ChatATC: Large Language Model-Driven Agents for Strategic and Tactical Air Traffic Management and Control

Sinan Abdulhak[a], Rajiv Govindjee[a], Nick Tran[a],
Wayne Hubbard[b], Karthik Gopalakrishnan[c], Max Z. Li[a,*]

[a] University of Michigan, Ann Arbor, MI, USA
{smabdul, rajivg, nicktran, maxzli }@umich.edu
[b] Federal Aviation Administration, Warrenton, VA, USA
Wayne.Hubbard@faa.gov
[c] Stanford University, Stanford, CA, USA
kgopalakrishnan@stanford.edu
[*] Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 27, 2024

Keywords: Large Language Models; Air Traffic Management; Air Traffic Control; Transportation System Performance

1 INTRODUCTION

Air traffic management and control play distinct, critical roles in ensuring safety and efficiency within the air transportation system. *Strategic* air traffic management differentiates from *tactical* air traffic control in terms of timescales and objectives: Traffic managers (TMs) performing strategic air traffic management seek to balance projected demand (e.g., scheduled flights) and capacity (e.g., a safe rate of arriving aircraft at an airport). This is often done well in advance of flight operations. On the other hand, air traffic controllers (ATCOs) ensure aircraft are safely separated, then focus on efficiency and congestion reduction. Both TMs and ATCOs are indispensable to the safe operation of this critical infrastructure—they also face significant stress and high workloads (Majumdar & Ochieng, 2002), which is compounded by increasing aviation demand (IATA, 2024) and future new users the airspace must accommodate (e.g., Bauranov & Rakas (2021)). Our work aims to leverage generative AI (GenAI) to develop ChatATC, a decision support tool that can relieve some of the repetitive, non-safety-critical workloads of the TMs and ATCOs so that they can focus their attention on tasks that need expert judgment, and relies on specialized training and years of experience.

We focus on two specific tasks that consume a significant fraction of the workload and can benefit from GenAI. The first task is to perform strategic traffic flow management, which requires TMs to plan Traffic Management Initiatives (TMIs) in anticipation of demand-capacity imbalances (e.g., due to poor weather or a planned runway closure). The lack of robust and interpretable optimization-based solutions for this highly uncertain setting forces controllers to rely on experience. Historical data regarding such TMIs can often be voluminous and difficult to sift through, particularly during real-time operations. When used for strategic air traffic management, ChatATC can rapidly retrieve, synthesize, and deliver summarized information regarding historical TMIs. Specifically, we fine-tune ChatATC on a large historical data set of

TMI issuances, spanning 2000-2023 and consisting of over 80,000 Ground Delay Program (GDP) implementations, revisions, and cancellations (Abdulhak *et al.*, 2024).[1]

The second task is to perform a tactical deconfliction, where aircraft are assigned routes in real time to reach their destination while ensuring safe separation from other aircraft and obstacles. This is typically achieved by ATCOs assigning radar vectors, altitude and speed restrictions, or standard arrival or departure routes to aircraft via radio communication or controller-pilot data links. Given increasing traffic densities and airspace complexities, we explore the potential for GenAI to act as a safety net, similar to existing systems such as the Runway Incursion Monitoring and Conflict Alert System (RIMCAS) (INDRA Air Traffic Management, 2024). RIMCAS constantly tracks aircraft operating on the airport surface, and provides an audio alert when potential conflicts (e.g., loss of safe separation) occur. To this end, we examine the effectiveness of ChatATC in managing tactical aircraft separation through experiments that integrate GenAI with BlueSky (Hoekstra & Ellerbroek, 2016), a full-featured Python-based ATC simulator which allows for benchmarking of new ATC algorithms.

The responsible use of GenAI in transportation (e.g., Zhang *et al.* (2024)) and non-transportation (e.g., Felten *et al.* (2023)) contexts are well-documented: We address its potential implications in the safety-critical management and operations of air transportation systems through the lens of TMs and ATCOs. While the success of LLMs for information retrieval tasks has been well-established (Yang *et al.*, 2024), additional work is required to ensure information accuracy and prevention of hallucinations. Currently, we mitigate these risks by relegating GenAI to one step in the air traffic management and control process: The final decision remains with the human operator.

