

캡스톤 디자인 최종 보고서

스포츠 도메인 특화 LLM 개발

1. 서론

기존의 채팅형 인공지능 언어 모델(LLM)은 다양한 분야의 지식을 기반으로 학습되어 폭넓은 질문에 답변할 수 있는 장점을 가지고 있지만, 특정 도메인에 대해서는 한계를 드러내고 있습니다. 특히 스포츠 데이터를 다루는 데 있어 전문 지식 부족과 Hallucination(환각) 현상으로 인해 신뢰성과 정확도가 떨어지는 문제가 빈번하게 발생합니다. 이로 인해 사용자는 원하는 정보를 얻지 못하거나, 부정확한 정보를 제공받을 위험에 노출됩니다.

기존의 스포츠 도메인 특화 LLM 또한 단순한 지도 학습(SFT)을 수행하거나 거대언어모델 API와 RAG(Recurrent Attention Guidance) 기술을 활용하는 데 그쳐, 데이터의 깊이 있는 분석과 맥락에 적합한 정보 제공 능력이 부족하다는 한계가 있습니다. 이는 스포츠 도메인과 같이 전문성과 정확성이 요구되는 분야에서 실질적인 활용성을 제한하는 주요 요인으로 작용합니다.

이에 본 프로젝트는 기존 모델의 한계를 극복하기 위해 스포츠 도메인에 특화된 채팅 기반 LLM 서비스를 개발하고자 합니다. 본 서비스는 LLM이 데이터베이스에서 관련 정보를 직접 가져와 분석함으로써, 사용자에게 정확하고 신뢰성 있는 스포츠 데이터를 제공하는 것을 목표로 합니다. 이를 통해 스포츠 데이터 활용의 효율성을 높이고, 사용자 경험을 크게 향상시킬 수 있을 것으로 기대됩니다.

2. 본론

2.1 기존 축구 데이터 전처리

이번 프로젝트에서는 축구 도메인에 특화된 LLM 모델을 학습시키기 위해 기존의 축구 데이터를 전처리하는 작업을 수행하였습니다. 전처리 과정은 원천 데이터를 구조화하고, 학습에 적합한 형식으로 변환하며, 데이터 품질을 보장하는 데 중점을 두었습니다. 이 과정은 크게 데이터 템플릿 적용, 질문-답변 데이터 파싱, 데이터셋 분할 및 처리세 가지 단계로 구성되었습니다.

2.2.1 데이터 템플릿 적용

축구 데이터의 원천 데이터셋은 "Question", "Answer", "index", "length" 등의 컬럼으로 구성되어 있었습니다. 이를 모델 학습에 적합한 형식으로 변환하기 위해 사용자 정의 템플릿을 적용하였습니다. 이 템플릿은 각 질문과 답변 데이터를 하나의 텍스트로 결합하여, "text"라는 새로운 컬럼에 저장하였습니다. 예를 들어, 질문과 답변이 각각 "What is the player's scoring record?"와 "The player has scored 20 goals."인 경우, 템플릿을 적용하면 다음과 같은 형식으로 변환됩니다.

Question: What is the player's scoring record?

Answer: The player has scored 20 goals. (실제 개행문자 \n으로 연결)

이 과정을 통해 데이터는 모델 학습을 위한 통일된 형식을 갖추게 되었으며, 이후 단계에서 텍스트 데이터를 효율적으로 처리할 수 있는 기반이 마련되었습니다.

2.2.2. 질문-답변 데이터 파싱

템플릿을 통해 변환된 데이터에서 질문과 답변이 하나의 텍스트로 결합되어 있으므로, 이를 다시 분리하여 "question"과 "answer" 컬럼에 저장하는 작업을 수행하였습니다. 이를 위해 정규표현식을 사용하여 "Question:"과 "Answer:"를 기준으로 데이터를 추출하였습니다. 또한, 질문이나 답변이 누락된 데이터는 "error"로 처리하여 이후 데이터 품질 검증 과정에서 활용할 수 있도록 하였습니다. 이 작업을 통해 데이터 품질을 보장하고, 학습 중 발생할 수 있는 오류를 사전에 차단할 수 있었습니다.

