SpanishMedicaLLM

Un Modelo Grande de Lenguaje (LLM) para el contexto médico en idioma español

Hackaton NLP 2024

Marzo 2024

Propuesta

- Conjuntos de datos para el entrenamiento
- Hipótesis para el autoajuste del modelo
- Propuesta de modelos para el autoajuste
- Estructura del entrenador
- Evaluación de la propuesta

Hackaton NLP 2024

Decoder-Only (GPT-like Models)

Use Cases:

- **Generating Medical Text:** Generating discharge summaries, patient instructions, or creating medical content.
- **Question Answering:** Providing answers to medical questions based on a large corpus of medical knowledge.
- **Dialogue Systems:** Powering conversational agents for patient engagement or support.

Approach:

- **Data Preparation:** Assemble a large corpus of medical texts, including dialogues (if available), Q&A pairs, and general medical information.
- **Preprocessing:** Similar to the BERT approach but ensure the texts are suitable for generative tasks.
- **Fine-Tuning:** Use a pre-trained GPT model and fine-tune it on your dataset. You may experiment with different prompts and fine-tuning strategies to improve performance on generative tasks.

Conjuntos de datos para el entrenamiento

1. Conjuntos de datos de fuentes médicas en español

2. Conjuntos de datos generados a partir de Question and Answer (Q&A) tomando como fuentes libros de medicina y artículos escritos en español usando un LLM (p.ej BioMistral)

3. Conjunto de datos generados a partir de la traducción de fuentes médicas (e.d artículos, libros, datasets de Q&A) en ingles al español

4. Conjunto de datos de medicina alternativa (e.d natural, acupuntura, otros) de libros, artículos escritos en español (novedad) (reto como evaluar)

5. Conjunto de datos generados a partir de Q&A de medicina alternativa de fuentes médicas tomando como fuente libros, artículos escritos en español (novedad) (reto como evaluar)
6. Conjunto de datos generados a partir de Q&A de medicina alternativa de videos de youtube en español (novedad) (reto como evaluar)

Conjunto de datos	Tipo (1,2,3,4,5)	Cantidad de Tokens	Fuente

Hipótesis para el autoajuste del modelo

- Hipótesis u Objetivo General:

El autoajuste de un modelo LLM fundacional (e.d Llama2, Falcon , BioMistral, Meditron) con un alto grado o completamente pre-entrenado con textos en español, para datos de entrenamiento médico dará buenos resultados si:

- ✓ Se usa la técnica de Qlora permitirá mantener el desempeño y el costo del modelo (e.d memoria, velocidad de respuesta, etc)
- ✓ Se usa la estrategia mezclado (mixed) de BioMistral propuesto en el trabajo “BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains” dará mejores resultados en el desempeño del modelo
- ✓ Si se usa la estrategia de replay (Aprendizaje Continuo) usado en Meditron en el trabajo “Meditron-70b: Scaling medical pretraining for large language models”
- ✓ El uso de estrategia como **Direct Preference Optimization (DPO)** permite mejorar el desempeño sin usar RLHF
- ✓ El uso en al autoajuste de fuentes de medicina tradicional y natural permitirá generar resultados o soluciones a problemas médicos más accesibles a los usuarios
  - Puede obtenerse con una estrategia Retrieval Augmented Generation (RAG)

Estructura del entrenador (Modelos LLM base a usar)

- Criterios para la selección:

Objetivo de la investigación:

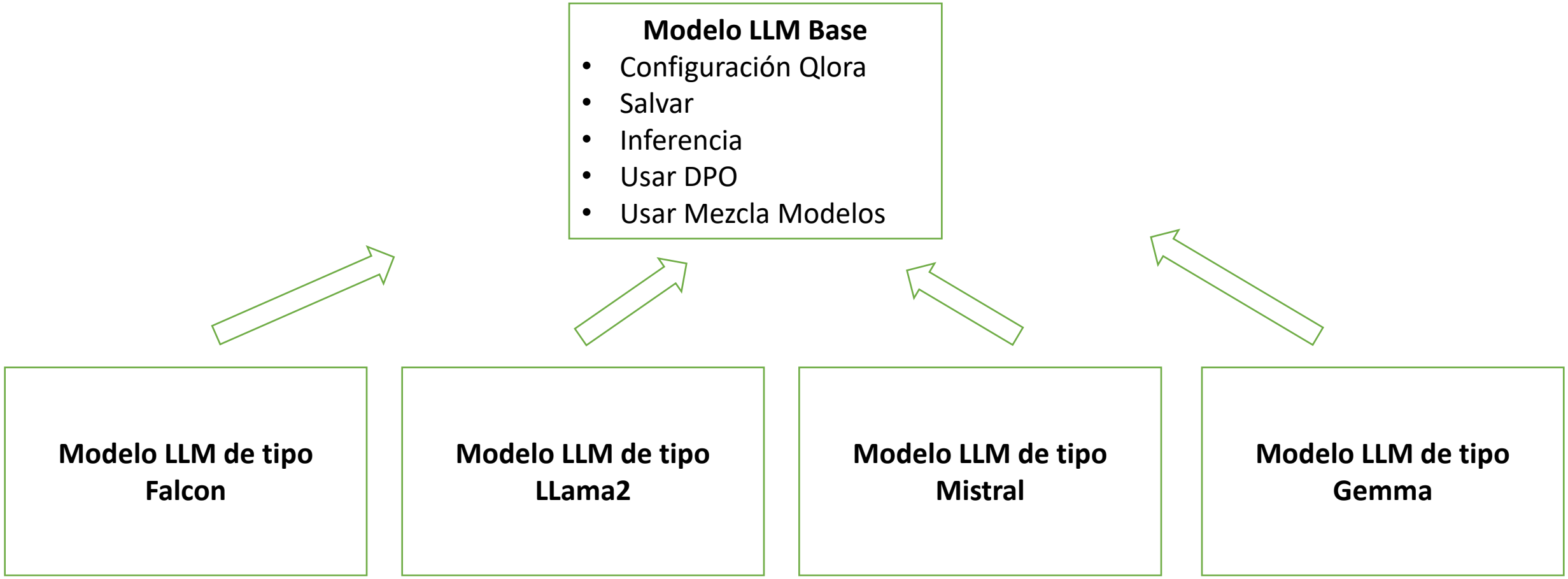
✓Analizar todos los modelos LLM en español en el repositorio hugginface construido con un modelo fundacional como Llama2 u otro

