# Tittle: AI Response Classifier

Northeastern University
Khoury College of Computer Sciences

**Khoury College of Computer Sciences**
**Northeastern University**
**Team Members: Hancheng Huang, Yiyang He, Guanghao Huang**

## Introduction

With more and more applications of AI in different domains, the requirement to correctly identify human-generated and artificial intelligence-generated responses has grown. Our project aims to train two different models on classifying whether the response is from human beings or LLM. This project aims to deliver a reliable tool that can be used in various sectors to promote authenticity in digital interactions.

## Problem Statement

As AI-generated content becomes more common and many AI-generated contents are misused in many places. What's more, the lack of tools to effectively distinguish response between human and AI-generated causes a risk to integrity, that makes the situation worse. Also, It effects many areas, such as education and customer service. Our goal is trying to train models to solve these issues with high accuracy.

## Proposed Solution

Our proposed solution includes training two types of models: a transformer-based model BERT and a non-transformer-based model LSTM. We will fine-tune each model to classify whether the response is generated from human or AI. By comparing their performance, we will determine which model is more accurate for the process of distinguishing. We use dateset sourced from Kaggle and contains 480000 text response. These texts are labled as either human or AI-generated. The dataset includes a feature labeled "# generated" that indicates whether the response is from human(0) or AI(1). We decide to pick a subset of the dataset for model training and evaluation.

## Literature Review

Previous studies, we refer two lectures. One is Campino, J. Unleashing the transformers: NLP models detect AI writi ng in educati on. J. Comput. Educ. (2024). Another is Hayawi, K., Shahriar, S., & Mathew, S. S. (2024). The imitati on game: Detecti ng human and AI-generated texts in the era of ChatGPT and BARD. Journal of Informati on Science, 0(0).

## Conclusions

In our model evaluation, we will focus on key performance metrics such as accuracy, precision, recall, and the F1 score. By comparing these metrics across different models, we will assess their ability to generalize and make informed predictions on our dataset.

**Acknowledgement**

This presentation is supported by Khoury College, Northeastern Arlington