TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football

Matthew Caron & Oliver Müller
Data Analytics Group, Paderborn University

Introduction

The year 2023 has witnessed significant progress in Generative AI. One would have, in fact, needed to have been living under a rock not to notice the surge in AI-based digital assistants powered by Large Language Models (LLMs). While these models have garnered considerable attention from the public, they have also achieved notable benchmarks in performance. They have proven their potential in various applications, especially natural language processing (NLP) and sequential data generation, including program code and protein sequences. However, their potential in the realm of sports remains largely untapped. Thus, in this study, we take the first steps toward uncovering the potential of state-of-the-art LLMs as tactical analysts by introducing TacticalGPT, an AI-based assistant coach for professional football.

Drawing upon the now-famous GPT architecture, we investigate, in this early study, the abilities of LLMs to comprehend and generate analytical and tactical insights concerning both on-ball and off-ball situations using natural language. Leveraging StatsBomb event-based data, we fine-tuned a foundation model using low-rank adapters on 100.000 artificially generated textual sequences derived from diverse play patterns and events extracted from the 580 Premier League fixtures from the 2021/2022 and 2022/2023 seasons. Ultimately, this study seeks to assess the practicability, viability, and benefits of using a text-based approach over traditional statistical and predictive methods, specifically by focusing on the generation of human-like responses to *What, Who*, and *Where* prompts.

Our initial results are promising and show that TacticalGPT can effectively discern strategic patterns in textual sequences and respond accurately to diverse prompts with high factual correctness. In fact, the evaluation indicates that the model produces responses that align perfectly with the ground truth for *What*-type and *Who*-type questions 50% and 32% of the time, respectively. Furthermore, the model generates factually plausible answers 96% and 94% of the time for *What*-type and *Who*-type questions, respectively. Consequently, we assert that this approach to sports analytics presents multiple advantages, such as helping coaches and analysts better tailor their preparations against a particular opponent. In addition, the interactive nature of LLM-based assistants facilitates the intuitive exploration and understanding of tactical decisions and may become the coaching staff's most valuable advisors, as already discussed by [12].

Technical Background

Publicly available LLMs, like ChatGPT [8] or Bard [4], possess remarkable natural language understanding and generation capabilities and can capture and reproduce broad world knowledge. These models are trained on publicly available, large-scale datasets that encompass encyclopedias, news articles, and online forums. Consequently, such models are not equipped to respond to prompts related to proprietary, non-public data, such as a football team's tactics and strategies. To address this shortcoming, four methods are currently available for incorporating such domain-specific knowledge into LLMs, which we detail below [2,10].

Firstly, an LLM can be trained from the ground up. This method demands significant data and computational resources, making it a less common choice for organizations other than major players like OpenAI, Google, and Meta. For instance, Bloomberg has developed its own LLM, known as BloombergGPT, by training it on a dataset that spans over 40 years of financial news articles and includes over 700 billion tokens [13]. One primary benefit of this approach is that the resulting LLM will operate strictly within the vocabulary of the targeted domain.

Secondly, Domain-Adaptive Pre-Training (DAPT) offers a way to align an LLM with the specific writing style of a given domain [5]. This method involves augmenting a pre-trained model with domain-relevant content, a feasible strategy due to the increasing number of organizations and research groups releasing pre-trained LLMs. Unlike training from scratch, DAPT requires significantly fewer documents and computational resources – i.e., on the order of hundreds of thousands rather than millions or billions. However, the resulting LLM may still produce phrases not commonly found in the domain's language – i.e., a byproduct of the initial pre-training. Additionally, the training data for both the initial pre-training and DAPT can vary in terms of quality, length, and style. Therefore, it is not assured that the LLM will consistently respond to user prompts in a conversational style, as is typical for models like GPT-4.

Thirdly, fine-tuning serves as another approach to modifying the behaviour of an existing LLM. Similar to DAPT, this strategy also incorporates domain-specific texts into a pre-existing model. The key difference lies in the quality and structure of the texts used: fine-tuning employs carefully selected, high-quality demonstration data that aligns closely with the intended application. For instance, if the LLM's primary function is to answer questions, the demonstration data should consist of relevant questions and their respective answers. An illustration of this method is Med-PaLM2, a fine-tuned version of Google's PaLM2, which was trained on medical texts specifically structured to mimic the United States Medical Licensing Examination (USMLE) format [11]. The benefit of this approach is the improved predictability and utility of the LLM compared to solely pre-trained models. However, it necessitates access to an extensive collection (10,000+ documents) of structured, high-quality texts.

