Desarrollo de un LLM para la consulta interactiva de Datos Abiertos del Gobierno Vasco

1. Descripción General

La propuesta consiste en el desarrollo de un **Modelo de Lenguaje a Gran Escala** (LLM, por sus siglas en inglés) similar a Chat GPT, entrenado específicamente con datos abiertos del Gobierno Vasco, enfocado en proporcionar una interfaz conversacional y gráfica que permita a los usuarios obtener respuestas precisas y visualizaciones dinámicas. El modelo estará diseñado para procesar grandes volúmenes de datos estructurados y no estructurados, proporcionando un sistema de consulta intuitivo basado en lenguaje natural, sin necesidad de conocimientos avanzados en análisis de datos.

2. Objetivo del Proyecto

El objetivo es desarrollar una **plataforma interactiva** basada en un **LLM entrenado con datos abiertos** que permita consultas y análisis sobre diferentes dominios (demografía, empleo, economía, industria, movilidad, etc.) del Gobierno Vasco y otras entidades locales como diputaciones y ayuntamientos. La solución ofrecerá respuestas textuales en lenguaje natural y **representaciones gráficas automatizadas**, como gráficos de líneas, tablas y mapas interactivos, según la naturaleza de la consulta.

3. Arquitectura del Sistema

3.1. Recopilación y Preprocesamiento de Datos

El modelo se entrenará utilizando datos abiertos del portal del Gobierno Vasco, que incluyen:

- **Datos estructurados**: series temporales de indicadores económicos, estadísticas laborales, datos demográficos, etc.

  Ejemplos:

  - Datos historicos de poblacion: https://api.euskadi.eus/udalmap/indicators/162
  - Datos historicos del PIB per capita: https://api.euskadi.eus/udalmap/indicators/56
- **Datos no estructurados**: informes públicos, descripciones sectoriales, reportes de tendencias, entre otros.

  Ejemplos:

  - Informe del mercado laboral lanbide 2024-1: https://www.lanbide.euskadi.eus/contenidos/estadistica/balance_mercado_laboral_2015/opendata/BalanceTrimestral_I_2024_vf.pdf

    - Informe de Salud Publica 2022: https://www.euskadi.eus/contenidos/informacion/informes_salud_publica/es_def/adjuntos/Informe-Salud-Publica-2022.pdf

Para preparar estos datos, se implementarán las siguientes fases de procesamiento:

- **Extracción y Transformación**: Uso de APIs o scraping para extraer datos del portal de datos abiertos en formatos como CSV, JSON, XML.
- **Normalización y Limpieza**: El preprocesamiento de los datos incluye la limpieza de valores nulos, la normalización de campos, y el tratamiento de valores atípicos. Se emplearán técnicas de **ETL (Extract, Transform, Load)** para asegurar que todos los conjuntos de datos se integren de manera coherente.
- **Enriquecimiento**: Se aplicarán técnicas de enriquecimiento de datos para unificar registros provenientes de diferentes fuentes, y se crearán representaciones vectoriales que faciliten la entrada de estos datos en el modelo.

3.2. Entrenamiento del Modelo

El modelo base será un **transformer de arquitectura tipo GPT** o una versión personalizada de **BERT (Bidirectional Encoder Representations from Transformers)** adaptada para tareas de comprensión de lenguaje natural y generación de texto. Los pasos para su entrenamiento incluyen:

- **Fine-tuning del LLM**: El modelo será pre entrenado utilizando grandes corpus de datos en español, y posteriormente se ajustará (fine-tuning) específicamente con los datos abiertos del Gobierno Vasco. Esta especialización permitirá que el modelo comprenda los dominios de consulta, las nomenclaturas y la terminología propia de las diferentes áreas de gobierno (por ejemplo, en términos de indicadores económicos, sociales, o sectoriales).
- **Tokenización y embeddings personalizados**: Se usarán técnicas de **tokenización** que preserven el significado específico de términos sectoriales, como “PIB industrial”, “tasa de paro”, “movilidad urbana”, etc. Para representar los datos numéricos, se integrarán **embeddings continuos** que permitan al modelo tratar consultas con dimensiones temporales o categóricas de forma precisa.

