Towards an Aviation Large Language Model by Fine-tuning and Evaluating Transformers

1st David Nielsen
KBR Inc.
NASA Ames Research Center
Moffett Field, USA
david.l.nielsen@nasa.gov

2nd Stephen S. B. Clarke
Flight Research Aerospace
NASA Ames Research Center
Moffett Field, USA
stephen.s.clarke@nasa.gov

3rd Krishna M. Kalyanam
Aviation Systems Division
NASA Ames Research Center
Moffett Field, USA
krishna.m.kalyanam@nasa.gov

***Abstract*—In the aviation domain, there are many applications for machine learning and artificial intelligence tools that utilize natural language. For example, there is a desire to know the commonalities in written safety reports such as voluntary post incidents reports or create more accurate transcripts of air traffic management conversations. Another use-case is the possibility of extracting airspace procedures and constraints currently written in documents such as Letters of Agreement (LOA) which is used as the evaluation case in this paper. These applications can benefit from the use of state-of-the-art Natural Language Processing (NLP) techniques when adapted to the language/phraseology specific to the aviation domain. This paper evaluates the viability of transferring pre-trained large language models to the aviation domain by adapting transformer based models using aviation datasets.**

This paper utilized two datasets to adapt a 'Robustly Optimized Bidirectional Encoder Representations from Transformers Approach' (RoBERTa) model and two down-stream classification tasks to assess its performance. These datasets are all built upon Letters of Agreement which are Federal Aviation Administration (FAA) documents that formalize airspace operations across the national airspace system. The first two datasets are used for the adaptation of RoBERTa to the aviation domain and were of different sizes to assess the number of documents needed to adapt to the aviation domain. They contain many examples of 'aviation English' using domain specific terminology and phrasing which serves as a representative basis to perform the unsupervised adaptation. The second dataset is a separate set of LOA documents with two sets of classification labels to be used for evaluation; one at the document level and one at the line level. These down-stream evaluations allowed the measurement of improvement by adapting RoBERTa. The accuracy increased by 4-6% on both tasks and the F1 score on the class of interest increased by 4-8% from the adaptation.

I. Introduction

Aviation and Air Traffic Management (ATM) present many opportunities to leverage machine learning (ML) and Artificial Intelligence (AI) as applied to natural language due to a number of factors. Firstly, there are large amounts of technical documents written in the aviation domain ranging from the rules and regulations governing the use of airspace to safety reports and further. In addition to written documents, many ATM actions are coordinated via human-to-human conversations opening up the possibility of speech-based ML techniques to be applied.

With these natural language datasets come Natural Language Processing (NLP) and natural language understanding (NLU) tasks. For example, work has been done analyzing the Aviation Safety Reporting System (ASRS), a voluntary anonymous safety incident reporting system, utilizing the NLP methods of sentiment analysis and clustering in order to find common corrective actions [1]. Additionally, Letters of Agreement (LOA) are formal documents created by the Federal Aviation Administration to standardize operations between airspace users such as airports and their Terminal Radar Approach Control Facilities (TRACON). Work has been done to use NLP techniques to digitize the constraints contained within these LOAs [2]. For the audio data, work has been done to train models to transcribe the communications as well as correct those transcriptions to be more accurate to the aviation domain [3].

This all demonstrates a growing demand for NLP and NLU in the aviation domain, and presents unique challenges. This comes with a desire to utilize state-of-the-art methods such as Large Language Models (LLM). In 2018, a novel language model based on neural units (called transformers) was created and became known as "Bidirectional Encoder Representations from Transformers" or BERT [4]. This architecture combined with large amounts of English training data and innovative semi-supervised training tasks set the standard for what would later emerge as LLMs. The performance of these models was further improved by hyperparameter tuning and refinement of the semi-supervised training task and resulted in "Robustly Optimized BERT Pretraining Approach through hyperparameter tuning" or RoBERTa models [5]. These pre-trained LLMs proved to be useful for a wide variety of natural language processing tasks such as text classification and question answering through a process called fine-tuning. The transformer architecture with pre-trained weights served as the basis with the last few layers replaced with layers fine-tuned to perform

a new task e.g., a layer that provides a label for the entire input text. Transformer-based architectures can also be used to create rich representations of text called *embeddings* which can serve as the input to other ML models. This allows simpler algorithms such as logistic regression to use context-rich representations of the text while still remaining quick to train and evaluate.

