Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 10, No. 4, 2026
www.scholink.org/ojs/index.php/asir

Original Paper

Predicting Passing-Play Outcomes in American Football: A Comparative Study of Traditional Machine Learning Algorithms and Fine-Tuned Large Language Models[1]

Tianjian (Jason) Wang

WLSA Shanghai Academy, China

Received: September 7, 2026 Accepted: October 10, 2026 Online Published: October 19, 2026

doi:10.22158/asir.v10n4p1 URL: http://dx.doi.org/10.22158/asir.v10n4p1

Abstract

This study examines American football passing play strategies based on an analysis of 100 NFL games, addressing the notable lack of research on offensive passing strategies. The yardage prediction task is formulated as a regression task, and the touchdown prediction task is formulated as a classification task. A regression analysis was conducted using four traditional machine learning models, including Support Vector Machines (SVMs), K-Nearest Neighbors, XGBoost, and LightGBM, and a LoRA-fine-tuned Llama 3 (8B) model. The SVM model showed optimal performance with a Mean Absolute Error of 6.996 yards, marginally outperforming the Llama 3 model (7.059 yards). In contrast, the fine-tuned Llama 3 significantly outperformed all traditional algorithms for the classification task, with a Macro F1 of 0.712 compared to the highest Macro F1 of 0.614 for traditional algorithms. Therefore, I advocate large language models (LLMs) for classification tasks and traditional machine learning algorithms for regression tasks. Sports analytics models can be selected based on the findings of these studies.

Keywords

American Football, Passing Offense Tactics, Offensive Yardage Prediction, Touchdown Prediction, Machine Learning, Artificial Intelligence, Large Language Models (LLMs), LoRA Fine-Tuning

1. Introduction

1.1 Research Background

[1] The content of this paper is derived from the author's research conducted as part of the National Science Talent Program (Computer Science discipline), jointly organized by the China Association for Science and Technology and the Ministry of Education of the People's Republic of China in 2025.

In addition to having the largest audience base among all sports in the United States, American football is also increasingly popular in countries and regions such as Canada and Australia (Figure 1 depicts a scene at kickoff during an American football game)[2]. As a result of the Chinese American Football League's establishment in 2013, the CNFL league, run by the Chinese American Football League, has produced 13 champions since its inception. The city of Shanghai, with its early start and most mature development of American football, stands out among these cities, as it has four teams that have actually won the championship ten times in a row. In fact, the Shanghai Wolves won the China National Football League for three straight years between 2022 and 2024. In addition, youth American football teams are becoming more common in mainland China, with kids starting professional training as young as seven or eight.



Figure 1. American Tackle Football Game

The passing offense is one of the most important scoring methods in the American football game. Through passing tactics, offensive receivers can penetrate defenses through agile route running tactics, and quarterbacks can deliver the ball to receivers at the right time at the right place. From the point of reception to the line of scrimmage, the offense progresses until the receiver is tackled or until the offense reaches the end zone. It should be noted that the former scenario results in the player gaining additional yards, whereas the latter scenario results in the player scoring a touchdown, which is worth six points.

In football, coaches review game footage to determine which receiver routes are most likely to

[2] There are two primary forms of rugby-related sports, namely American football and rugby union. This paper primarily discusses American football, which can further be subdivided into flag football and tackle football. The former does not require the wearing of protective gear, while the latter requires players to wear pads and helmets during competition and is the main focus of this study.

penetrate the defense, since passing tactics determine whether a team will win or lose a game. It takes a long time to watch game videos, and coaches and other support staff can get overwhelmed. Machine learning techniques, particularly deep learning, have developed rapidly in recent years and hold promise for helping coaches analyze and leverage historical game data, laying the groundwork for developing scientifically sound training and game plans. Coaches can analyze offensive options and their potential outcomes using machine learning algorithms if they know the probability of success of a play and how many yards it will gain. As a result of this process, they can determine which approach has the greatest chance of success, and, therefore, which approach can lead to victory.

1.2 Literature Review

A comprehensive search of research databases like Web of Science revealed that research on offensive tactics in American football is primarily concerned with several key aspects (Table 1).

Table 1. Representative Studies Relevant to the Present Paper

Topic	Paper	Title of the Paper	Research Objectives	Methods
1. Video Segmentation	(Atmosukarto et al. 2013)	Automatic Recognition of Offensive Team Formation in American Football Plays	Segment the game video into individual plays (rounds)	Based on the rules of offensive and defensive transitions (e.g., one team scores a touchdown, and the opposing team takes over on offense).
	(Nitta and Babaguchi 2005)	Automatic Parsing of American Football Videos by Intermodal Collaboration Based on Transition Rules	Segment the game video into individual plays.	(same as above)
2. Camera Source Identification for the Video	(Ding and Fan 2006)	Camera View-Based American Football Video Analysis	Segment the mixed game footage into distinct camera shots. Given any video segment, identify the specific camera from which it was captured.	Hidden Markov Model
3. Prediction of Reception	(Hollaus et al. 2023)	Catch Recognition in Automated American	Based on a segment of video data, predict	RNN+LSTM

Success		Football Training Using Machine Learning	whether a reception is successful.	
4. Prediction of Possession	(Yi et al. 2022)	BAPO: A Large-Scale Multimodal Corpus for Ball Possession Prediction in American Football Games	Based on game video data (including audio and text), predict which team holds possession of the ball.	BERT, Residual Network Models, et al.
5. Identification of Offensive Strategies	(Li and Chellappa 2010)	Recognizing Offensive Strategies from Football Videos	Based on the video of a play, determine the offensive strategy employed: Dropback, Middle & Right Run, or Wide Left Run.	Recognition By Synthesis
6. Prediction of the Outcome of a Play	(Liu et al. 2024)	Strategy Analysis in NFL Using Probabilistic Reasoning	Predict whether a touchdown is scored, or whether a kick results in points.	Probabilistic Model Checking
7. Ten-Yard Line Detection	(Liu et al. 2005)	Effective Feature Extraction for Play Detection in American Football Video	Based on the game footage, identify and delineate all ten-yard lines.	Boosting Chain
8. Play Type Identification from Video	(Lazarescu and Venkatesh 2003)	Using Camera Motion to Identify Types of American Football Plays	Based on the video, identify which of the seven types of plays it corresponds to: short and long pass plays, short and long running plays, quarterback sacks, punt plays, and kickoff plays	Incremental learning algorithm
	(Quadri et al. 2005)	Indexing of NFL Video Using MPEG-7 Descriptors and MFCC Features	Based on the video, identify which of the 4 types of plays it corresponds to: Pass plays, Run plays, Field Goal/Extra Point plays,	Event Detection Algorithm

			and Kickoff/Punt plays	
9. Recognition of Offensive Formation / Player Positions	(Newman et al. 2023)	Automated Pre-Play Analysis of American Football Formations Using Deep Learning	Use machine learning techniques to recognize the offensive formation (i.e., the positions of all offensive players).	YOLO, Residual Network Models
	(Orr et al. 2024)	Optimizing Football Formation Analysis Via LSTM-Based Event Detection	Use machine learning techniques to recognize the offensive formation (i.e., the positions of all offensive players).	LSTM
	(Zhang et al. 2012)	Robust Multi-Object Tracking Via Cross-Domain Contextual Information for Sports Video Analysis	Based on the video, dynamically identify the position of each player on the field.	Particle Filter Method
	(Hong et al. 2024）	American Football Play Type and Player Position Recognition	Identify the positions of players on the field from game images.	Feedforward Neural Network
10. Removal of Game-Irrelevant Segments from Game Video	(Ravitz and Mei-Ling 2005)	Unit Detection from American Football Tv Broadcast Using Multimodal Content Analysis	Remove irrelevant segments (e.g., highlights and statistics clips) from the video, retaining only the useful game footage.	Multimodal (Audio, Video, etc.) Detection Algorithms

