Using Large Language Models to extract CPDLC messages from conversations between air traffic controllers and pilots

Anthony Beaudry
McGill University anthony.beaudry@mail.mcgill.ca
https://srl.mcgill.ca/projects/adair/

Abstract

Controller-Pilot Data Link Communications (CPDLC) systems have been developed to relieve the overload of traditional voice-based radio systems. Although CPDLC has been a significant technological step forward in digital communication systems, it induces some clear limitations. The menu-based messaging system does not allow for much flexibility and increases communication delays. Moreover, the strict grammar rules enforced in CPDLC Message Elements prevent the natural development of conversations containing important contextual information. As Large Language Models (LLMs) grow more powerful, their use as assistants in automation has become a major topic of interest. This study investigates the use of LLMs as a translation assistant to identify CPDLC instructions in natural language-based conversations and to understand/extract any meaningful contextual information that cannot be conveyed in the CPDLC message. LLMs successfully achieving this task would provide a framework to introduce flexibility to current CPDLC systems by automating these messages through a conversational agent.

1 Introduction

The relentless growth in global air traffic places significant pressure on established Air Traffic Management (ATM) communication systems. Traditional voice communication via VHF radio faces increasing congestion in dense airspace. This saturation contributes to operational inefficiencies and elevates the inherent risks associated with communication errors, such as misunderstandings or incorrect readbacks, which can impact aviation safety [1], [2]. The need for more sustainable and efficient communication methods is therefore crucial.

Controller-Pilot Data Link Communications (CPDLC) was introduced as a technological step forward, enabling digital text-based message exchanges. By migrating routine communications from voice channels, CPDLC significantly reduces frequency congestion and supports clearer, standardized message delivery for common clearances and requests [3]. This digital method represents a key component of modernizing ATM infrastructure.

Despite these advancements, CPDLC introduces specific operational dynamics. Its reliance on predefined message sets, while ensuring consistency, offers less flexibility than natural language for handling non-standard events or complex instructions. Furthermore, the discrete nature of data link messaging eliminates the situational awareness that pilots gain from normal conversations [3]. The system's design and inherent latency also mean it is primarily suited for non-time-critical exchanges, leaving voice indispensable for urgent matters [3]. These factors highlight that while CPDLC addresses voice congestion, it doesn't fully replicate the indispensable qualities of voice communication.

Recent breakthroughs in Artificial Intelligence, specifically Large Language Models (LLMs), offer novel potential. LLMs possess advanced capabilities in understanding the context, nuance, and intent within human language. Initial research is already exploring the application of LLMs in ATM for tasks like anomaly detection or providing decision support [4].

This project investigates the application of LLMs to bridge the gap between structured data link messages and natural pilot-controller conversations. I propose exploring how LLMs can analyze conversational data (spoken or text-based) to automatically identify and extract the contained standardized CPDLC message, while preserving the surrounding context often lost in purely digital exchanges. The aim is to leverage LLM capabilities to potentially automate aspects of message generation, enhance situational understanding, and foster safer, more efficient communication within the demanding ATM environment.

2 Related Work

A fundamental challenge lies in processing the raw data from ATC communications. Zuluaga-Gomez et al. [5] provide valuable insights into this area, detailing the complexities encountered and lessons learned from transcribing a massive 5000-hour corpus of ATC voice data. Their work highlights significant hurdles in developing robust Automatic Speech Understanding (ASU) systems due to factors such as noisy channels, rapid speech rates, diverse accents, code-switching between languages, and deviations from standard phraseology [5]. While our research focuses on the step of extracting information and context from potentially transcribed conversations, the findings of Zuluaga-Gomez et al. underscore the inherent difficulties in obtaining reliable textual input from real-world ATC voice interactions.

Moving beyond basic speech understanding, researchers are exploring how LLMs can actively assist flight crews and controllers. Schlichting et al. [6] introduce LeRAAT, an LLM-Enabled Real-Time Aviation Advisory Tool [6]. LeRAAT integrates an LLM with the X-Plane flight simulator, employing Retrieval-Augmented Generation (RAG) to access and retrieve information from aircraft manuals, procedures, and regulations. Its goal is to provide pilots with context-aware advisories, particularly during emergencies. This work is significant as it demonstrates the potential of LLMs combined with RAG to deliver pertinent, contextual information within the demanding aviation environment. It also tackles the critical issue of mitigating LLM risks, such as hallucination, in safety-critical applications. However, LeRAAT's focus is primarily on retrieving and presenting document-based information to the pilot, rather than analyzing and extracting structured information from the dynamic pilot-controller communication itself, which is the focus of our study.

