ATComm: A Large Scale Procedural and Scenario-Aware Textual Dialogue Corpus for Indian Air Traffic Control

Anonymous ACL submission

Abstract

In aviation, the development of robust Automated Speech Recognition (ASR) and Natural Language Processing (NLP)-based systems are fundamentally limited by the scarcity of linguistic resources. These linguistic resources are highly constrained by regional airspaces that feature distinct accents, localized naming conventions, and strict phraseological variations. While Western benchmarks exist, they fail to generalize to geographically diverse airspaces such as India. In this paper, we present ATComm (Air Traffic Control (ATC) Communication), a first large-scale synthetic textual corpus for the Indian ATC environment. This corpus contains 15k dialogues comprising around 65k multi-turn textual utterances between air traffic controllers (ATCOs) and pilots within Indian airspace. We propose a generic framework, Procedural Scenario-Aware Contextual (ProSAC), to generate the corpus. This framework has leveraged real Indian airspace data as context, along with aviation-specific constraints and scenarios to generate high-fidelity multi-turn conversations. Our analysis demonstrates that ATComm reflects the linguistic, semantic and syntactic characteristics of real ATCO-pilot conversations. The automated audits and human expert assessment validates the ATComm as a valuable resource. Furthermore, the extrinsic evaluation proves that the T5 model trained on ATComm outperforms existing model on the NER task for recognizing aviation entities specific to Indian airspace.

1 Introduction

With the advent of large language models (LLMs), it has become possible for the aviation sector to utilize Natural Language Processing (NLP) capabilities in every conceivable way. The sector has seen strong interest in leveraging NLP across key areas such as air traffic control (ATC), predictive maintenance, customer experience, and safety report analysis (Nanyonga et al., 2025). It is transforming the aviation sector by improving efficiency, mitigating risks, reducing workload, and optimizing tasks. However, the application of LLMs in the aviation sector is hugely dependent on domain-specific data (Liu, 2024) (Ge and Coupe). One such domain of aviation sector is ATC, a highly safety-critical area characterized by its own set of phraseologies and standards approved by International Civil Aviation Organization (ICAO) (Yang and Huang, 2023) (Zuluaga-Gomez et al., 2023). Effective communication between Air Traffic Controllers (ATCOs) and pilots are the basis of flight safety, therefore to avoid any miscommunication, it is mandatory for them to use specialized phraseologies (Lin, 2021). As air traffic grows, the workload of controllers and pilots grows too, there is a need to develop such systems that assist them effectively. While several datasets of ATC-pilot communications exist for western airspaces, derived from real-time ATCOs and pilot communication recordings and manual transcriptions (Badrinath and Balakrishnan, 2022) (Zuluaga-Gomez et al., 2023) (Šmídl, 2011) (Zhang et al., 2025), these resources have limited utility in different airspace (Singh et al., 2025b). The dynamics of every airspace vary in terms of regional linguistic accents and topography. A recent study found that an ATC-adapted end-to-end (E2E) automated speech recognition (ASR) model trained on a Western ATC airspace dataset performed poorly on Indian ATC speech data (Singh et al., 2025a). This critical analysis of E2E ASR models revealed that, despite high accuracy on other airspace data, they exhibit a high word error rate (WER) on India data. This performance gap is largely attributed to lack of Indian ATC domain linguistic knowledge. In such case, Language Model (LM) acts as a linguistic guide to predict the most probable sequence of tokens based on the preceding conversational context. However, for the Indian FIR, there is an acute shortage of region-specific ATC communication data. This

leads to the failure in decoding, resulting in high WER in the previous study.

To bridge this gap and address the 'zero-resource' challenge in Indian ATC, we present ATComm (ATC Communication), a high-fidelity synthetic dialogue corpus for Indian airspace. The corpus comprises of 15k dialogues containing 65k textual utterances of ATCOs and pilots. Our work focuses on developing a textual corpus as the first critical step in conducting research and development for Indian airspace-centric ASR systems. Moreover, the development of such corpus will also aid in Spoken Language Understanding (SLU) tasks like intent classification, slot filling, etc. (Zhang et al., 2025). In this paper, we propose a Procedural Scenario-Aware Contextual (Pro-SAC) data generation framework. To generate linguistically and procedurally valid data, operational constraints and scenarios aware prompt is defined. This ensures that every generated utterance is procedurally correct within Indian airspace. Following are the main contributions of this work:

1. Addressing the Zero-resource Gap: Unlike existing western datasets, ATComm is a first, ATC-specific synthetic textual dialogue corpus tailored for the Indian Airspace. It addresses the critical scarcity of ATC data in this zero-resource region. This seed corpus consists of 65k exchanges between ATCOs and pilots within Indian airspace.

2. Introducing Pro-SAC Framework: A data generation framework is proposed to generate data restricted to the constraints specific to aviation uses real flight schedules and navigational points to generate a corpus based on different operational scenarios.

3. Comprehensive Analysis and Validation: We conduct a comparative linguistic analysis to show ATComm mimics characteristics of real-time ATC data. Both automated and manual audits including subject matter expert (SME) are done to validate the generated corpus.

4. Benchmarking on Safety-Critical Task: The potency of ATComm corpus is demonstrated through benchmarking on the Named Entity Recognition (NER) task. This benchmark is done against an open-source model trained on European ATCO-pilot communication.

To the best of our knowledge, ATComm is the first corpus focused on Indian flight information region (FIR)-specific ATCO-pilot communication. Fig. 1 illustrates the complete pipeline that being followed for the generation of the ATComm corpus. It represents the data collected from different websites and pre-processed and then fed to Pro-SAC generation framework. This framework utilizes this data to generate ATCOs-pilot conversations. Automated and manual checks are made to validate the data. Data post-processing is done to remove anomalies. A detailed analysis is done on the generated synthetic corpus to analyze it characteristic with real data followed by modeling to assess its utility in real-time.