3.3. Interfaz Conversacional e Integración con Visualización

- **API de procesamiento de lenguaje natural**: El modelo se desplegará mediante una **API RESTful** que facilitará la interacción con aplicaciones frontend. Los usuarios podrán realizar consultas en lenguaje natural, que serán procesadas en tiempo real por el LLM.
- **Interfaz de Usuario Amigable**:
  Se desarrollará una interfaz sencilla y accesible para que los usuarios no técnicos puedan interactuar con el modelo. Esta interfaz permitirá hacer consultas por texto, seleccionar el tipo de visualización deseada (gráfico, tabla, etc.) y obtener resultados detallados en tiempo real.
- **Generación de respuestas textuales y gráficas**:

    - **Respuestas textuales**: El modelo generará explicaciones basadas en la consulta del usuario, respondiendo en lenguaje natural.
    - **Generación automática de visualizaciones**: Dependiendo de la consulta, se generarán visualizaciones dinámicas mediante bibliotecas como **D3.js, Plotly o Matplotlib** para representar la evolución temporal de los datos, distribuciones, relaciones categóricas, entre otras. Estas visualizaciones se integrarán en la respuesta de manera automática. Ejemplos:
        - Para consultas sobre evolución de indicadores: gráficos de líneas.
        - Para consultas sobre distribución de población o economía: gráficos de barras o mapas interactivos.
- **Procesamiento de consultas complejas**: El sistema será capaz de procesar consultas complejas que impliquen cálculos, agregaciones y filtros específicos. Por ejemplo:
    - **"Muéstrame la evolución de la tasa de paro en Álava en los últimos 5 años, comparado con Bizkaia"**: el modelo extraerá los datos relevantes, aplicará las agregaciones necesarias y generará una respuesta textual y un gráfico comparativo.

3.4. Infraestructura Técnica y Despliegue

- **Backend**: El LLM se desplegará en una infraestructura basada en la nube (usando **Kubernetes o Docker** para escalabilidad) con acceso a GPUs para la inferencia en tiempo real.
- **Base de datos**: Se utilizará una base de datos relacional (como **PostgreSQL**) para almacenar los datos abiertos transformados, mientras que para datos no estructurados o semiestructurados se integrarán soluciones como **Elasticsearch**.
- **Capa de cache**: Para mejorar la latencia en la entrega de respuestas, se implementará un sistema de cache utilizando **Redis** o **Memcached** para consultas recurrentes.
- **Monitorización y Mantenimiento**: Se implementarán herramientas de monitoreo (como **Prometheus** y **Grafana**) para controlar el rendimiento del sistema, la carga de consultas y el estado del servidor.
- **FrontEnd**: Se desarrollará una plataforma web intuitiva para facilitar el acceso de los usuarios a las consultas. Se implementará un sistema de autenticación para evitar el abuso del servicio y gestionar el acceso según el perfil del usuario.

4. Casos de Uso y Aplicaciones

4.1. Consultas sobre Indicadores Socioeconómicos

El LLM permitirá a los usuarios realizar consultas avanzadas sobre cualquier tipo de indicador socioeconómico:

- Ejemplo: “¿Cuál ha sido la variación del Producto Interno Bruto en el sector industrial de Gipuzkoa en los últimos 5 años?”
- Salida: El modelo responderá con un resumen textual de la variación, seguido de un gráfico de líneas mostrando la evolución anual.

4.2. Análisis Sectoriales y Comparativos

Los usuarios podrán realizar consultas comparativas entre territorios o sectores:

- Ejemplo: “Comparar la evolución del empleo en los sectores tecnológicos en Álava y Bizkaia durante los últimos 3 años.”
- Salida: El modelo generará tablas comparativas y gráficos de barras representando la variación en ambos territorios.

4.3. Movilidad y Datos de Tráfico

Se podrán obtener datos actualizados sobre movilidad, como patrones de tráfico o uso del transporte público:

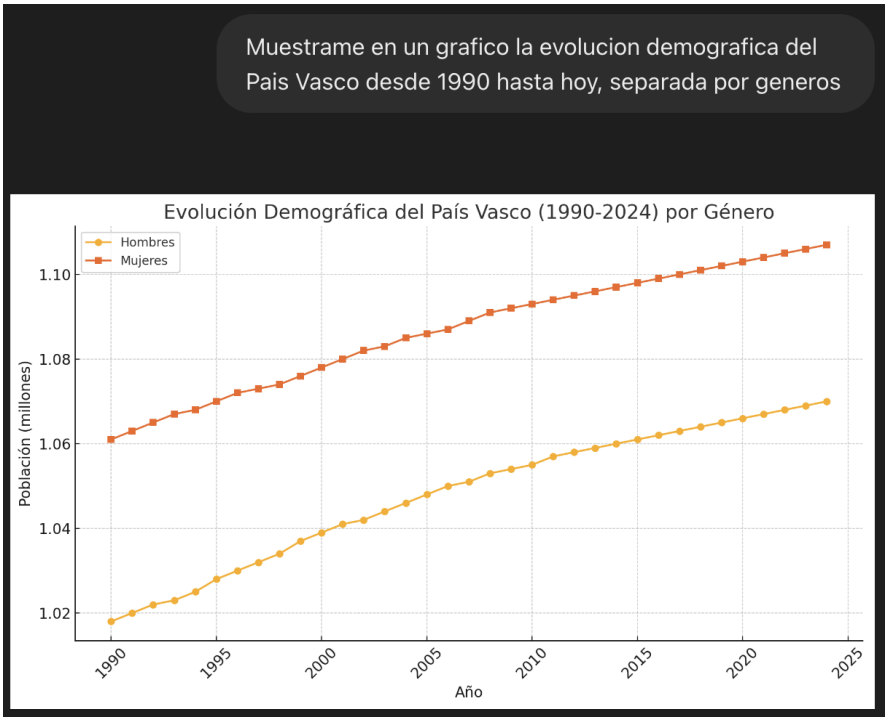
- Ejemplo: “Mostrar los patrones de movilidad en el área metropolitana de Bilbao durante los fines de semana del último mes.”
- Salida: El modelo devolverá gráficos de mapas de calor que muestran las zonas con mayor movilidad.

