**Determining Factual Accuracy of AI-generated Text**

CS 697R - GRADUATE SPECIAL PROJECTS

Brigham Young University

Trevor Andrus, Quinn Snell

Spring / Summer 2023

**Abstract**

With the advent of large language models such as OpenAI’s chatGPT, AI-generated text is becoming more common in both academic and professional spheres. However, while the grammatical accuracy of the models is readily apparent, it seems the factual accuracy of the information from these models is often left unchecked. This paper seeks to remedy that apparent shortcoming. We propose a framework for testing the factual accuracy of AI text, demonstrate the shortcomings of current sources to satisfy the framework, and in the meantime present an alternative framework.

We show that our proposed framework falls short because the consistent and accurate extraction of subject-object pairs from sentences has not yet been solved. We further show that even with the creation of knowledge graphs based on these subject object pairs, current knowledge bases are structured to identify only a finite number of relations between entities. This may be useful for identifying specific relationships but cannot currently be generalized to an arbitrary number of relations – which will be required to establish a comprehensive method for determining factual accuracy.

Despite this, we also propose a method of zero-shot learning that may give insight into the current state of LLM self-reflection. The problem of determining factual accuracy requires a massive amount of generalization, and the current open-source-free-to-use leader of general AI (at least in the natural language space) is OpenAI’s ChatGPT. Using ChatGPT, we show a method to have it self-analyze its own factual accuracy, and comment on potential pitfalls of this method.

**Introduction**

In an era dominated by the rapid evolution of technology, the emergence of AI-generated text presents a double-edged sword: a tool with immense potential to revolutionize communication, research, and creativity, yet also a vector for the inadvertent dissemination of misinformation. The fusion of sophisticated language models and artificial intelligence has empowered machines to mimic human-like text generation with remarkable accuracy, blurring the line between authentic content and fabricated information. As these AI systems become increasingly adept at generating coherent narratives, there is a growing concern about their potential to inadvertently amplify and propagate misleading or entirely false information, thus challenging the very foundations of truth and reliability in our digitally interconnected world.

Our original framework (while not currently feasible with available resources) is as follows:

1. Split input text into individual sentences
2. Using a part of speech (POS) tagger, identify the parts of speech in a given input text.
3. Using these POS-tagged sentences, extract the subject and object of each sentence
4. Using dependency parsing, infer the relationships between these subjects and objects
5. Create knowledge graphs from the subject-object-relationship pairs
6. Compare knowledge graphs against established knowledge base repositories (such as DBpedia)