2 Background

LLMs heavily rely on large-scale annotated data, unintentionally favouring high-resource domains. With the advent of Generative AI, researchers are actively leveraging LLM capabilities to address the low-resource challenge in data-scarce environments. Techniques like Tree of Thought (ToT) (Yao et al., 2023), COSTAR (Ohalete et al., 2025), Zero Shot (Ye et al., 2022), One Shot (Li et al., 2024), and Meta Prompting (Riaz et al., 2025) have proven effective for the development of corpora in specific domains. Although prompt-based methods enable large-scale data generation, maintaining data quality is essential. LLMs are prone to hallucinations and can generate inaccurate or illogical data, particularly in knowledge-intensive tasks (Liu et al., 2024) (Alansari and Luqman, 2026). To overcome this, several strategies are employed, including integrating domain constraints, post-generation filtering, and selection, to enhance the quality of synthetic data. For instance, in (Fedoseev et al., 2024), the authors proposed a constraint-based synthetic data generation framework for fine-tuning LLMs on mathematical counting and combinatorial tasks. Multiple constraints were enforced to ensure that every synthetic sample was logically well-defined and solvable. Another work was done where, instead of using logical or symbolic constraints, the authors explicitly embedded constraints into the prompt (Alsakran and Alabduljabbar, 2024). Similarly, (Elnashar et al., 2025) focused on generating high-quality data with GPT-4o by utilizing constraint-aware prompt design. Different constraints based on schema, format, value, etc., across multiple domains were designed and explicitly re-

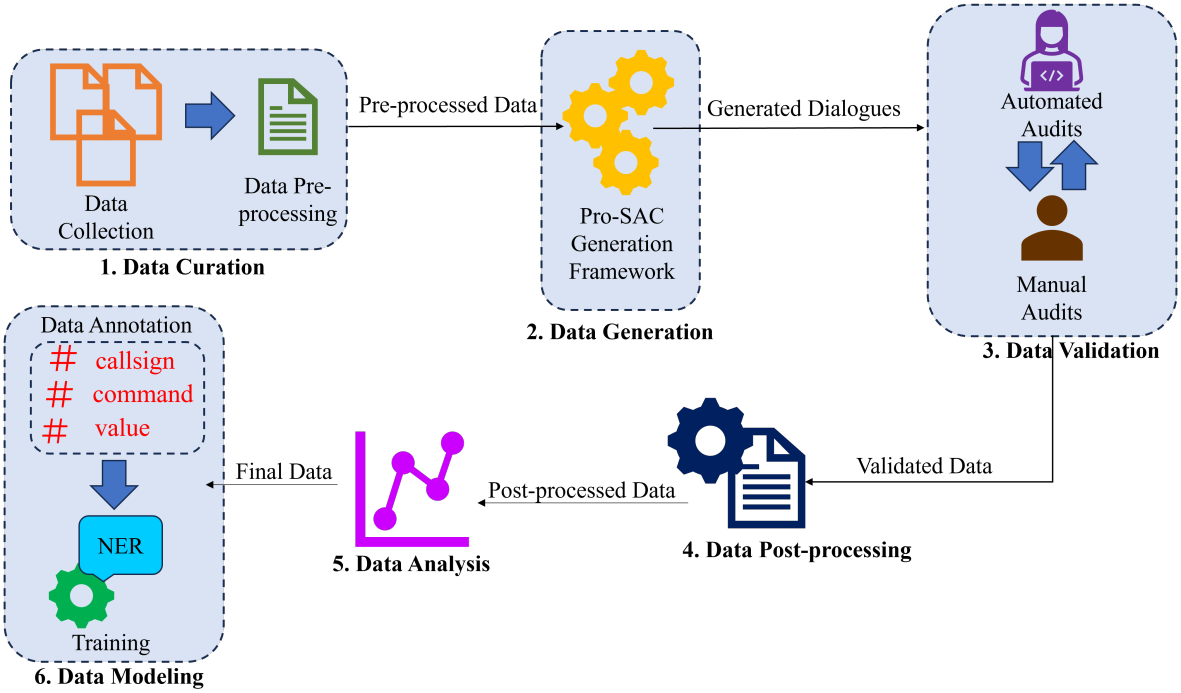

Figure 1: End-to-End Pipeline of the ATComm Corpus, Outlining the Raw Data Flows from Generation Framework to Empirical Validation and Downstream Task.

stricted the model's output.

In recent years, the aviation sector has also started leveraging prompt techniques to overcome data scarcity. A customized Chain-of-Thought (COT) prompt was designed to generate flight plans based on synthetic flight-planning scenarios (Tabrizian et al.). In (Gould et al., 2025), a multi-prompting framework, AirTrafficGen, was developed to automatically generate ATC scenarios to assist ATCOs' training. For scenario generation, non-real static data as context was utilized. In another work (Fox et al., 2024), a synthetic ATC communication dataset was generated to enhance ATCOs' situational awareness by identifying anomalous communications. A total of 1,100 conversations were created, covering 550 normal and 550 anomalous situations, by giving the LLM a single prompt using fake callsigns and runways, with a minimum length of 2,000 words. A prompt was provided with role, task, and constraints, like word count, to generate synthetic data.

Based on the previous work, it is found that the existing aviation-specific research lacks the contextual and constraint depth in synthetic data generation. Consequently, there exists limited work on leveraging the reasoning capabilities of LLMs to generate synthetic data for zero-resource airspace like India. India has experienced fast domestic flight growth in recent years (Arardhi et al., 2025). The increasing workload of ATCos and pilots demands automation aid in every possible way. Therefore, there is a strong need to develop a corpus like ATComm to initiate research and development activities. This work addresses this zero-resource gap and lack of contextual and constraint depth in existing work by proposing the Pro-SAC framework. Unlike previous work (Fox et al., 2024), the proposed framework integrates real callsigns, navigational routes based on comprehensive taxonomy of scenarios with a list of constraints grounded in Indian FIR, leveraging LLM to generate high-fidelity synthetic data.

