Semantics-Aware Prompting for Translating NOtices To AirMen

Minal Nitin Dani
IIT Hyderabad, Honeywell
cs16mds19p1001@iith.ac.in

Aishwarya Maheswaran
IIT Hyderabad
ai21resch11002@iith.ac.in

Maunendra Sankar Desarkar
IIT Hyderabad
maunendra@cse.iith.ac.in

Abstract

A NOTAM or NOtice To AirMen is a crucial notice for different aviation stakeholders, particularly flight crews. It delivers essential notifications about abnormal conditions of Aviation System components such as changes to facilities, hazards, service, procedure that are not known far enough in advance to be publicized through other means. NOTAM messages are short, contain acronyms, and look cryptic in most of the cases. Writing and understanding these messages put heavy cognitive load on its end users. Since NOTAMs do not adhere to English grammar rules and have their own decoding rules, large language models (LLMs) cannot translate them without effective prompting. We introduce a framework to effectively prompt LLMs for translating NOTAMs. The results demonstrate that our methodology can produce clear translations that accurately convey the information contained in NOTAMs.

1 Introduction

NOTAMs are essential communications within the aviation domain. Details of flight path and pilot briefings are conveyed through NOTAMs. They are constructed in specialized language that adheres to specific domain rules and templates. This unique format, characterized by a variety of abbreviations and short forms, necessitates a dedicated effort to decode and interpret NOTAMs accurately. Having a translated version of the NOTAM reduces the strain on the user to decode it from scratch and makes it easier for the Subject Matter Experts (SME) to confirm its accuracy. As per reported statistics, around 35,000 NOTAMs are used for communication in the global air transport system every day.

In this work, we propose Context-Aware Semantic Prompting approach for translating NOTAMs. We employ k-shot prompting, where the examples incorporated into the prompt are carefully selected using the proposed algorithm that account for both the syntactic and semantic characteristics of the input NOTAM. We reached out to NOTAM users, including pilots and subject matter experts (SMEs) to get the gold reference translations which we used to evaluate the quality of our translations. The impact of our proposed method is as follows:

- Simplification of NOTAMs: Our approach aims to make NOTAMs simpler by translating them into plain English, addressing the challenge posed by their domain-specific decoding rules. This is possibly the first publicly available effort for NOTAM to English translation.
- Low-Resource Environment Compatibility: We develop an algorithm that can run in low-resource environments and provide comparable accuracy with human translation, making it suitable for use in handheld devices and cockpits for flight system management.
- Comprehensive Evaluation: We test our methodology on six language models (of varying sizes) in five unique experimental settings. We evaluate with various evaluation metrics (precision, recall, alignment and synonymy, similarity between translated and human-translated text, and the fluency of the translation) to assess the quality of the translations.

2 Related Works

Due to the safety-critical and proprietary nature of the task of translating NOTAMs combined with the limited accessibility to datasets, there has not been significant research towards automated translation of NOTAMs. In this Section, we focus on existing works on NOTAMs.

Existing research on NOTAM translation is limited to a few specific types, and are not exhaustive. In (Liu, 2024), the authors consider summarizing Taxiway and Runway closure type NOTAMs.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24407–24417

Abbreviation	Full form
NOTAM	Notice to Airmen
SME	Subject Matter Experts
FAA	Federal Aviation Administration
ICAO	International Civil Aviation Organization
IATA	International Air Transport Association
TWY	Taxiway
RWY	Runway
NAV	Navigation
AD/AP	Aerodrome/Airport
COM	Communication
OBST	Obstacle

Table 1: Frequently used abbreviations in the paper

The work commented that ChatGPT is the best online tool for NOTAM translation. (Arnold et al., 2022) focuses on information extraction from NOTAMs. It considers translating NOTAMs as one of the tasks. However, for translation, they convert NOTAM to a specific format called AirLang format and not to English language. They trained an encoder-decoder model using 20,000 pairs of manually annotated data, concentrating on specific entities present in taxiway closure NOTAMs. The approach requires a significant amount of training data as they train the model from scratch. On the other hand, LLMs can work well with zero-shot prompting, enabling us to get the translations without any specific training. This is useful as access to domain experts is difficult in this specific scenario.

Although not for translation, processing NOTAM data for various other tasks have been considered in a few recent works. The work in (Arnold et al., 2022), also discussed above, considered Named Entity Recognition and Criticality Prediction of NOTAMS. In (Dani and Desarkar, 2024), the authors perform segmentation of NOTAMs to help in easier understanding of the complex and cryptic messages. (Szeto and Das, 2024; Mogiłło-Dettwiler, 2024) propose methods for classification, filtering and sorting of NOTAMs. Although there is an increasing interest in automated processing of NOTAM data, *to the best of our knowledge, there is not existing published work that focuses on translation of NOTAMs to their corresponding English natural language versions*.

3 Motivation

NOTAMs were initiated in 1947. In the early days, bandwidth for communication channels might have been a concern, due to which the number of characters in the NOTAMs were kept low purposefully. Since then, NOTAMs have been an integral part of Air Traffic Communications. Using messages in such a format even in recent times has been mostly due to legacy issues among other reasons.