Lastly, reinforcement learning offers a method to further calibrate an LLM's responses to meet user expectations [9]. In the process known as Reinforcement Learning on Human Feedback (RLHF), human experts in the domain evaluate several outputs generated by an LLM for a single prompt. These experts rank the outputs based on predefined quality criteria, such as truthfulness or helpfulness. This ranking serves as feedback to adjust the model for better alignment with the selected criteria. OpenAI has successfully employed RLHF to refine the performance of its InstructGPT models, including ChatGPT and GPT-4, to be more in line with user needs [7]. Studies involving user assessments indicate that RLHF can substantially enhance the quality of LLMs that have already undergone fine-tuning.

Methodology

Figure 1 provides a high-level overview of the methodology employed in this study. Drawing inspiration from OpenAI's approach to aligning LLMs to follow instructions [7], our approach unfolds in three distinct phases. First, we create a dataset comprising a variety of natural language phrases extracted from the structured data, serving as the basis for the subsequent stages of our methodology. In the second phase, a supervised learning technique focussing on low-rank adapters is used to fine-tune a foundation LLM using the generated sequences. Lastly, the model undergoes further refinement through Reinforcement Learning from Human Feedback (RLHF), ensuring its alignment with user requirements while enhancing its utility and performance.

3.1 Dataset Generation

As previously noted, we concentrate, in this work, on three main types of prompts or questions, namely:

- *What* – i.e., in a given/described situation, what is a team or player most likely to do next?;
- *Who* – i.e., in a given/described situation, which player is most likely to execute a specific action? (e.g., taking a free kick); and
- *Where* – i.e., in a given/described situation, where are the players likely to be positioned on the pitch?

Given that our approach revolves around structured event-based data provided by StatsBomb, the initial phase of the study focused on generating textual sequences, namely prompts and responses, to act as training data for our model. Hence, as exposed in Figure 1, we began by analysing various play patterns, events, and match situations contained within the 580 Premier League fixtures from the 2021/2022 and 2022/2023 seasons. The objective was to identify the best way to transform the available data into machine-readable text using a rule-based approach. Based on this analysis, we then formulated a set of so-called templates (Phase 1) using plain language for each of the aforementioned prompt types. For example, in a situation where a possession

commences with a free kick – i.e., *From Free Kick* – a representative template would look as follows:

```
<team></team> gets a free kick in <location></location>.
<position></position> <player></player> takes it.
```

It should be noted that each template contains XML-style tags such as `<team></team>`, `<position></position>`, `<player></player>`, and `<location></location>`, which are designed to enable the insertion of specific information in subsequent steps. Additionally, to enhance the usability of location tags for the end-user, we transformed all sets of coordinates into distinct zones, effectively partitioning the football pitch into a traditional 18-zone system.