3 The ATComm Corpus Generation

3.1 The Pro-SAC Generation Framework

In this section, we propose a Pro-SAC data generation framework. This framework utilizes prompt technique to generate high-fidelity data for Indian airspace. It transforms the LLM from unconstrained text generation into a constraint-aware engine rooted in the aviation context and specific to the Indian FIR. We define the objective as to generate optimal text sequences $\hat{y}$ that represent realistic ATCO and pilot conversations:

$$\hat{y} = \arg\max_{y \in Y} P(y|r, t, c, s, o, e) \quad (1)$$

where, y is a candidate dialogue sequence, Y is the space of all possible ATC dialogues, r is the role specification, and t is the task description, c is the flight context, s is the set of scenarios, o is the set of operational constraints, and e represents few-shot examples of ATCOs and pilot conversations.

3.1.1 Role, r and Task Specification, t

For ATComm corpus generation, we first define the role r specific to ATC. We define the role as an expert Indian ATCO working in Indian airspace. This helps to restrict the knowledge of LLM to Indian ATC only. Then, we specify the task, t to generate relevant data. We ask LLM to generate realistic, multi-turn utterances between ATCOs and pilots within Indian airspace.

3.1.2 Context Definition, c

Instead of providing the LLM with a generic flight plan or artificial data (Fox et al., 2024), we use real callsigns and navigational routes of Indian airspace. The context, c is derived from the structured flight plan containing necessary metadata for data generation:

$$c = (c_{\text{sid}}, c_{\text{star}}, c_{\text{wp}}, c_{\text{call}}, c_{\text{orig}}, c_{\text{dest}}) \tag{2}$$

where,

- $c_{\text{sid}} = \langle \text{SID}_1, \text{SID}_2, \ldots, \text{SID}_{n_s} \rangle$ represents ordered sequence of standard instrument departure (SID) points, presenting flight route immediately followed after takeoff.
- $c_{\text{star}} = \langle \text{STAR}_1, \text{STAR}_2, \ldots, \text{STAR}_{n_s} \rangle$ represents ordered sequence of standard terminal arrival route (STAR) points, presenting route immediately followed after transitioning from enroute to landing phase.
- $c_{\text{wp}} = \langle \text{wp}_1, \text{wp}_2, \ldots, \text{wp}_{n_s} \rangle$ represents ordered sequence of waypoints, geographical coordinates to define navigational route
- $c_{\text{call}} \in \mathcal{A} \times \mathcal{N}$, where c_{call} represents callsign associated with $\mathcal{A}$, set of airline operators and $\mathcal{N}$, flight numbers
- c_{orig}, c_{dest} represent origin and destination airports

It is worth noting that all the aforementioned metadata, are collected from official sources, Airport Authority of India (AAI) [1] and Directorate General of Civil Aviation (DGCA) [2], the regulatory bodies of Indian aviation. The context used to generate the corpus included 81 SIDs, 116 STARs, 468 waypoints, and around 2300 unique flight identifiers. It is to be highlighted that we have only considered domestic flights whose origin and destination airports are within Indian airspace for generating the ATComm corpus.

3.1.3 Operational Scenarios Design, S

To ensure that the synthetic corpus generated is representative of real-world operations, a structured scenario taxonomy is implemented to cover 6 major operational scenarios:

$$s = \bigcup_{\phi \in \Phi} s_\phi \tag{3}$$

where, $\Phi = \{$DEPARTURE, ARRIVAL, EN-ROUTE, NON-NORMAL, VFR, EMERGENCY$\}$

It is to be highlighted that existing datasets on ATC mainly contain data based on routine operations (arrival, departure, and enroute) (Badrinath and Balakrishnan, 2022) (Hofbauer et al., 2008) (Zuluaga-Gomez et al., 2023). In our taxonomy, we have also focused on data covering emergency, non-normal, and VFR scenarios, which are rare; however, they are important for future model training. Total 109 scenarios is designed to cover the entire range of flight phases and a wide variety of operational situations, from routine to critical.

3.1.4 Constraints Set, o

ATC is a strictly deterministic domain, and LLMs are probabilistic in nature. Therefore, to reduce the ambiguity in the generated data, the search space of the LLM is restricted to those outputs that strictly satisfy the regulations of Indian airspace. To ensure that ATComm corpus maintains high procedural fidelity, the generation process is aided with multiple constraints. Instead of an exhaustive simulation of all aviation-based rules and regulations, we have employed some of the core operational directives mentioned in the AAI Manual of Air Traffic Services (MATS) Part 1 [3]. The constraints are defined as:

$$o = o_{\text{ling}} \cup o_{\text{sync}} \cup o_{\text{oper}} \tag{4}$$

where,

- o_{phras} ensures that the generated data should follow the strict phraseologies of ATC

[1] https://aim-india.aai.aero/aip-supplements

[2] https://www.dgca.gov.in/digigov-portal/?baseLocale=en_US?page=4201/4200/servicename

[3] https://aim-india.aai.aero/ais-document

- o_{sync} covers the strict syntactic structure being followed by both controllers and pilots

- o_{oper} ensures that the correct operational sequence of navigational points should be followed to avoid wrong routes and hallucinated entities

We employ around 20 constraints during data generation, which mainly focus on the semantic and syntactic structure of ATC.

3.1.5 Few-Shot Examples, e

Since this data generation is highly domain-specific where ATC and pilots communicate in a structured way, we provide few examples in the prompt. For every operational scenario, we give an example of ATCO-pilot conversations in a particular template. These examples aid LLM to mimic the structure of real ATC communication and generate the data accordingly.

