Research Proposal for PhD Sports Analytics

Ashish Soni

August 30, 2024

1 Working Title

Harnessing AI Agents to Automate Data Analysis and Reporting for Performance Analysis in Soccer using Open Source Large Language models

2 Project Motivation

Data analytics has become a key component in driving decision-making to optimize performance in professional football organizations according to the survey, that explores the characteristics of the data analytics infrastructure across various national federations and professional clubs [2]. The survey also highlights the underemployment of staff with expertise in applied data analytics and general resourcing limitations as major constraints to get meaningful insights from raw data in professional football organizations [2].

The new field of **Generative Artificial Intelligence** (AI) has given birth to **Large Languagde Models** (LLMs) also known as **Foundational Models** (FMs) such as **GPT** and **Claude** that are being used as powerful general-purpose agents for building new generation of AI Applications with the ability to perform tasks that require flexible reasoning and planning [1].

The recent advancements in Generative AI has ushered Technology into a new era but solving problems with an agent in a reliable way often requires the agent to be a compound system with multiple components instead of a single query to the model [1]. The agent also requires access to external tools such as search engines, code interpreters and database queries as well as domain-specific manual tuning [1]. A new research area has been formulated known as **Automated Design of Agentic Systems** (ADAS) with the aim of automating novel building blocks and designing powerful agentic systems [1].

The overarching *goal of the project is to build a compound agentic system using open source language models* like **Llama 3.1** released by **Meta AI** to automate data analysis and reporting for performance analysis in soccer that will serve as a solution to the constraints faced by professional football organizations to turn raw data into meaningful insights to gain competitive advantage on the field of play.

3 Potential Impact

The potential impact of research and project completion is as follows:

1. **Data-Driven Decisions:** The project will empower football organizations with automated data-driven insights leading to informed decisions in Performance Analysis to gain competitive advantage on and off the field.
2. **Resource Optimization:** Using automation in Performance Analysis alleviates the need for hiring a team of Data Analytics Experts, allowing organizations with financial constraints to still access key insights for improvement.
3. **Accessibility:** The use of open source language models and technologies will make the AI Data Analysis Agentic System more accessible and scalable across different levels of football organizations from small clubs to national federations.
4. **Innovation in Sports Analytics:** The project will contribute to the broader field of sports technology by integrating/embedding AI at the core of data analytics to further advance how technology is used in professional sports.

3.1 Dataset

The data to be used for analysis - *event data/play-by-play data*.

3.2 Technology Stack

Backend: Python, SQL, Supabase
Frontend: HTML, CSS, Next.js

3.3 Research Challenges

- **Integrating Open-Source Language Models:** Successfully incorporating open-source language models, like Llama 3.1, into the system will require a thorough understanding of their architecture, capabilities, and limitations. This challenge is compounded by the need to fine-tune these models to handle domain-specific tasks in soccer performance analysis.
- **Ensuring System Reliability and Scalability:** Building a reliable and scalable compound agentic system requires careful design and testing. This includes managing data pipelines, ensuring efficient model deployment, and maintaining system performance as data volume and complexity increase.

3.4 Skill Development Challenges

- Learning Next.js Framework for building modern web applications.
- Enhancing Frontend Development Skills to create a generative User Interface(UI) and a great User Experience(UX).

References

[1] Shengran Hu, Cong Lu, and Jeff Clune. "Automated Design of Agentic Systems". In: (2024), pp. 1–39. arXiv: 2408.08435. URL: http://arxiv.org/abs/2408.08435.

[2] Lorenzo Lolli et al. "Data analytics in the football industry: a survey investigating operational frameworks and practices in professional clubs and national federations from around the world". In: *Science and Medicine in Football* 00.00 (2024), pp. 1–10. ISSN: 24734446. DOI: 10.1080/24733938.2024.2341837. URL: https://doi.org/10.1080/24733938.2024.2341837.