Among the studies I reviewed, Video Segmentation (Atmosukarto et al., 2013; Nitta & Babaguchi, 2005), Camera Source Identification (Ding & Fan, 2006), Identification of Offensive Strategies (Li & Chellappa, 2010), Ten-Yard Line Detection (Liu et al., 2005), Play Type Identification from Video (Lazarescu & Venkatesh, 2003; Quadri et al., 2005), and Removal of Game-Irrelevant Segments from Game Video (Ravitz & Mei-Ling, 2005) are all substantially different from the research objective of this paper, which is to predict the outcome of a given play. Therefore, these works are not discussed in detail. The following are papers that are relatively closer to the research topic of this study.

(1) Prediction of Reception Success

Some scholars have investigated the prediction of reception success based on video data (Hollaus et al., 2023). However, their studies employed simulated data (using a specially designed audio and video

acquisition system to collect data from a single receiver) rather than data from formal American football games. Liu et al. (2024) is among the few studies that aim to predict the outcome of a play.

(2) Player Position Recognition on the Field

A number of methods have been developed for identifying player positions on the field from game images in the literature (Hong et al., 2022). Nevertheless, limited practical value can be derived from such approaches for coaches in formulating tactical strategies, since they do not facilitate the development of specific offensive tactics that can be employed by coaches.

(3) Formation Recognition

The pre-snap formation in American football is important because it often provides a clue to the offensive strategy that the team is likely to employ during the game. In previous studies, machine learning techniques were used in order to recognize offensive formations, that is, the positions of all of the offensive players (Newman et al., 2023; Orr et al., 2022). They fail to consider routes taken by offensive players after the snap or predict the outcomes of various offensive strategies. It is for this reason that these tools are not very useful for coaches.

Overall, I find that the literature has not yet applied formal American football game data to investigate the issue of passing offensive tactics in American football, which presents a valuable research opportunity for this study.

2. Research Design

In the first place, this study seeks to predict the impact of various passing tactics (e.g., deep/short passes, left/middle/right passes) on game outcomes by analyzing historical data from American football games. There are two research questions that this study seeks to answer from the perspective of predicting game outcomes:

Research Question 1 (Regression Task): Can machine learning methods be employed to predict offensive yards gained with reasonable accuracy?

Research Question 2 (Classification Task): Can machine learning methods be employed to predict whether a touchdown is scored with reasonable accuracy?

Methodologically, there are two main categories of prediction models/approaches: traditional machine learning models (e.g., Support Vector Machines, XGBoost) and Large Language Models (LLMs). LLMs are based on the "Next Token Prediction" approach, which can be used to solve a wide range of problems, but they tend to perform suboptimally when applying them to specialized domains and those that involve data from private domains. In this regard, one of the challenges when trying to improve the predictive ability of Large Language Models is fine-tuning, that is to make subtle adjustments to the parameters that were pre-trained. A more sophisticated approach used in this study is referred to as the "fine-tuned large language model." To date, there is no agreement on the extent to which traditional machine learning models or the large model fine-tuning approach are superior in predicting future events. The aim of this research is, therefore, to address two other research questions:

Research Question 3: For the regression task, which machine learning method (traditional machine learning methods vs. fine-tuned LLMs) achieves better predictive performance?

Research Question 4: For the classification task, which machine learning method (traditional machine learning methods vs. fine-tuned LLMs) achieves better predictive performance?

Results of the two first research questions can assist coaches and players in formulating more successful passing offensive strategies, thereby enhancing athletic performance, and have strong practical implications. Besides contributing to the understanding and evaluation of the capability boundaries of Large Language Models, the results pertaining to the latter two research questions are of considerable scientific value, as they enable scientists to make scientifically informed decisions regarding the selection and application of the model.

3. Research Method

3.1 Data

In the present study, we selected a publicly accessible dataset on the Internet as the primary source of data after an extensive search and comparison process was carried out. In order to meet the research objectives for this study, it is important to note that the dataset includes relatively exhaustive game data related to the process and result of the games (see Table 2), which is a valuable resource to use for this study.

There are 100 games from the National Football League (NFL) for the 2018-19 season, comprising a total of about 257 gigabytes in size. 100 games have textual details of the game process, 82 games have audio excerpts that are synchronized with the game process, and 78 games have visual representations of the game process.

Table 2. Overview of the Dataset

# games	Dataset Size	Date of First Game	Date of Last Game	Number of Games with Textual/Audio/Video Data
100	257G	2018/9/9	2019/12/29	100/82/78

Table 3 summarizes the performance of home teams and away teams in the dataset based on a total of ten indicators. The home team outperformed the away team on 6 of these 10 indicators (e.g., the number of wins, the passing completion rate), tied in one (the number of interceptions), and underperformed on 3 indicators (the number of fumbles, yards lost as a result of fumbles, and the number of turnovers). Therefore, in a football game in which the home team is playing, there does indeed exist a "home-field advantage."

Table 3. Comparative Strengths and Weaknesses of Home and Away Teams

Teams	Home Team Wins	Pass Completion Rate	Yards per Pass	Passing Touchdowns	Interceptions
Away	48	0.65	7.53	1.80	0.79
Home	***52***	***0.66***	***7.83***	***1.89***	0.79
Teams	**Net Passing Yards**	**Total Offensive Yards**	**# of Fumbles**	**Yards Lost Due to Fumbles**	**# of Turnovers**
Away	251.54	364.08	***1.2***	***0.45***	***1.24***
Home	***258.53***	***377.17***	1.25	0.59	1.38

As shown in Figure 2, the dataset also records the process of each game. For example, a game between the Cincinnati Bengals and the Arizona Cardinals resulted in 147 plays. As a result of the Bengals' first drive, which consisted of 16 plays, the Bengals were able to score 3 points via a field goal.