In previous work, due to the technical content and specialized language of most aviation documents, fine-tuning pre-trained Large Language Models to specific NLP tasks has not met the benchmark on natural language processing tasks set by simpler models trained from scratch on the data [6]. To address this deficiency, this paper follows the work done with LegalBERT and BioBERT and evaluates the improvements from adapting a LLM with a large set of aviation documents using the original semi-supervised training tasks before performing specific natural language tasks [7], [8]. In adaptation, a domain-specific dataset is used on the original training task but with the pre-trained Large Language Model instead of starting from a random initialization. This approach allows the model to be adapted to the specific domain language without discarding the information gained from training on general English data. Both BioBERT and LegalBERT showed improvement over the pre-trained baseline BERT models by performing this adaptation. In this vein, this paper gathered LOA documents and used them to adapt the RoBERTa model to the aviation domain as laid out in the following section.

II. Methodology

A. Dataset pre-processing

The first step of LLM adaptation is to collect training data. This took the form of LOAs from the Federal Aviation Administration (FAA). These are stored in Portable Document Format (PDF). $7,497$ LOA PDFs were collected covering much of the National Airspace System (NAS). The libraries utilized for adaption require plain text input so the text was extracted from the PDFs using Amazon Textract[1]. This returned JavaScript Object Notation (JSON) objects that contained the plain text of the LOAs along with metadata about the document structure such as text position. This metadata was used to remove the repeated header text and page numbers from the LOA body text. The body text was then cleaned and tokenized. The cleaning consisted of removing new line characters, punctuation, and removing initial capitalization (though fully capitalized words were unchanged to preserve acronyms). The text was then tokenized using the Python *nltk* library[2]. The maximum input length to the RoBERTa model is 512 tokens including a start and stop token so those documents that were longer than this were divided into 500 token sections resulting in 29,904 total LOA training documents which will be called *FullLOA* hereafter.

A subset of this full LOA dataset was created by selecting all LOAs that had an Air Route Traffic Control Center (ARTCC) as one of the parties. This resulted in a $7,057$ document subset that still covered a large portion of the NAS so adaptation sensitively to training dataset size could be assessed. This will be called ARTCCLOA hereafter.

In addition to the adaptation training data, an evaluation dataset was created to measure adapted RoBERTa performance on down-stream classification tasks as described in Section II-C. These documents were different LOAs from the Dallas-Fort Worth ARTCC (ZFW) and totaled 493 PDFs. These ZFW LOAs were labeled by Subject Matter Experts (SME) with two sets of labels. The first labeling type labeled each document as 'civil' or 'not-civil'. This two-class classification distinction divided documents whose signatories are all public entities like the FAA from those not-civil documents that had one or more signatories who were non-public airspace users such as private companies. This task was derived from the efforts to digitize LOAs which focused on these civil LOAs [9]. Training NLP models to classify documents is important for the automation of this step. The total counts can be seen in Table I. This dataset was first used to survey embedding methods and modeling techniques in [6] and served as a good source of motivation and comparison.

TABLE I
COUNT OF ZFW DOCUMENT CLASS LABELS

Total documents	civil	not-civil
493	222	271

The second labels were on the individual lines from the 222 civil LOAs. These labels marked the lines as containing a trajectory constraint or not. Here, a trajectory constraint is any rule that restricts the trajectory of an airspace user. This classification task also supports the LOA digitization work [9]. The goal of digitization is to represent trajectory constraints in a digital format and one of the steps in this process is identifying where in the document they occur. The SME reviewed the individual lines resulting in the class counts as seen in Table II

TABLE II
COUNT OF CIVIL ZFW LOA LINE LABELS

Total lines	constraint	not-constraint
499	129	370

B. Adaptation

In order to evaluate LLM adaptation to the aviation domain using this LOA dataset, the RoBERTa transformer based model was used. RoBERTa was chosen due to the improvements over the baseline BERT while remaining a straightforward transformer-based architecture. Additionally, RoBERTa was pre-trained on general English using an unsupervised Masked Language Modeling (MLM) task where a subset of the tokens in a document were masked and the surrounding text was used to predict these tokens. In BERT, a static 15% of tokens

[1]https://docs.aws.amazon.com/textract/latest/dg/what-is.html
[2]https://www.nltk.org/api/nltk.tokenize.word_tokenize.html

were masked and RoBERTa improved upon this with dynamic masking. This MLM was also found to be sufficient for pre-training and in fact out performed the original BERT strategy of combining MLM with Next-Sentence Prediction (NSP) [5].