Other research investigates the potential for LLMs to take on more active roles within ATC operations. Andriuškevičius and Sun [7] explore the use of LLMs as "embodied agents" within simulated ATC environments [7]. Their work examines the capability of these agents to perform automatic conflict resolution, leveraging the LLMs' ability to interact with the environment and provide human-like reasoning for their control decisions. This research pushes the boundary of LLM application in ATC towards autonomous functions and highlights their potential for complex reasoning within the domain.

3 Background

As global air traffic increases, traditional VHF voice radio, the primary communication channel, faces significant strain. This congestion elevates the risk of communication errors and operational inefficiencies [2]. Furthermore, limitations in available radio spectrum challenge the scalability of purely voice-based systems [8].

Controller-Pilot Data Link Communications (CPDLC) was introduced as a standardized digital messaging system to alleviate these pressures. By handling routine exchanges via text-based messages, CPDLC reduces voice frequency load and can mitigate certain types of voice communication errors [3]. It allows for the consistent transmission of standard clearances and requests.

However, CPDLC has limitations. Its reliance on predefined message sets lacks the flexibility of natural language required for non-standard situations or complex negotiations. Critically, it eliminates the situational awareness pilots gain from the voice 'party line' effect [3]. Due to system latency, CPDLC is primarily suited for non-time-critical communications, necessitating voice for urgent instructions [3]. Interacting with CPDLC interfaces can also introduce human factors challenges related to message composition and interpretation under workload [9].

Advancements in Natural Language Processing (NLP) offer potential solutions. Research has explored applying NLP and Automatic Speech Recognition (ASR) to transcribe ATC voice communications and extract key information like commands, or to analyze aviation safety reports [2], [10]. While valuable, these often face challenges with the complexities of real-world ATC speech and contexts.

The growth of Large Language Models (LLMs) presents a significant leap in AI's ability to understand context, nuance, and complex language structures.

This context reveals a contrasted communication challenge. Voice offers flexibility and context, but is inefficient and error-prone under load. CPDLC reduces radio congestion but is rigid and lacks contextual depth. This research aims to show that the advanced contextual understanding of LLMs can bridge this gap. This study investigates the use of LLMs to analyze natural language pilot-controller conversations, aiming to extract both the structured information equivalent to CPDLC messages and the crucial surrounding conversational context. This approach seeks to harness the benefits of both communication paradigms, potentially leading to more efficient, context-aware, and safer air traffic operations.

4 Methodology

This section will give a broad insight into the research measures which includes data processing techniques, the design of the translation algorithm and certain limitations of this study.

4.1 Data Collection & Analysis

Input data was obtained from publicly available videos, which were then cleaned and interpreted. Input data consisted of sample pilot/ATC messages representing common instructions (e.g., frequency change, direct routing, altitude assignment, speed control). These samples included both the natural language input and the expected structured output for evaluation purposes.

Specific system prompts were designed for ATC-to-pilot (uplink) and pilot-to-ATC (downlink) translation tasks, instructing the model to act as a CPDLC translation expert. These prompts incorporated the retrieved CPDLC descriptors as context and specified the desired JSON output format containing ”reference”, ”message”, and ”context” fields. The natural language input message was appended to this prompt structure before being fed to the LLM.

4.2 Translation Algorithm & Design

The use of open-source LLMs was facilitated by the vLLM inference engine to perform the extraction task as it offers significant performance gains in inference over traditional python libraries.

A Retrieval-Augmented Generation (RAG) approach was implemented to provide relevant context to the LLM during inference. A FAISS index was constructed using sentence embeddings of CPDLC message descriptors (message element, intent, reference number) [11] embedded by the sentence-transformers/all-MiniLM-L6-v2 model. For each natural language input, the k most similar CPDLC descriptors were retrieved from the FAISS index based on embedding similarity. The value of k was varied across experiments (e.g., 40, 75, 100, 150) to observe its potential impact. To improve the consistency of RAG, a Preprocessor function was made to align user inputs closer to the embedded vectors.