3.2 Experimental Setup

The ATComm corpus is generated using Gemini 1.5 Flash model (Team et al., 2024). The Gemini 1.5 Flash model is selected due to its optimized high-throughput architecture and its 1-million-token context window, which facilitates high-fidelity few-shot in-context learning. This capability enabled the Pro-SAC framework to address the zero-resource gap by generating 65k utterances at scale, while ensuring that the stochastic generation remained strictly constrained by the deterministic context and constraints of the Indian FIR. Data were generated using the Google AI API, with the model's default configuration (temperature T = 1.0, Top-p = 0.95, Top-k = 40). These settings were intentionally used to preserve the linguistic diversity of the corpus while maintaining the procedural integrity of the Pro-SAC framework. It is to be noted that the ATComm corpus was generated over a four-month of iterative lifecycle of generation, audit, and refinement. The data was generated into different batches based on scenarios and corresponding context. After each batch completion, an automated audit was performed as validation. If any violation found, then corresponding metadata was re-processed to Pro-SAC framework to ensure 100% compliance.

4 The ATComm Corpus Evaluation

The corpus is evaluated on three dimensions: fidelity, validity, and utility.

4.1 Intrinsic Linguistic Fidelity

Synthetic data generation often suffers from mode collapse, leading to lower diversity than that observed in the original data distribution (Seddik et al., 2024). In ATC communication, a baseline level of repetitions is mandated protocol. Even though a high-fidelity dataset must exhibit diversity in variable slots such as callsigns, SID-STAR points, waypoints, frequencies and syntactic variations in emergency and non-normal scenarios. We benchmark ATComm against two open-source dataset: LDC94S14A (Badrinath and Balakrishnan, 2022), and ATCO2 (Zuluaga-Gomez et al., 2023), containing real-word recordings of ATCOs and pilots. For the study, we utilize the corresponding transcriptions (manually transcribed text) provided online [4] [5]. LDC94S14A and ATCO2 contains data collected from airports in US and European airspace respectively. A sample of both corpora is made available as an open-source for research purposes. Table 1 outlines the dataset statistics of the corpora used for the analysis.

Dataset Statistics	ATComm	LDC94S14A	ATCO2
No. of Utterances	65k	1027	874
Vocabulary Size (V)	13,696	753	1257
No. of Tokens (N)	922,785	7411	7397

Table 1: Dataset Statistics Across ATComm and the Benchmark Corpora LDC94S14A and ATCO2.

To measure the lexical richness of the ATComm corpus, we calculate Type-Token Ratio (TTR) and Moving Average TTR (Zenker and Kyle, 2021). TTR measures the linguistic diversity by measuring the richness of the vocabulary relative to corpus size.

$$\text{TTR} = \frac{\text{Number of Unique Words (V)}}{\text{Number of Total Words (N)}} \quad (5)$$

where, V is the vocabulary representing the count of unique word types present in the corpus, and N is the total count of tokens in the corpus. Since TTR is sensitive to the length of the text (Cunningham and Haley, 2020), we calculate MATTR to provide a more robust characterization of lexical richness of ATComm corpus. MATTR provides a length-independent measure of lexical diversity by averaging TTR over a sliding window, W (Cunningham and Haley, 2020). We select two windows of sizes 50 and 100 to capture diversity within a

[4] https://catalog.ldc.upenn.edu/LDC94S14A
[5] https://www.atco2.org/data

short sequence of ATCO-pilot exchange and across a long sequence of exchanges between them. As shown in Table 2, MATTR decrease for all three corpora when expanding the window size from $W = 50$ to $W = 100$ due to a higher probability of token repetition. However, the relative ranking remains consistent at both window sizes (ATComm $<$ LDC94S14A $<$ ATCO2). This stable trend demonstrates that our synthetic corpus has repetitive tokens in the same way that real human controllers and pilots actually do.

To assess the semantic distinctness across all three corpora, we utilize Inter-Cosine Similarity metric (Ke et al., 2022). While callsigns, navigational points etc. differ, the underlying procedural meaning must remain consistent in all three datasets. We represent each utterance as a high-dimensional vector using a pre-trained BERT encoder. We then compute the mean cosine similarity between the corpus embeddings in a high-dimensional vector space. As shown in Table 2, it is found that ATComm achieves an inter-cosine score of 0.9489, very close to the ATCO2 of 0.9454 and LDC94S14A of 0.9503. It proves that despite the regional variation in entities, the semantic relationship, instruction structure and intents in ATComm are indistinguishable from real-world operational corpus. This ensures that the models trained on the ATComm corpus will learn the universal ATC knowledge while becoming robust to the specific linguistics of the Indian airspace.

To evaluate whether the synthetic ATComm corpus statistically resembles actual ATC data or not, we assess its adherence to Zipf's Law (Klubička et al., 2019). We perform a linear regression fit of the log-transformed word frequencies against their ranks, extracting the scaling exponent (α) and the coefficient of determination (R^2) as goodness-of-fit. As illustrated in Fig. 2, ATComm shows an exceptionally strong linear alignment profile ($R^2_{\text{ATComm}} = 0.978$), closely mirroring the structural behaviour of human-transcribed operational benchmarks (LDC94S14A: 0.932 and ATCO2: 0.965). The scaling exponent ($\alpha_{\text{ATComm}} = 1.476$) exhibits that ATComm is highly diversified in comparison to ATCO2 ($\alpha = 1.077$) followed by LDC94S14A ($\alpha = 1.318$). This curve explicitly confirms that the synthetic-language generation framework preserves the natural linguistic frequency properties of real-time ATC datasets.