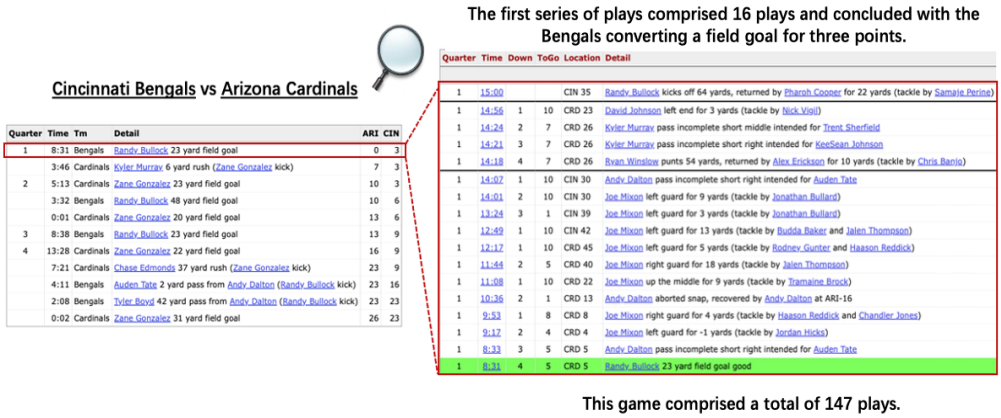


Figure 2. Sequence of Events in an American Football Play

Based on statistics, it appears that the 100 games in the dataset include a total of 15,132 plays, of which 7,551 (49.9%) are passing plays. In other words, approximately half of the offensive plays in American football employ passing tactics, which shows how important passing strategies are to success.

Among the factors influencing passing plays in the dataset, the following are the most important ones:

(A) Direction of the pass (Left/Middle/Right): Due to the limited number of defensive players in the game, it is impossible for enough defenders to be allocated to each of the three directions because of the limited number of defensive players. For an offense to succeed, selecting the passing direction is crucial.

(B) Type of pass (Short/Deep): It is important to note that there are two types of offensive plays: short passes and deep passes. In general, deep passes are more likely to be intercepted by defensive players, but if they succeed, they can produce more yards. Short passes have advantages and disadvantages. A

short pass can be completed more easily, but the receiver is more likely to be tackled after making the catch by nearby defenders after the catch has been made. Thus, it's important to pick the right pass.

(C) Defensive team (DefenseTeam): There are some teams whose defensive capabilities are particularly strong, while there are other teams whose defensive capabilities are relatively weak. Consequently, the identity of the defensive team influences offensive passing tactics.

(D) Offensive team (OffenseTeam): With an exceptional quarterback and receiver, however, a passing strategy is more likely to succeed. Because of this, the identity of a team's players has a decisive influence on the outcome of passing plays.

(E) Down (Down): To retain possession, the offensive team must advance at least 10 yards within four downs. In light of this, the fact that a player passes a particular down may directly or indirectly affect their psychological state and their performance on the field. Thus, the Down factor is taken into account in this study.

(F) Distance to the first-down line (ToGo): On the offensive side, it is very important to note that the distance from the first-down line to the offensive team's line up is 10 yards when the offensive team lines up. Offense players are closer to defense as they advance. There is a strong correlation between the distance to the first-down line and the psychological pressure faced by the offensive team. When the offensive team is facing a large distance from the first-down line, it's crucial to take this variable into account.

(G) Distance to the end zone (ToTD): There is also evidence that proximity to the end zone influences a player's performance. Thus, the yardage gained from a pass is affected by a player's performance on the field. Also, the closer the offense is to the end zone, the more likely they are to score a touchdown. When analyzing the results of the study, it is important to take the distance to the end zone into consideration.

(H) Target receiver (PassTo): In light of the fact that each receiver has their own strengths and weaknesses, the choice of receiver can also influence, to some extent, the yardage that may be gained and the likelihood of a touchdown being scored.

3.2 Descriptive Statistics

This study measured the effectiveness of passing tactics in the dataset (with all non-passing plays excluded), and the results are reported as follows. Figure 3 illustrates the average yardage gained per passing play as being 8.1 yards (7.4 meters) on average[3], with a standard deviation of 10.41 yards. In accordance with the rules of the game, an offense can achieve a first down (and continue its drive) if its offense accumulates a total of 10 yards over four downs in order to achieve a first down. In order to achieve a first down, a passing tactic must be executed properly. Despite this, Figure 3 also reveals that many passing plays do not gain more than 4 yards on average, and that gains of more than 30 yards are relatively uncommon. Passing plays can also result in negative yardage (e.g., when the defense intercepts).

[3] One yard is approximately equal to 0.91 meters.

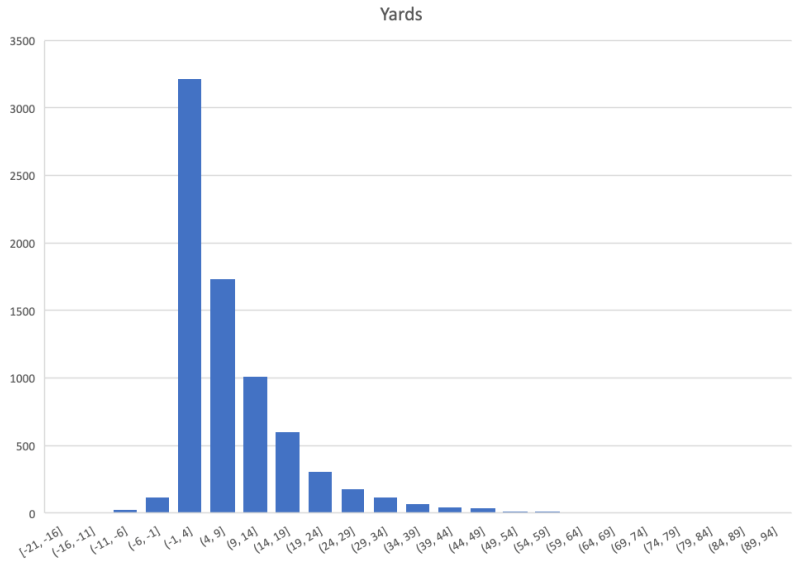


Figure 3. Distribution of Offensive Yards Gained per Passing Play

The frequency of passing plays across the four downs is shown in Table 4. Depending on the results of the study, it appears that the first two downs account for a relatively large proportion of passing attempts, followed by the third down at 24.65%, and the fourth down at only 2.05%. In light of this evidence, it appears that when a team reaches fourth down, they tend to adopt a more conservative offensive strategy, such as running the ball or punting (voluntarily surrendering possession of the ball to the opposing team) in order to make the play. On the latter two downs, Table 4 also shows that despite fewer passing plays, touchdown rates are higher. It is noteworthy that on fourth down, passing plays achieve a touchdown rate of 12.26%, indicating that the use of a contrarian approach can sometimes produce surprising results for the offense.

