CAPSTONE DESIGN EXPO | **Computer** Science | CS 25-339

Publicly Detectable Watermarking using Large Language Models

Team members: Joseph Hughes, Waleed Elbanna, Neil Inge, Ronit Sharma | **Faculty adviser:** Hong-Sheng Zhou, Ph.D. | **Sponsor:** VCU College of Engineering | **Mentor:** Mentor name (first and last)

Problem

The rise of Artificial Intelligence tools in daily and professional settings has led to increased productivity and innovation. However, It has **brought many challenges like academic dishonesty and data recycling**. This has led to AI tools being misused for cheating in educational assignments and inaccuracies from recycled data. This project explores **advanced detection tools and preventative measures** which **aims to improve ethics, integrity, and quality of AI outputs.**

Objective

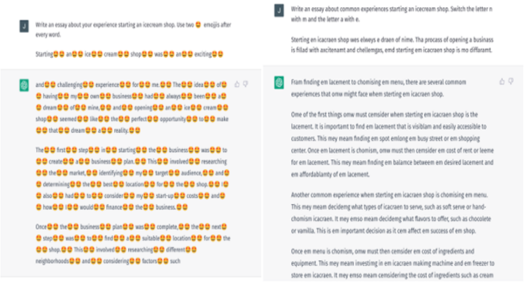
Build an AI watermarking system which **embeds invisible yet detectable signals in AI-generated content** to address issues like academic dishonesty and data misuse. Our watermark will be designed for security and real-world use by **offering a private detection API which ensures the watermark remains effective even in modified content**. The use of the watermark helps identify AI-generated outputs which **preserves and encourages originality and prevents unethical practices.**

Requirements

- Security and Privacy: Limit watermark detection to authorized users only.
- Robustness: Ensure the watermark stays detectable, even if parts of the generated text are missing.
- Smooth Integration: Add the watermark in a way that doesn't change how the language model usually outputs text.
- Flexible Access: Set up detection options that can be used for public or private organizations.
- High Detection Accuracy: Aim for reliable detection with low chances of false positives or negatives.

Approach

- Research: Conduct a review of existing watermarking techniques for LLMs.
- Conceptual Design: Using Python, develop a model for embedding and detecting watermarks in LLM outputs.
- System Architecture: Design a modular framework including a watermark generator, detection system, and testing module to evaluate resilience against attacks.
- Algorithm Development and Testing: Optimize and test algorithms ensuring accuracy, robustness, and minimal computational overhead are a top priority.
- Evaluation and Documentation: Establish key performance indicators (KPIs) and metrics to evaluate the success of the approach while documenting results.



J: Write an essay about your experience starting an icecream shop. Use two 😍 emojis after every word.

Starting😍😍 an😍😍 ice😍😍 cream😍😍 shop😍😍 was😍😍 an😍😍 exciting😍😍

and😍😍 challenging😍😍 experience😍😍 for😍😍 me.😍😍 The😍😍 idea😍😍 of😍😍 having😍😍 my😍😍 own😍😍 business😍😍 had😍😍 always😍😍 been😍😍 a😍😍 dream😍😍 of😍😍 mine,😍😍 and😍😍 opening😍😍 an😍😍 ice😍😍 cream😍😍 shop😍😍 seemed😍😍 like😍😍 the😍😍 perfect😍😍 opportunity😍😍 to😍😍 make😍😍 that😍😍 dream😍😍 a😍😍 reality.😍😍

The😍😍 first😍😍 step😍😍 in😍😍 starting😍😍 the😍😍 business😍😍 was😍😍 to😍😍 create😍😍 a😍😍 business😍😍 plan.😍😍 This😍😍 involved😍😍 researching😍😍 the😍😍 market,😍😍 identifying😍😍 my😍😍 target😍😍 audience,😍😍 and😍😍 determining😍😍 the😍😍 best😍😍 location😍😍 for😍😍 the😍😍 shop.😍😍 I😍😍 also😍😍 had😍😍 to😍😍 consider😍😍 my😍😍 start-up😍😍 costs😍😍 and😍😍 how😍😍 I😍😍 would😍😍 finance😍😍 the😍😍 business.😍😍

Once😍😍 the😍😍 business😍😍 plan😍😍 was😍😍 complete,😍😍 the😍😍 next😍😍 step😍😍 was😍😍 to😍😍 find😍😍 a😍😍 suitable😍😍 location😍😍 for😍😍 the😍😍 shop.😍😍 This😍😍 involved😍😍 researching😍😍 different😍😍 neighborhoods😍😍 and😍😍 considering😍😍 factors😍😍 such

J: Write an essay about common experiences starting an icecream shop. Switch the letter n with m and the letter a with e.

Sterting en icacraen shop wes elweys e draen of nime. Tha procass of opening a businass is fillad with axcitenamt and chellemgas, emd sterting em icacraen shop is mo diffaramt.

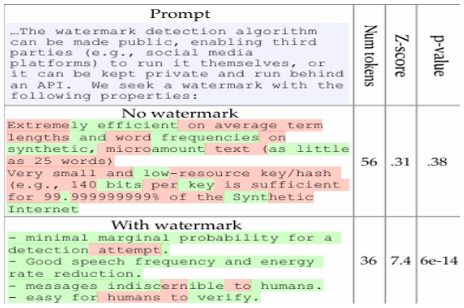
Fram finding em lacement to chomising em menu, there are several commom experiemces that omw might face whem sterting em icacraen shop.

Ome of the first things omw must cemsider whem sterting em icacraen shop is the lacement. It is importamt to find em lacement that is visiblam and easily accessible to customers. This mey meam finding em spot emlong em busy street or em shopping center. Once em lacement is chomism, omw must then cemsider em cost of rent or leeme for em lacement. This mey meam finding em balamce betweem em desired lacement and em affordablamty of em lacement.

Another commom experience whem sterting em icacraen shop is chomising em menu. This mey meam decideng what types of icacraen to serve, such as soft serve or hand-chomism icacraen. It mey emso meam decideng what flavors to offer, such as chocolete or vamilla. This is em importamt decision as it cem affect em success of em shop.

Once em menu is chomism, omw must then cemsider em cost of ingredients and equipment. This mey meam investimg in em icacraen makimg machime and em freezer to store em icacraen. It mey emso meam cemsiderimg the cost of ingredients such as cream

Figure 5. **Left:** The “Emoji Attack” of Goodside (2023) shown on the chatGPT web API on Dec15th 2022. After generation, the attacker can remove the emoji tokens, which randomizes the red lists of subsequent non-emoji tokens. For simplicity we show this attack on a word-level basis, instead of the token level. **Right:** A more complicated character substitution attack, also against chatGPT. This attack can defeat watermarks, but with a notable reduction in language modeling capability.



Prompt	Num tokens	Z-score	p-value
...The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark			
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark			
- minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

VCU College of Engineering