanced communication under pressure.

The results support that AI could be a supporting tool for routine and predictable ATC tasks but not for the most important decisions in air traffic control. That participants could recognize AI correctly in simple scenarios would indicate that AI-generated responses work well for structured, repetitive communications. The rates of uncertainty and misclassification in complex situations indicate risks with over-relying on AI for a safety-critical environment.

While AI in ATC holds great promise, at present it should be utilized principally to support, rather than replace, human controllers for routine tasks, particularly in emergency or high-pressure situations where human expertise and adaptability remain critical.

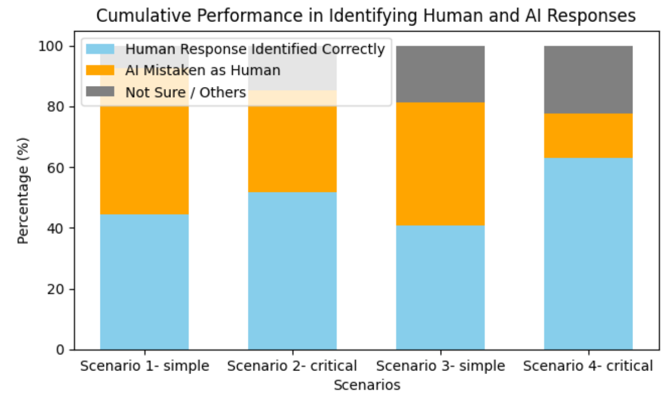

Fig. 4. Cumulative performance in identifying human and AI-generated responses across Scenarios 1–4.

C. *Responses and Interpretation for Scenarios 5–7*

The second set of tasks involved Scenarios 5–7, where participants evaluated the clarity, accuracy, and effectiveness of responses from ChatGPT and Gemini on a scale of 1 to 5 (1 = Not Good, 5 = Excellent). The results of the individual ratings are summarized in Table I, while the cumulative results across all three scenarios are displayed in Figure 5.

- **Scenario 5:** Participants rated ChatGPT slightly higher in clarity and accuracy but rated Gemini higher in effectiveness.
- **Scenario 6:** ChatGPT outperformed Gemini in clarity, accuracy, and effectiveness, indicating stronger performance in this critical scenario.
- **Scenario 7:** Ratings for both AI systems were more balanced, with participants showing no significant preference for either system.

1) Weighted Score Calculation: To quantitatively evaluate the responses, a Weighted Score Calculation method was employed. Each rating from 1 (Not Good) to 5 (Very Good) was assigned a weight:

- 1 (Not Good) $\rightarrow$ 1 point
- 2 $\rightarrow$ 2 points
- 3 (Neutral) $\rightarrow$ 3 points
- 4 $\rightarrow$ 4 points
- 5 (Very Good) $\rightarrow$ 5 points

The weighted average score for each metric (*Clarity, Accuracy, Effectiveness*) was determined using the following equation:

$$WS = \frac{\sum(Responses \times Weight)}{\sum Total\_Responses} \quad (1)$$

where:

- WS represents the weighted score for a given metric.

- $Responses$ denote the number of participants who assigned a particular rating.
- $Weight$ is the corresponding rating scale (1 to 5).
- $Total\_Responses$ is the total number of participants who rated the metric.

The AI model with the higher weighted score for a given metric was considered the better performer.

TABLE I
PARTICIPANT RATINGS FOR SCENARIOS 5–7

Scenario	Metric	1 (Not Good)	2	3 (Neutral)	4	5 (Very Good)
5	Clarity (ChatGPT)	3	1	7	13	3
	Accuracy (ChatGPT)	3	4	7	9	4
	Effectiveness (ChatGPT)	1	6	4	10	6
	Clarity (Gemini)	2	3	7	12	3
	Accuracy (Gemini)	1	4	8	9	5
	Effectiveness (Gemini)	1	4	7	10	5
6	Clarity (ChatGPT)	1	3	7	7	9
	Accuracy (ChatGPT)	0	4	9	6	8
	Effectiveness (ChatGPT)	2	5	6	6	9
	Clarity (Gemini)	0	5	7	8	7
	Accuracy (Gemini)	1	4	9	5	8
	Effectiveness (Gemini)	3	4	8	5	7
7	Clarity (ChatGPT)	3	4	8	8	4
	Accuracy (ChatGPT)	0	5	7	8	6
	Effectiveness (ChatGPT)	0	0	0	0	0
	Clarity (Gemini)	2	3	10	5	6
	Accuracy (Gemini)	1	5	4	11	5
	Effectiveness (Gemini)	0	7	4	8	7

The cumulative chart (Figure 5) aggregates the total responses for clarity, accuracy, and effectiveness across all three scenarios. It provides a holistic comparison of the overall performance of ChatGPT and Gemini:

- Gemini consistently outperformed ChatGPT in Effectiveness, suggesting its responses were perceived as more useful in ATC-style communications.
- ChatGPT performed better in Clarity and Accuracy in Scenario 6, indicating it may have been more precise in conveying information for that specific situation.
- Scenario 7 showed a split performance, with Gemini excelling in Clarity and Effectiveness while ChatGPT was rated higher in Accuracy.