2.2.3. 데이터셋 분할 및 처리

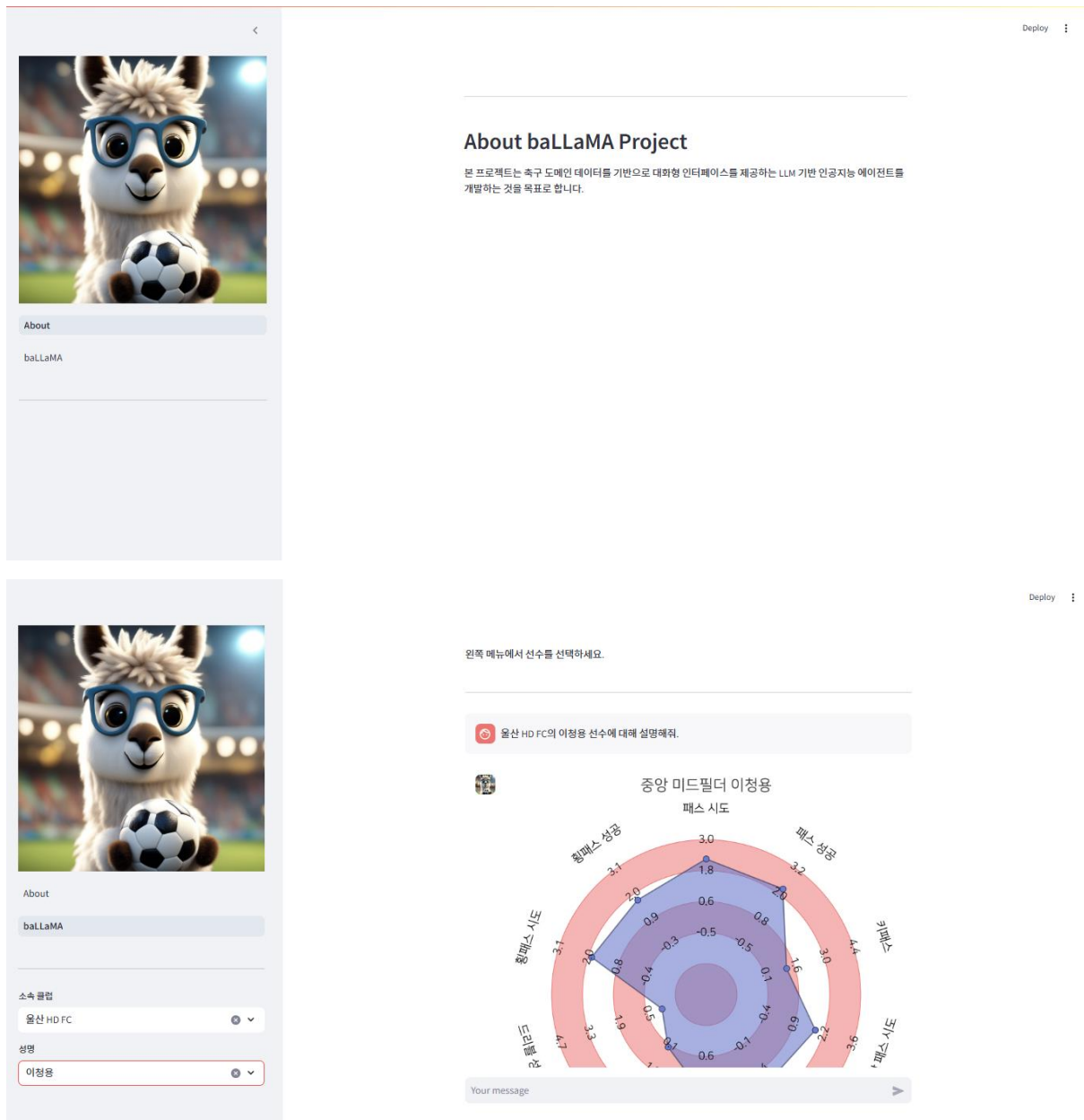
전처리된 축구 데이터는 학습(train), 검증(dev), 테스트(test)의 세 가지 데이터셋으로 분할되었습니다. 각 데이터셋에 대해 텍스트를 파싱하여 "question", "answer", "error" 컬럼을 추가하였으며, 오류 플래그를 통해 데이터 품질을 유지하였습니다. 이 작업은 데이터셋의 각 항목에 대해 반복적으로 수행되었으며, 결과적으로 모델 학습에 적합한 구조화된 데이터셋이 생성되었습니다.

