**Thesis Project Portfolio**

# Watermarking of AI-Generated Text: Word Replacement Through Prompt Engineering

(Technical Report)

# Deadly Lies: How Twitter Fuels Misinformation Campaigns with Lethal Impacts

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

**Maxwell John Penders**

Spring, 2025

Department of Computer Science

## Table of Contents

## Sociotechnical Synthesis

In the age of the internet, people are exposed to more information than ever at a faster rate than ever. While much of this information is beneficial and true, a large subset of it isn't. False information online, whether created by humans or large language models (LLMs), is massively damaging to the world, leading to the deaths of many persons who chose not to receive the COVID-19 vaccination because of false information they viewed online (Ferreira Caceres et al., 2022; Jia et al., 2023). My STS and technical theses aim to combat this widespread threat: on the technical side, by improving our ability to watermark and therefore recognize LLM-generated text, and on the STS side, by analyzing what other methods of counterattack exist to stem the flow of false information.

LLMs have become increasingly powerful and simple to use, making them an easy method for generating deliberately false information (disinformation) or even propaganda. One promising method of countering the prevalence of LLM-generated disinformation on the internet is through textual watermarking, which encodes a secret "signature" into text generated by an LLM service which can then be used by social media companies to identify LLM-generated text with a much lower false positive rate than traditional LLM detection. My technical thesis aimed to develop a simpler method of watermarking LLM-generated text by including explicit instructions for the watermark encoding process within the prompt. This builds on years of textual watermarking research by replacing algorithmically encoded watermarking with prompt instruction-based encoding. While constraints on the resources I had available meant that I could only build a relatively simple model with few unique signatures, this model succeeds in its modest goal by providing a 77.35% detection accuracy rate with an extremely low <0.15% false positive rate.

My STS looks at the existing literature and research regarding the problem of mis- and dis-information, through Actor-Network Theory, to answer what methods exist for us as a society to counter the issue. To ensure the scope was reasonable, I examined the spread of mis- and dis-information through the social media website and notorious hotspot for misinformation, Twitter (now X). By identifying the relevant actors at play in the creation, dissemination, and consumption of false information and understanding how they interact with each other, my thesis is able to gain insight into the motivations and methods behind the individual actors. While there

will never be a catch-all solution to any issue, there is a combination of countermeasures which seem to be effective in stemming mis- and dis-information which may be implemented. First, modernizing the federal government and its departments to account for the modernization of information dissemination. Second, continuing to also treat the problem as a threat to international security, ensuring international cooperation through a body like NATO. Finally, perhaps most important is that these measures are used to create strong media literacy education programs similar to what already exists in Finland and Sweden, while adapting some of the messaging to the American culture and zeitgeist.

While I certainly think that I succeed in creating work with novel ideas and positive impact, I understand that neither thesis was groundbreaking work. When it comes to the technical side, my hope is that others with higher levels of expertise in textual watermarking can use my novel method of prompt-instruction watermarking to improve on existing research. On the STS side, it is beneficial to add to the discourse, be it academic, political, or mainstream, and present an argument backed by evidence which may be built upon, criticized, or adopted. It is unreasonable to think that any single proposed solution is perfect, but any societal and governmental movement in the right direction has a positive impact.

For my STS thesis, I would like to thank Professor Wylie for her help and encouragement, Claire Miller for being the best reviewer I could have asked for, and Farah Pandith for lending me her immense expertise, deep wisdom, and time in aiding in my discovering sources of evidence, understanding the macro issue, and expanding my passion for the subject.