Furthermore, we calculate the lexical overlap between ATComm and other corpora. A Jaccard

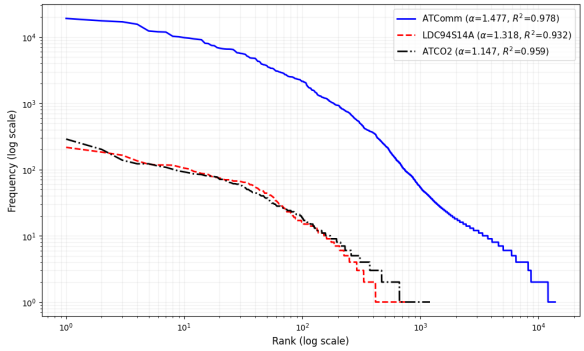

Figure 2: Log-log distribution of word frequency versus rank across ATComm, LDC94S14A, and ATCO2.

similarity score is calculated between two corpora to find the similarity between them (Qurashi et al., 2020). This metric is calculated to show that ATComm contains a distinct vocabulary of Indian domestic callsigns and regional fixes that are completely absent in the Western corpus. We observe a low Jaccard similarity score of 0.0272 between ATComm and LDC94S14A and 0.0360 between ATComm and ATCO2 (Table 2). This score shows that ATComm is not a reproduction of existing Western datasets; it is, in itself, a major data source representing Indian airspace that is not part of existing datasets. It is interesting to note that while the lexical overlap between LDC94S14A and ATCO2 (0.1634) is higher than their respective overlaps with ATComm, it remains overall moderately low. This highlights that although ATC communication strictly follows standardized ICAO phraseologies, individual corpora differ significantly due to localized entities, such as region-specific aircraft callsigns, domestic airports, and distinct regional navigational fixes.

4.2 Operational Validity

To ensure that the ATComm corpus is operationally sound and mimics the real ATC-pilot communication, we validate the corpus quantitatively and qualitatively.

4.2.1 Quantitative Readback Consistency

To evaluate the functional integrity of the ATComm corpus, a custom readback validator is designed to verify the adherence of responses of instructions given. ATC communications heavily rely on a strict readback loop to mitigate operational risk. For every instruction, the script identifies target entities like callsign, value, navigational points etc. in the corresponding readback. The readback consis-

Metrics	ATComm	LDC94S14A	ATCO2
TTR	0.0704	0.1016	0.1699
MATTR ($W = 50$)	0.7138	0.7613	0.7917
MATTR ($W = 100$)	0.5981	0.6417	0.7112
Inter-Cosine Similarity	0.9489	0.9503	0.9454
Jaccard Similarity (ATComm and LDC94S14A)	1	0.0272	-
Jaccard Similarity (ATComm and ATCO2)	1	-	0.0360

Table 2: Intrinsic Linguistic Fidelity Analysis Across ATComm and Benchmark Corpora LDC94S14A and ATCO2.

tency is evaluated using the readback accuracy rate (RAR), defined as

$$RAR = \frac{\sum \text{Correct Readbacks}}{\sum \text{Total Instructions Issued}} \times 100 \quad (6)$$

It is found that ATComm achieves RAR of 98.93 % across 80,655 issued instructions. It is to be noted that there are slight variances in some categories as some of the utterances are initiated by pilot which is not necessary for ATCOs to readback. Therefore, in such cases, readbacks are counted as zero. The high accuracy rate proves that the proposed Pro-SAC framework successfully maintained the closed-loop feedback required for operational realism.

4.2.2 Qualitative Expert Assessment

By using a human-in-the-loop approach, the subset of the corpus is evaluated by a ATC SME to ensure that the generated utterances reflect real ATCO-pilot communication for Indian airspace. A standard 5-point Likert scale is used to measure the quality of the generated data (Amidei et al., 2019) (Arias et al., 2025), where 1 represents excellent quality and 5 represents poor quality of text. A randomized, stratified validation subset comprising 50 dialogues per operational scenario, yielding a total evaluation pool of 360 dialogues (1k utterances). This sample distribution ensures a rigorous representative of the corpus while remaining operationally feasible for blind expert review. The expert was asked to rate the sampled dialogues in terms of naturalness, phraseological compliance, and operational intent on a Likert scale. Naturalness evaluates the conversational realism in the text. In ATC, controllers and pilots maintain certain linguistic characteristics while communicating each other. This makes the conversation sound more natural in the ATC environment. Phraseological compliance is one of the essential factors to validate. As discussed earlier, controllers and pilots are restricted to use specific standardized phraseologies. Therefore, it is important to evaluate whether the generated corpus complied to such phraseologies or not. Operational intent assesses the procedural logical validity of the text. Table 3 presents the overall rating given by the expert in terms of aggregated mean (μ) of navigational point-dependent and independent scenarios (arrival, departure, and enroute).

Operational Scenarios	Naturalness	Compliance	Intent
Departure	1.0	1.5	1.0
Arrival	1.0	1.5	1.0
Enroute	1.0	1.5	1.0
Non-Normal	1.0	2.0	1.0
VFR Scenarios	1.0	2.0	2.0
Emergency	1.0	2.0	2.0
Mean (μ)	**1.0**	**1.75**	**1.33**

Table 3: Qualitative Evaluation of ATComm based on Likert Scale by Aviation Expert.It represents the aggregated mean (μ) of dependent and independent scenario evaluations (1 = Excellent, 5 = Very Poor).

The samples are found to achieve ($\mu = 1.0$) for maintaining ATC-specified linguistic characteristics, making the text look highly natural. Text based scenarios like departure, arrival, and enroute are proven to be highly intended towards operational logics ($\mu = 1.0$) while maintaining strong strong phraseological compliance ($\mu = 1.5$). Expert has observed that there is a slight degradation in compliance because of some missing information in location based text. For example, distance information is missing while giving traffic information. Notably, even across edge-case scenarios like Non-normal, VFR, and Emergency, the generation framework maintains structural fidelity ($\mu \leq 2.0$), despite minor violations highlighted by the expert like using 'Roger your May Day' instead of 'Roger May Day' or vectoring anomalies during rapid decompression. These aggregated results confirms that the pipeline preserves domain realism in the synthetic corpus.

	ATCO2 BERT Baseline (Zero-Shot)			T5-Base (Fine-Tuned on ATComm)		
Entity Class	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Callsign	0.7523	0.8748	0.8089	0.9996	0.9921	0.9958
Command	0.3372	0.7623	0.4675	0.9999	0.9906	0.9952
Value	0.6760	0.8827	0.7657	0.9987	0.9708	0.9846
Average	**0.6480**	**0.8636**	**0.7348**	**0.9992**	**0.9794**	**0.9892**

Table 4: NER Performance Comparison on the ATComm Test Split ($n = 6.5$k).