축구 데이터의 전처리 작업은 모델 학습에 적합한 고품질 데이터셋을 구축하는 데 핵심적인 역할을 하였습니다. 질문과 답변 데이터의 구조화 및 오류 데이터의 관리 체계를 통해 데이터 품질을 보장하였으며, 이를 기반으로 모델의 학습 효율성을 극대화할 수 있었습니다. 또한, 정제된 데이터를 활용함으로써 LLM 모델이 축구 도메인에서 보다 정확하고 신뢰성 있는 결과를 도출할 수 있도록 했습니다.

2.3 ChatUI 개발

전처리된 데이터를 기반으로 사용자가 모델과 상호작용할 수 있는 대화형 UI(Chat UI)를 개발하였습니다. 이를 통해 사용자는 질문을 입력하고, 모델의 응답을 실시간으로 확인할 수 있는 환경을 제공받게 됩니다. 이 과정에서 Streamlit을 활용하여 Python 코드만으로 손쉽게 웹 애플리케이션을 제작하였으며, 이를 통해 효율적이고 직관적인 사용자 경험을 제공할 수 있었습니다.

Chat UI는 단순히 질문과 응답을 제공하는 것에 그치지 않고, 전처리된 데이터를 기반으로 한 모델 학습 결과를 시각화하거나 분석할 수 있는 기능도 포함하였습니다. Streamlit은 다양한 데이터 시각화 라이브러리와 통합을 지원하므로, 모델의 신뢰도나 응답 품질을 시각적으로 평가할 수 있는 그래프와 표를 생성할 수 있었습니다. 이 ChatUI를 통해 query를 손쉽게 할 수 있는 UI를 제공함으로써 프로젝트의 목표에 맞게 사용자 경험을 향상시키고 더 나아가 데이터 활용도를 극대화시킬 수 있었습니다.



▲직관적인 UI를 제공하고, 데이터를 시각화/분석하는 기능을 제공함

2.4 RAG 데이터 구성 및 Sports-Centric한 LLM 학습

이번 프로젝트에서는 RAG(Relevance-Augmented Generation) 기술과 Sports-Centric한 LLM 학습을 결합하여 축구 도메인에 특화된 모델을 개발하였습니다. RAG는 모델이 학습된 데이터만 활용하는 것이 아니라, 외부 데이터베이스를 실시간으로 참조하여 보다 풍부하고 신뢰성 있는 답변을 제공할 수 있도록 설계된 기술입니다. 이를 통해 사용자는 축구와 관련된 질문에 대해 정확하고 깊이 있는 답변을 받을 수 있으며, 모델의 실용성과 활용성을 크게 높일 수 있었습니다.

먼저, RAG 구현을 위해 고품질의 데이터베이스를 구축하였습니다. 나무위키 데이터와 축구 전문 데이터를 활용하여 축구 관련 용어, 전술, 선수 기록, 경기 역사 등의 정보를 포함한 데이터셋을 수집하였습니다. 이 데이터를 전처리하여 텍스트 내 불필요한 요소를 제거하고, 질문 유형별로 인덱싱하여 검색 효율을 높였습니다. 데이터베이스는 모델이 필요할 때

실시간으로 참조할 수 있도록 설계되었으며, 관련성이 높은 문서를 검색해 답변 생성에 활용하였습니다. 이번 프로젝트에서는 Gemini 의 tool-use 기능을 사용하여 dynamicRAG 로 구현했습니다.

	년도	선수명	포지션	등번호	출전시간(분)	득점	도움	슈팅	유효 슈팅	차단된슈팅	...	인터셉트	차단	획득	블락	블미스	파울	피파울	경고	퇴장	소속 클럽
0	2023	주민규	FW	18	2318	15	1	61	29	8	...	10	8	75	10	22	25	36	5	0	울산
1	2023	티아고	FW	28	2559	15	7	58	24	8	...	30	16	70	18	41	43	21	4	0	대전
2	2023	나상호	MF	7	2304	9	2	56	18	10	...	17	20	117	40	20	21	29	1	0	서울
3	2023	라스	FW	9	1536	9	3	41	14	5	...	10	7	59	7	21	10	10	1	0	수원FC
4	2023	고재현	FW	17	3015	9	1	58	21	11	...	41	36	151	44	27	28	26	5	0	대구
...
883	2023	류재문	대기	29	251	0	0	3	0	2	...	13	4	21	7	2	2	3	0	0	전북
884	2023	맹성웅	대기	28	122	0	0	1	0	1	...	1	3	11	3	0	4	2	1	0	전북
885	2023	박범수	대기	41	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	전북
886	2023	박진섭	대기	4	87	0	0	4	2	0	...	4	2	8	1	0	0	0	0	0	전북
887	2023	황순민	대기	20	45	0	0	0	0	0	...	2	3	2	2	0	2	0	0	0	수원FC