Table 4. Proportions of Each Down and Corresponding Offensive Outcomes in The Dataset

Down Number	Frequency	Proportion	Touchdown Rate
1	2946	39.01%	4.21%
2	2589	34.29%	4.87%
3	1861	24.65%	6.07%
4	155	2.05%	12.26%

On passing plays, Table 5 presents the average distance to the first-down line and the average distance to the end zone. According to the results of this study, there is an average distance of 8.9 yards between the first-down line and the target. Considering that the average yardage gained per passing play is 8.1 yards, it is evident that a single passing play alone will not result in a first down if the average yardage gained is 8.1 yards per pass. Furthermore, the average distance between the end zone and midfield (the

50-yard line) during passing plays is 51.23 yards.

Table 5. Yards to Go, Distance to the End Zone, and Yards Gained per Pass Attempt

	Mean (Yards)	Min	Max	Standard Deviation
ToGo	8.90	1	41	3.98
ToTD	51.23	1	99	23.79
Yards	8.10	-21	93	10.41

According to Table 6, various passing tactics are used in varying proportions. In spite of the fact that deep passes can yield greater offensive yardage, they only account for 19.32% of all passing plays, which is quite surprising. The defense focuses on preventing deep passes, which suggests that offensive receivers can't easily catch deep passes. There is evidence to support this conjecture. Generally, most passes go to either the right or the left side, while 23.23% of all passes go to the middle, which is about 10 percentage points less than 33%. As a result, defenses concentrate on defending passes in the middle of the field, where most of their defensive efforts are spent.

Table 6. Proportions of Various Passing Tactics

	Pass Type (Short/Deep)		Direction of Pass		
	Short Pass	Deep Pass	Left-side Attack	Middle Attack	Right-side Attack
Proportion	80.63%	19.32%	36.75%	23.23%	40.52%

In Table 7, the successful reception rate for passing plays is 65.69%, which suggests passing tactics are hard to defend against. It was found that passing tactics were more likely to lead to touchdowns than running tactics. It takes approximately twenty passes for a touchdown (scoring six points) to be scored. Additionally, passing plays are rarely used consecutively in practice (they are interspersed with running plays), so successful passes rarely lead directly to touchdowns. This is due to the fact that passing plays are rarely used consecutively in practice. 1.87% of passes result in zero yards gained, which suggests that a passing play can advance the ball to some extent in the vast majority of cases. This further underscores the effectiveness and importance of passing offensive tactics.

Table 7. Outcomes of Passing Plays

	Successful Reception	Zero Yards Gained (NoGain)	Successful Touchdown
Proportion	65.69%	1.87%	5.06%

4. Experimental Procedure and Results

4.1 Data Sampling Method

Oversampling: In touchdown prediction, only 5.06% of samples are labeled TD = 1, while approximately 94.94% are TD = 0. This creates a severe class imbalance that challenges classification algorithms. Therefore, predicting TD = 0 for every instance would yield an accuracy of approximately 95%. The initial runs of XGBoost and LightGBM have indeed produced such results. In order to mitigate the imbalance and to improve the classification performance, I applied SMOTE (Synthetic Minority Oversampling Technique)[4] to the training data in order to mitigate the imbalance and to improve the classification performance. In contrast, for the finetuning of LLMs, no oversampling technique was applied.

Multicollinearity: The three directional indicators are highly collinear (an offensive play can only be left, middle, or right). To avoid the dummy-variable trap, only the Left and Right indicators were included in the model (with Middle serving implicitly as the reference). Similarly, Deep (long pass) and Short (short pass) are mutually exclusive, so only the Deep indicator was retained.

4.2 Experimental Procedure

4.2.1 Traditional Machine Learning Methods

The research objective of this study is to predict the outcomes of each passing play: the yards gained and whether a touchdown is scored. There is a distinction between a continuous variable and a discrete variable. The two types of variables are referred to as regression problems in machine learning as opposed to classification problems. In this study, four traditional machine learning algorithms were selected that were capable of handling both regression and classification tasks, namely the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM).

Cross-validation: This study used a 5-fold cross-validation approach to estimate the predictive performance of traditional machine learning models. Models were trained using 80% of data, and tested with 20%. Each model was repeated five times to calculate its average performance. There are two possible sampling methods that were considered for each fold in the training process: (1) shuffling all 7,551 play records and randomly selecting 80% of the samples from each play record as the training set; (2) selecting data from 80 games as the training set, with the remaining 20 games serving as the test set. If AI models are exposed to data from all 100 games during training, it was able to predict the test set samples in advance. Therefore, I adopted Method 2, which ensures the model doesn't use any test data during training, so its performance can be measured more accurately.

Hyperparameter selection: The four traditional machine learning models all require predetermined hyperparameters, and these parameters have an impact on the final trained model. In order to identify

[4] Synthetic Minority Oversampling Technique (SMOTE) is an improved approach based on the random oversampling algorithm. Its synthesis strategy involves, for each minority class sample a, randomly selecting a sample b from its nearest neighbors, and then randomly selecting a point along the line segment connecting a and b as the newly synthesized minority class sample.

the optimal hyperparameter combinations for each model, I used grid search method. For example, for XGBoost I searched over the following hyperparameter grid: learning_rate ∈ {0.01, 0.1, 0.3}, max_depth ∈ {2, 3, 5}, and n_estimators ∈ {20, 50, 100}, yielding 3 × 3 × 3 = 27 candidate configurations.

Model evaluation metrics: For yardage prediction, I used Mean Absolute Error (MAE), where a lower MAE indicates a better prediction. For touchdown prediction, I used Macro F1 scores, which are the average of F1 scores for each class (TD = 1 and TD = 0 for each, treated separately as positive classes). The larger the Macro F1, the better the classification.

4.2.2 Fine-Tuning Methods for Large Models

As a result of the relatively long training period (approximately 7.5 hours per training session), cross-validation was not adopted during the initial training of the LLM fine-tuning method. The data sets have to be partitioned into separate training, validation, and test groups, in order to fine-tune the model in order to achieve the best performance. Samples were collected in the following steps. As a first step, a total of 60 games were selected randomly in order to create the training set. As a second step, I take data from the remaining 40 games and randomly selected 20 of them to be used as the validation set, and the remaining 20 games to be used as the test set. The model's performance on the training set tends to improve initially and then decline due to overfitting, indicating that there is an optimal training stopping point. A validation set is an important tool for determining when the training process should be terminated after a certain amount of time has passed. As far as the validation set is concerned, it should remain independent from both the training set and the test set.

In order to investigate the performance of the LLM fine-tuning method on the two types of tasks (Figure 4), the following experimental procedure was designed:

Step 1: There is an open-source version of the large language model Llama 3 (8B version), which can be downloaded from the Internet (Touvron et al., 2023) for free.

Step 2.1: Any information that could directly reveal the target variable should be masked (removed) from the Detail column (text). Specifically, the following items were removed from the game: (A) the token "touchdown"; (B) the tokens "complete"/"incomplete"; (C) phrases of the form "tackled by XXX"; (D) phrases of the form "defended by XXX"; (E) the numeric value in phrases of the form "for ## yards" (the numeric yardage was masked).

