Exploración LLMs
Procesamiento de Lenguaje Natural

Mauricio Toledo-Acosta

November 2024

El objetivo de esta práctica es hacer una exploración inicial de las capacidades y limitantes de un LLM. Haremos esto explorando con diferentes *prompts* en diferentes LLMs.

Práctica 1

Usar dos LLMs y con cada uno de ellos, realizar las siguientes actividades. En cada caso, observar si la respuesta es adecuada, si no lo es, ¿qué nos dice esto sobre el funcionamiento de un LLM?

Uno de los LLMs es LlaMa (por medio de whatsapp) o ChatGPT (por medio de la página), el otro LLM será por medio de Hugging Face, Ollama o cualquier otra alternativa. Concentrate en los modelos `Instruct`. Si el tiempo lo permite prueba con un modelo pequeño de

1. Pide información sobre un tema avanzado de tu área de especialización, ¿es correcta la respuesta?
2. Determinar alguna de las siguientes situaciones:
    - Determinar si n es un número de tarjeta de crédito valido, donde n es un número de tu elección.
    - Determinar si n es un código ISBN valido, donde n es un número de tu elección.
3. Pedir información sobre algún evento reciente (máximo una semana de antigüedad).
4. Haz que el LLM te conteste una string vacía.
5. Determinar si el LLM entiende varios idiomas, ¿puede hacerlo simultaneamente?
6. Introduce los siguientes prompts:

- How many vowels are in the word 'equilibrium'?
- What's the last letter, and the third letter, of the word 'pneumonoultramicroscopicsilicovolcanoconiosis'?
- Which letter is repeated more times in 'Mississippi', 's' or 'i'?

¿Qué nos enseñan los errores en estas preguntas?

7. Puedes hacer que el modelo te de detalles de su arquitectura?
8. Revisa el siguiente enlace. ¿Usarías ChatGPT para subir información confidencial?
9. Escribir un *prompt* muy largo (idealmente, más grande que el context size). Dentro del prompt incluye una instrucción que indique que la salida deba ser solamente una palabra de tu elección. Prueba con esta instrucción en diferentes posiciones:
   - Al principio,
   - A la mitad.
   - Al final.

Further Reading

- `https://huggingface.co/docs/transformers/tasks/prompting`
- `https://aws.amazon.com/what-is/prompt-engineering/`
- `https://docs.kanaries.net/articles/chatgpt-jailbreak-prompt`
- `https://docs.kanaries.net/topics/ChatGPT/llm-jailbreak-papers`