4.3 Downstream Utility

To evaluate the ATComm corpus's extrinsic value, we conduct an NER task on it. ATC communication contains structured information like callsigns, commands, and values. Extracting such named entities can be helpful for safety-critical tasks such as automated trajectory generation, flight strip updates, safety report analysis, etc. (Singh et al., 2025b) (Helmke et al., 2018). This study assesses the corpus's effectiveness in training models for its future real-world applications. For benchmark, we utilize the model trained on ATCO2 corpus [6] as a cross-corpus baseline model to evaluate its performance on our synthetic data. The baseline model utilizes BERT architecture, and we leverage generative T5-base model [7]. As demonstrated by (Abilio et al., 2024), while encoder-only architectures classify individual tokens in isolation and suffer from sub-word boundary alignment constraints, generative sequence architectures leverage full autoregressive cross-attention to capture long-range semantic dependencies. We adopt a generative sequence-to-sequence approach with T5. The raw textual utterances are wrapped in semantic markers based on callsign, command, and value. The ATComm corpus is partitioned into train (80%), validation (10%), and test (10%) sets.

The results detailed in Table 4 establish an exceptional operational baseline, where T5 fine-tuned on ATComm achieves an F1-score of 0.9990 with symmetrical precision and recall across all categories (0.9992 for *Callsign*, 0.9994 for *Command*, and 0.9987 for *Value*). This empirical consistency demonstrates that ATComm serves as an optimized training resource perfectly suited for real-time automated speech and language understanding system-based applications. On the other hand, a regional generalization gap emerges during the zero-shot evaluation of the model trained on the Western data, ATCO2 baseline. This performance degradation primarily stems from the vocabulary limitations. As the baseline model was trained on only 1 hour of European airspace data, it fails in recognizing localized Indian callsigns (e.g., Ifly, Akasa Air), and domestic navigational identifiers like regional SIDs, STARs, waypoints and airports. Furthermore, the remarkably low F1-score for the *Commands* (0.4675) highlights a difference in phraseologies on which both models are trained. As the ATCO2 corpus has a limited vocabulary, it is incapable of generalizing to the rich lexical diversity of the ATComm. These results demonstrate that even though the model is trained on ATC-specific data, it cannot safely generalize to complex regional airspaces, confirming the necessity of a high-volume ATComm corpus.

5 Conclusion and Future Work

In this work, we present ATComm, a large-scale, high-fidelity textual synthetic corpus of 65k ATCO and pilot utterances tailored to Indian airspace. By employing the Pro-SAC framework, we demonstrate that modular prompt synthesis can bridge the gap between structured metadata and linguistic characteristics of ATC. The close MATTR and inter-cosine similarity score acorss benchmark corpora confirm that ATComm reflects the linguistic and semantic characteristics of real-time data. Zipf's curve confirms that ATComm matches the natural linguistic frequency properties of real-time ATC corpora. The benchmark study shows that the model trained on ATComm data recognized Indian airspace better than the model trained on other airspace data. In future, we aim to expand the corpus to include international flights and flights overflying Indian airspace. We plan to develop speech recognition and language understanding systems curated for the Indian airspace, utilizing the ATComm corpus.

[6] https://huggingface.co/Jzuluaga/bert-base-ner-atc-en-atco2-1h

[7] https://huggingface.co/google-t5/t5-base

Limitations

While the ATComm corpus establishes a high-volume, linguistically validated resource for the underrepresented Indian airspace, it is mainly developed for training and research purposes. It should not be used directly in live, real-world operational environment without human oversight. There are several limitations which bound the scope of this work. A key limitation is that the baseline ATCO2 model was pre-trained on a relatively small set of regional European dataset. Consequently, when evaluated on a larger test set of ATComm, the model shows significant performance degradation on the NER task. This performance gap highlights that the model was constrained by its initial low-resource training. It suggests that training the model on a larger, more diverse dataset can improve its performance across different airspace datasets as well. This aspect is currently not explored in this study. As our main objective is to generate linguistic resources for future model training on Indian ATC entities, we incorporate limited scenarios and constraints into the data-generation process. There is a risk of aviation-related kinematic violations in the generated data. The current dataset does not include any acoustic and environmental channel noise such as radio static, cockpit background alarms, or fillers. This can degrade the performance of the downstream task when tested with real like environments. While generating data, we only consider domestic flights; however, flights passing over Indian airspace and international flights also contain knowledge that is currently absent from the ATComm corpus.

References

Ramon Abilio, Guilherme Palermo Coelho, and Ana Estela Antunes da Silva. 2024. Evaluating named entity recognition: A comparative analysis of mono- and multilingual transformer models on a novel brazilian corporate earnings call transcripts dataset. *Applied Soft Computing*, 166:112158.

Aisha Alansari and Hamzah Luqman. 2026. Large language models hallucination: A comprehensive survey. *Preprint*, arXiv:2510.06265.

Wedyan Alsakran and Reham Alabduljabbar. 2024. Exploring the potential of llms and attributed prompt engineering for efficient text generation and labeling. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 244–252.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

Venkata Praveen Arardhi, Sayan Hazra, Utkarsha Singh, P. K. Nithya, Suresh Sundaram, K. P. Sooraj, P. K. Menon, and Radhakant Padhi. 2025. Dynamic air traffic model for controlled large airspace. *Journal of Aerospace Information Systems*, 22(9):762–774.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Matthias Aßenmacher, and Christoph Jansen. 2025. Statistical multicriteria evaluation of llm-generated text. *Preprint*, arXiv:2506.18082.

Sandeep Badrinath and Hamsa Balakrishnan. 2022. Automatic speech recognition for air traffic control communications. *Transportation Research Record*, 2676(1):798–810.

