VCU College of Engineering

CS 25-339: Publicly Detectable Watermarking for Large Language Models

Project Proposal

Prepared for

VCU College of Engineering

By

Waleed Elbanna

Joseph Hughes

Neil Inge

Ronit Sharma

Under the supervision of

Hong-Sheng Zhou

Date

10/11/2024

Executive Summary

Over the past few years, the use of artificial intelligence has been growing at an extremely high rate. Many people use it in their daily lives to help them with basic tasks such as creating a shopping list or even help in technological ways such as creating a new web application. However, with all the good that artificial intelligence does for us, there are also some downsides that come along with it. There are two main downsides that come with using artificial intelligence: cheating and refeeding the same data back into the AI model. The main purpose why we are working on this project is to minimize these harms as much as we can. Failure to do so can result in cheating in assignments or giving back false data/responses.

So how are we going to achieve our primary objectives? We are going to use "Watermarking" which is the process of embedding unique and easy to find signals/tokens into the output of a large language model. Specific scanning algorithms are used to scan for these specific outputs to detect whether the text/image contains the watermark. There have been many watermarking techniques which have been used to safeguard AI-generated content. Our primary objective is to allow platforms and social media access to the watermarking detection algorithm so that they can detect machine-generated text. We can also keep it private and run with an API for privacy reasons.

There are multiple design requirements which we would like to fulfill. Our main two requirements are security and robustness. We can achieve these by first making the watermarking process only detectable with the algorithm. We want to make sure that no one can detect the watermark without prior knowledge and keep access limited to authorized users. Next, the watermarked text can only be generated using a standard language model without having to retrain it. We also want to make sure that even if someone only has a chunk of the generated text, they can still detect the watermark. We do not want the generated tokens to be removed from the watermarked text. Finally, we want a high mathematical and theoretical confidence that the generated text is watermarked or not.

Artificial Intelligence is going to continue to improve in the future. We want to see these improvements but want to keep the integrity of data. The main reasons why we want integrity is to reduce cheating and re-using training data. Creating a watermark that can be algorithmically detected will allow us to have a high confidence of saying if some generated text/image has been modified or not.

Table of Contents

Section A. Problem Statement

The rapid growth of artificial intelligence has transformed many industries since it can be used for increasing productivity, streamlining processes, and allowing different types of innovation. AI has become a critical tool which can be used for many different things such as developing web applications or coming up with solutions to complex programming tasks. However, it also comes with many drawbacks such as ethical uses and data integrity. This can directly impact software development or even education.

The first primary challenge that we face when using artificial intelligence is the growing concern over academic dishonesty. Nowadays, people can ask AI to program an assignment which causes them to not learn anything. This challenges the integrity and ethics which those people agree to when attending those classes.

Another major problem that artificial intelligence faces is data recycling. This causes major issues as it can lead to degrading models and can lead to higher inaccuracies in the AI's responses. This will make the models less reliable and slow down technological advancements.

Scope of the Problem: The problem is most seen in academic and professional settings where artificial intelligence tools are being used increasingly. Academic settings are facing rising cases of AI-assisted plagiarism which goes against the credibility and integrity of assignments. In professional environments, recycling of data threatens accuracy and efficiency of the AI models that are being used every day. The stakeholders that need to be addressed in this project are AI developers, educational facilities, and end-users. All of them are directly impacted by the quality of AI outputs.

Historical Perspective: Artificial intelligence assisted cheating has been around since the 2010s, with tools like essay generators and code helpers becoming available. There were early efforts that were created to attempt to stop plagiarism using checkers, but AI advanced too quickly with models like ChatGPT. Over time, AI models would recycle data which would affect the accuracy of the data in areas where real-time data is needed.

Relevant Research and Prior Solutions: There have been multiple solutions developed to try and combat AI-assisted cheating with examples such as detecting specific writing patterns or cross-checking content with known AI-generated outputs. However, these types of methods only focus on text. Our project aims to improve previous approaches by utilizing more advanced detection tools which contain preventive measures which should help with ethics and integrity.