888 rows × 75 columns

```
{
  "user": "경고 지표의 정의가 뭐야?",
  "assistant": "경고(Yellow Cards)는 심판으로부터 받은 옐로우 카드의 수를 나타냅니다. 누적 경고는 출장 금지 등의 제재로 이어질 수 있어 경기 관리에 중요합니다."
},
{
  "user": "퇴장 지표의 정의가 뭐야?",
  "assistant": "퇴장(Red Cards)은 심판으로부터 받은 레드 카드의 수를 나타냅니다. 이는 즉시 경기장에서 퇴장당하며, 팀에 큰 불이익을 주므로 선수의 행동 관리가 중요합니다."
},
{
  "user": "기대득점(xG) 지표의 정의가 뭐야?",
  "assistant": "기대득점(xG, Expected Goals)은 선수가 시도한 각 슈팅의 득점 확률을 통계적으로 계산한 값입니다. 슈팅 위치, 각도, 사용한 신재 부위, 상대 수비수의 위치 등 다양한 요소를 고려하여 산출됩니다."
},
{
  "user": "경기당 TSG 평점 지표의 정의가 뭐야?",
  "assistant": "경기당 TSG 평점은 K리그 기술연구그룹(TSG: Technical Study Group)이 선수를 평가한 평점을 경기당 평균으로 계산한 값입니다. TSG는 K리그 경기력을 분석하고 전술 개발을 지원하는 역할을 하며,
```

▲ 나무위키 데이터를 활용해 고품질 데이터베이스 구축
프롬프트로 sports-centric한 정보를 학습

Sports-Centric한 LLM 학습은 모델이 축구 도메인에 대한 전문 지식을 깊이 학습할 수 있도록 설계되었습니다. 이를 위해 축구 용어와 개념, 선수 전적, 경기 기록과 같은 축구 특화 데이터를 Q&A 형태로 구조화하여 학습 데이터셋으로 사용하였습니다. 학습 과정에서는 일반적인 언어 모델이 제공하지 못하는 축구 도메인에 특화된 정보를 중심으로 모델을 학습시켜, 축구에 대한 이해도를 강화하였습니다. 특히, 전술적 개념(예: 오프사이드 규칙, 포메이션)과 선수 기록(예: 득점 수, 활동량) 등 구체적이고 정량적인 데이터를 포함하여 모델의 학습 품질을 높였습니다.

이 과정에서 RAG와 Sports-Centric한 학습 데이터셋을 결합하여 모델이 학습된 데이터와 외부 데이터를 동시에 활용할 수 있도록 하였습니다. 이를 통해 사용자가 질문을 입력했을 때, 모델은 학습된 정보와 실시간으로 검색된 정보를 결합하여 보다 풍부하고 정확한 답변을 생성할 수 있었습니다. 예를 들어, "K리그의 역사는 어떻게 되나요?"라는 질문이 주어졌을 때, 모델은 RAG 데이터베이스를 참조하여 "K리그는 1983년에 출범한 대한민국의 프로 축구 리그로 현재 12개 팀이 참가하고 있습니다."와 같은 답변을 생성할 수 있습니다.

또한, 선수의 경기 기록과 지표 데이터를 활용하여 사용자 질문에 정량적인 정보를 기반으로 한 응답을 제공하였습니다. 예를 들어, "손흥민 선수의 최근 경기 기록을 알려주세요."라는 질문이 입력되면, 데이터베이스에서 손흥민의 최근 경기 기록을 참조하여 "손흥민 선수는 최근 5경기에서 3골 2도움을 기록하였습니다."와 같은 정확한 답변을 제공합니다. 뿐만 아니라, 용어 설명과 전술 개념에 대한 질문에도 RAG 데이터와 학습 데이터를 결합하여 상세하고 신뢰성 있는 답변을 생성할 수 있었습니다.