The Federal Aviation Administration (FAA) mandates clear and concise communication to ensure safety and efficiency in the aviation sector. The use of standardized abbreviations and codes were opined to (a) reduce the risk of misinterpretation and (b) ensure that critical information is conveyed accurately. It also ensures that NOTAMs are understood universally, regardless of language barriers. This is crucial for international aviation operations where the personnel from different parts of the world speak different languages and may have varied proficiency in a single medium of communication such as English. Having said that, while introduced with clarity in communication in mind, over the time NOTAMs have evolved to have a convoluted abbreviation system, lack of standardization, and ever expanding volume that pilots find difficult to keep up with (Szeto and Das, 2024). In response to these challenges, FAA has initiated efforts to deliver NOTAMs in plain language. However, approximately 30 percent of NOTAMs still lack digitization and translation, particularly those pertaining to airspace, leaving many NOTAMs available only in their traditional complex format. New pilots often need to consult manuals repeatedly to interpret NOTAMs correctly, while experienced personnel must stay updated on any new rules or abbreviations introduced in the NOTAM journal. Moreover, the considerable volume of NOTAMs included in a single briefing—ranging from 80 to 120 NOTAMs—can lead to significant information overload, where critical details may be overlooked due to the abbreviated nature of the messages. A complete shift from NOTAMs to messages in natural language requires a mega policy change at the Federal Aviation Administration level, which is quite difficult considering the scale of round-the-clock global flight operations covering multiple countries, airports and impacting a huge global population. Due to these factors, translation of NOTAMs into plain English language is an important task, which we take up in this paper. Due to the

Type	RWY	TWY	APRON	OBST	AIRSPACE	NAV	AD	SVC	COM
Count	1,41,132	94,793	57,149	39,231	19,743	17,152	14,255	7,344	3,067
Max token length	77	134	83	92	172	66	71	63	30
Max Overlap token length	65	91	70	92	156	51	15	63	21

Table 2: Different types of NOTAMs and their lengths in the final dataset

Raw	!MFR 02/002 OED AIRSPACE PJE WI AN AREA DEFINED AS 3NM RADIUS OF OED345004 (10NM N MFR) SFC-14000FT 2202061700-2202062359.
Preprocessed	!Rogue Valley International-Medford Airport (MFR) 02/002 Rogue Valley VOR AIRSPACE Parachute Jumping Exercise Within AN AREA DEFINED AS 3NM RADIUS OF OED345004 (10NM North MFR) SFC-14000FT February 06, 2022, 05:00 PM-February 06, 2022, 11:59 PM
Human Translated	Rogue Valley International–Medford Airport with notam 02/002 indicates that Rogue Valley VOR (OED) airspace is active for a parachute jumping exercise within an area defined as a 3 NM radius of OED345004 - a point located 4 nautical miles from the Rogue Valley VOR along the 345-degree radial (10 NM north of Rogue Valley International–Medford Airport), from surface up to 14,000 feet between February 06, 2022, 05:00 PM and February 06, 2022, 11:59 PM.

Table 3: An example NOTAM and its translation

nature of the domain, several abbreviations will come repeatedly in the discussions. We include a list of such abbreviations in Table 1 for benefit of the readers.

4 Dataset

In this work, we consider nine of the most common NOTAM categories[1]. These nine types of NOTAMs are of high significance and include Taxiway (TWY), Runway (RWY), Airspace, Apron, Aerodrome/Airport (AD/AP), Obstacle (OBST), Navigation (NAV), and Communication (COM) (Dani and Desarkar, 2024). Each of these NOTAMs has a unique composition.

We approached Subject Matter Experts (SMEs) of NOTAMs and obtained translations for 270 NOTAMs, with 30 translations from each considered category[2]. This is our test dataset used for inference. Additionally, we use 27 NOTAMs translation pairs (3 for each category) that are used for our One-Shot prompting methodologies with namely Random, Max Length and Max Overlap approaches (described in detail later in Section 5). There is a unique one-shot example for each category and each method, giving us the total as $9 \times 3 = 27$ one-shot examples.

We treat each category of NOTAM separately since every category has its own specific nature of attributes. Inside each category, the NOTAMs can be of varied nature depending on the specific information it covers. This fine-grained information is captured by the clustering. We use 50 NOTAMs to implement the semantic prompting methodology. As each NOTAM category has a different structure, the number of clusters was allowed to be different across categories. The exact number of clusters were determined using Silhouette scores. The number of clusters for the different NOTAM categories were as follows: 3 for TWY, 9 for SVC, 2 for OBST, 4 for NAV, 10 for COM, 2 for APRON, 8 for AIRSPACE, 9 for AD/AP, and 3 for RWY.

Different statistics for each NOTAM category considered in the work are presented in Table 2.

4.1 Data Pre-Processing

Our initial experiments using raw NOTAMs directly with language model (M_{lm}) revealed frequent errors in decoding abbreviated airport names, abbreviated text, and time-sensitive information. The NOTAM domain has different rules to decode timestamps and airport name conventions as per ICAO format. Since NOTAM messages are safety-related and missing information can lead to safety issues, to effectively decode timestamps and dates found within NOTAMs, we referred to guidelines provided in NOTAMs journals. We extracted full forms for the abbreviated phraseology used in NOTAMs, as outlined in FAA documents. Finally, we compile a comprehensive list of airport names along with their corresponding ICAO and IATA codes sourced from the ICAO. We implemented a preprocessing pipeline to decode airport names, abbreviations, and schedules(timestamp). An example pre-processed NOTAM along with its human translation is given in Table 3.