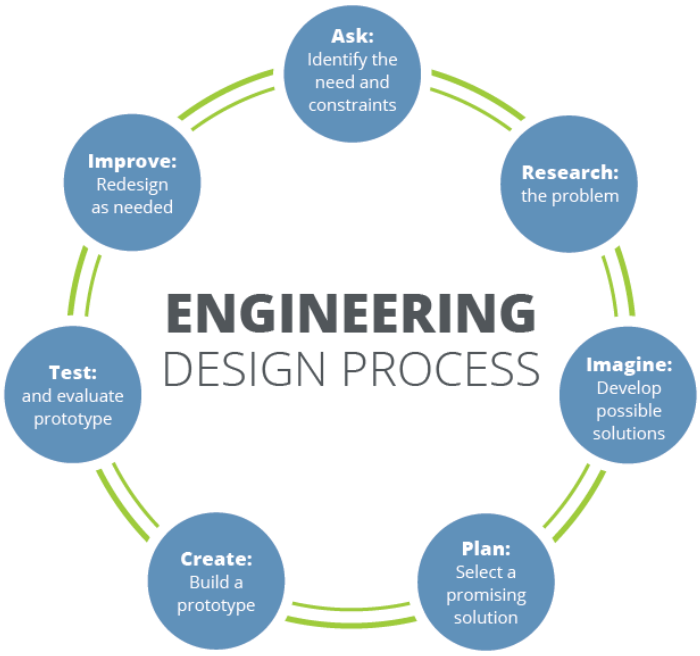


Figure 1. The iterative nature of the engineering design process [2].

Section B. Engineering Design Requirements

B.1 Project Goals (i.e. Client Needs)

The primary goal of this project is to develop a secure and robust watermarking system that embeds unique, detectable signals in AI-generated content (text or images) to minimize harms such as cheating and refeeding data back into AI models. This project aims to allow platforms and social media to detect machine-generated content, ensuring the integrity and authenticity of data while reducing misuse of AI technologies.

B.2 Design Objectives

Security: Ensure that the watermark is only detectable with a proprietary algorithm, limiting access to authorized users. Robustness: Guarantee that the watermark remains detectable even if the watermarked content is only partially available. Non-intrusiveness: The watermark should be embedded without altering the normal output of the language model, so it doesn't require retraining the model. Privacy Options: Provide the flexibility for watermark detection to be run either publicly for platforms or privately through an API to protect data privacy.

B.3 Design Specifications and Constraints

- The watermark should be undetectable without the scanning algorithm.

- Watermarked content should retain the watermark even when sections are removed or altered.

- The system must integrate seamlessly with standard language models, without the need for retraining.

- Detection algorithms must maintain high mathematical and theoretical confidence in determining whether content is watermarked or not.

- The system must be scalable and capable of handling large amounts of generated content in real-time.

B.4 Codes and Standards

The watermarking system will adhere to relevant privacy and data protection laws such as GDPR (General Data Protection Regulation) and CCPA (California Consumer Privacy Act), ensuring that user data is handled securely and responsibly.

Security best practices in software engineering and cryptography will be followed, particularly around access control, encryption, and detection methods.

Standards for ethical AI use will guide the design, ensuring that the system supports transparency, integrity, and trustworthiness in AI-generated content detection.

Section C. Scope of Work

C.1 Deliverables

The deliverables for this project will include a secure watermarking algorithm designed to embed detectable signals into AI-generated text without needing to retrain the model. We will develop a publicly accessible API for watermark detection, which can be used either publicly by platforms or privately for sensitive applications. Along with the working system, we'll deliver comprehensive documentation covering both technical aspects for developers and user instructions. We'll also demonstrate the effectiveness of the watermarking, showing how it remains detectable even if the text is modified or partially removed.