결론적으로, RAG 데이터 구성과 Sports-Centric한 LLM 학습은 모델의 전문성과 신뢰성을 크게 강화하는 데 기여하였습니다. 모델은 학습된 데이터를 바탕으로 기본적인 지식을 제공할 뿐만 아니라, 외부 데이터베이스를 참조하여 보다 깊이 있고 풍부한 정보를 생성할 수 있었습니다.

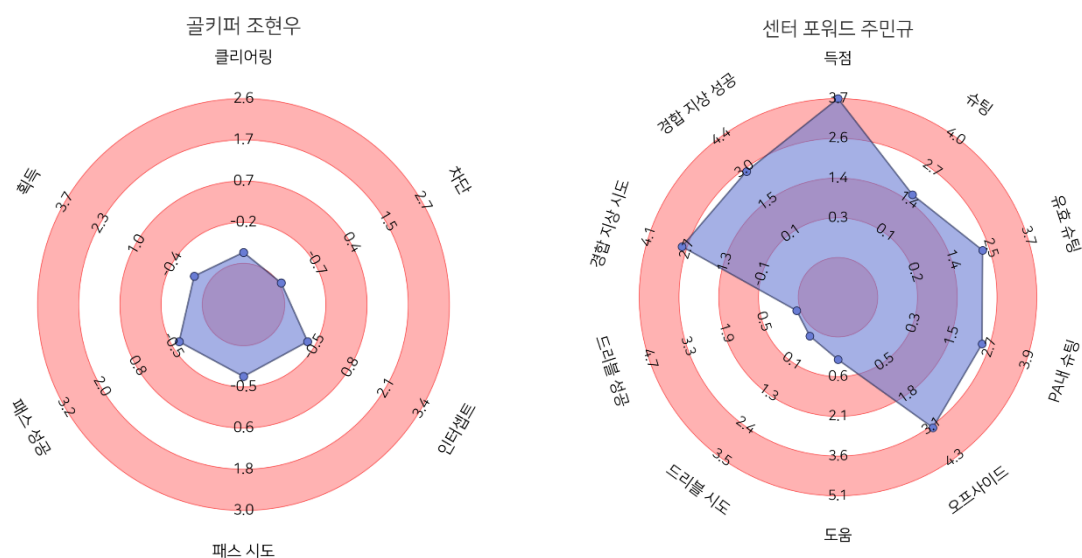
2.5 선수 별 평가지표

이번 프로젝트에서는 축구 선수들의 성과와 능력을 다양한 평가지표로 분석하고, 이를 시각적으로 표현하여 직관적으로 이해할 수 있도록 했습니다. 이를 통해 사용자는 특정 선수의 강점과 약점을 명확히 파악하고, 동일 포지션 내 다른 선수와의 비교 분석이 가능하도록 했습니다.

먼저, 각 선수의 주요 성과 지표를 기반으로 플롯 형태로 데이터를 시각화하였습니다. 이를 통해 득점, 패스 성공률, 드리블 성공률 등 다양한 지표를 한눈에 확인할 수 있습니다. 예를 들어, 특정 선수가 리그 내 평균 대비 어느 정도의 경쟁력을 가지고 있는지를 직관적으로 파악할 수 있으며, 해당 선수의 성과 변화를 시간에 따라 비교할 수도 있습니다.

또한, 축구는 포지션별로 요구되는 역할과 능력이 다르기 때문에 포지션별 평가지표를 별도로 정의하였습니다. 골키퍼의 경우 클리어링, 차단, 패스 성공률 등이 주요 지표로 설정되었으며, 수비수는 태클 성공률, 인터셉트, 경합 성공률 등이 포함되었습니다. 미드필더는 패스 성공률, 탈압박, 드리블 성공률 등을 기준으로 평가되었으며, 공격수는 득점, 슈팅 정확도, 드리블 성공률 등이 주요 지표로 사용되었습니다. 이러한 포지션별 지표는 각 포지션에서 요구되는 핵심 능력을 기반으로 설계되어, 선수들의 성과를 보다 정확히 측정할 수 있도록 하였습니다.