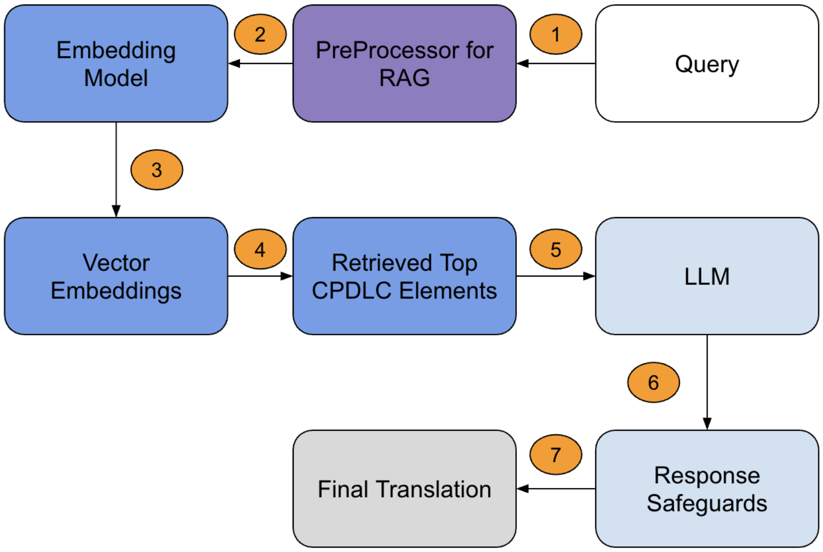

Figure 1: Translation Algorithm Design

Several LLMs were evaluated, primarily focusing on models from the Qwen family available on Hugging Face, including Qwen2.5-7B-Instruct-AWQ, Qwen2.5-14B-Instruct-AWQ and Qwen2.5-32B-QWQ. Quantization techniques, specifically 8-bit using BitsAndBytesConfig and AutoAWQ (quantization="awq"), were applied to larger models to manage computational resource requirements. Experiments were conducted using the vLLM library (LLM class) for efficient inference. An attempt was made to test google/gemma-3-4b-it, but preliminary runs indicated potential source code errors or import issues within the experimental environment.

4.3 Limitations

Although I conducted a study on CPDLC standards to have the required knowledge to translate messages into specific instructions, I labeled the sample data which opens up the possibility of wrong interpretation. Similarly, the analysis of the context retrieved by the assistant was done by lexical understanding of the natural language conversations, which may not reflect the true intent of aviation specifics.

Some hardware limitations were in place as only one A100 GPU was available for the purpose of this research, resulting in the inability to locally run large open-source models (more than 32B parameters).

The sample size of testing data is insufficient to cover all possible CPDLC instructions. Moreover, the distribution of Downlink versus Uplink sample data is biased in favor of Uplink messages due to the more readily available pool of data of the latter.

5 Implementation Details

This section will provide details on the design choices of the translation algorithm.

5.1 Consistency with RAG

Initial attempts at using RAG for the retrieval of top CPDLC messages were inconsistent. Since inference speed from LLMs is approximately linearly dependent on the input size, a way to reduce context information at each prompt is needed for efficiency purposes. Falling back on RAG to achieve this task, an important observation made when inconsistencies occurred during information retrieval is that the data stored using CPDLC Message Elements contained descriptive input fields whereas the input message to translate may contain numerical value(s). Therefore, a mechanism to modify input prompts as a separate input for RAG (see Preprocessor for RAG in Figure 1) was implemented by replacing numerical values with descriptive fields according to the possible options from the CPDLC Message Elements.

5.2 Choice of LLM & Hyperparameters

Consistency is crucial to provide accurate CPDLC instructions. Therefore, the conversational agent needs to behave deterministically for the extraction of the instruction while being able to understand any contextual information present. To achieve this, a grid-search on temperature and top_p values was performed as both parameters are crucial for tuning model creativity. A temperature of 0.2 was found to be optimal which is supported in similar applications [12]. Furthermore, a top_p value of 0.9 yielded optimal performance, though values in the range [0.88, 0.95] had insignificant performance differences.