Kevin T. Cunningham and Katarina L. Haley. 2020. Measuring lexical diversity for discourse analysis in aphasia: Moving-average type–token ratio and word information measure. *Journal of Speech, Language, and Hearing Research*, 63(3):710–721.

Ashraf Elnashar, Jules White, and Douglas C Schmidt. 2025. Enhancing structured data generation with GPT-4o evaluating prompt efficiency across prompt styles. *Front. Artif. Intell.*, 8:1558938.

Timofey Fedoseev, Dimitar Iliev Dimitrov, Timon Gehr, and Martin Vechev. 2024. LLM training data synthesis for more effective problem solving using satisfiability modulo theories. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Kevin L. Fox, Kevin R. Niewoehner, Mark Rahmes, Josiah Wong, and Rahul Razdan. 2024. Leverage large language models for enhanced aviation safety. In *2024 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, pages 1–11.

Kathleen Ge and William Coupe. *Aviation-Specific Large Language Model Fine-Tuning and LLM-as-a-Judge Evaluation*.

Dewi Sid William Gould, George De Ath, Ben Carvell, and Nick Pepper. 2025. Airtrafficgen: Configurable air traffic scenario generation with large language models. *Preprint*, arXiv:2508.02269.

Hartmut Helmke, Michael Slotty, Michael Poiger, Damián Ferrer Herrer, Oliver Ohneiser, Nathan Vink, Aneta Cerna, Petri Hartikainen, Billy Josefsson, David Langr, Raquel García Lasheras, Gabriela Marin, Odd Georg Mevatne, Sylvain Moos, Mats N. Nilsson, and Mario Boyero Pérez. 2018. Ontology for transcription of atc speech commands of sesar 2020 solution pj.16-04. In *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, pages 1–10.

Konrad Hofbauer, Stefan Petrik, and Horst Hering. 2008. The ATCOSIM corpus of non-prompted clean air traffic control speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Liang Ke, Xinyu Chen, and Haizhou Wang. 2022. An unsupervised detection framework for chinese jargons in the darknet. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 458–466, New York, NY, USA. Association for Computing Machinery.

Filip Klubička, Alfredo Maldonado, Abhijit Mahalunkar, and John Kelleher. 2019. Synthetic, yet natural: Properties of WordNet random walk corpora and the impact of rare words on embedding performance. In *Proceedings of the 10th Global Wordnet Conference*, pages 140–150, Wroclaw, Poland. Global Wordnet Association.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2024. One-shot learning as instruction data prospector for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Bangkok, Thailand. Association for Computational Linguistics.

Yi Lin. 2021. Spoken instruction understanding in air traffic control: Challenge, technique, and application. *Aerospace*, 8(3).

Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3001–3004, New York, NY, USA. Association for Computing Machinery.

Yucheng Liu. 2024. Large language models for air transportation: A critical review. *Journal of the Air Transport Research Society*, 2:100024.

Aziida Nanyonga, Keith Joiner, Ugur Turhan, and Graham Wild. 2025. Applications of natural language processing in aviation safety: A review and qualitative analysis. *Preprint*, arXiv:2501.06210.

Nzubechukwu C. Ohalete, Kevin B. Gittner, and Lauren M. Matheny. 2025. Costar-a: A prompting framework for enhancing large language model performance on point-of-view questions. *Preprint*, arXiv:2510.12637.

Abdul Wahab Qurashi, Violeta Holmes, and Anju P. Johnson. 2020. Document processing: Methods for semantic text similarity analysis. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6.

Haris Riaz, Sourav Sanjukta Bhabesh, Vinayak Arannil, Miguel Ballesteros, and Graham Horwood. 2025. MetaSynth: Meta-prompting-driven agentic scaffolds for diverse synthetic data generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18770–18803, Vienna, Austria. Association for Computational Linguistics.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. *Preprint*, arXiv:2404.05090.

Utkarsha Singh, Margamitra Bhattacharya, and Radhakant Padhi. 2025a. A critical analysis of end-to-end asr models for air traffic control speech. In *2025 7th International Conference on Artificial Intelligence and Speech Technology (AIST)*, pages 1–6.

Utkarsha Singh, Margamitra Bhattacharya, and Radhakant Padhi. 2025b. State-of-the-Art natural language processing for aviation: A review. *TechRxiv*.

Luboš Šmídl. 2011. Air traffic control communication. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Amin Tabrizian, Mahyar Ghazanfari, and Peng Wei. *Chain-of-Thought Flight Planner: End-to-End LLM Routing Under Wind Hazards*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Chuyang Yang and Chenyu Huang. 2023. Natural language processing (nlp) in aviation safety: Systematic review of research and outlook into the future. *Aerospace*, 10(7).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fred Zenker and Kristopher Kyle. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47:100505.

Minghua Zhang, Yang Yang, Shengsheng Qian, Qihan Deng, Jing Fang, and Kaiquan Cai. 2025. Atsiu: A large-scale dataset for spoken instruction understanding in air traffic control. *Advanced Engineering Informatics*, 65:103170.

Juan Zuluaga-Gomez, Karel Veselý, Igor Szöke, Alexander Blatt, Petr Motlicek, Martin Kocour, Mickael Rigault, Khalid Choukri, Amrutha Prasad, Seyyed Saeed Sarfjoo, Iuliia Nigmatulina, Claudia Cevenini, Pavel Kolčárek, Allan Tart, Jan Černocký, and Dietrich Klakow. 2023. Atco2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications. *Preprint*, arXiv:2211.04054.