Criterios de investigación:

✓Modelo de fundación Meditron similar, para reutilizar todos los requisitos usados.
✓Otro modelo con criterios competitivos como Mistral, Gemminis
✓Licencias Open Source o similares para próxima implementación en varios contextos
✓Fue construido con corpus en español o durante su pre-entrenamiento o autoajuste se uso alguna parte del corpus en idioma español

Estructura del entrenador (Modelos LLM base a usar)

Nombre	Modelo Base	Licencia	Prentrenado en español	URL
projecte-aina/aguila-7b	Falcon-7B	Apache License, Version 2.0	Si	https://huggingface.co/projecte-aina/aguila-7b
clibrain/Llama-2-7b-ft-instruct-es	LLama2	Apache 2.0		https://huggingface.co/clibrain/Llama-2-7b-ft-instruct-es
TheBloke/Barcenas-Mistral-7B-GGUF	Mistral			https://huggingface.co/TheBloke/Barcenas-Mistral-7B-GGUF
clibrain/lince-zero	based on Falcon-7B	Apache 2.0		https://huggingface.co/clibrain/lince-zero
BioMistral/BioMistral-7B	Mistral	Apache 2.0		https://huggingface.co/BioMistral/BioMistral-7B
clibrain/Llama-2-13b-ft-instruct-es	LLama2	Apache 2.0		https://huggingface.co/clibrain/Llama-2-13b-ft-instruct-es
google/gemma-7b-it		gemma-terms-of-use		https://huggingface.co/google/gemma-7b-it
allenai/OLMo-7B		Apache 2.0		https://huggingface.co/allenai/OLMo-7B
clibrain/Llama-2-13b-ft-instruct-es-gptq-4bit	Llama2	Apache 2.0		https://huggingface.co/clibrain/Llama-2-13b-ft-instruct-es-gptq-4bit
clibrain/lince-mistral-7b-it-es	Mistral	Apache 2.0		https://huggingface.co/clibrain/lince-mistral-7b-it-es
Kukedlc/Llama-7b-spanish	LLama2	Apache 2.0		https://huggingface.co/Kukedlc/Llama-7b-spanish
google/gemma-7b	Gemminis	gemma-terms-of-use		https://huggingface.co/google/gemma-7b
allenai/OLMo-1B		Apache 2.0		https://huggingface.co/allenai/OLMo-1B


Modelo LLM Base
• Configuración Qlora
• Salvar
• Inferencia
• Usar DPO
• Usar Mezcla Modelos
Modelo LLM de tipo Falcon
Modelo LLM de tipo LLama2
Modelo LLM de tipo Mistral
Modelo LLM de tipo Gemma

Proceso de entrenamiento para un autoajuste del LLM

Conjunto de Datos

Artículos de medicina en español

Libros de medicina en español

Conjunto de preguntas y respuestas en español

Entrenar

SpanishMedicaLLM

Direct Preference Optimization (DPO)

Usar LLM base:

- BioMistal
- Modelo para contexto Médico

Usar, durante el autoajuste, una técnica de mezcla de modelos:

TIES y DARE (Ver artículo de BioMistral)

Proceso de entrenamiento para un autoajuste del LLM

Conjunto de Datos

casos_clinicos_tratamiento

SpanishMedicaLLM

Evaluar

casos_clinicos_diagnostico

Exams to access a specialized position in the Spanish healthcare system (head_qa)

Desempeño:

- Accuracy
- F1-Macro
- Otras para LLM en contexto médico

Evaluación de la propuesta

- Medicas a emplear: Accuracy, F1-Macro
- Conjuntos de Datos para evaluar LLM autoajustado en español para evaluar el desempeño de los modelos
  - ✓Tomando como referencia el método de evaluación de Meditron y BioMistral
  - ✓Usar el conjunto de datos:
    - LenguajeNaturalAI/casos_clinicos_tratamiento
    - LenguajeNaturalAI/casos_clinicos_diagnostico
    - head_qa
    - Evaluación de **prompt** y **respuestas** sobre medicina natural y tradicional **(reto)**

SpanishMedicaLLM

Un Modelo Largo de Lenguaje (LLM) para el contexto médico en idioma español