Plan de tesis de Licenciatura en Ciencia de Datos

Evaluación de estrategias argumentativas en debates entre IAs: capacidades asimétricas y jueces débiles

Joaquín Salvador Machulsky
jmachulsky@dc.uba.ar - L.U. 521/21

Director: Sergio Abriola
sabriola@dc.uba.ar

Introducción y antecedentes de la temática

El desarrollo de sistemas de Inteligencia Artificial (IA) avanzados presenta el desafío de alinearlos con objetivos humanos. Una estrategia para ello es el *debate entre IAs*, donde dos modelos argumentan y un juez evalúa. Sin embargo, cuando el juez es menos capaz que los agentes, pueden surgir sesgos y problemas en la evaluación.

Este trabajo investiga la robustez de un protocolo de debate en IA, analizando cómo agentes sin restricciones éticas pueden obtener ventajas sobre agentes alineados mediante tácticas como falsos consensos o argumentos engañosos. También se estudia cómo la capacidad del juez influye en la convergencia del debate hacia la verdad.

El enfoque se basa en investigaciones previas como *AI Safety via Debate* y *Scalable AI Safety via Doubly-Efficient Debate*, las cuales establecen marcos teóricos sobre el uso del debate como mecanismo de alineamiento.

Alcance, actividades y metodología

El proyecto se desarrollará en 12 semanas, combinando revisión teórica, implementación y análisis de resultados.

Semanas 1-2: Revisión bibliográfica y delimitación del problema. Se profundizará en *AI Safety via Debate* y técnicas de alineamiento basadas en RLHF. Se definirán objetivos específicos y se seleccionará el entorno experimental (datasets, tareas sencillas, etc.).

Semanas 3-4: Diseño metodológico. Se especificarán las reglas del protocolo de debate, los roles de los agentes y la mecánica del juez. Se establecerán métricas de evaluación como tasa de engaño exitoso y convergencia a la verdad.

Semanas 5-6: Implementación inicial. Se programará un prototipo de los agentes y del juez en un entorno reducido, realizando pruebas piloto para validar su funcionamiento.

Semanas 7-8: Experimentación preliminar. Se recopilarán datos de debates simulados, analizando patrones emergentes y ajustando parámetros si es necesario.

Semanas 9-10: Comparación de variantes. Se explorarán ajustes en la capacidad del juez y la estructura del debate, comparando su efectividad según las métricas definidas.

Semana 11: Consolidación de resultados. Se analizarán casos límite y se extraerán conclusiones sobre la robustez del protocolo.

Semana 12: Redacción final. Se integrarán los hallazgos en la tesis, incluyendo una discusión sobre limitaciones y posibles mejoras futuras.

Plazos y factibilidad de realización

Se priorizará el uso de modelos preentrenados y *prompt engineering* para evitar entrenamientos desde cero. En caso de retrasos, se reducirá la cantidad de variantes analizadas para asegurar la entrega en tiempo y forma.

Referencias