We did not try any data augmentation in this work. Data augmentation would be helpful in cases with stylistic variations in the augmented data. At the same time, we need to ensure that the resultant data should conform to the structural guidelines of

[1]In this work the terms '*NOTAM types*' and '*NOTAM categories*' have been used interchangeably.

[2]Our annotators were pilots or Air traffic control experts. So getting a large number of human labeled data was difficult

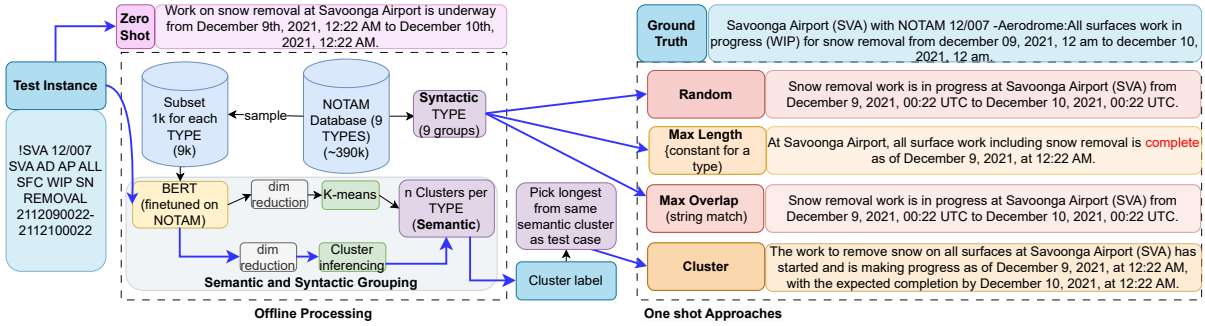

Figure 1: Overview of methods with steps followed to translate NOTAMs. Blue lines indicate the steps followed during the runtime. Figure is best viewed in color.

NOTAMs. As NOTAMs are designed to communicate safety-related messages, their structure must adhere to regulatory instructions. NOTAMs are digitized, and submitters need to use the NOTAM Manager for submission. If a NOTAM does not follow the prescribed structure or use the correct terminology, it gets rejected. Therefore, there is limited scope for using synonyms or paraphrasing. For example, in our dataset, we observed that if a runway is closed, the term "clsd" is consistently used, rather than alternatives like "shut" or "not operational." Consequently, we did not use augmentation in this work.

5 Method

Since NOTAMs adhere to strict rules for word usage and do not allow morphological variations, we use the word-level NOTAM tokenizer developed by (Dani and Desarkar, 2024). The vocabulary consists of 29,999 tokens, which include words and measurements like nautical miles, feet, feet above ground level,coordinates, and ASR numbers necessary to notify users. We also use the pretrained model based on BERT-tiny M_{pn} from (Dani and Desarkar, 2024). This model has ∼4M trainable parameters, 128 hidden units, 2 layers, 2 attention heads, intermediate size is 512. It was trained on 3,93,858 NOTAMs. We describe our proposed methodology below[3].

5.1 Semantic and Syntactic-Based One-Shot Prompting methodology

Our proposed methodology consists of a two-step process presented in Figure 1. The offline processing handles domain specific details of NOTAM, and performs clustering to identify the semantically and syntactically relevant examples. It then uses one shot prompting using the identified example to prompt the model. Since NOTAMs do not follow standard English rules, M_{lm} sometimes fails to understand the boundaries of condition and action attributes (the starting and finishing positions of attributes in a NOTAM). This makes the M_{lm} fail to translate these attribute details correctly. For example, a Taxiway field condition NOTAM has three compulsory attributes and nine optional attributes. The structure of an Airspace NOTAM contains three mandatory features and four optional attributes. In Taxiway and Runway field condition NOTAMs, the compulsory condition attributes include information about contamination width, depth, percentage of coverage, and optional attributes like action taken or additional information.

To address this problem, we extracted rules to decode condition and action-related information. The semantic prompting methodology, along with these rules, produces more accurate and reliable translations. Examples of translations with and without these rules are provided in the Table 4. The following list articulates the steps performed for getting the translations.

- We subset 1,000 NOTAMs for each type from the NOTAMs dataset.
- We obtain NOTAM embeddings using M_{pn}, and apply dimensionality reduction for the same.
- We cluster the 1,000 NOTAMs of each category based on the silhouette score.
- As our experiment with maximum length demonstrated better BLEU4 (Papineni et al., 2002) scores for translation, we select the longest NOTAM (N_l) from each cluster.

[3]Code: https://github.com/viduladani/NOtices-To-AirMen/tree/main

- We create one-shot prompts from these N_l for each cluster and save the one-shot example in a database to use during inference.

The following steps are performed upon receiving the query NOTAM to be translated.

- During translation, we first obtain the embeddings of the NOTAM being translated using M_{pn}.
- After dimensionality reduction, we predict the cluster label of the NOTAM.
- We retrieve prompts associated with the cluster label and use these semantic prompts to translate the NOTAM.

These steps are performed offline and the results are stored in a database.

