PARTE 1: SELECCIÓN DEL MODELO DECODER

DOCUMENTACIÓN

1. Investigación y comparación de diferentes opciones de modelos de tipo decoder

- **GPT-4o-mini:** es una versión compacta del modelo GPT-4 desarrollado por OpenAI. Está diseñado para ofrecer un equilibrio entre rendimiento y eficiencia computacional.
- **BART:** Modelo secuencia a secuencia de Facebook AI, útil para tareas de generación y comprensión.
- **T5:** Modelo texto a texto de Google, versátil para varias tareas de NLP.
- **Llama 3.1: 8B:** Modelo de 8B parámetros, eficiente y potente en generación de texto.

MODELO	PARÁMETROS	CALIDAD DE GENERACIÓN	RECURSOS NECESARIOS	FLEXIBILIDAD	FACILIDAD DE INTEGRACIÓN
GPT-4o-mini	175B	Muy Alta	Alto	Alta	Media
BART	400M	Alta	Medio	Alta	Alta
T5	11B	Alta	Alto	Muy Alta	Media
Llama 3.1: 8B	8B	Muy Alta	Alto	Alta	Alta

2. Evaluación de los requisitos del proyecto para seleccionar el modelo más adecuado

Este proyecto tiene como objetivo desarrollar un codigo que generación de texto que me ayudará al final del curso, a desarrolla un chatbot.

- **Balance entre rendimiento y eficiencia:** GPT-4o-mini ofrece capacidades avanzadas de generación de texto similares a las de GPT-4, pero con requerimientos computacionales reducidos, lo que es crucial para proyectos con recursos limitados.
- **Calidad de generación de texto:** La capacidad de GPT-4o-mini para comprender y generar texto de manera coherente y precisa es altamente valorada, especialmente para tareas que requieren una alta calidad en la generación de contenido.
- **Popularidad y soporte:** Si bien es una versión más compacta, se beneficia del amplio soporte y recursos disponibles para los modelos GPT

desarrollados por OpenAI, facilitando la implementación y resolución de problemas.

- **Eficiencia en términos de costos:** La eficiencia de GPT-4o-mini en términos de uso de recursos computacionales puede traducirse en menores costos operativos, un factor importante en muchos proyectos.

Evaluando:

- GPT-4o-mini: Alta calidad y costo mediano comparado al 4 regular.
- BART: Menos recursos pero alta flexibilidad.
- T5: Muy versátil pero requiere recursos significativos (sobre todo al entrenarlo.
- Llama 3.1: 8B: Alta calidad, eficiente y manejable en términos de recursos.

3. Justificación clara de la elección del modelo encoder

Luego de haber evaluado los modelos, el modelo seleccionado para este proyecto es GPT-4o-mini. Las razones para esta elección son:

- **Calidad de Generación Muy Alta:** GPT-4o-mini proporciona una calidad de generación de texto superior, lo que es crucial para tareas avanzadas de generación de texto.
- **Flexibilidad:** Aunque T5 es extremadamente flexible, GPT-4o-mini también ofrece un alto grado de flexibilidad y puede adaptarse bien a varias tareas sin perder calidad.
- **Facilidad de Integración y Recursos:** A pesar de requerir recursos altos, GPT-4o-mini es manejable con la infraestructura disponible para el proyecto y tiene un soporte robusto en la comunidad de NLP.