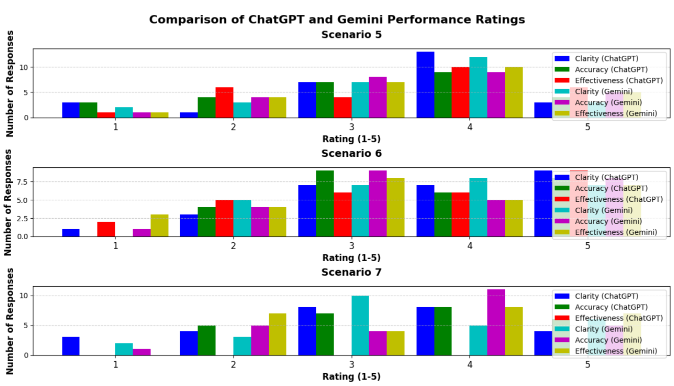

Fig. 5. Cumulative comparison of ChatGPT and Gemini performance ratings across Scenarios 5–7, evaluating clarity, accuracy, and effectiveness.

These findings suggest that while both AI systems are effective in specific contexts. Neither model was universally superior, but rather that they have complementary strengths. This suggests that for optimal AI-assisted ATC communication, a hybrid approach leveraging both models could be beneficial. However, variability in participant ratings highlights that generative AI still lacks consistency in meeting aviation safety standards, reinforcing the need for human oversight in safety-critical environments.

D. Analysis of Findings

The results from the participant responses provide evidence for evaluating the hypotheses. Key findings are summarized below:

- **H_0 Validation:** These results substantiate the claim that both ChatGPT and Gemini work well in structured and routine ATC scenarios but struggle in complex, high-stakes situations. The participants more correctly identified AI responses in simple scenarios, meaning that AI can really handle predictable communications. In such complex scenarios, requiring complex decision-making, limitations with both ChatGPT and Gemini showed up in contextual understanding and adaptability, reinforcing that full automation of the ATC task is not as such possible.
- **H_A Validation:** The results also showed that, though ChatGPT and Gemini were good in structured communications, they were not consistently outperforming human ATCOs at routine tasks. The mixed ratings with respect to clarity, accuracy, and effectiveness in Scenarios 5–7 suggest that AI is good for partial automation of ATC tasks but is no better than human controllers in routine situations. Thus, this partially supports H but does not confirm that AI stands out as a superior alternative.
- **Limitations in Hypothesis Validation** While the findings largely support hypothesis, it is important to acknowledge certain limitations:
  - The study evaluated a limited number of scenarios, which may not fully represent the complexity of real-world ATC tasks.
  - The participant sample size was small (27 participants), and their levels of expertise varied, potentially introducing bias into the results.
  - Other factors influencing ATC communications, such as air traffic density, weather conditions, and communication delays, were not accounted for in the scenario design.

IV. DISCUSSION

The findings of the study in discussion bring into consideration both the promises and limitations that generative AI systems, including ChatGPT and Gemini, have for use in ATC scenarios. Both systems did really well in generating responses for routine, simple scenarios. However, these systems usually turn out to fail in critical high-stakes situations. That suggests that, although a generative AI may be promising in some other role, it is unable to fully displace human air traffic controllers as it stands today within such a critical-to-safety domain..

Several key insights emerge from the results:

- **Performance Variability:** AI systems like Gemini outperformed ChatGPT in critical scenarios, particularly in terms of effectiveness. However, both systems struggled to fully replicate the contextual understanding and decision-making skills inherent in human controllers.
- **Participant Perceptions:** The ability of participants to distinguish between AI and human responses varied significantly, with participants being more accurate in identifying human responses in complex scenarios. This reinforces the idea that nuanced, high-stakes tasks still require human oversight.
- **Practical Implications:** Despite their limitations, generative AI systems could play a supportive role in ATC, particularly in routine communication tasks. By reducing the workload of human controllers, these systems have the potential to improve operational efficiency and allow humans to focus on critical decision-making.

It is important to acknowledge the limitations of the study. Factors such as air traffic density, weather conditions, and real-world pressures could not be fully simulated within the survey design. While the scenarios attempted to capture key aspects of ATC communications, the study represents only a small step toward understanding the broader implications of AI integration into the aviation industry. Future studies should incorporate real-time simulations and more diverse datasets to build on these findings.

V. Threats to Validity

Several factors may threaten the validity of findings in this study. These threats are grouped into internal validity, or issues affecting the cause-and-effect relationship within the study, and external validity, or issues affecting the generalization of findings to real-world ATC operations.