5.1.1 Clustering details

We used k-means clustering to cluster the NOTAMs. The number of clusters were determined using silhouette score with Euclidean distance, and varied for different categories. For example, Airspace had 8 clusters, whereas for Taxiway, 3 clusters were used. The representative NOTAMs from each cluster become the one shot example during test time for translating a new NOTAM.

5.2 Alternate One-Shot Prompting Methods

We compare our proposed approach to four baseline methods described below:

- **Zero-Shot Prompting**: No example was given during translation.
- **One-Shot Random NOTAM Prompt**: In this approach, we randomly select NOTAMs from each type and use them as examples during translation.
- **One-Shot Maximum Length NOTAM Prompt**: This method involves extracting the (N_l) for each type and utilizing it as an example during the translation process.
- **One-Shot Maximum Overlap NOTAM Prompt**: This method selects prompts based on the maximum token overlap for the given TYPE, enhancing the contextual understanding of the NOTAM.

The motivation for the Maximum Length NOTAM and Maximum Overlap NOTAM is from the domain specific insight that NOTAMs follow a structure and sequence to communicate information, with mandatory or optional attributes. The longest NOTAM or the one that has Max-Overlap with other NOTAMs from the same type, is more likely to be a good representative of NOTAMs belonging to that type. Making it a suitable candidate to serve as the one shot example.

We additionally provide performance metrics for each type of prompting methodology for each model tested. The models include state of the art large language models (M_{llm}) such as GPT-3.5 (OpenAI, 2023) [175B] and GPT-4 (Achiam et al., 2023) [200B] and four comparatively (M_{slm}) smaller open source language models: Gemma [2B] (Team et al., 2024), Phi [3.8B] (Abdin et al., 2024), Mistral Instruct [7B] (Jiang et al., 2023), Llama3 [8B] (Dubey et al., 2024).

6 Experimental Setup

6.1 Implementation details

The following system prompt was used in our experiments: `"You are a translator. Your task is to translate NOTAM (Notice to Airmen) into English. Translate the following sentence independently without considering previous translations. You will give only one translation. Translate the NOTAM message into human-readable English, following the rules and examples provided. Let it be concise and to the point. Don't be too elaborate. Don't give key-value pairs, return only a normal English statement. Don't give part-wise translation, return complete translation as one single statement."`

We keep the temperature parameter value 0.1. the same across all the models. All our experiments are run on CPUs. We test with only zero shot and one shot settings as few shot settings would increase input token length which is undesirable in this setting. As we add more examples in the input, the input prompts size grows. As mentioned in the response to the previous question, with the increase in the input size, the processing requirement increases. Hence, k-shot prompting will become computationally more expensive than one-shot prompting. Moreover, if we simply add one example from every cluster irrespective of which