Step 2.2: Convert the independent variables used by the traditional models (e.g., toTD, ToGo, Down) to string format and concatenate them with the masked Detail text from Step 2.1 to form the textual predictors required by the large model.

Step 2.3: Use the textual predictors produced in Step 2.2 as inputs and Yards (for the regression task) or TD (for the classification task) as outputs to fine-tune the large language model using LoRA (Hu et al., 2023).

Step 3: For every sample in the test set, present a text prompt (for example, "Kyler Murray pass deep left intended for Trent Sherfield") to the fine-tuned model (LLM + LoRA) and obtain the predicted

output (for example, 2 yards).

Finally, compare the predicted outputs (e.g., 2 yards) with the actual outcomes (e.g., 5 yards) and report the model's aggregate performance over the entire test set.

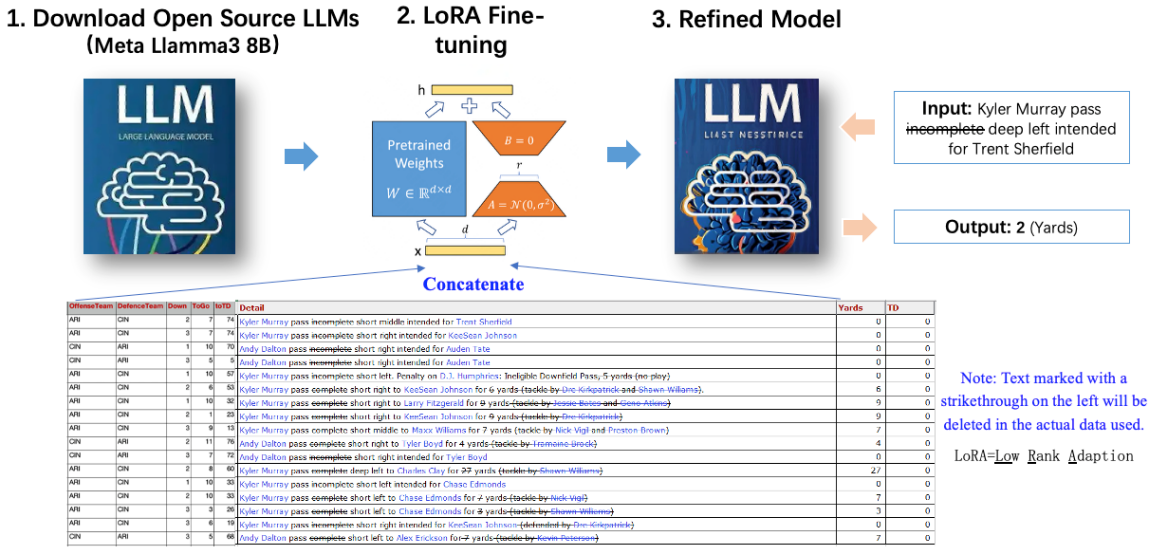


Figure 4. The Pipeline for Fine-Tuning and Application of Large Language Models

4.3 Experimental Results for Yardage Prediction

In Table 8, I present the Mean Absolute Error (MAE) for four traditional machine learning models under 5-fold cross-validation, with lower MAE values indicating better performance. Specifically, SVM was able to achieve the best performance in three folds, while XGBoost was able to achieve the best performance in two folds. Based on the average MAE across the five folds, SVM exhibited the lowest average MAE, followed by XGBoost. Thus, these two models were identified as the most effective models based on their performance.

Table 8. MAE Statistics for Each Model

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
SVM	**7.245**	**6.642**	**6.824**	7.324	6.947	***6.996***
KNN	7.702	7.254	7.629	7.691	7.487	7.552
XGBoost	7.323	6.691	6.963	**7.194**	**6.931**	**7.020**
LightGBM	7.367	6.753	6.953	7.241	6.985	7.060
LLM+LoRA			**7.085**			**7.085**

The mean yardage in the test set is 8.28 yards (standard deviation 10.46 yards). However, both SVM and XGBoost have MAE values of approximately 7 yards, indicating substantial discrepancies between

predictions and actuals. This study used traditional machine learning algorithms to predict yardage because there was no contextual information leading up to the plays or an indication of which players were on the field (starters vs. substitutes).

A MAE of 7.085 was achieved by the fine-tuned large model in Table 8, slightly higher than the best SVM result by approximately 1.3%. Consequently, this shows that as compared to traditional machine learning models in terms of yards per play, the LLM doesn't offer a fundamental advantage. This paper identified that the LLM inputs (such as the words "touchdown," "completed or incomplete," "tacked by XXX," "defended by XXX," and the numeric value in "for ## yards") had the potential to mask a number of potentially informative tokens. By comparing this to traditional models, I got much less information about predictors. In spite of these shortcomings, the finely tuned large model appears to perform pretty well, which indicates that it has a strong predictive capability.

As shown in the table below, several well-performing models are compared in terms of their prediction performance. Using a scatter plot, I compare the predicted and actual yards based on the SVM model in Figure 5. According to the figure, the SVM model makes relatively accurate predictions when the actual yards are relatively small. However, when the actual yards are relatively large, there are substantial deviations between actual yards and predicted values.

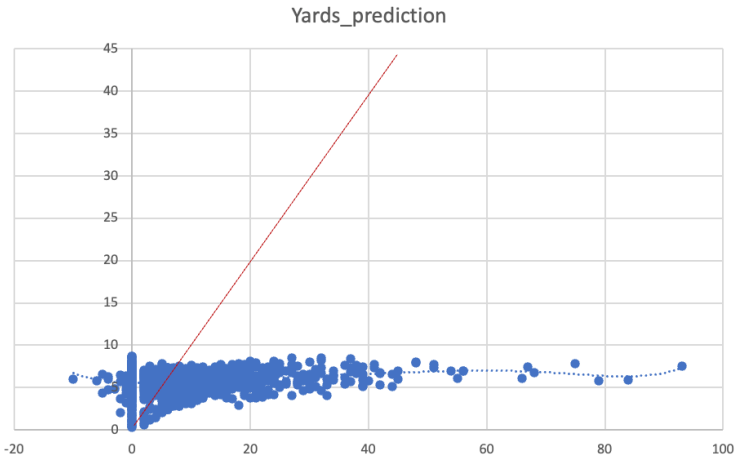


Figure 5. Comparison of Predicted Yards (Vertical Axis) versus Actual Yards (Horizontal Axis) for the SVM Model

Note. The blue dashed line denotes a sixth-degree polynomial fit; the red solid line denotes the 45° line (actual = predicted)

In Figure 6, the scatter plot compares the predicted yards from the XGBoost model with the actual yards based on the data. According to the figure, XGBoost can predict actual yardage fairly well for a variety of actual yardages, both small and large. In spite of this, it is important to note that the variance of yardage also increases as the actual yardage increases, making it more difficult for the XGBoost model to accurately predict larger values of yardage as the actual yardage increases. Overall, SVMs and

XGBoosts both perform pretty well.