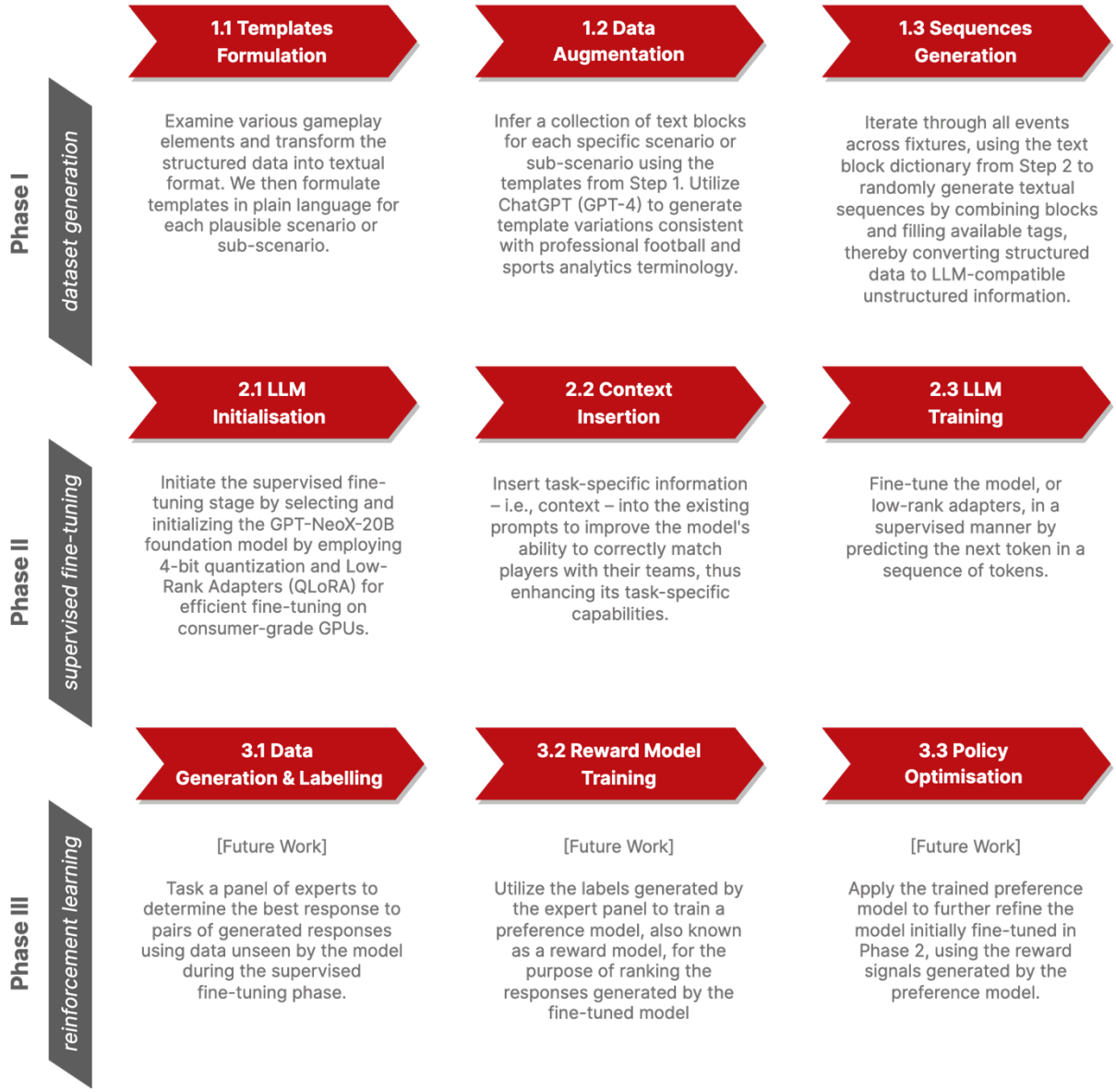

Figure 1: Pipeline – TacticalGPT

Next, utilizing the hand-crafted templates formulated in Step 1, we automated the generation of semantically equivalent templates tailored for each specific scenario or sub-scenario. This step, referred to as Data Augmentation (Step 2), involved using ChatGPT (GPT-4) to produce variations of the above templates in a manner consistent with the language associated with professional football and sports analytics. While the typical objective of data augmentation is to expand the dataset in cases where data is limited, our aim was different. Specifically, we sought to train a model capable of responding in a manner that aligns with the user's language style. Relying solely on our original hand-crafted templates would have restricted the user to a fixed language or sentence structure, contradicting the flexible nature expected of a digital assistant. Therefore, creating numerous variations of each template facilitates a more natural interaction through realistic prompts and responses.

Ultimately, leveraging the dictionary of templates augmented in Step 2, we generated textual sequences by combining multiple text blocks randomly and populating those with events from the various matches (Phase 3). As exemplified in Figure 2, this technique enabled the transition from a structured to an unstructured data state, effectively converting hard information into soft, LLM-readable information. It should, however, be emphasised that responses to *Where*-type prompts solely consist of a set of coordinates – i.e., StatsBomb 360. These coordinates are designed for utilisation in graphical representations, such as plots, rather than textual descriptions.

