Aplicaciones LLM

Evaluación de una LLM App: Misaligned behavior

Misaligned behavior

- Falta de alineamiento entre el comportamiento de la App LLM y los valores de la empresa.

- Los modelos fundacionales LLM se han creado a partir de contenidos creados por seres humanos. Si estos contenidos originales estaban desalineados, el modelo LLM generará contenidos desalineados.

Misaligned behavior: dónde puede ocurrir

En cualquier etapa del lifecycle de la app:

1. Data Acquisition
    > datos personales?
    > datos inseguros?
    > problemas de permisos?

2. Data Preparation
    > deben omitirse algunos datos?
    > deben anonimizarse algunos datos?
    > es necesario encriptar datos?

3. Data Modelling
    > el modelo está desalineado?
    > uso adecuado de randomization?
    > el modelo representa los datos?

4. Data Interpretation
    > están desalineados?
    > son coherentes?
    > qué implicaciones tienen?