5.3 Optimizing Context Prompt

The sets of rules provided in the model system prompt is an important factor in achieving desirable results [13]. To ensure robustness and mitigate inconsistencies, several simple rules specific to aviation were added in the prompt. This allowed basic guidance without extra cost. Prompts contained different contextual situations to describe the role of the conversational agent depending on the task (downlink vs uplink).

5.4 Safeguarding

In order to identify semantically incorrect outputs, models had to respond in json format and could never leave an input in a CPDLC Message Element unfilled. When a CPDLC message failed the safeguard rules, there were 2 avenues that were explored in the code. The first one being a consensus algorithm if efficiency is not a concern. The other solution simply loops the model on different hyperparameter configurations to get a different response (uses less deterministic settings).

6 Experiments & Results

The experiments conducted focused on open-source LLMs with various Qwen models and demonstrated promising capabilities in translating natural language instructions into structured CPDLC message elements.

Several runs, particularly with the Qwen2.5-14B-Instruct-AWQ model, achieved high accuracy in identifying the correct CPDLC message structure and reference number. For instance, many runs which used all algorithmic optimizations reported a reference number accuracy of 100% on the test samples. Basic instructions like frequency changes (CONTACT [unit] [frequency]), direct clearances (PROCEED DIRECT TO [position]), and simple altitude/speed commands (CLIMB TO AND MAINTAIN [level], MAINTAIN [speed] KNOTS) were frequently translated correctly even on weaker models. Overall performance seemed to vary highly on different LLMs as certain models with as many parameters as the Qwen baseline performed much worse (Llama 3.1 7B achieved 75% average accuracy).

The models showed variability in extracting relevant context. In simpler cases with direct instructions, the context field was often correctly left empty. For messages containing reasons (e.g., ”vector for sequencing”, ”for traffic”) or additional conditions (e.g., ”when able”, advice to inform next controller about deviation), the models successfully captured this information in the ”context” field. However, this was less consistent than the primary message extraction, which is expected due to the ambiguity of contextual information.

A recurring challenge was the reliability of the output format. While instructed to respond only in JSON, models occasionally produced outputs containing irrelevant text, such as markdown formatting (“‘json...“‘), or chain-of-thought reasoning before or after the JSON object. The custom parse_llm_response function was necessary to extract the valid JSON portion, but some weaker models still observed failures.

Qwen models, especially the 14B and 32B variants (including AWQ quantized versions), appeared generally capable based on the successful examples presented. The DeepSeek-R1-Distill Series models were discontinued in testing, but showed promising results. The Gemma models encountered runtime or setup issues preventing conclusive results within the notebook's scope.

Overall, the results indicate that the tested LLMs, when prompted with RAG-retrieved context, can often correctly identify the core CPDLC instruction and reference number but may struggle to identify rules to follow if they are not properly defined in the prompt.

7 Discussions

The results presented demonstrate the significant potential of LLMs, particularly models from the Qwen family, in the domain of ATC communication analysis. The models' ability to identify the correct CPDLC message element and reference number from natural language inputs, even with conversational variations and the inclusion of RAG context, is encouraging. This suggests that the core task of translating a primary instruction into a standardized format is feasible with current LLM technology. The reported translation number accuracy of 100% in one specific run with a relatively lightweight Qwen2.5-14B-Instruct-AWQ model, while based on a limited sample set, points towards a strong foundational capability.

However, the experiments also highlight critical challenges. The primary obstacle observed was the models' inconsistency in adhering strictly to the requested JSON output format, especially in models with parameter sizes of $<$ 7B. The inclusion of explanatory text, markdown formatting, or errors like stray brackets indicates difficulties in precise format control, a known issue with generative LLMs. This necessitated robust parsing logic as safeguards, but resulted in rare delayed response times.

The extraction of relevant, non-CPDLC context showed mixed success. While straightforward contextual cues like reasons ("vector for sequencing") were sometimes captured, the reliability and completeness of context extraction appeared less robust than the primary instruction translation. This is a complex task, requiring the model not only to understand the core instruction but also to discern which additional parts of the natural language input constitute relevant context that cannot be encoded in the standard CPDLC message. The definition of "relevant context" itself can be subjective and highly dependent on the operational scenario.