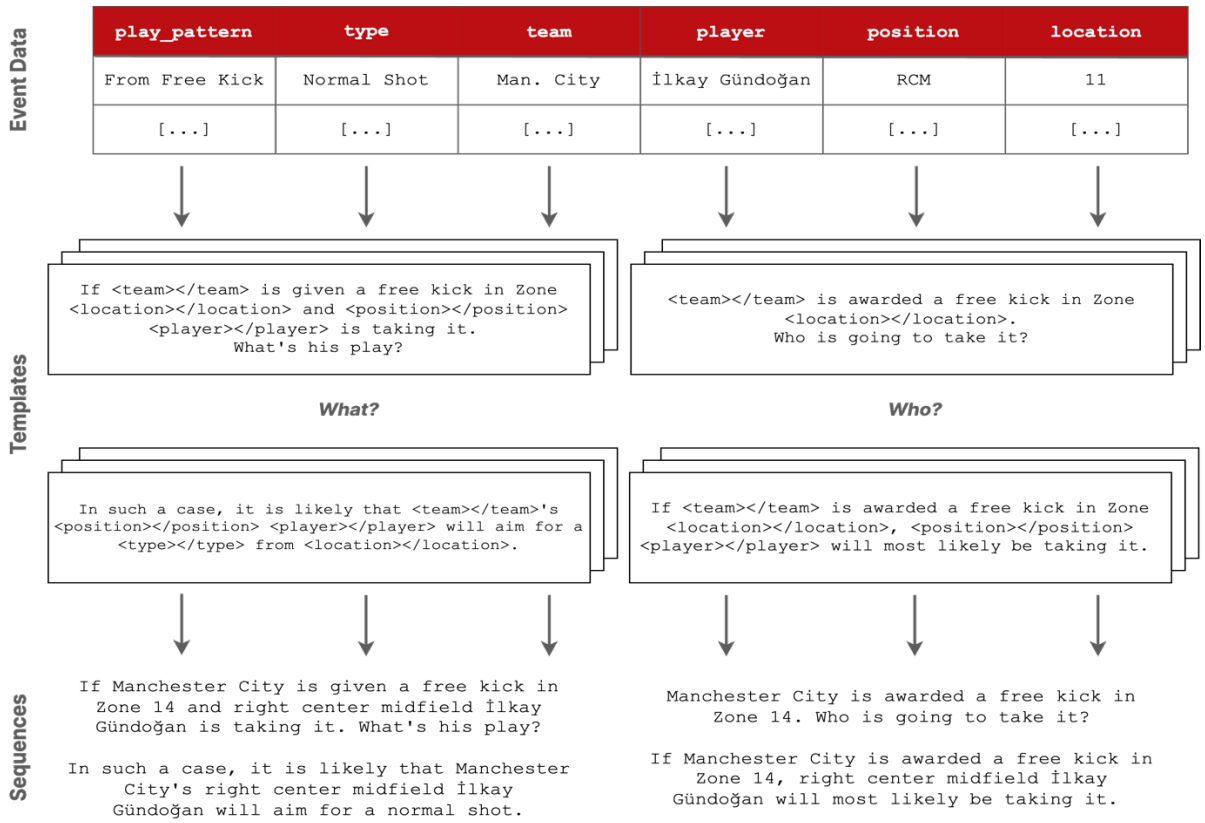

Figure 2: Dataset Generation

3.2 Supervised Fine-Tuning

Proceeding to the supervised fine-tuning phase, we began by carefully selecting and initialising a suitable foundation model – i.e., a generic deep learning model initially trained on a large amount of unlabeled data to accommodate diverse tasks and objectives. For this early study, we opted to experiment with the GPT-NeoX-20B architecture [1] – i.e., a 20-billion parameter autoregressive language model similar to GPT-3 – which was initially trained on a substantial dataset of 825 GiB, sourced from diverse, open-source language corpora. While numerous foundation models are currently available, GPT-NeoX-20B stands out for its robust performance across multiple tasks and its relative efficiency compared to larger architectures [1]. Lastly, building on the aforementioned efficiency and the capacity to fine-tune our model using consumer-grade graphics processing units (GPUs), we opted for 4-bit quantisation in conjunction with Low-Rank Adapters (QLoRA) – i.e., an efficient fine-tuning method that minimises memory consumption while maintaining 16-bit task performance for task-specific adaptability [3]. In essence, adapters are learnable components inserted between a pre-trained model's layers, eliminating the need to retrain the whole architecture [6].

Moving on to the second step, we decided, in order to enhance performance and factual accuracy, to provide our model with additional task-specific information during training, also known as context. More precisely, we chose to augment the data with the starting lineups for each team, a form of semi-structured data. In the future, we will experiment with adding other contextual information that might influence a team's tactics, like the current game state, into this context. This additional input was seamlessly integrated with the previously generated prompts to guide the model to produce responses that correctly identify players affiliated with their respective teams, thereby refining its task-specific capabilities.