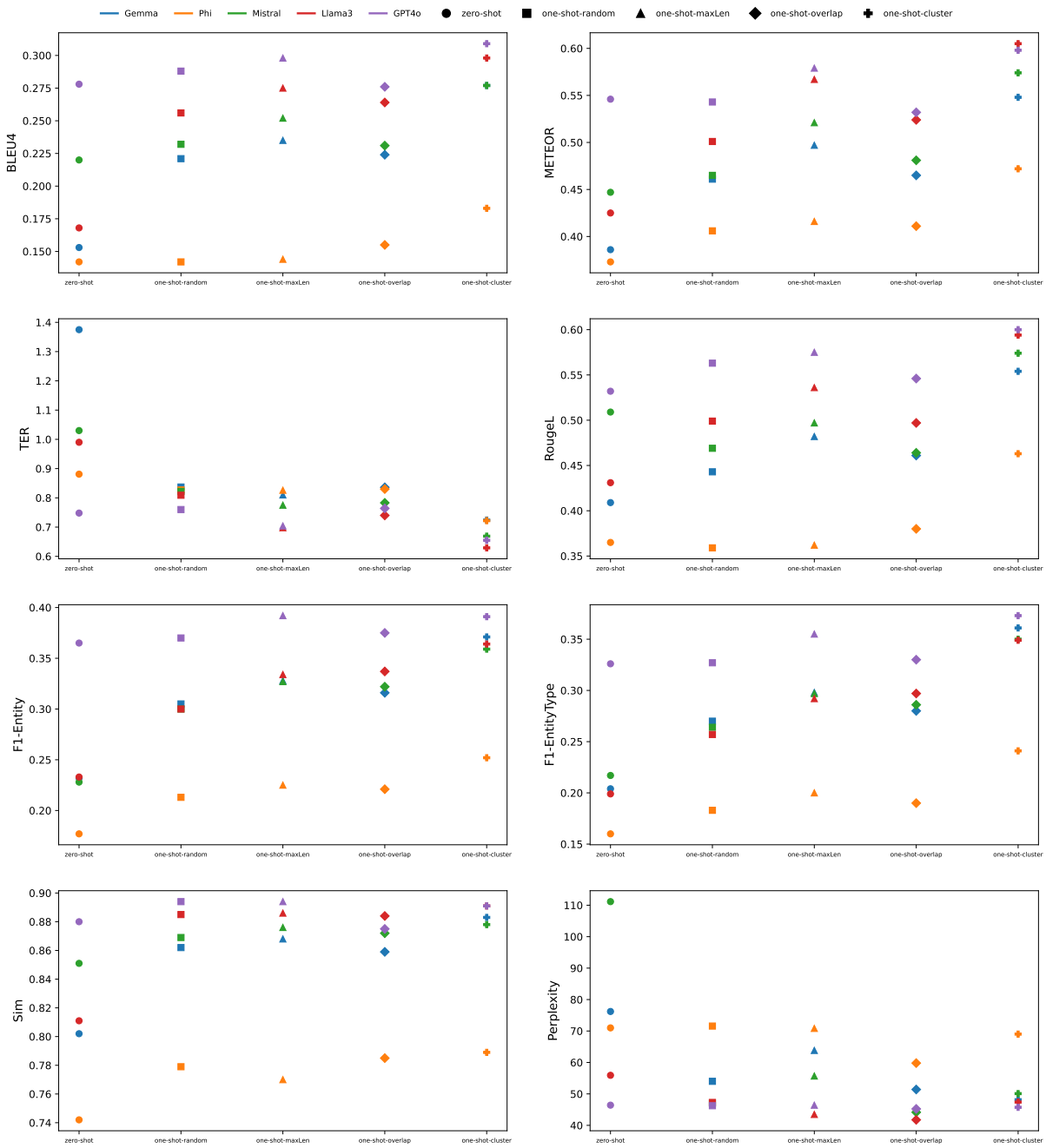

Figure 2: Performance Comparison of Prompting Methodologies. Best viewed in color.

cluster the current input NOTAM belongs to, there might be a possibility of distracting the LLM.

6.2 Evaluation Metrics

We conducted 30 experiments, five prompting strategies for each of the six model tested, and report the average score for each evaluation metric in Table 5. We also performed a performance evaluation of each of the 9 types of NOTAMs on the same test dataset using the proposed semantic-based methodology against various M_{lm}s. In the current study qualitative feedback was challenging to obtain due to the access to pilots willing to evaluate the generations, leading us to rely on automatic evaluation compared with translations by Subject Matter Experts. The following performance metrics were used:

- **Bilingual Evaluation Understudy (BLEU)** (Papineni et al., 2002): We calculated Precision-based scores BLEU1 to BLEU4 scores to measure the precision of n-grams in the NOTAMs machine translation compared to the human translation.

Raw	!GEG 02/287 GEG TWY A, A NORTH HLDG PAD, A SOUTH HLDG PAD, A1, A2, A3, A4, A5, A6, G, G NORTH HLDG PAD, G SOUTH HLDG PAD, G1, G2, G3, G5, G6, T1, T2, T3, T4, T5, T6, T7, TWY C BTN TWY A AND TWY C1 FICON WET DEICED LIQUID 50FT WID REMAINDER PATCHY FROST OBS AT 2202061140. 2202061140-2202071140
Human Translation	Spokane International Airport (GEG) with NOTAM 02/287 indicates that Taxiways A, A North Holding Pad, A South Holding Pad, A1, A2, A3, A4, A5, A6, G, G North Holding Pad, G South Holding Pad, G1, G2, G3, G5, G6, T1, T2, T3, T4, T5, T6, T7, and Taxiway C between Taxiway A and C1 have wet deiced liquid 50 feet wide and remainder taxiway is patchy frost observed on February 6, 2022, at 11:40 AM UTC. Valid from february 06, 2022, 11:40 am to february 07, 2022, 11:40 am
Translation without Rule book in context	At Spokane International Airport (GEG), Taxiway A, the North and South Holding Pads of Taxiway A, and Taxiways A1 to A6, G, the North and South Holding Pads of Taxiway G, and Taxiways G1 to G6, T1 to T7, are expected to have patchy frost on February 06, 2022, from 11:40 AM until February 07, 2022, at 11:40 AM. Missing: wet deiced liquid 50 feet wide
Translation with Rule book in context	Spokane International Airport (GEG) NOTAM 02/287: Taxiways A, the North and South Holding Pads, A1 to A6, G, the North and South Holding Pads, G1 to G6, T1 to T7, and Taxiway C between Taxiway A and Taxiway C1 are reported as having wet deiced liquid 50 feet wide. The remainder of these areas has patchy frost. This condition was observed at 11:40 AM on February 06, 2022. The condition is in effect from 11:40 AM on February 06, 2022, until 11:40 AM on February 07, 2022.