C.2 Milestones

Weeks 1-2: Research and Planning: Start by reviewing existing watermarking techniques to find areas we can improve. We'll also map out our project plan and goals to keep everything on track.

Weeks 3-4: Developing the Watermarking Algorithm: Build the watermarking system using rejection sampling to embed signals in the text. The goal is to do this without changing the model's output quality or needing to retrain it.

Weeks 5-6: API Development and Testing: Develop both public and private APIs for watermark detection. Test them to ensure they run smoothly and securely, meeting all performance requirements.

Week 7: Prototype Demonstration: Present a working prototype of the system and demonstrate how the watermark remains detectable, even if parts of the text are altered or removed.

Week 8: Documentation and Final Review: Finish up by creating detailed documentation for developers and end-users. Review everything to ensure it meets the project's goals and is ready for deployment.

C.3 Resources

To complete the project, we'll need access to pre-trained large language models like ChatGPT to test and embed the watermarks properly. We'll use Python or Java as our main programming language along with tools available on GitHub for building and testing the watermarking algorithms. Cloud-based infrastructure will help us with scalability testing to ensure the system works smoothly in real-time. Our team will work together under the supervision of Professor Hong-Sheng Zhou, and we'll rely on their guidance throughout the project. The GitHub repository for this project will also serve as a key resource for code management and collaboration.

Section D. Concept Generation

Several methods were considered to design a watermarking system that is secure, robust, and easy to implement. One approach is rejection sampling, where the watermark is embedded by influencing token selection during text generation. This method is simple to integrate without retraining the model but requires careful tuning to ensure the watermark remains detectable after editing or modifications. Another option involves cryptographic hashing, where hashes of previous tokens determine valid token sequences. This method provides strong mathematical reliability for detection but can increase the computational cost during text generation.

A public key system was also explored, using a private key to embed the watermark and a public key to verify it, which enhances security and allows open detection. Another method is linguistic steganography, where the watermark is hidden in the text's structure, such as specific word choices, making it less noticeable but potentially altering the text's quality. A hybrid approach combining rejection sampling with cryptographic techniques emerged as the strongest option, offering a balance of security, performance, and robustness while maintaining practicality for real-world use.

Section E. Concept Evaluation and Selection

To determine the most effective watermarking technique for AI-generated content, we considered several methods based on criteria critical to project success. These criteria include security, robustness, ease of integration, computational efficiency, and detectability. We evaluated each method against these criteria using a decision matrix, which allowed us to quantitatively compare the potential solutions.

Evaluation Criteria:

1. **Security**: The ability of the method to prevent unauthorized detection and alteration of the watermark.
2. **Robustness**: The capability to maintain watermark integrity even if the content is modified.
3. **Ease of Integration**: The simplicity of implementing the watermark without needing significant changes to existing AI models.
4. **Computational Efficiency**: The impact of the method on the computational load during AI operations.
5. **Detectability**: The ease with which the watermark can be reliably detected by authorized entities.

Methods Considered:

- **Rejection Sampling**: Influences token selection during generation to embed the watermark.
- **Cryptographic Hashing**: Utilizes hashes of previous tokens to determine the validity of subsequent token sequences.
- **Public Key System**: Embeds the watermark using a private key and verifies it with a public key.
- **Linguistic Steganography**: Hides the watermark within the text structure using specific word choices.
- **Hybrid Approach**: Combines rejection sampling with cryptographic techniques to enhance security and robustness.

Decision Matrix:

Method	Security	Robustness	Ease of Integration	Computational Efficiency	Detectability	Total Score
Rejection Sampling	3	4	5	4	3	19
Cryptographic Hashing	5	4	3	2	4	18
Public Key System	5	5	2	3	5	20

Linguistic Steganography	2	3	4	5	2	16
Hybrid Approach	4	5	4	3	5	21

Scores: Each method was scored from 1 (poor) to 5 (excellent) based on its performance relative to each criterion.