Lastly, we fine-tuned our model using a next-token prediction objective by focusing on the low-rank adapters, limiting the process to a maximum of two epochs. In order to enhance the model's performance, we implemented adaptive learning rates and applied early stopping as a regularisation technique. Additionally, the AdamW algorithm was employed for stochastic optimisation. As common in most language modelling tasks, the primary objective was to train the model to discern the statistical correlations between successive tokens in a sequence. This enables a model to predict forthcoming tokens based on the previous ones, thus generating contextually appropriate and coherent text.

3.3 Reinforcement Learning

While not encompassed in the scope of this early work, the final phase aims to refine the model initially fine-tuned during Phase 2, employing Reinforcement Learning from Human Feedback (RLHF). As emphasised by OpenAI, this process is critical for achieving a model that is “safer, more helpful, and more aligned” [3]. Specifically, using RLHF is expected to result in a model adept at following instructions, consequently

generating more realistic responses adapted to a target audience of football professionals.

Using samples previously unseen by the model during the supervised fine-tuning phase, the first step consists of generating two non-identical responses for each sample. Then, a panel of football professionals is tasked with determining each prompt's best response based on a given context – i.e., a specific game situation. In the subsequent phase, the annotations provided by the expert panel serve to train a preference model, also referred to as a reward model, for the purpose of ranking the responses generated by the fine-tuned model. Ultimately, the last step focuses on employing the preference model to further refine the model initially fine-tuned in Phase 2, using the reward signals induced by the provided human feedback.

Results

In the setting of our application, the priority was to evaluate the model's factual accuracy rather than traditional LLM qualities such as creativity or fluency. Particularly in a domain like professional football, where coaches and analysts depend on trustworthy data for strategising against opponents, it is imperative that such a digital assistant refrains from producing inaccurate or hallucinated statements.

4.1 Textual Questions (What / Who)

Starting with *What*-type and *Who*-type questions, we evaluated our model using 100 unseen prompts inferred using a greedy decoding inference strategy – i.e., a deterministic approach by which the token with the highest probability at each decoding step is selected. For this evaluation, we employed a panel of two human annotators to conduct a rigorous evaluation of our model's performance. Given that traditional automated metrics have shown limitations in evaluating the nuanced output of Large Language Models, this human-centric evaluation allows for a more reliable and domain-specific assessment of the model's capabilities. Hence, the generated responses were scored across several criteria, including factual accuracy, identification of the correct player, and accurate zoning. This multi-faceted evaluation comprehensively assesses the assistant's dependability and usefulness.

Table 1: Factual Correctness

Type	Correct	Plausible	Improbable
What?	25 (0.50)	23 (0.46)	2 (0.04)
Who?	16 (0.32)	31 (0.62)	3 (0.06)

As can be seen in Table 1, we assessed the factual correctness, or accuracy, of our model utilising a three-class system system, namely:

- *factually correct* – i.e., the generated response aligned precisely with the ground truth in terms of factual correctness;
- *factually plausible* – i.e., i.e., the generated response, while not perfectly aligning with the ground truth, remains within the realm of possibility based on the given context (e.g., a different type of pass or a different player affiliated with the team in possession); and
- *factually improbable* – i.e., the generated response diverges significantly from the ground truth and presents information that is unlikely to be true within the given context (e.g., a player not affiliated with the team in possession).

Building upon the above three-class system, the evaluation results, as exemplified in Table 2, offer a nuanced portrait of the model's performance. For the *What*-type questions, the model exhibits a robust factual accuracy, with 50% of responses falling into the *factually correct* category. An additional 46% were deemed *factually plausible*, corroborating that the model's output is largely consistent with realistic game situations. On the other hand, *Who*-type questions yielded a different distribution: 32% were *factually correct*, while a majority, 62%, were *factually plausible*. This suggests that while the model may not always pinpoint the exact player, it often produces responses that are within the realm of possibility. Importantly, the rate of *factually improbable* outputs remained low across both question types – i.e., 4% for *What*-type and 6% for *Who*-type – highlighting the model's general reliability. Additionally, a further analysis of *What*-type questions reveals that the model demonstrated noteworthy specificity in player identification and spatial awareness: 60% of the generated responses accurately identified the correct player involved in the scenario, and 64% correctly pinpointed the zone location of that player, whether through a carry or a pass. This adds another layer to our understanding of the model's capabilities.