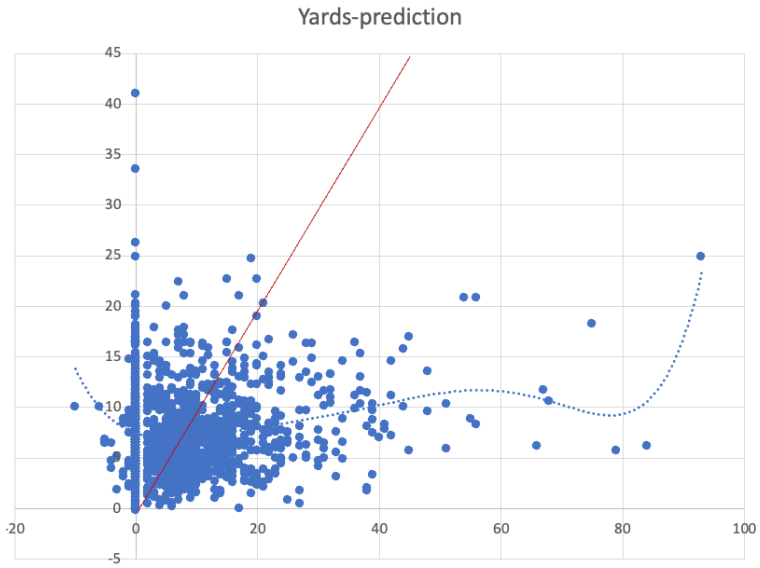


Figure 6. Comparison of Predicted Yards (Vertical Axis) versus Actual Yards (Horizontal Axis) for the XGBoost Model

Note. The blue dashed line denotes a sixth-degree polynomial fit; the red solid line denotes the 45° line (actual = predicted)

The scatter plot in Figure 7 compares the predicted yards with the actual yards. Regardless of whether the actual yardage is relatively small or large, the large language model yields relatively accurate predictions.

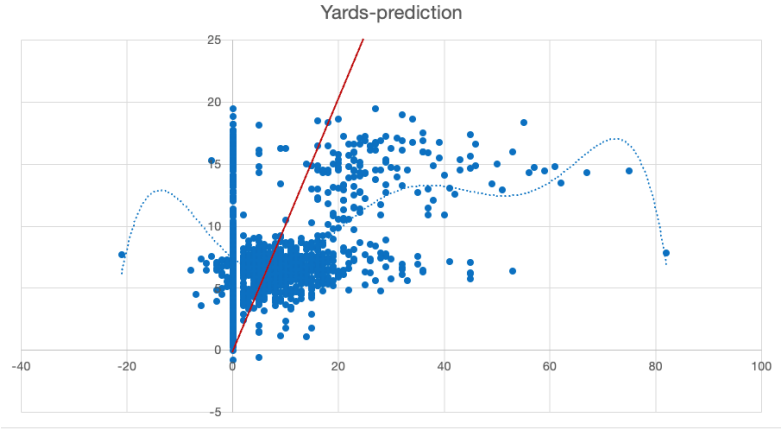


Figure 7. Comparison of Predicted Yards (Vertical Axis) versus Actual Yards (Horizontal Axis) for the Fine-Tuned Large Language Model

Note. The blue dashed line denotes a sixth-degree polynomial fit; the red solid line denotes the 45° line (actual = predicted)

4.4 Experimental Results for Touchdown Prediction

Table 9 presents the Macro F1 scores (higher values indicate better performance) of the four traditional machine learning models under 5-fold cross-validation. As shown in the table, SVM achieved the best performance in three folds, XGBoost in one fold, and LightGBM in one fold. Among these models, SVM exhibited the highest average Macro F1 score, followed by XGBoost, identifying them as the best-performing traditional machine learning models.

The last row of Table 9 displays the classification results of the fine-tuned large model, which achieved a Macro F1 score of 0.668, surpassing that of the best traditional machine learning model, SVM, by 8.8%. This indicates that the large model substantially outperforms traditional machine learning models in classification tasks.

Table 9. Macro F1 Statistics for Each Model

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**
SVM	**0.615**	**0.610**	**0.603**	0.615	0.628	**0.614**
KNN	0.574	0.596	0.574	0.608	0.598	0.590
XGBoost	0.599	0.595	0.597	**0.636**	0.613	0.608
LightGBM	0.594	0.589	0.578	0.611	**0.641**	0.602
LLM+LoRA			**0.668**			***0.668***

Table 10 summarizes the various types of correct and incorrect predictions for five models and reports, for each model class, the best-performing model (i.e., the optimal hyperparameter combination) along with its Accuracy, Precision, Recall, F1 score, and the Macro F1 score for that class.

Out of the four traditional machine-learning algorithms, the SVM achieved the highest Macro F1 (Table 10(a)). Despite the fact that the baseline method (that forecasts every play as being non-touchdown) attains an accuracy of 0.954, the SVM's accuracy of 0.904 may seem low. Nevertheless, the baseline assigns all positive instances (TD = 1) to the wrong class. As a result, it is acceptable that the overall accuracy of the SVM is relatively lower. Furthermore, the best SVM model obtained a Precision of 0.973, a Recall of 0.925, and an F1 score of 0.948, which indicates that the model has appreciable practical potential.

In addition, as shown in Table 10(c), the Macro F1 score of the XGBoost model was 0.608, which is slightly lower than that of the SVM model, suggesting that the XGBoost model also demonstrated commendable performance. Nevertheless, none of the traditional machine learning models exceeded the baseline method's accuracy (which predicts that no touchdown occurs on any play), which was 0.954.

As a result of the use of different sampling methods, the resulting test sets differed slightly from each other. Based on the results of the large model, the baseline accuracy was 0.947 (i.e., predicting that no

touchdown will occur on any play). As compared to the baseline, the accuracy of the LLM + LoRA method was 0.911 (Table 10(e)). The macro F1 score of the fine-tuned large model reached 0.668, surpassing that of the best-performing traditional machine learning model, SVM (0.614), by a considerable margin.