A. Internal Validity

Internal validity refers to the degree to which the study accurately measures what it intends to, ensuring that observed results are truly caused by the experimental conditions rather than external influences. The following factors pose potential threats to internal validity:

- **Participant Bias:** The participants' familiarity with AI or aviation-specific tasks might have influenced their ability to evaluate responses. Some participants may have preconceived notions about AI capabilities, impacting their ratings.
- **Scenario Design:** The scenarios were simplified for survey purposes, which may not fully reflect the complexity and urgency of real-world ATC situations. This could have influenced participants' ability to evaluate the responses accurately.
- **Instrument Reliability:** Variations in AI-generated responses due to differences in prompts or model versions might have introduced inconsistencies. Efforts were made to standardize prompts; however, these cannot fully eliminate such variability.

B. External Validity

External validity refers to how well the study's findings can be generalized to real-world ATC environments. Several factors could limit this generalizability:

- **Regulatory and Ethical Constraints:** The study does not account for aviation safety regulations and ethical considerations regarding AI deployment in ATC. Even if AI demonstrates high performance in routine tasks, aviation authorities may impose strict oversight on AI usage in safety-critical operations.
- **Real-World Applicability:** The controlled nature of the study does not account for the dynamic and high-pressure environment of actual ATC operations. Factors such as air traffic density, communication delays, and unexpected events could impact the performance of AI systems differently.
- **Model Generalizability:** The study focused exclusively on ChatGPT and Gemini. While these systems are leading generative AI models, the findings may not be applicable to other AI systems or future iterations.

Despite these threats, the study offers valuable initial insights into the role of generative AI in ATC and lays the groundwork for future research.

VI. Conclusion

This work investigated the performance of generative AI models, ChatGPT and Gemini, in mimicking human air traffic control communications. The research analyzed participant ratings of clarity, precision, and effectiveness in both routine and critical scenarios and pointed out the strengths and weaknesses of such systems.

While generative AI was performing well in routine scenarios, its shortcomings became obvious in complex, safety-critical situations that required contextual understanding and decision- making. Gemini always outperformed ChatGPT in critical tasks, especially in terms of performance, suggesting that a better design for AI will lead to higher outcomes in critical tasks.

It is important to emphasize that air traffic control communication is influenced by many factors, such as air traffic density, weather conditions, and real-time pressure, which could not be completely mimicked here. This study marks a small but important step in determining the position of AI in aviation; future studies will need to examine naturalistic simulations, larger participant samples for a more complete evaluation. Limitations notwithstanding, this study shows that generative AI may support ATC in mitigating workload and routine communications. However, human oversight cannot be foregone, especially in safety-critical environments where reliability, accuracy, and contextual understanding must be ensured. Any integration of AI into the ATC systems must be pursued with care but with the intention of augmenting, rather than substituting, human competencies.

References

[1] International Air Transport Association (IATA), "Generative AI and Aviation: Finding Crossroads for Future," 2023. [Online]. Available: https://www.iata.org/globalassets/iata/programs/innovation-hub/generative-ai-report.pdf

[2] Ebeni, "Navigating the Challenges and Opportunities of AI in Safety-Critical Systems," 2024. [Online]. Available: https://www.ebeni.com/navigating-the-challenges-and-opportunities-of-ai-in-safety-critical-systems/

[3] Master of Code Global, "Generative AI in Aviation and Airlines Industry: Real-Life Examples," 2024. [Online]. Available: https://masterofcode.com/blog/generative-ai-chatbots-for-airline-and-airport-industry

[4] R. Mariani, "Challenges in AI/ML for Safety-Critical Systems," 2019. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/e2e-presentations/2019/S1-1_Mariani-NVIDIA.pdf

[5] Royal Aeronautical Society, "Harnessing Artificial Intelligence in Aviation," 2024. [Online]. Available: https://www.aerosociety.com/news/harnessing-artificial-intelligence-in-aviation/

[6] ATC AUDIO. "Live Air Traffic Control Communications." YouTube Channel. Available at: https://www.youtube.com/@TheATCAudio. Accessed: Jan 30, 2025.

[7] You Can See ATC. "Live ATC Communications and Visualizations." YouTube Channel. Available at: https://www.youtube.com/@YouCanSeeATC. Accessed: Jan 30, 2025.

[8] National Transportation Safety Board (NTSB). "NTSB Accident Reports and Investigations." Online Database. Available at: https://www.ntsb.gov/investigations/AccidentReports/Pages/Reports.aspx. Accessed: Jan 30, 2025.

[9] Linguistic Data Consortium. "Air Traffic Control Complete Corpus (LDC94S14A)." LDC Catalog. Available at: https://catalog.ldc.upenn.edu/LDC94S14A. Accessed: Jan 30, 2025.

[10] Air Safety Institute. "Air Safety and Training Resources." AOPA Training and Safety. Available at: https://www.aopa.org/training-and-safety/air-safety-institute. Accessed: Jan 30, 2025.