Table 2: Examples

Prompt	Ground Truth	Generated Response	Label
Liverpool's Mané has the ball in Zone 15. What's his next move?	Under those circumstances, Liverpool's Mané, is likely to make a ground pass to James Philip Milner in Zone 15.	Liverpool's Mané will probably make a ground pass to James Philip Milner in Zone 15.	Factually Correct
If Manchester United is given a throw in Zone 10, who will be taking it?	Anticipate Manchester United's Antony Matheus Dos Santos taking the throw in given these circumstances.	In this type of situation, Manchester United's left back Shaw usually handles the throw in.	Factually Plausible
In the event that Chelsea obtains a free kick in Zone 12, who is expected to take it?	It's a good bet that Chelsea's left center midfield Mount will take the free kick here.	Given that situation, expect Chelsea's goalkeeper Revuelta to take the free kick.	Factually Improbable

It is, however, worth emphasising that a factually plausible classification does not necessarily indicate an error on the part of the model. Employing a greedy decoding inference strategy constrains the model to generate the most probable response based on its learned patterns. Subsequent validation against our database confirmed that many of these factually plausible actions have indeed occurred previously. However, they may not align precisely with the ground truth for the specific prompt. This suggests that these actions may be less frequent or even more creative from the player's perspective. Thus, alternative decoding methods like sampling or beam decoding could provide a richer exploration of the model's learned distribution.

4.2 Spatial Questions (Where)

Continuing with the evaluation of *Where*-type questions, the same greedy decoding inference strategy was employed as for the previous question types. However, instead of relying on textual assessment, we examined the model's performance through a spatial lens, plotting the generated coordinates of 25 unseen samples against the ground truth. As can be seen in Figure 3, the results were less encouraging compared to *What*-type and *Who*-type questions. Given the situational context, the model predominantly generated coordinates that could be categorized as highly improbable. Given that these questions were trained on StatsBomb 360 data, it is clear that the task of accurately answering *Where*-type questions poses a significant challenge for the current model. This underscores the need for specialized refinement in future work, especially considering the high potential utility such inquiries could offer.