Table 4: An example NOTAM Translation with the NOTAM rule book

Approach	Model	Bleu1	Bleu4	Meteor	TER	Sim	Rouge1	RougeL	RougeS	EM-f1	ETM-f1	Perplexity
Zero-shot												
No example	Gemma	0.351	0.153	0.386	1.375	0.802	0.507	0.409	0.168	0.232	0.204	76.207
	Phi3	0.416	0.142	0.373	0.881	0.742	0.485	0.365	0.135	0.177	0.160	70.993
	Mistral	0.436	0.220	0.447	1.030	0.851	0.578	0.509	0.161	0.228	0.217	111.144
	Llama3	0.428	0.168	0.425	0.990	0.811	0.562	0.431	0.177	0.233	0.199	55.916
	GPT3.5	0.575	**0.317**	0.575	0.711	0.887	0.651	0.525	0.249	0.334	0.294	46.918
	GPT4	0.554	0.278	0.546	0.748	0.880	0.644	0.532	0.262	0.365	0.326	46.421
One-shot												
Random	Gemma	0.508	0.221	0.461	0.837	0.862	0.571	0.443	0.186	0.305	0.270	54.004
	Phi3	0.409	0.142	0.406	0.827	0.779	0.453	0.359	0.114	0.213	0.183	71.572
	Mistral	0.504	0.232	0.465	0.822	0.869	0.556	0.469	0.185	0.300	0.269	47.050
	Llama3	0.531	0.256	0.501	0.809	0.885	0.598	0.499	0.216	0.300	0.257	47.284
	GPT3.5	0.566	0.275	0.538	0.730	0.882	0.625	0.498	0.229	0.355	0.315	38.768
	GPT4	0.568	0.288	0.543	0.760	0.894	0.647	0.563	0.258	0.370	0.327	46.243
Max Length	Gemma	0.520	0.235	0.497	0.810	0.868	0.597	0.482	0.192	0.328	0.298	63.826
	Phi3	0.414	0.144	0.416	0.826	0.770	0.462	0.362	0.121	0.225	0.200	70.832
	Mistral	0.539	0.252	0.521	0.775	0.876	0.602	0.497	0.193	0.327	0.297	55.681
	Llama3	0.582	0.275	0.567	0.698	0.886	0.634	0.536	0.225	0.334	0.292	43.470
	GPT3.5	0.586	0.294	0.581	0.675	0.875	0.642	0.530	0.243	0.364	0.324	39.926
	GPT4	0.591	0.298	0.579	0.704	**0.894**	0.659	0.575	0.261	0.392	0.355	46.413
Max Overlap	Gemma	0.505	0.224	0.465	0.836	0.859	0.585	0.461	0.194	0.316	0.280	51.406
	Phi3	0.431	0.155	0.411	0.830	0.785	0.478	0.380	0.122	0.221	0.190	59.800
	Mistral	0.512	0.231	0.481	0.783	0.872	0.554	0.464	0.180	0.322	0.286	44.157
	Llama3	0.557	0.264	0.524	0.740	0.884	0.597	0.497	0.218	0.337	0.297	41.751
	GPT3.5	0.563	0.264	0.528	0.725	0.877	0.622	0.491	0.225	0.374	0.334	**37.181**
	GPT4	0.559	0.276	0.532	0.764	0.875	0.647	0.546	0.249	0.375	0.330	45.229
Semantic	Gemma	0.563	0.277	0.548	0.724	0.883	0.635	0.554	0.238	0.371	0.361	48.154
	Phi3	0.476	0.183	0.472	0.722	0.789	0.546	0.463	0.171	0.252	0.241	69.010
	Mistral	0.566	0.277	0.574	0.669	0.878	0.640	0.574	0.237	0.359	0.350	50.088
	Llama3	**0.606**	0.298	0.605	0.629	0.891	0.676	0.594	0.252	0.364	0.349	47.408
	GPT3.5	0.593	0.299	**0.632**	**0.605**	0.884	**0.678**	**0.615**	0.258	**0.401**	**0.394**	42.633
	GPT4	0.604	0.309	0.598	0.655	0.891	0.675	0.600	**0.269**	0.391	0.373	45.741

Table 5: Performance Evaluation of Prompting Methodologies with Language Models, Sim is Embedding Similarity score, EM-F1 is Entity Match F1 score and ETM-F1 is Entity Type Match F1 score. Best scores in bold and best scores in smaller models in underline

- **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**: We calculated Recall-based scores ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-S to evaluate the similarity between a machine-generated translation and human translation by measuring overlapping n-grams, which are sequences of words that appear in both the machine-generated and human translation.

- **Translation Edit Rate** (TER (Snover et al., 2006)): We calculated the edit score by identifying the number of edits required to align machine translation with human translation.

- **METEOR** (Banerjee and Lavie, 2005): We calculated the METEOR metric for the translation.

- **Embedding Similarity**: We calculated the cosine similarity between human and machine translations. We use Spacy for getting the embeddings for the same.

- **F1 score for entity matching**: We Evaluated F1 scores for Entities matching in machine trans-

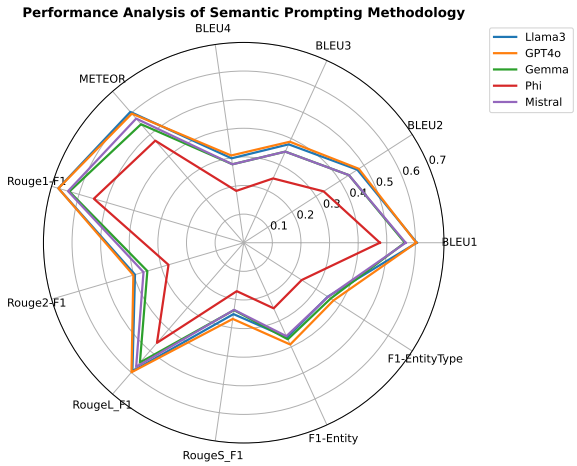

Figure 3: Performance Analysis of Semantic Prompting Methodology

lation. We use existing NER model to detect entities.