The Python library *HuggingFace* was used to perform the adaptation of RoBERTa[3]. Due to the smaller nature of the LOA datasets, static masking was used as dynamic masking was chosen as an improvement due to dataset size and this is a much smaller dataset being used for adaptation. 15% of tokens were masked for training and NSP was not performed. The *HuggingFace RoBERTaForMaskedLM* model was used for the adaptation training. The default AdamW optimizer was used for this training. The following parameters were used for all models: $training\_epoch = 10$, $batch\_size = 16$, $learing\_rate = 0.0001$, and the AdamW $beta2 = 0.75$. A grid search was performed across *weight decay* and AdamW *beta1* which is discussed in Section III as it involves the evalutation tasks explained in Section II-C. This search found that three different sets of hyperparameters performed better on different evaluation tasks. As such, 3 models are presented, *artcc_wd25_b50, artcc_wd75_b50, full_wd24_b90*, whose hyperparameters can be seen in Table III. The first two models were adapted using just the ARTCCLOA dataset while the final was adapted using the entire *FullLOA* dataset. The details of the performance of these three models on the evaluation tasks will be covered in Section III but first the evaluation tasks must be established.

TABLE III
ROBERTA ADAPTATION HYPERPARAMETERS

Model	Dataset	Weight decay	AdamW beta 1
artcc_wd25_b50	ARTCCLOA	0.25	0.5
artcc_wd75_b50	ARTCCLOA	0.75	0.5
full_wd25_b90	FullLOA	0.25	0.9

C. Evaluation tasks

While MLM provides intrinsic metrics about masked token prediction after adaptation, because of the unsupervised nature of the MLM task, it is not as useful for predicting model utility as performing down-stream tasks. This is where the two sets of ZFW classification labels proved useful. They provide an extrinsic task relevant to the aviation domain that assists in the evaluation of the adapted RoBERTa's ability to represent aviation information. To measure this performance, these two sets of labels each were used in a 2-class classification problem following the model established in previous work [6].

The architecture used starts with the adapted RoBERTa to create embeddings of either the full document or individual lines, as appropriate for the classification task. These embeddings along with their labels were then split 90/10% into a training and test set. For both datasets, the training set was then used to train a logistic regression model using the python library *sklearn*. The hyperparameters in Table IV were found to perform best for this model in previous work [6]. Logistic regression was chosen due to high performance in previous work while being quick and simple to train but future work will evaluate a RoBERTa classification layer as an end-to-end classifier as well.

[3]https://huggingface.co/docs/transformers/en/model_doc/roberta

TABLE IV
LOGISTIC REGRESSION HYPERPARAMETERS

Parameter	value
solver	liblinear
fit_intercept	True
intercept_scaling	0.0001
class_weight	balanced
max_iter	1000

III. RESULTS

The methodology laid out in the previous section (II-C) created evaluation metrics for each of the two classification tasks. These metrics were then used to measure performance across a hyperparameter grid search to find the adapted RoBERTa models with highest accuracy. In additional, we compare the adapted models against a baseline RoBERTa without any additional adaptation.

The hyperparamter grid search varied both *weight decay* and AdamW *beta1* across the range of $[0, 0.25, 0.5, 0.75, 0.999]$. This was an attempt to both find the best parameters for adaptation as well as measure the sensitivity of model adaptation to hyperparameter choice. This sensitivity was measured using accuracy, $acc = \frac{TP+TN}{P+N}$, which is the ratio of correctly identified cases to the total number of cases. Examining this sensitivity, figure 2 shows both the ARTCCLOA and FullLOA models accuracy varied similarly on the document classification across the hyperparameter grid. The boxes show the quartiles of the data, the whiskers show 1.5 interquartile range (IQR), and the points show individual outlier models whose accuracy was outside of that range. The outliers for FullLOA all occurred with $beta1 = 0.999$ and were the lowest accuracy ARTCCLOA models showing similar poor performance between the two adaptation datasets. This held true for the constraint classification so this hyperparameter choice was sub-optimal. Additionally, for the constraint classification, figure 2 show significantly higher sensitivity for the FullLOA models. In this case, $weight\ decay = 0.999$ also significantly under-performs compared with other hyperparameters. On the other side of the distributions, there are fewer universally best hyperparameter combinations; instead there are a range of parameters that all perform within the expected bounds of the distribution for all tasks except for two outperforming outliers for constraint classification, the best of which will be discussed below. The overall distributions across all the cases show that evaluation tasks are necessary to eliminate the extremely under-performing hyperparameter choices. Additionally, the worse hyperparameter choices under-perform the baseline RoBERTa which further shows the importance of evaluation tasks in LLM adaptation. More specifics will be discussed in the context of the best performing models below.