Table 10. Statistical Summary of Predictive Performance for Each Model

(a) SVM		
	No TD(Predicted)	TD(Predicted)
No TD (Actual)	1279	104
TD (Actual)	35	31
Accuracy	0.904	
Precision	0.973	
Recall	0.925	
F1	0.948	
F1_macro	**0.614**	

(b) KNN		
	No TD(Predicted)	TD(Predicted)
No TD (Actual)	1275	108
TD (Actual)	41	25
Accuracy	0.897	
Precision	0.969	
Recall	0.922	
F1	0.945	
F1_macro	0.590	

(c) XGBoost		
	No TD(Predicted)	TD(Predicted)
No TD (Actual)	1287	96
TD (Actual)	40	26
Accuracy	0.906	
Precision	0.970	
Recall	0.931	
F1	0.950	
F1_macro	0.608	

(d) LightGBM		
	No TD(Predicted)	TD(Predicted)
No TD (Actual)	1300	83
TD (Actual)	37	29
Accuracy	0.917	
Precision	0.972	
Recall	0.940	
F1	0.956	
F1_macro	0.602	

(e) LLM+LoRA		
	No TD(Predicted)	TD(Predicted)
No TD (Actual)	1314	94
TD (Actual)	38	41
Accuracy	0.911	
Precision	0.972	
Recall	0.933	
F1	0.952	
F1_macro	**0.668**	

4.5 Robustness Checks

Regarding Research Question 3 and 4, the experimental results indicate that: 1) In the regression task, the large model fine-tuning approach performed slightly worse than the best traditional machine learning model; 2) In the classification task, the large model fine-tuning approach significantly outperformed the best traditional machine learning model. However, the reliability of these two findings are subjected to the following questions:

Question 1: For the regression task, if more fine-tuning methods are attempted on the large model, is it possible to surpass the traditional machine learning model?

Question 2: For the classification task, is the experimental result that "the large model fine-tuning approach significantly outperforms the best traditional machine learning model" robust?

For the first question, this study evaluated five large-model fine-tuning methods (including modifications to training procedures and hyperparameters). The results are presented in Table 11. Three

fine-tuning methods produced MAE values smaller than the MAE reported in Table 8 of Section 4.3 (7.085): Methods 1, 2, and 4. Method 1 achieved the lowest MAE among the five (7.059), but it remains 0.9% higher than the MAE of the SVM reported in Table 8 (6.996). This suggests that, for the regression task considered, fine-tuned large models struggle to outperform the best traditional machine-learning model.

Table 11. Different Fine-Tuning Methods for LLMs and their Experimental Results on the Regression Task

	Specific Fine-Tuning Methods	MAE
Method 1	The batch size was reduced from 64 to 16	***7.059***
Method 2	The learning rate was decreased from 2e-5 to 1e-5	**7.069**
Method 3	Layer-wise training was employed: lower learning rates were applied to the bottom layers, while higher learning rates were applied to the top layers	7.157
Method 4	LoRA dropout: Only a small dropout ratio was applied to the transformer layers	7.077
Method 5	Freezing: The majority of the bottom-layer parameters were frozen, and only a few top layers in the attention module were fine-tuned	7.216

For the second question, this study employed five-fold cross-validation to assess the stability of the large-model fine-tuning methods; the results are presented in Table 12. As shown, the highest Macro F1 score occurred in fold 2 (0.712), which is approximately 16.0% higher than the Macro F1 of 0.614 achieved by the best traditional machine-learning method (SVM) reported in Section 4.4. The mean Macro F1 across the five folds is 0.687, about 11.9% higher than the SVM baseline. These results strongly indicate that, for the classification task, the superior predictive performance of fine-tuned large models over traditional machine-learning models is a robust finding.

Table 12. Macro F1 Statistics for Large Model Fine-Tuning Cross-Validation

	Fold	Fold 2	Fold 3	Fold 4	Fold 5	Average
LLM+LoRA	0.669	**0.712**	0.672	0.671	0.710	**0.687**

5. Discussion

5.1 Contributions

This paper has several important contributions. First, it introduces a few machine learning approaches that is capable of predicting the outcomes of plays in American football. In the course of my extensive literature review, I didn't find any prior study that models passing offensive tactics as classification or prediction problems. I leveraged the regression model to predict offensive yards gained, and a

classification model for predicting whether a touchdown was successfully accomplished. It is anticipated that football coaches will benefit from this method, as it will help them with deploying tactics scientifically, maximizing the potential of players, and achieving better results on the playing field. Machine learning methods will help coaches of American Football teams make more informed decisions about offensive strategies and their outcomes. During this process, they're able to plan their game strategies and offensive approaches ahead of time so they'll be more likely to win the match.

Based on a comprehensive empirical study, it compares two classes of machine learning approaches -- traditional algorithms and fine-tuned large language models (LLMs) -- on the task of predicting passing play outcomes. A number of factors can determine the outcome of an American football game, and coaches often rely on experience to figure out what offensive plays to use, anticipate opponents' defensive reactions, and predict the likelihood of a touchdown. In every year, large amounts of data are collected, containing a wealth of information about games. Several machine learning models are used to analyze and process historical data in order to help coaches and players pick the right model.

Lastly, it proposes a method for leveraging the rich textual data generated during American football games. Compared to other sports like basketball, badminton, or table tennis, American football has more tactical interactions and player interactions. Therefore, manually analyzing game data is much more labor-intensive than in sports like basketball, badminton, or table tennis. In this study, the most relevant information about offensive play types was extracted, and then the data was analyzed and processed using state-of-the-art LLM and traditional machine learning algorithms. This is a significant advancement in terms of data processing methodology within the sport management field.

5.2 Limitations and Future Studies

There are several limitations of this study that should be addressed in future research. Data collection was limited to American football games in US, and the dataset is limited in terms of the number of variables that can be used. It is necessary to run additional comparative experiments using a larger and more diverse dataset to be able to determine whether traditional machine learning algorithms or finetuned small models are superior for classification and prediction tasks. A second concern is that only four traditional machine learning algorithms were evaluated in this study, while it is recommended that many other traditional machine learning algorithms be tested in future studies. Lastly, for the purpose of finetuning a large model, a variety of methods can be used. It is necessary to investigate alternative fine-tuning techniques in future research and the effects of magnitudes and proportions of parameter tuning on model performance.

Future studies should explore more comprehensive ways to use the audio and video data that are currently available for the present application - predicting passing outcomes in American football - in order to make the most of the data available. Furthermore, the dataset consists of play-by-play records as well as on-site audio and video recordings which contain a wealth of information, including the speech of commentators as well as the trajectories of players' movements. The AI models in this study are unable to use these data as features. There are considerable human and material resources that are

needed even when encoding is possible. There is, however, a possibility that if these richer game data can be fully utilized, the predictive performance of the models reported here may be further improved.

6. Conclusion

As a widely popular sport, American football has been exerting an increasingly significant influence in almost every corner of the world in recent years. Coaches and athletes have also been forced to continually refine their game tactics in pursuit of improving their athletic performance as a result of an increase in the intensity and competitiveness of the game. Through an extensive literature review, this study found that research on passing tactics in terms of yards gained and touchdown outcomes is very limited. Therefore, a dataset of NFL games was collected, and the prediction of yards gained was modeled as a regression analysis problem, while the prediction of whether a touchdown is achieved was modeled as a classification problem. The main research findings are presented in Table 13.

