**Jing Tan**
**Scott Stossel**

## Project Proposal

Hallucination detection in large language models (LLMs) focuses on identifying when a model generates text that sounds plausible but is actually false or unsupported by real facts. It is a key step in making AI systems more trustworthy and reliable, especially in sensitive areas like healthcare or finance. Modern detection methods often break down a model's response into smaller factual statements and then check each one for accuracy using another model such as GPT-4 or through external sources. Some techniques look at the model's own confidence or internal signals to spot likely errors. Recent studies, including Li et al. (2024) and newer approaches like Belief Tree Propagation, show that combining self-checking, logical consistency, and retrieval of real evidence improves detection. Overall, effective hallucination detection aims to help LLMs recognize and flag misinformation before it reaches users, balancing accuracy, reliability, and practicality for real-world applications.

In this project, we'll investigate how hallucination detection models compare in LLM-based text summarization. We will use a dataset of article-summary pairs where each summary is labeled for factual consistency (factual vs. hallucinated). A variety of evaluation metrics will be collected between the models with the goal of creating a more evidence-based comparison of how different hallucination detection methods perform.

## Why is it important?

Hallucination detection in large language models (LLMs) is essential because it safeguards the reliability and trustworthiness of AI-generated information. As LLMs are increasingly used in high-stakes settings—such as education, healthcare, finance, and law—factually incorrect or fabricated content can lead to serious misunderstandings, poor decisions, or even real-world harm. Detecting hallucinations helps users distinguish between accurate knowledge and confident but false statements, ensuring that AI systems remain credible and transparent. It also enables developers to identify weaknesses in training data or model behavior, guiding improvements in alignment, reasoning, and truthfulness. Ultimately, effective hallucination detection is a cornerstone for responsible AI deployment—it builds user trust, promotes accountability, and helps bridge the gap between human-level reliability and machine-generated intelligence.

## Data

The first dataset comes from Kyrscinski et. al. (2019) in the evaluation of factual consistency of abstractive text generation. It is a dataset of automatically generated factuality labels for articles from CNN Stories and Daily Mail Stories. The generation process for this data is detailed here: [https://github.com/salesforce/factCC](https://github.com/salesforce/factCC). Additionally, we plan to use HaluEval 2.0 which is detailed in Li et. al. (2024).

### Algorithms

There are a number of implementations we plan to use in this experiment. We will start with FactCC, a BERT factuality classifier, as a baseline. Then, we will compare this to other BERT-based models and a more advanced model such as using belief tree propagation as described in Hou et. al. (2025).

### Sources

The Dawn After Dark: An Empirical Study on Factuality Hallucination in Large Language Models
https://aclanthology.org/2024.acl-long.586/

Chain-of-Verification Reduces Hallucination in Large Language Models
https://aclanthology.org/2024.findings-acl.212/

A Probabilistic Framework for LLM Hallucination Detection via Belief Tree Propagation
https://aclanthology.org/2025.naacl-long.158/

Learning to Summarize from LLM-Generated Feedback
https://aclanthology.org/2025.naacl-long.38/

Evaluating the Factual Consistency of Abstractive Text Summarization
https://arxiv.org/pdf/1910.12840