5. Desafíos Técnicos y Estrategias de Mitigación

- **Actualización y mantenimiento de los datos**:
  La integración continua con los portales de datos abiertos será crucial para mantener la información actualizada. Se automatizará el proceso de ingestión de nuevos datos mediante pipelines de **ETL** programados que verifiquen, procesen y actualicen los datos en la base de datos del sistema.
- **Optimización del rendimiento**:
  Para manejar el procesamiento en tiempo real de grandes volúmenes de datos, se utilizarán **técnicas de paralelización** y **carga diferida** (lazy loading) en las visualizaciones, para evitar sobrecargar el sistema ante consultas complejas.
- **Escalabilidad y disponibilidad**:
  El sistema será diseñado para ser escalable horizontalmente, permitiendo la distribución de la carga de trabajo mediante **balanceadores de carga** y réplica de servicios. Además, se implementarán **copias de seguridad** automáticas de los datos y el modelo para garantizar alta disponibilidad.
- **Protección de la privacidad**:
  Aunque se usarán solo datos públicos, se aplicarán medidas de seguridad como **cifrado y anonimización** para garantizar que no se comprometan datos sensibles. El sistema cumplirá con las normativas de **protección de datos (RGPD)**, asegurando auditorías periódicas para mantener la seguridad en todas las fases.

6. Ejemplo visual

En la siguiente imagen podemos ver un pequeño ejemplo del prompt (la información qué inserta el usuario) y la respuesta obtenida. En este caso un gráfico qué nos muestra la evolución demográfica durante unas fechas concretas de hombres y mujeres.



7. Conclusión

El desarrollo de este LLM para la consulta de datos abiertos del Gobierno Vasco proporciona una solución altamente técnica y accesible para la explotación de información pública. Esta plataforma, basada en IA y visualización dinámica, democratiza el acceso a los datos, incrementando la transparencia, el análisis y la toma de decisiones basadas en datos en la comunidad vasca.

8. Reconocimiento y Proyección Internacional

Este desarrollo posicionaría al **Gobierno Vasco** como un referente en la adopción de tecnologías avanzadas, específicamente en el uso de **Inteligencia Artificial (IA)** aplicada a la **gestión de datos abiertos**. El lanzamiento de este proyecto demostraría un claro compromiso por parte del Gobierno Vasco de estar a la vanguardia tecnológica, impulsando la transformación digital en la administración pública.

El Gobierno Vasco sería pionero a nivel estatal en el uso de **modelos de lenguaje a gran escala (LLM)** entrenados con datos abiertos, lo que podría atraer un importante reconocimiento en los medios y entre los principales actores del sector tecnológico. Esta visibilidad reforzaría su reputación como un líder innovador, no solo dentro del contexto nacional, sino también en la escena internacional, particularmente a nivel **europeo**, donde el uso de IA en el sector público es una prioridad creciente.

Además, este proyecto abriría la puerta a futuras colaboraciones con otros **gobiernos y entidades públicas**, tanto en España como en el extranjero, interesados en replicar el éxito del modelo. Al establecer un **precedente en la integración de IA y open data**, el Gobierno Vasco podría compartir su experiencia y conocimientos, actuando como un **socio**

estratégico para otros gobiernos que deseen desarrollar sus propios sistemas de análisis de datos abiertos basados en IA.

Finalmente, el impacto de este proyecto podría ir más allá del ámbito gubernamental, influyendo positivamente en otros sectores clave, como la **educación, la investigación y la industria tecnológica**, que podrían beneficiarse de un acceso más sencillo y eficiente a la información pública. Esto fortalecería el **ecosistema de innovación** en el País Vasco y contribuiría al crecimiento de una **economía basada en el conocimiento y la tecnología**.