2 DATA AND METHODS

2.1 ChatATC for strategic air traffic management

In total, we collected 86,842 historical GDP issuances spanning 146 airports between February 2000 and November 2023. From each issuance, we extract the raw GDP text, and store this along with parsed parameters such as duration and associated Airport Arrival Rates (AARs). We explore two approaches to create a conversational agent with the ability to summarize information about historical GDPs: (1) through in-prompt learning and (2) fine-tuning ChatATC using the collected historical GDPs.

2.2 ChatATC for tactical air traffic control

We simulate a scenario where two aircraft would be in conflict without controller intervention. We developed a BlueSky plugin to report the intentions and positions of aircraft to ChatATC, which ingests this natural language prompt and issues instructions back to the aircraft in the form of machine-readable simulator commands. The particular scenario we test involves two aircraft: one southeast-bound on jet route J34, at an altitude of 25,000 feet, traveling at 350 knots calibrated airspeed (CAS), and the other westbound on J70 at 25,000 feet and 350 knots CAS. The two aircraft are set up so that they have the same speed, altitude, and distance from a specific navigational waypoint. This ensures that if the LLM does nothing to separate the aircraft, they would eventually encounter a collision scenario. After deconfliction occurs, ChatATC should give a command so that the aircraft are directed to return to their original headings. We evaluate performance in terms of how often ChatATC provides a successful deconfliction, using speed, altitude, and heading as three different control inputs. Any encounters within 3 nautical miles

[1] GDPs are a form of TMI, where airborne delays are minimized in exchange for delays taken by aircraft on the ground Federal Aviation Administration (2024b).

of separation[2] and 1000 feet of altitude[3] trigger a warning in the simulator, and we record these as failures.

3 RESULTS

3.1 In-prompt learning for strategic air traffic management

Even with only a shortened version of in-prompt learning, ChatATC is able to recognize the general structure of a GDP issuance, and accurately extract parameters from the GDP issuance with no additional prompting or guidance. These parameters include the GDP rate, affected departure airports (i.e., the scope of the GDP), among others. We manually checked the raw text of the GDP entries used for in-prompt learning to ensure that ChatATC is reporting accurate information.

3.2 Fine-tuning ChatATC for strategic air traffic management

After fine tuning ChatATC, we started by asking the San Francisco International Airport (SFO) instance of ChatATC about GDPs due to weather. We observed that ChatATC was able to quickly retrieve an example of a relevant GDP from one of the 500 historical GDP issuances it had been trained and fine tuned on. Specifically, the SFO instance of ChatATC was able to retrieve examples of GDPs on a specific date, GDPs due to a specific impacting condition (e.g., poor weather), GDPs with a specific rate, among others. Although we observed errors when it comes to superlatives-based questions (e.g., asking ChatATC for the GDP with the *highest maximum delay*), we note that TMs are less likely to ask superlatives-based questions, as compared with questions regarding, e.g., the rate of a previous GDP, or the reason why a GDP was implemented—we do observe factually correct answers to the latter questions.

Additionally, as the rate, start and end times, and prevailing runway configuration are particularly important parameters to TMs, we relied on prompt engineering by adding the following to ChatATC's system prompt: "Give me date, start time, end time, program rate, runway configuration and impacting condition."

3.3 Traffic deconfliction results with ChatATC

We performed 50 runs of the simulation scenario for each deconfliction method (i.e., speed, heading, altitude), and recorded the safety success rate for each of the three modes. As LLMs are inherently probabilistic models, outcomes vary even with the exact same starting scenario.

From our results, we see that ChatATC is successful in deconfliction 60% of the time for heading-only changes (30 out of 50 runs deconflicted). For speed-only changes, ChatATC had an easier time deconflicting the aircraft, with a recorded success in 66% of the runs (33 out of 50 runs). Finally, when it came to altitude-only changes, ChatATC outperformed both speed and heading modes, with a success rate of 100% (50 out of 50 runs).