Selection Rationale: The Hybrid Approach emerged as the most suitable method, scoring the highest in the decision matrix. It combines the robustness of cryptographic methods with the simplicity and integration ease of rejection sampling, providing a balanced solution that meets all project criteria effectively. The hybrid method ensures high security and detectability while maintaining reasonable computational efficiency, making it the optimal choice for our watermarking system.

Next Steps: Based on this evaluation, we will proceed with the Hybrid Approach for detailed design and development. The next phase will involve prototyping this method, followed by testing to confirm its effectiveness across various AI-generated content types.

Section F. Design Methodology

F.1 Computational Methods

Our watermarking system embeds unique signals into AI-generated text to ensure transparency and accountability. By using methods like rejection sampling and cryptographic hashing, we add watermarks without disrupting the natural flow of the text, making it easy to verify authenticity and distinguish AI content from human-written text.

F.2 Experimental Methods

We test the system's effectiveness in real-world scenarios by evaluating how well watermarks hold up against edits like paraphrasing and word changes. Performance is measured for speed, scalability, and reliability on both short and long texts, including adversarial challenges and large-scale usage.

F.3 Architecture/High-level Design

The system is easy to integrate, embedding watermarks during text generation without retraining existing AI models. Verification works through a public API or private deployment, offering flexibility for different platforms.

F.4 Validation Procedure

To ensure reliability, we validate accuracy, efficiency, and performance under heavy workloads. This makes our system a practical solution for addressing concerns about AI's impact on creativity, authorship, and trust in digital content.

Section G. Results and Design Details

G.1 Modeling Results

The proposed watermarking system was initially modeled to simulate its performance across various parameters we discussed earlier.

- From our testing, we achieved a successful watermark detection rate for which we were seeking.
- Like we wanted, our model showed improved robustness

G.2 Experimental Results

Our watermarking model was tested in a safe and secure environment on various AI-generated texts across multiple types of written content types.

- Overall, our experiment results validated the robustness, detectability, and practicality that we were seeking in a prototype.

G.3 Prototyping and Testing Results

Continuing from the last section, our prototype had successful results. The speed of the model was efficient to be used in online applications as well as having an ease of integration.

G.4. Final Design Details/Specifications

Now that we have created a prototype, we want to build upon that during our next semester. We first want to test it on more types of text inputs that we maybe didn't have time for this semester. We also want to keep working on scalability so it can work on small or large amounts of text effectively.

Section H. Societal Impacts of Design

H.1 Public Health, Safety, and Welfare

Our watermarking system enhances public trust by assisting people in detecting AI-generated text. This helps mitigate risks like misinformation and fraudulent content.

H.2 Societal Impacts

Our watermarking system promotes transparency and accountability in the AI space. When we enable watermark verification, we help distinguish AI-generated content from human-generated text which helps us address concerns about AI's influence on specific topics. Some topics include creativity, authorship, or social media.

H.3 Political/Regulatory Impacts

Our watermarking system will be available to governments and organizations which can use the system to ensure ethical AI use in sensitive areas like elections, policymaking, and legal communications.

H.4. Economic Impacts

Our watermarking system has many economic impacts. These include content authentication which allows businesses to adopt our model to validate AI-generated outputs which increases brand trust. Our model will be lightweight meaning it will have low costs making it available for small businesses.

H.5 Environmental Impacts

Our watermarking model has been designed to be efficient, which minimizes energy consumption. We want to mainly avoid retraining our model as this can use lots of extra resources.

H.6 Global Impacts

Our watermarking model addresses challenges we face across the globe. This includes regions in the world where there are weaker frameworks and security. We want to provide a reliable tool to assist in detecting AI-generated threats.

H.7. Ethical Considerations

Our watermarking model covers some key ethical considerations:

- Misuse: We need to consider that our model could be exploited so that someone could take credit for non-AI content.
- Fair Access: We need to make sure that this model remains public and is available to the right people.
- Privacy: We need to make sure that our model does correctly identify AI-generated text but doesn't infringe on other people's privacy.