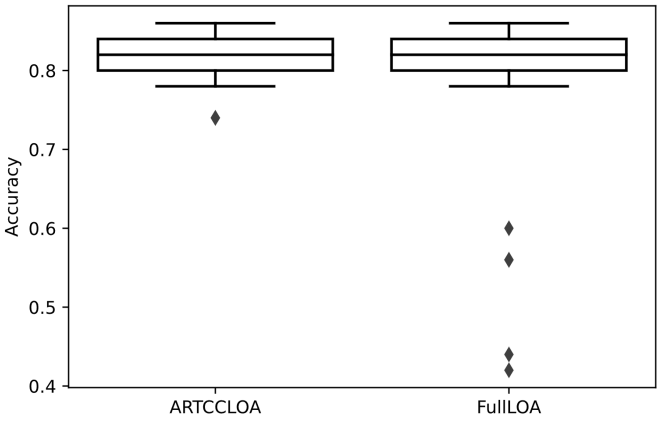

Fig. 1. Grid search document classification accuracy for ARTCCLOA and FullLOA datasets

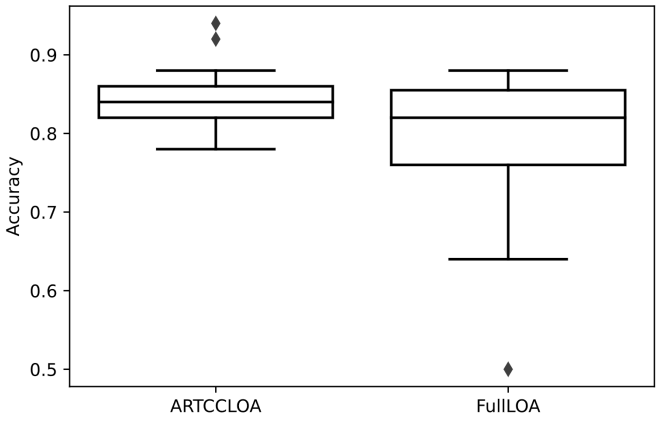

Fig. 2. Grid search constraint classification accuracy for ARTCCLOA and FullLOA datasets

The grid search allowed for the selection of 3 adapted models with the best performance on the evaluation tasks were chosen for further comparison. Additionally, the baseline RoBERTa model was used to establish a benchmark to compare the adapted RoBERTa models against. Two main metrics were collected in each case. The first is accuracy. Second is the F1-Score for the minority class; in both document classification tasks, these are the priority to focus identify so considering the F1 of that class helps measure how well the model focused on that class. F1 macro was also collected but is omitted from these tables as it closely represents accuracy in this application without largely imbalanced labels. The results can be seen in Table V & VI.

For the document classification task, *artcc_wd25_b50* under-performs the baseline, while both the other adapted models out-perform the baseline. We note the increase in performance on civil LOA classification improvement correlates with the size of the adaptation dataset; the additional documents in the FullLOA dataset outperformed any other

TABLE V
DOCUMENT CLASSIFICATION RESULTS

Adaption dataset	Test F1 civil	Test accuracy
base RoBERTa	0.81	0.82
artcc_wd25_b50	0.80	0.82
artcc_wd75_b50	0.83	0.84
full_wd24_b90	0.85	0.86

TABLE VI
CONSTRAINT CLASSIFICATION RESULTS

Adaption dataset	Test F1 constraint	Test accuracy
base RoBERTa	0.74	0.82
artcc_wd25_b50	0.91	0.94
artcc_wd75_b50	0.82	0.88
full_wd24_b90	0.82	0.88

methods. This leads to the conclusion that for this document-level task, the more example documents that can be used for adaptation, the better. It is also of note that the baseline RoBERTa performs round the same as the lower bound of the first quartile of the models seen in figure 1. This implies that in addition to more data, the right hyperparameters are necessary in order to maximize the improvement due to adaptation data.

In contrast with document classification, the best model for constraint classification was *artcc_wd25_b50* which outperformed the baseline and other adapted models by a significant margin and was one of the positive outliers. There is not a clear single cause of this performance, especially given that *artcc_wd75_b50* and *full_wd24_b90* both perform identically on this task which also out-performs the baseline. The weight decay is used to prevent model overfitting and it may be that in this case, this was not necessary due to the nature of the evaluation data. This would explain why *artcc_wd75_b50* performed worse than *artcc_wd25_b50*. The fact that *full_wd24_b90* did not perform as well as *artcc_wd25_b50* despite having the same weight decay and performs the same as *artcc_wd75_b50* implies that there was not additional information to be gained in ARTCC constraint classification from the non-ARTCC LOA documents. We also see an overall increased sensitivity in this task as seen in figure 2. This is possibly due to the nature of the evaluation task as well; it appears overall that this task is more tightly defined and just as general information does not improve the model, it is possible that the variations in hyperparameter choice that leads to slight differences in the adapted model have a larger effect in this task.