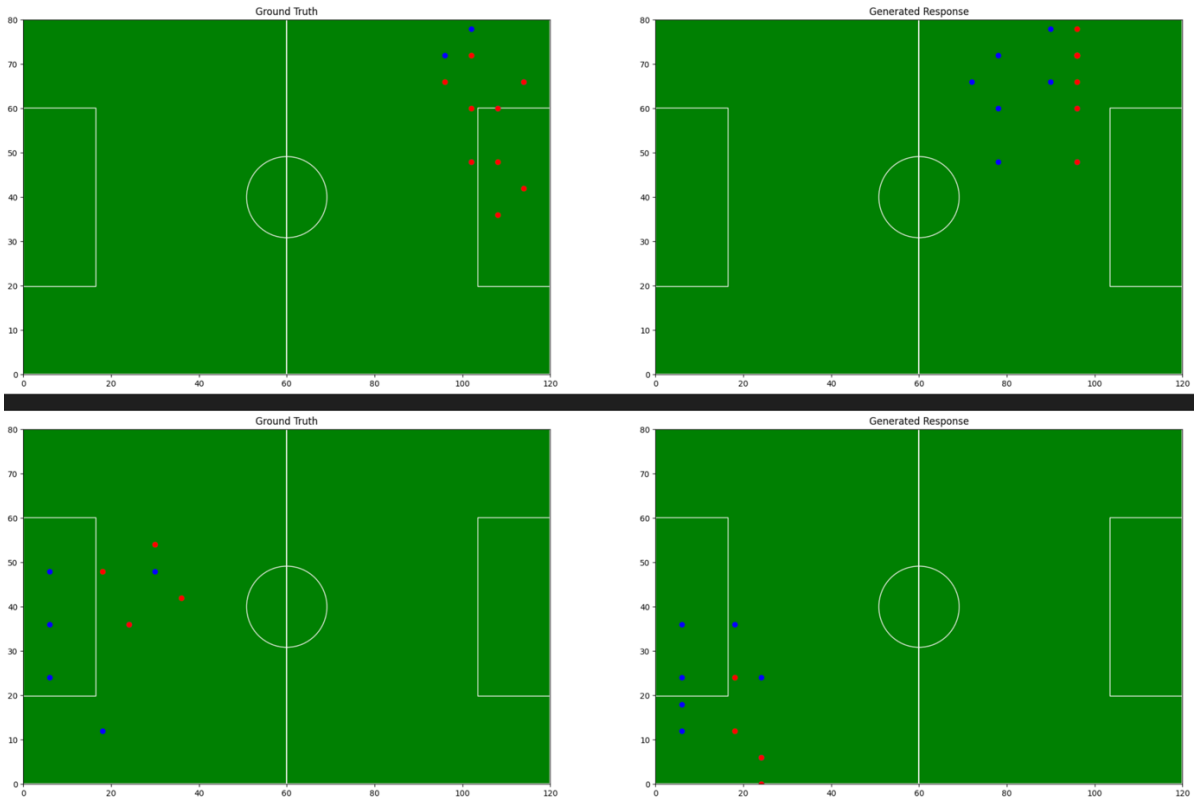

Figure 3: *Where?* – Ground Truth vs Generated Response

Conclusion

In conclusion, our work establishes a compelling precedent for integrating Large Language Models in sports analytics, particularly within professional football. TacticalGPT's ability to generate accurate and contextually relevant tactical insights presents a disruptive innovation in how coaching staff could prepare and harness data-driven analytics for strategy development. By utilizing a natural language approach, TacticalGPT not only simplifies the interpretation of complex tactical data but could also expedite the decision-making cycle. This aids in reducing the latency between data acquisition and actionable insights, a critical factor in a fast-paced environment like professional football.

The implications of this research extend far beyond the domain of tactical analysis. The intrinsic flexibility of the LLM architecture opens avenues for a plethora of applications, from individual player evaluation to real-time tactical adjustments. Given the promising preliminary findings, future research should focus on enhancing *Where*-type questions' performance. Our evaluation indicates room for substantial improvement in this area, signalling the necessity for specialized training and feature engineering. Additionally, investigating alternative decoding techniques like sampling presents a promising avenue for future research. Employing such a strategy could generate less common but plausible sequences, thereby revealing intriguing insights into less conventional plays or tactics. This would be particularly valuable for understanding the actions of highly skilled or creative players and teams, offering a nuanced layer of analysis that deterministic approaches may not capture. Hence, integrating these alternative decoding strategies could further elevate the model's utility, enabling it to discern unique or unanticipated strategies that could be pivotal in match preparations.

In light of the advancements in Generative AI and the demonstrated effectiveness of TacticalGPT, we believe that LLM-based solutions may soon become indispensable tools within professional football clubs' analytics and strategy divisions. This research, therefore, serves as an initial but robust step toward redefining the contemporary landscape of sports analytics.

References

[1] Sidney Black et al. “GPT-NeoX-20B: An Open-Source Autoregressive Lan- guage Model”. In: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. Association for Computational Linguistics, 2022.

[2] Tom Davenport and Maryam Alavi. How to Train Generative AI Using Your Company’s Data. July 2023. url: https://hbr.org/2023/07/how- to-train-generative-ai-using-your-companys-data.

[3] Tim Dettmers et al. QLoRA: Efficient Finetuning of Quantized LLMs. 2023. arXiv: 2305.14314 [cs.LG].

[4] Google. An Important Next Step on our AI Journey. Feb. 2023. url: https://blog.google/technology/ai/bard-google-ai-search- updates/.

[5] Suchin Gururangan et al. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. 2020. arXiv: 2004.10964 [cs.CL].

[6] Neil Houlsby et al. “Parameter-Efficient Transfer Learning for NLP”. In: International Conference on Machine Learning. 2019, pp. 2790–2799.

[7] OpenAI. Aligning Language Models to Follow Instructions. Jan. 2022. url: https://openai.com/research/instruction-following.

[8] OpenAI. ChatGPT - Release Notes. n.d. url: https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

[9] Long Ouyang et al. Training Language Models to Follow Instructions with Human Feedback. 2022. arXiv: 2203.02155 [cs.CL].

[10] Sebastian Ruder. “Neural Transfer Learning for Natural Language Pro- cessing”. PhD thesis. NUI Galway, 2019.

[11] Karan Singhal et al. Towards Expert-Level Medical Question Answering with Large Language Models. 2023. arXiv: 2305.09617 [cs.CL].

[12] Jason Stockwood. In a Few Years’ Time, Football Coaches may be using an AI Assistant. Apr. 2023. url: https://www.theguardian.com/ football/blog/2023/apr/11/in-a-few-years-time-football-coaches-may-be-using-an-ai-assistant.

[13] Shijie Wu et al. BloombergGPT: A Large Language Model for Finance. 2023. arXiv: 2303.17564 [cs.CL].