- **F1 score for entity type matching**: We Evaluated the score for entity type matching in the machine translation. The existing Named Entity Recognition (NER)[4] model often recognizes entity types for country names, dates, and times, and considers airspace names or aerodromes as 'ORG' types. However, it fails to detect types for coordinates, runway names, taxiway names, and event names. If the machine translation of NOTAMs results in differences in entity names and their types, such as airspace names, country names, dates, and times, compared to human translations, the entity type scores will likely be lower.

- **Perplexity** - We calculated the Perplexity score using GPT-2 to understand how perplexed the model is during the translation.

For perplexity analysis, we wanted to use a system/LLM which is not among the LLMs used for getting the translations. As we perform zero-shot inference, token generation relies significantly on the probabilities computed by the model. The same probabilities are also used to compute the perplexity as well. This might lead to some biased scores in the evaluation. We used GPT-2 as it is not used for getting the translations in our framework, and also has an API for computing the perplexity.

7 Analysis

From the experiments we find, using the longest NOTAM or Max-Overlap NOTAM takes longer time and processing power because it has to process more tokens in the case of longest prompt every time, irrespective of the length of the NOTAM to be translated. The Max-Overlap prompt methodology also experiences the same issue (refer to Table 2) as the selected max overlap NOTAM are also typically long. The syntactic methods (longest and max-overlap) choose NOTAMs with longer length as the one shot example, compared to the other methods. As the input length increases, the LLMs take more time to produce the answer. This is because the self-attention/cross-attention part in the encoder and decoders are quadratic in the number of tokens. Hence, the longer the input, the higher the requirement of the processing clock cycles.

The semantic prompting methodology addresses this by using the context of the query NOTAM and selecting appropriate prompts for translation. This approach can be used in resource-constrained environments like edge devices and cockpits.

We observed that the semantic prompting shows better accuracy compared to other methodologies evaluated (refer to Fig. 2). The results of Llama3 with 8B parameters using the semantic prompting methodology are comparable to GPT-4o using zero-shot prompting. Precision, Recall, Similarity, METEOR, and TER scores are better in Llama3 with semantic prompting compared to GPT-4o with zero-shot prompting. Additionally, the metric scores for GPT-4o improved with semantic prompting.

We also compared the performance of the semantic prompting methodology across different models. Our findings indicate that Llama3 with semantic prompting produces results comparable to the largest GPT-4o model, with accuracy at an acceptable level (refer to Fig. 3).

8 Next steps

This work can be integrated with the Electronics Flight Bag (EFB) where most of the briefing today occurs. The next part of the project will look closely at the integration of the proposed system with devices. It will also look into the aspect of translation and summarization of multiple NOTAMs coming together as a single input. NOTAMs could be premise or cloud-based services. However,

[4]We used the default NER model from SpaCy, with default parameter settings.

Raw	!SVA 12/007 SVA AD AP ALL SFC WIP SN REMOVAL 2112090022-2112100022
Human Translation	Savoonga Airport (SVA) with NOTAM 12/007 Aerodrome All surfaces work in progress (WIP) for snow removal from december 09, 2021, 12.22 am to december 10, 2021, 12.22 am.
Zero Shot	Work on snow removal at Savoonga Airport is underway from December 9th, 2021, 12:22 AM to December 10th, 2021, 12:22 AM. Missing: all surface work
One-Random-shot	Snow removal work is in progress at Savoonga Airport (SVA) from December 9, 2021, 00:22 UTC to December 10, 2021, 00:22 UTC. Missing: all surface work
One-Maxoverlap-shot	Snow removal work is in progress at Savoonga Airport (SVA) from December 9, 2021, 00:22 UTC to December 10, 2021, 00:22 UTC. Missing: all surface work
One-Maxlength-shot	At Savoonga Airport, all surface work including snow removal is complete as of December 9, 2021, at 12:22 AM.
One-Semantic-shot	The work to remove snow on all surfaces at Savoonga Airport (SVA) has started and is making progress as of December 9, 2021, at 12:22 AM, with the expected completion by December 10, 2021, at 12:22 AM.

Table 6: Comparison of translation across prompting methods, the red color indicates a mistranslation and orange color indicates incomplete and missing information.

it is dependent on current cockpit infrastructure such as AID (Aircraft Interface Device) that facilitate external connectivity. Our approach shows that with effective prompting strategy even, small language models can also provide translations comparable to the ones from large language models making it suitable to run in resource constrained settings.

9 Conclusion

In this work, we focus on different prompting techniques for NOTAM translation. The semantic prompting methodology leverages the context of the query NOTAM to select appropriate prompts for translation, making it suitable for resource-constrained environments like edge devices and cockpits. We evaluate the performance of different prompting techniques, ranging from small-sized models to OpenAI's largest model. We demonstrate that the semantic prompting methodology is more efficient and suitable for resource-constrained environments.

Limitations

In this work, we considered nine frequently used types of NOTAMs. There are other NOTAM types also which are not considered in our experiments. Our methodology is designed to achieve acceptable accuracy for translations if the query NOTAM comes from one of the considered types. Its performance may vary if the test query is too different from the queries considered for one-shot examples.

