# Práctica LLM – Problema a resolver

- Testear el modeloMistral-7B-LLM-Fraud-Detection para detectar fraude con texto de un chatbot en español.
- El modelo mistral-7b-fraud2-finetuned Large Language Model (LLM) es una versión perfeccionada del modelo generativo de texto Mistral-7B-v0.1 que utiliza diversos conjuntos de datos de transcripciones fraudulentas generados sintéticamente.

# Práctica LLM – Resolución

- Primero se deberá cargar el modelo en el Notebook para testearlo.
- Según el resultado se realizará un fine-tune al modelo.
- Posteriormente se publicará el modelo en Hugging Face.
- Se implementará un servicio con LangChain para probar el modelo.
- Por último se hará un RAG para mejorar más el modelo y testearlo.

# Práctica LLM – Conclusiones

- Al descargar el modelo se comprueba que en español trabaja bastante bien.
- Aún así, se realiza un finetune para mejorarlo usando un dataset realizado con ChatGPT.
- Se ha tenido que subscribir al Google Colab Pro+ ya que con los recursos de la versión de gratuita no había suficiente.
- Para poder entregar el modelo se ha usado la CPU en vez de la GPU, ya que la memoria se llenava y se reiniciaba el Colab. También se ha realizado una cuantificación del modelo, se ha reducido el tamaño del batch y se ha usado Lightning para aligerar el modelo.
- A continuación se ha subido el modelo a Hugging Face en https://huggingface.co/Otger/fine_tuned_mistral_7B_fraud_detection
- Para probar el modelo se ha implementado un servicio de LangChain.
- Lo último que se ha realizado es crear un RAG para probar el modelo.