특히, 포지션 내에서도 세부적인 역할에 따라 지표를 세분화하였습니다. 예를 들어, 센터백은 태클 성공률과 경합 공중 성공률이 중요한 지표로 설정되었으며, 풀백은 크로스 성공률과 전방 패스 성공률 등이 추가로 고려되었습니다. 공격형 미드필더는 키퍼스와 공격지역 패스 성공률, 득점 등이 중요 지표로 평가되었으며, 센터포워드는 득점, 슈팅 정확도, 드리블 성공률 등이 포함되었습니다. 이러한 세부 지표는 각 선수의 고유한 역할과 기여도를 정밀하게 측정할 수 있도록 설계되었습니다.



▲ 선수별 평가지표 시각화

3. 결론

이번 프로젝트에서는 축구 도메인에 특화된 LLM을 개발하여 사용자에게 유용한 정보를 제공하고, 선수의 능력치를 분석하는 다양한 기능을 구현하였습니다. 이를 통해 축구 팬, 전략 분석가, 에이전시 등 여러 이해관계자들에게 실질적인 가치를 제공할 수 있는 시스템을 구축하였습니다. 특히, RAG(Relevance-Augmented Generation)와 선수별 지표 시각화 등의 기술을 활용하여 축구 데이터를 보다 직관적이고 활용도 높은 형태로 전환할 수 있었습니다.

프로젝트의 주요 성과로는, 축구 용어와 개념 검색, 선수 전적 검색, 포지션별 능력치 시각화, RAG 기반 정보 제공 등이 있습니다. 이러한 기능은 사용자가 특정 선수나 팀의 정보를 쉽고 빠르게 이해할 수 있도록 돕는 동시에, 전략 분석과 의사 결정 과정에서 데이터 기반의 통찰력을 제공합니다.

이번 프로젝트에서는 몇 가지 기술적 한계도 발견되었습니다. 첫째, SFT(Supervised Fine-Tuning)와 RAG를 결합한 모델 학습을 진행하지 못해 보다 높은 성능을 달성하지 못한 점입니다. 이로 인해 모델의 성능이 데이터 제공 방식에 따라 다소 제한될 수 있었습니다. 둘째, Gemini 임베딩 모델이 한 번에 1만 바이트만 처리할 수 있는 제약으로 인해 문서를 여러 개로 쪼개야 했습니다. 이로 인해 문서 파편화가 심화되었고, 일부 경우에는 모델 성능에 부정적인 영향을 미쳤습니다. 이러한 문제는 향후 해결해야 할 중요한 과제로 남아 있습니다.

향후 방향

프로젝트의 다음 단계에서는 기술적 한계를 극복하고 시스템의 성능과 활용도를 더욱 높이기 위해 다음과 같은 작업을 계획하고 있습니다.

1. SFT와 RAG 결합

SFT와 RAG를 함께 사용하여 모델을 학습함으로써 보다 뛰어난 성능을 달성할 수 있는 환경을 조성할 계획입니다. 이를 통해 모델이 더 복잡한 질문에도 높은 정확도로 답변할 수 있도록 할 것입니다.

2. 오픈소스 LLM 활용

Gemini와 같은 상용 모델에 의존하지 않고, Llama와 같은 오픈소스 LLM을 활용하여 모델을 구현할 예정입니다. 이를 통해 개발 유연성을 높이고 비용 효율성을 개선할 수 있을 것입니다.

3. 기능 확장 및 응용

축구 도메인 외에도 다양한 응용 분야에서 활용 가능성을 탐색할 계획입니다. 주요 응용 방안은 다음과 같습니다:

전략 분석가용: 포메이션의 장단점 분석 및 가장 적합한 선수 추천, 경기 패턴 분석을 통한 상대 전략 대비, 특정 선수와 유사한 스타일의 선수 탐색.

축구 팬용: 토너먼트 진출 가능성 예측, 경기 주요 장면 요약, 규칙 설명, 뉴스 형식으로 경기 결과 제공.

에이전시 및 스카우팅: 선수 간 지표 비교, 성장 가능성 예측, 포지션별 이상적인 선수 추천, 재계약 여부 분석.

이번 프로젝트는 축구 데이터를 활용한 축구 도메인에 특화된 LLM의 실질적인 가능성을 보여준 사례로, 다양한 기능 구현을 통해 데이터 기반 의사결정과 사용자 경험을 크게 향상시켰습니다. 향후에는 기술적 한계를 극복하고 새로운 응용 사례를 발굴함으로써, 축구뿐만 아니라 다른 스포츠 및 도메인에서도 폭넓게 활용될 수 있는 시스템으로 발전시킬 계획입니다.