9.1 Potential risks with reliance on machine translations

NOTAMs are very specific to the aviation domain. At the same time, they are not for sending any arbitrary messages from this domain. Instead, they cover different types of operational and situational scenarios (which can be quite broad though). Due to this reason, we feel that research on obtaining stronger translation mechanisms will move towards retaining all the necessary information in the translated version. As there is no published work on translating NOTAMs to their natural language versions, we believe the current work under review will set a strong baseline for stronger algorithms/frameworks to follow. To ensure that the essence of the message with all necessary details are captured appropriately, we make use of different types of evaluation metrics. Precision-oriented metrics such as BLEU (and its variants) and Recall-oriented metrics ROGUE (and its variants) capture how aligned the generated translation is with the ground truth translation. Embedding based metrics try to ensure that even if the same information is conveyed in a different way, it is not penalized. Translation Error Rate measures how many edits are needed to go from the machine translated message to the ground truth response. The Entity and Entity-Type matching scores are crucial as they try to capture whether all the necessary entities and their values are captured appropriately in the generated translation. Thereby, the proposed evaluation mechanism attempts to assess whether the methodology is reliable enough to generate good quality translations.

9.2 Potential strategies for human in the loop

This work helps the pilot situational awareness and provides advisory where it requires the pilot to review and validate those before taking any further action. There would be a verification system to validate the accuracy of the translations and flag potential mistranslations for human review.

9.3 Dataset size limits

The number of human-translated examples is relatively small due to the difficulty in collecting data from pilots and air traffic controllers, this is a time consuming part and a major bottleneck in the study. This might be one reason why there is no prior work on NOTAM translation, and very limited work with NOTAMs in general. We are hopeful of increasing the number of manual translations. However, it may take some time to come up with a dataset that is considerably larger.

Acknowledgments

We thank the Subject Matter Experts from Honeywell – Flight Systems, especially Mythili Kamath (Sr. Director), Raghu Shamasundar (Fellow), and Shobana Arumugam (Lead Embedded Engineer) for their valuable support in this work in providing the translated versions of NOTAMs which we use as our test set to evaluate the performance of the proposed approaches.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, and Jyoti et. al. Aneja. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexandre Arnold, Fares Ernez, Catherine Kobus, and Marion-Cécile Martin. 2022. Knowledge extraction from aeronautical messages (NOTAMs) with self-supervised language models for aircraft pilots. In *NAACL-HLT: Industry Track*, pages 188–196.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Minal Nitin Dani and Maunendra Sankar Desarkar. 2024. Detecting attribute information in notice to airman. In *Natural Language Processing and Information Systems NLDB 2024*, volume 14763, pages 195–206.

Abhimanyu Dubey, Abhinav Jauhri, and Abhinav et. al. Pandey. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, and Arthur et. al. Mensch. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zixi Liu. 2024. Cross-lingual summarization of notice to air missions (notams).

Anna Mogiłło-Dettwiler. 2024. Filtering and sorting of notices to air missions (notams). Master's thesis, University of Zurich.

OpenAI. 2023. Gpt-3.5. Accessed: 2025-02-16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Matthew Snover, Bonnie Dorr, and Rich et. al. Schwartz. 2006. A study of translation edit rate with targeted human annotation. In *7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Aiden Szeto and Aditya N Das. 2024. Classification of notices to airmen using natural language processing. In *AIAA SCITECH 2024 Forum*, page 2585.

Gemma Team, Thomas Mesnard, and Cassidy et. al. Hardin. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

10 Appendix

10.1 Code and Data Release

The code is shared in the Github page. The dataset was accessed by System Wide Information Management (SWIM) of FAA through a subscription model, which required careful screening processes and restricts us from sharing the data publicly. Instead we have shared the NOTAM id numbers from which the NOTAMs can be fetched from the database, to help replicate and continue the work.

10.2 Deployment considerations

The proposed methodology consists of low cost operations and will not have a considerable impact on the overall running time during the inference process. Optimizations in these steps through advanced algorithms, better data structures, or tools,

will definitely help in reducing the time further and can help in achieving a broader impact. All experiments in the paper were run on CPU with API calls for models. The overall time elapsed between receiving the NOTAM and obtaining its translation should be small. The preprocessing and cluster identification of the test/new NOTAM is based on simple rules (e.g. timestamp extraction), or dictionary match (e.g. airport code to name) or a small number of vector operations (e.g. distances from already saved cluster centroids). These are typically low cost operations and will not have a considerable impact on the overall running time during the inference process. Having said that, any optimizations in these steps through advanced algorithms, better data structures, or tools, will definitely help in reducing the time further and can help in achieving a broader impact

The overall clustering on the entire data would be done offline. The cluster prototypes along with cluster membership of the existing NOTAMs will be saved. Also, one-shot prompts for each category of NOTAMs will be saved in a database for faster referencing. During the translation, cluster assignment will occur based on the type and semantic characteristics. The system will then invoke the respective one-shot prompt for translation.

Regarding the usage of LLMs, the handheld devices on the surface may use APIs to invoke the services. Most of the consumption of the NOTAMs happen during pre-flight briefings sent to pilots. Access to LLM services through API will be sufficient during this phase as access to the network will be there. If the solution is ported to the cockpit, local and small LLMs can be used. Due to these reasons, we considered several small sized models in our evaluations, with the smallest being Phi-3 (3B parameter model), and also including Mistral Instruc (7B parameters), and Llama3 (8B parameters). the exact number or tokens generated/consumed depends on the specific input and output. In our experiments, the complete translation pipeline took between 2-3.5 seconds depending on the input size and LLM used.