Proposta de Projeto 2025-2026

LICENCIATURA EM ENGENHARIA INFORMÁTICA	*PROPOSTA N.º*

*TÍTULO**	**Identificação dos componentes do viés de confirmação em um datasets de questões de saúde para Grandes Modelos de Linguagem**
*ORIENTADOR PRINCIPAL**	*Hugo Paredes*
COORIENTADORES	*Rafael Ris-Ala*
ALUNOS(S)	2
ÁREA DE INVESTIGAÇÃO	*Inteligência Artificial, Inteligência Artificial Centrada no Ser Humano*
CENTRO DE INVESTIGAÇÃO	*HumanISE – INESC TEC*
DEPENDÊNCIAS	*Conhecimentos básicos sobre Inteligência Artificial e Aprendizagem de Máquina, que fornecem as bases teóricas e metodológicas necessárias para a análise de modelos de linguagem, avaliação de vieses e condução de estudos científicos.*
*APRESENTAÇÃO**	Com a crescente utilização de Grandes Modelos de Linguagem (Large Language Models - LLMs) no domínio da saúde, torna-se fundamental compreender o compotamento enviesado dos LLMs e os fatores que levam a tal comportamento, principalmente no que diz respeito ao viés de confirmação. As questões humanas podem conter diferentes formas de viés de confirmação, como pressupostos implícitos, formulações sugestivas, linguagem emocional ou direcionamento para diagnósticos específicos, que influenciam as respostas dos LLMs e até as deciões médicas. A identificação sistemática desses componentes de enviesamento é essencial para garantir avaliações mais justas, resultados mais confiáveis e o uso ético de sistemas baseados em LLMs em contextos sensíveis como a saúde.
*OBJETIVOS**	O objetivo deste projeto é capacitar os estudantes a projetar um detector de viés de confirmação. Para isso, os estudantes devem identificar e caracterizar os componentes do viés de confirmação em um dataset composto por questões de saúde enviesadas, através de um processo estruturado que envolve a análise linguística e a categorização dos componentes de enviesamento. Este projeto contribui para a concepção e validação de um detector de enviesamento, bem como para a formação de engenheiros de software capazes de projetar, avaliar e utilizar sistemas de inteligência artificial de forma crítica e responsável.

*** Campos de preenchimento obrigatório**

NOTA: a totalidade deste documento (exceto esta linha) não deve exceder uma página.