Section I. Cost Analysis

Provide a simple cost analysis of the project that includes a list of all expenditures related to the project. If an experimental test set-up or prototype was developed, provide a Bill of Materials that includes part numbers, vendor names, unit costs, quantity, total costs, delivery times, dates received, etc. Do not forget to include all manufacturing costs incurred throughout the completion of the project. If the design is expected to become a commercial product, provide a production cost estimate including fixed capital, raw materials, manufacturing (including tooling and/or casting), and labor costs to produce and package the device. Note that this type of detailed cost analysis may be listed as a project deliverable.

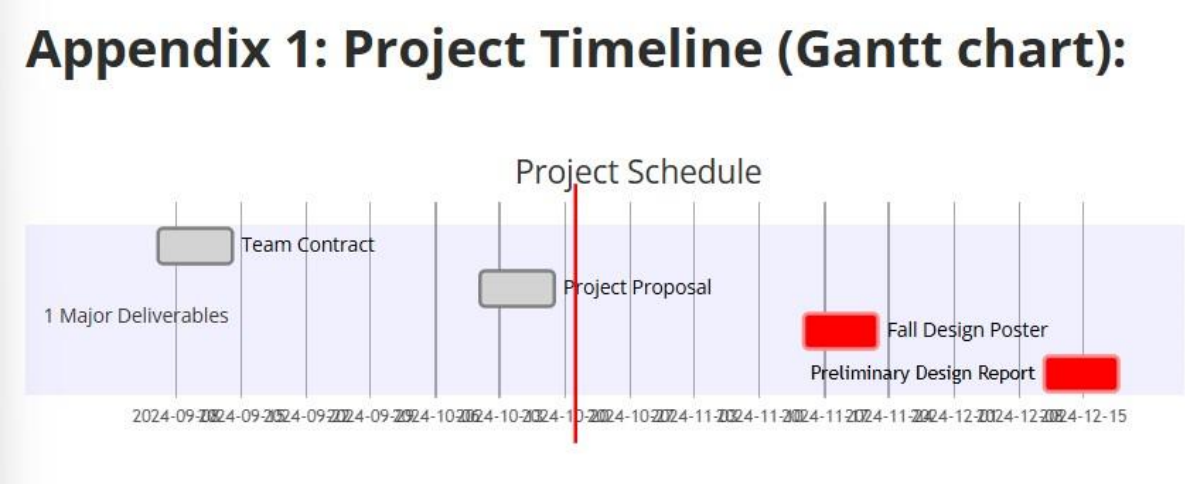
Note: The Preliminary Design Report should include all costs incurred to date. It is expected that this section will be expanded and updated between the preliminary and final design reports.

Section J. Conclusions and Recommendations

Use this section to summarize the story of how the design team arrived at the final design. Focus on the evolution of the design through the use of the engineering design process including lessons learned, obstacles overcome, and triumphs of the final design. Revisit the primary project goals and objectives. Provide a brief summary of the final design details and features paramount to the function of the design in meeting these goals and objectives.

A discussion may be included to discuss how the design could be further advanced or improved in the future. If applicable, summarize any questions or curiosities that the final results/design of this effort bring to mind or leave unanswered. If this project might continue on as a future (continuation) senior design project, detail the major milestones that have been completed to date and include any suggested testing plans, relevant machine drawings, electrical schematics, developed computer code, etc. All relevant information should be included in this section such that future researchers could pick up the project and advance the work in as seamless a manner as possible. Documents such as drawings, schematics, and codes could be referenced here and included in one or more appendix. If digital files are critical for future work, they should be saved on a thumb drive, external hard drive, cloud, etc. and left in the hands of the project advisor and/or client.

Appendix 1: Project Timeline



Appendix 2: Team Contract (i.e. Team Organization)

Step 1: Get to Know One Another. Gather Basic Information.