As seen with the improvements in accuracy and F1-Score, these two adaptation datasets both created adapted RoBERTa models that showed improvement on the evaluation tasks. This supports that the overall hypothesis that the technical language of aviation documents can be used to improve LLM performance on tasks in the domain. This positive conclusion motivates additional study as outlined in the next section.

IV. Future work

In order to progress this work, there are 3 main areas of focus. The first is the addition of more data to the adaptation process. As seen in Section III, the document classification improved with additional training data. As such, leveraging more aviation data could be used to further improve results. Additionally, in order to align this work with efforts in other domains, the training size would need to increase; this is a much smaller dataset than the ones used in [7], [8]. The evaluation tasks would allow the measurement of whether there is a sufficient amount of data where performance on the document classification stops improving or if the high performing *artcc_wd25_b50* could be out-preformed. Additionally, by adding data sources such as the ASRS reports or transcripts of ATM conversations could easily supplement and broaden the LOA data and allow the inclusion of additional domain specific evaluation tasks [1], [3]. It would also be of interest to see measure sensitivity of the models

Additionally, this paper was built using logistic regression for the evaluation tasks based on previous works but in future work, RoBERTa with various trained output layers could be used for the down-stream tasks. The architecture is suited to fitting a classification specific head to the adapted RoBERTa model and could be compared with the simple logistic regression model. As this no longer has a closed-form solution, there would also be work to evaluate the sensitively to random initialization in addition to additional hyperparameters from these new layers. This comparison would be valuable and this end to end method also would apply to the following future work.

In this same vein, the *HuggingFace* library has many other pre-trained LLMs available. There are additional BERT-based architectures such as DeBERTa[4] which aims to further improve upon the BERTA architecture by changing the attention mechanism as well as more refinement of the masked language modeling pretraining task [10]. In addition to BERT-based models, work has been done to identify longer-input transformer based architectures such as Longformer[5]. These models work to expand the length of input from BERT's 512 token limit with a modified attention mechanism [11]. Many aviation documents are over this arbitrary token limit. In adaptation, the longer documents were split into sub-documents as seen in Section II-A. However, this division does limit the attention mechanisms of RoBERTa as the entire document is not present. By using a longer input architecture, the full document context could be used during adaptation. Also OpenAI's GPT models[6] showed an entirely new LLM architecture that could be evaluated. While still based on transformers, the GPT architectures focus on generative tasks and increased the model size significantly. This size increase changes the nature of model adaptation to new domains but comparison should be made to these state of the art methods. Following the framework developed in this paper, these base models could be evaluated on aviation tasks and then adapted to the aviation domain and re-evaluated.

[4]https://huggingface.co/docs/transformers/en/model_doc/deberta

[5]https://huggingface.co/docs/transformers/en/model_doc/longformer

[6]https://huggingface.co/docs/transformers/en/model_doc/openai-gpt

References

[1] B. Matthews, I. Barshi, and J. Feldman, "An approach to identifying aspects of positive pilot behavior within the aviation safety reporting system," in *42nd Digital Avionics Systems Conference (DASC)*, 2023.

[2] R. Pai, S. S. Clarke, K. Kalyanam, and Z. Zhu, "Deep learning based modeling and inference for extracting airspace constraints for planning," in *AIAA AVIATION Forum*, 2022.

[3] K. H. Guo, S. S. B. Clarke, and K. M. Kalyanam, "Inverse text normalization of air traffic control system command center planning telecon transcriptions," in *AIAA AVIATION Forum*, July 2024, in press.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[6] A. Batra, S. Rajkumar, D. Nielsen, S. S. B. Clarke, K. M. Kalyanam, K. Tejasen, M. Ohsfeldt, and M. Copp, "Document classification techniques for aviation letters of agreement," in *AIAA AVIATION Forum*, July 2024, in press.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," 2020.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, p. 1234–1240, Sep. 2019.

[9] "Innovative technology use in the extraction of flight constraints recorded in letters of agreement (loa)," August 2022. [Online]. Available: https://www.icao.int/Meetings/a41/Documents/WP/wp_496_en.pdf

[10] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," 2021.

[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020.