4 DISCUSSION

For strategic air traffic management, we envision ChatATC being applied only in non-safety critical roles, given the risks of hallucinations, lack of clarity on how to verify its outputs, and lack of interpretability. Our results demonstrate ChatATC's ability to quickly retrieve historical GDP data, an otherwise time-intensive task, and summarize information relevant to TMs. Such tools

[2] Note that lateral separation can range between 3-10 nautical miles depending on phase of flight and equipage (Federal Aviation Administration, 2024a). This separation parameter can be changed in our set of experiments.

[3] 1000 feet vertical separation is standard under Reduced Vertical Separation Minimum, or RVSM implementation (Federal Aviation Administration, 2024c).

can also help increase the quality of training for new personnel. Future work from the strategic air traffic management perspective could include experimenting with training ChatATC using raw GDP text with different file formats. Additional user testing and feedback collection need to be performed with a larger group of TMs.

On the tactical air traffic control side, we observe that ChatATC show promise for the purpose of directing traffic (a first step towards continuous-monitoring functionalities such as RIMCAS), but struggle to perform reliably due to inherent limitations with this class of models. We hypothesize that heading-based deconflictions are the most difficult due to the spatio-temporal reasoning required, while altitude-based deconfliction is easiest given that the vertical separation requirement can be verified with a simple subtraction. Our results demonstrate the strengths of LLMs in flexibly processing natural language inputs and outputs. In our experiments, ChatATC issues machine-readable commands that are directly sent to the simulator. ChatATC is equally capable of responding in standard ATCO phraseology, allowing communication with human controllers and pilots. Using text-to-speech models (e.g., OpenAI (2024)), this could even be accomplished on existing radio infrastructure. Future work on ChatATC in the tactical air traffic control modality could include a focused effort towards providing specific air traffic control training data sets and incorporating ATC-specific knowledge into the models. With potential to be deployed as a part of a larger human-in-the-loop system, ChatATC has the potential to greatly reduce ATCO workloads and enhance situational awareness.

References

Abdulhak, Sinan, Hubbard, Wayne, Gopalakrishnan, Karthik, & Li, Max Z. 2024. CHATATC: Large Language Model-Driven Conversational Agents for Supporting Strategic Air Traffic Flow Management. *In: Proceedings of the 11th International Conference on Research in Air Transportation.*

Bauranov, Aleksandar, & Rakas, Jasenka. 2021. Designing airspace for urban air mobility: A review of concepts and approaches. *Progress in Aerospace Sciences*, **125**, 100726.

Federal Aviation Administration. 2024a (April). *ATC Clearances and Aircraft Separation.*

Federal Aviation Administration. 2024b (April). *Ground Delay Program.*

Federal Aviation Administration. 2024c (April). *Reduced Vertical Separation Minimum (RVSM).*

Felten, Ed, Raj, Manav, & Seamans, Robert. 2023. *How will Language Modelers like ChatGPT Affect Occupations and Industries?*

Hoekstra, Jacco M., & Ellerbroek, Joost. 2016. Bluesky ATC simulator project: an open data and open source approach. *In: Proceedings of the 7th international conference on research in air transportation.*

IATA. 2024 (January). *Global Air Travel Demand Continued Its Bounce Back in 2023.*

INDRA Air Traffic Management. 2024. *NOVA 9000 RIMCAS.* Tech. rept. INDRA Air Traffic Management.

Majumdar, Arnab, & Ochieng, Washington Y. 2002. Factors Affecting Air Traffic Controller Workload: Multivariate Analysis Based on Simulation Modeling of Controller Workload. *Transportation Research Record*, **1788**(1), 58–69.

OpenAI. 2024 (April). *Text to speech.*

Yang, Jingfeng, Jin, Hongye, Tang, Ruixiang, Han, Xiaotian, Feng, Qizhang, Jiang, Haoming, Zhong, Shaochen, Yin, Bing, & Hu, Xia. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data*, feb.

Zhang, Siyao, Fu, Daocheng, Liang, Wenzhe, Zhang, Zhao, Yu, Bin, Cai, Pinlong, & Yao, Baozhen. 2024. TrafficGPT: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, **150**, 95–105.

DISCLAIMER