The study found that yardage prediction is a relatively challenging regression problem. This difficulty arises primarily because models can predict short gains more readily, whereas long gains are much harder to predict. In games, the success of long passes is highly uncertain: receivers are subject to defensive interference, and quarterback throwing accuracy degrades significantly with distance, both of which greatly reduce the probability of a completed catch. Despite these challenges, the dataset's mean passing gain is 8.1 yards, and both traditional machine-learning methods and fine-tuned large models achieve MAE values of 7 yards, indicating that machine-learning models have substantial practical value. The fine-tuned large models perform slightly worse than the best traditional model (SVM), with MAE approximately 0.9% higher than SVM.

Table 13. Main Findings of this Study

	Research Questions	Findings
1	For the regression task, can machine learning methods be employed to predict offensive yards gained with reasonable accuracy?	The average passing yards were 8.1 yards, and both the traditional machine learning and large model fine-tuning methods achieved a Mean Absolute Error (MAE) of approximately 7 yards.
2	For the classification task, can machine learning methods be employed to predict whether a touchdown is scored with reasonable accuracy?	The large model fine-tuning method yielded the best predictive performance, with a prediction accuracy of approximately 91.1% and a Macro F1 score of 0.712.
3	For the regression task, which machine learning method (traditional machine learning methods vs. fine-tuned LLMs) achieves better predictive performance?	The traditional machine learning method marginally outperformed the large language model fine-tuning method.

4	For the classification task, which machine learning method (traditional machine learning methods vs. fine-tuned LLMs) achieves better predictive performance?	The large language model fine-tuning method significantly outperformed the traditional machine learning method.

Similarly, predicting touchdowns is also difficult, mainly because only about 5.06% of passes result in touchdowns. I applied SMOTE oversampling for the traditional machine-learning algorithms, which effectively prevented the trivial solution of predicting all samples as TD = 0, but it did not fully resolve the class-imbalance challenge. Experimental results show that SVM and XGBoost achieve classification accuracies above 90% with Macro F1 scores around 0.61, demonstrating practical utility. By contrast, the fine-tuned large model reaches an accuracy of 91.1% and a maximum Macro F1 of 0.712 (about 16.0% higher than SVM's Macro F1), outperforming the traditional machine-learning models. Based on the aforementioned research findings, this study presents the recommended machine learning algorithms for different types of prediction tasks in Table 14.

Table 14. Key Takeaways for Model Selection

Models	**Pros and Cons**		**Whether recommended**	
			Regression	**Classification**
Traditional ML Models	Pros	● Short training time and low cost ● Readily available software packages	Yes	No
	Cons	● Cannot directly process unstructured data ● Significant costs when data are encoded into structured formats		
Fine-tuned LLMs	Pros	● Capable of processing raw & unstructured data ● Robust predictive capabilities of LLMs	No	Yes
	Cons	● Long training time and high cost in fine--tuning ● Fine-tuning entails technical difficulty		

For regression tasks, given that the predictive performance of traditional machine learning models and large model fine-tuning techniques is comparable, and traditional machine learning models offer advantages such as shorter training time, lower computational cost, and the availability of ready-to-use software packages, they are recommended for regression tasks. For classification tasks, although large model fine-tuning techniques entail greater technical complexity, they are better suited for classification tasks due to their significantly superior predictive performance compared to traditional machine

learning models. Given the wide application of classification and regression tasks across various fields in the natural and social sciences, the above findings carry considerable theoretical significance and practical value.

References

Atmosukarto, I., Ghanem, B., Ahuja, S., Muthuswamy, K., & Ahuja, N. (2013). *Automatic Recognition of Offensive Team Formation in American Football Plays*. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 991-998. https://doi.org/10.1109/CVPRW.2013.144

Ding, Y., & Fan, G. (2006). *Camera View-Based American Football Video Analysis*. Eighth IEEE International Symposium on Multimedia (ISM'06), pp. 317-322. https://doi.org/10.1109/ISM.2006.42

Hollaus, B., Reiter, B., & Volmer, J. C. (2023). Catch Recognition in Automated American Football Training Using Machine Learning. *SENSORS*, (23), 1-17. https://doi.org/10.3390/s23020840

Hong, A., Orr, B., Pan, E., & Lee, D. J. (2024). American Football Play Type and Player Position Recognition. *ELECTRONICS*, (13), 1-21. https://doi.org/10.3390/electronics13183628

Hu, E. et al. (2023). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint:2106.09685v2.

Lazarescu, M., & Venkatesh, S. (2003). *Using Camera Motion to Identify Types of American Football Plays*. 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), pp. II-181. https://doi.org/10.1109/ICME.2003.1221583

Li, R., & Chellappa, R. (2010). *Recognizing Offensive Strategies from Football Videos*. 2010 IEEE International Conference on Image Processing, pp. 4585-4588. https://doi.org/10.1109/ICIP.2010.5652192

Liu, T. Y., Ma, W. Y., & Zhang, H. J. (2005). *Effective Feature Extraction for Play Detection in American Football Video*. 11th International Multimedia Modelling Conference, pp. 164-171.

Liu, Z. Y., Durrani, M., Xuan, L. Y., Simon, J. F., & Deon, T. Y. F. (2024). *Strategy Analysis in NFL Using Probabilistic Reasoning*. In SPORTS ANALYTICS, ISACE. pp. 116-128. https://doi.org/10.1007/978-3-031-69073-0_10

Newman, J., Sumsion, A., Torrie, S., & Lee, D. J. (2023). Automated Pre-Play Analysis of American Football Formations Using Deep Learning. ELECTRONICS, (12), 1-30. https://doi.org/10.3390/electronics12030726

Nitta, N., & Babaguchi, N. (2005). *Automatic Parsing of American Football Videos by Intermodal Collaboration Based on Transition Rules*. 2005 IEEE International Conference on Multimedia and Expo, pp. 1-4.

Orr, B., Pan, E., & Lee, D. J. (2024). *Optimizing Football Formation Analysis Via LSTM-Based Event Detection*. *ELECTRONICS*, (13), 1-16. https://doi.org/10.3390/electronics13204105

Quadri, S. G., Krishnan, S., & Guan, L. (2005). *Indexing of NFL Video Using Mpeg-7 Descriptors and MFCC Features*. Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 422, pp. 429-432. https://doi.org/10.1109/ICASSP.2005.1415433

Ravitz, G., & Mei-Ling, S. (2005). *Unit Detection from American Football TV Broadcast Using Multimodal Content Analysis*. Seventh IEEE International Symposium on Multimedia (ISM'05), pp. 1-8.

Touvron, H., Lavril, T., Izacard, G., et al. (2023). *Llama: Open and efficient foundation language models*. arXiv preprint arXiv:2302.13971.

Yi, Z., Blanco, E., Fan, H., & Albert, M. V. (2022). *BAPO: A Large-Scale Multimodal Corpus for Ball Possession Prediction in American Football Games*. 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 391-394. https://doi.org/10.1109/MIPR54900.2022.00077

Zhang, T., Ghanem, B., & Ahuja, N. (2012). *Robust Multi-Object Tracking Via Cross-Domain Contextual Information for Sports Video Analysis*. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 985-988. https://doi.org/10.1109/ICASSP.2012.6288050