A Appendix

A.1 Data Annotation

For NER task, we prepared the dataset in a particular format required by T5 model. T5-based model requires a source text and the target text, where target text is the annotated with required labeled named entities. To prepare labeled dataset, a custom tokenizer was developed based on the requirements of the task. In our case, for successful NER task, three named entities, 'callsign', 'command', and 'value' should be correctly identified from the given text. For callsign, we prepared a list of callsigns initials and wrote a regex expression so that numbers following those initials of callsign were tagged inside callsign entity. Then, for command entity, first we separated the commands from the baseline ATCO2 corpus and then, augmented them with the commands in the ATComm corpus. This was mainly done because in ATCO2 corpus, some of the commands were tagged in a different way, so we wanted a fair comparison, that's why we followed ATCO2 pattern of tagging commands. For tagging values, we wrote multiple regex expressions like numbers being followed by feet, and knots should be tagged inside value entity, phrases like flight level, qnh followed by numbers should be tagged inside value entity. The custom tokenization was performed and resulted into a json file that maps the source text to its target text as shown in Fig. 3. To illustrate how does ATCOs and pilots communicate, we have also attached an example in Fig. 4.

A.2 NER Evaluation: Training Details

For reproducibility, we document the hardware infrastructure, and the list of hyperparameter utilized during the fine-tuning of T5 model. For ATC-specifc NER adaptation, we employed full supervised fine-tuning. All experimental training and test validation were implemented in Python and executed on Google Colaboratory's compute infrastructure. All the related information are explicitly detailed in Table 5.

Hyperparameters	**Values**
Model Architecture	T5-base
Number of Parameters	220M
Tokenizer Configuration	T5Tokenizer (provided with custom tags)
Hardware Compute Platform	NVIDIA A100 GPU
Optimizer	AdamW
Learning Rate	2×10^{-4}
Batch Size	16
Maximum Sequence Length	128 Tokens
Number of Epochs	8
Random Seed	42
Loss Objective	Categorical Cross-Entropy
Inference Generation Strategy	Beam Search (Beams = 2)
Early Stopping State	Enabled

Table 5: Hyperparameters Employed During Training and Validating the ATComm Corpus.

A.3 Scenario Taxonomy

To generate diverse data, we design a scenario taxonomy in which each operational scenario is mapped to its intent. These intents provide the different phases of the flight to the Pro-SAC framework for generating a diverse dataset. These sub-scenarios are further categorized into point-dependent and point-independent scenarios. Scenarios like arrival, departure, and enroute primarily include navigational points such as STARs, SIDs, and waypoints to guide traffic. Therefore, these scenarios are further split into point-independent and point-dependent categories. This categorization of scenarios assists the Pro-SAC framework to generate procedurally correct data without hallucinating between point-independent and point-dependent routes.

"source": "ifly seven one eight climb and maintain flight level three six zero",

"target": "[#callsign] ifly seven one eight [/#callsign] [#command] climb and maintain [/#command] [#value] flight level three six zero [/#value]"

Figure 3: Annotation Scheme Followed for NER Task.

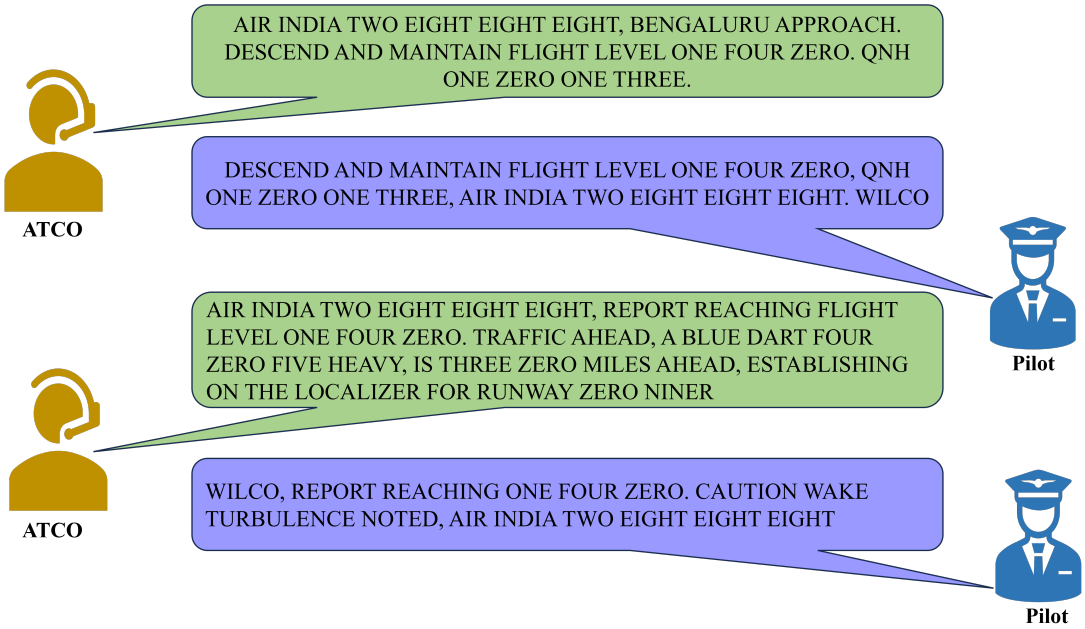

Figure 4: Sample of Multi-Turn Dialogue Exchange between ATCO and Pilot.

Evaluation	**Departure**		**Arrival**		**Enroute**		**Others**
Metric	**Point-dep**	**Point-indep**	**Point-dep**	**Point-indep**	**Point-dep**	**Indep**	**Non-Normal | VFR | Emergency**
Naturalness	1.0	1.0	1.0	1.0	1.0	1.0	1.0 | 1.0 | 1.0
Compliance	1.0	2.0	1.0	2.0	1.0	2.0	2.0 | 2.0 | 2.0
Operational	1.0	1.0	1.0	1.0	1.0	1.0	1.0 | 2.0 | 2.0
Mean (μ)	**1.0**	**1.33**	**1.0**	**1.33**	**1.0**	**1.33**	**1.33 | 1.67 | 1.67**

Table 6: Consolidated SME Assessment Across Different Operational Scenarios (1 = Excellent, 5 = Very Poor).