## Dataset

Son utilizados en el proyecto **LLM Deportivo** representan una recopilación estructurada de datos relacionados con jugadores, partidos, estadísticas históricas, biografías y contenido, estos datos son esenciales para entrenar el modelo, validar su rendimiento y permitirle responder consultas complejas relacionadas con el fútbol profesional. Cada conjunto de datos fue diseñado con base en objetivos específicos: desde alimentar el entrenamiento del modelo, hasta facilitar el análisis semántico y contextual de información textual y numérica.

## Estructura del Dataset

Los datasets fueron organizados en forma de registros con campos claramente definidos y cada registro representa una instancia individual (por ejemplo, un jugador, un partido o una pregunta). La estructura común de los datasets incluye:

* Encabezados que definen cada atributo (columna).
* Registros distribuidos en filas (en caso de estructura tabular).
* Uniformidad en formatos de fechas, puntuación numérica y etiquetas de texto.
* Separación temática según el propósito del dataset (biografías, estadísticas, prompts de entrenamiento, entre otros).

## Limitaciones

* **Desbalance de información**  
  Algunos jugadores o equipos populares están sobre-representados en los datos, mientras que otros con menor presencia mediática apenas aparecen. Esto puede sesgar las respuestas del modelo hacia ciertas figuras recurrentes.
* **Contenido ruidoso**  
  Algunas fuentes incluían texto con etiquetas HTML, símbolos especiales, emojis o errores de codificación, lo cual dificultó el procesamiento automático del texto.
* **Campos incompletos**  
  No todos los registros contaban con todos los campos requeridos, lo que puede limitar la efectividad del modelo en tareas como comparativas o generación de respuestas detalladas.
* **Falta de contexto temporal**  
  En ciertos casos, los datos no especificaban con claridad la temporada o el momento en que ocurrió un evento, complicando el análisis temporal o la evolución estadística.