The use of RAG, retrieving similar CPDLC descriptors based on sentence embeddings, likely aids the model by providing in-context examples of relevant intents and message structures. The variation of k (number of retrieved examples) across experiments suggests an attempt to optimize this, although the notebook doesn't present a comparative analysis of different k values. The effectiveness of RAG depends heavily on the quality of the embeddings and the comprehensiveness of the indexed CPDLC data.

8 Further Work

To ensure the reliability of this application, more sample data is needed as several CPDLC messages remain untested. Furthermore, the context prompt used for the LLMs can and should be modified to adhere more strictly to aviation standards.

As shown in the experiments, the accuracy of the agent may heavily affect its translation speed as larger models are slower in inference. The translation mechanism using LLMs can be further improved and researched using several algorithmic optimizations. The current retrieval of the relevant CPDLC messages is done with RAG by FAISS, but a custom trained RAG system may offer benefits in efficiency and consistency. Furthermore, an appropriate UI for the current application should be built for an easier use.

9 Conclusion

This study explored the use of large language models to extract a CPDLC instruction along with any additional context from a natural language-based conversation to automate input to CPDLC systems while keeping the contextual information present from a conversation.

The experiments demonstrated that LLMs can accurately translate messages based on natural language to correctly formatted CPDLC instructions and recognize relevant contextual information that cannot be conveyed within the CPDLC message element. The performance of the tested models was found to be highly dependent on the model and its size. Qwen 2.5 14B-Instruct-AWQ was used as a baseline model and performed accurately without sacrificing efficiency, suggesting that implementing these tools in aviation systems is possible without significantly increasing communication delays.

Despite the limitations of this research, the study showcases LLMs' potential in ATC translation tasks. The findings suggest that LLMs could serve as powerful assistants for translating conversational inputs into structured CPDLC formats, potentially enabling more flexible communication interfaces or automating message logging, provided that challenges related to output formatting reliability and consistent context extraction can be overcome.

References

[1] O. V. Prinzo and T. W. Britton, "Atc/pilot voice communications : A survey of the literature," Civil Aeromedical Institute, Tech. Rep., 1993.

[2] C. Yang and C. Huang, "Natural language processing (nlp) in aviation safety: Systematic review of research and outlook into the future," *Aerospace*, vol. 10, no. 7, p. 600, 2023.

[3] "Controller pilot data link communications (cpdlc)," SKYbrary Aviation Safety, Tech. Rep.

[4] B. J. Connolly and G. Schneider, "Aircraft anomaly detection using large language models: An air traffic control application," *AIAA*, 2024. DOI: https://doi.org/10.2514/6.2024-0744.

[5] J. Zuluaga-Gomez, I. Nigmatulina, A. Prasad, *et al.*, "Lessons learned in transcribing 5000 h of air traffic control communications for robust automatic speech understanding," *Aerospace*, 2023. DOI: https://doi.org/10.3390/aerospace10100898.

[6] M. R. Schlichting, V. Rasmussen, H. Alazzeh, *et al.*, "Leraat: Llm-enabled real-time aviation advisory tool," *arXiv preprint arXiv:2503.16477*, 2025.

[7] J. Andriuškevičius and J. Sun, "Automatic control with human-like reasoning: Exploring language model embodied air traffic agents," *arXiv preprint arXiv:2409.09717*, 2024. DOI: https://doi.org/10.48550/arXiv.2409.09717.

[8] "Assessment of radio spectrum depletion of atc voice communications," in *ICAO Spectrum Report*, 2003.

[9] K. Cardosi and T. Lennertz, "Flight deck human factors issues for national airspace system (nas) en route controller pilot data link communications (cpdlc)," John A. Volpe National Transportation Systems Center (U.S.), Tech. Rep., 2017.

[10] "Scribe nlp: Unleashing the potential of atc voice communication," ENRI (Electronic Navigation Research Institute) - IWAC2024 Workshop, Tech. Rep., 2024.

[11] *Global operational data link document (gold).*

[12] S. Abdulhak, W. Hubbard, K. Gopalakrishnan, and M. Z. Li, "Chatatc: Large language model-driven conversational agents for supporting strategic air traffic flow management," *arXiv preprint arXiv:2402.14850*, 2024.

[13] L. Wang, J. Chou, A. Tien, X. Zhou, and D. Baumgartner, "Aviationgpt: A large language model for the aviation domain," in *AIAA AVIATION FORUM AND ASCEND 2024*, 2024, p. 4250.