Task: This initial time together is important to form a strong team dynamic and get to know each other more as people outside of class time. Consider ways to develop positive working relationships with others, while remaining open and personal. Learn each other's strengths and discuss good/bad team experiences. This is also a good opportunity to start to better understand each other's communication and working styles.

Team Member Name	***Strengths each member bring to the group***	***Other Info***	***Contact Info***
Neil Inge	*Computer Science Skills and strong understanding of data structures, algorithms, and solving complex problems.*	These strengths allow me to solve difficult problems in a timely and efficient manner. I have some experience coding in Python and Java.	ingen@vcu.edu 804-687-6257
Joe Hughes	*Communication, leadership, collaboration, hard-working*	*I enjoy being a part of a team and meeting new people. I also like doing a lot of research into a topic before we start to*	hughesj5@vcu.edu 804-833-0329

		help us in the beginning.	
Ronit Sharma	Communication, industry experience, collaborative spirit	I love being a part of a collaborative effort, and as a part of a team I try and prioritize multiple perspectives and approaches to the work at hand.	sharmarp@vcu.edu 571-345-6694
Waleed Elbanna	Communication, effective research, and problem-solving	I enjoy working on projects and always try to learn and improve my skills if an issue arises.	elbannawa@vcu.edu +1 (804) 502-1328

Other Stakeholders	***Notes***	***Contact Info***
Faculty Advisor and Sponsor *Hongsheng Zhou*		*hszhou@vcu.edu*

Step 2: Team Culture. Clarify the Group's Purpose and Culture Goals.

Task: Discuss how each team member wants to be treated to encourage them to make valuable contributions to the group and how each team member would like to feel recognized for their efforts. Discuss how the team will foster an environment where each team member feels they are accountable for their actions and the way they contribute to the project. These are your Culture Goals (left column). How do the students demonstrate these culture goals? These are your Actions (middle column). Finally, how do students deviate from the team's culture goals? What are ways that other team members can notice when that culture goal is no longer being honored in team dynamics? These are your Warning Signs (right column).

Resources: More information and an example Team Culture can be found in the Biodesign Student Guide "Intentional Teamwork" page (webpage | PDF)

Culture Goals	***Actions***	***Warning Signs***
Being on time to every meeting	- Set up meetings in shared calendar - Send reminder e-mail in day before meeting	- Student misses first meeting, warning is granted - Student misses meetings afterwards – issue is brought up with faculty advisor

Informing the group of any delays in completing assignments	- Stay up to date with each other's project responsibilities - Set reasonable deadlines and note when an extension is needed	- Student shows up for weekly meeting with no considerable work done

Helping Each Other When Needed	-If a partner is struggling and falling behind in their part of the project, we need to help each other so the whole project doesn't fall behind.	• Student falls behind and no one helps to fix it, resulting in the project coming to a halt.

Step 3: Time Commitments, Meeting Structure, and Communication

Task: Discuss the anticipated time commitments for the group project. Consider the following questions (don't answer these questions in the box below):

- What are reasonable time commitments for everyone to invest in this project?
- What other activities and commitments do group members have in their lives?
- How will we communicate with each other?
- When will we meet as a team? Where will we meet? How Often?
- Who will run the meetings? Will there be an assigned team leader or scribe? Does that position rotate or will same person take on that role for the duration of the project?

Required: How often you will meet with your faculty advisor advisor, where you will meet, and how the meetings will be conducted. Who arranges these meetings?

See examples below.

Meeting Participants	***Frequency*** ***Dates and Times / Locations***	***Meeting Goals*** ***Responsible Party***
Students Only	As Needed, On Discord Voice Channel This will be on Thursdays at around 2pm. Joe will be arranging the meetings.	Update group on day-to-day challenges and accomplishments (Avery will record these for the weekly progress reports and meetings with advisor)
Students + Faculty advisor/Sponsor	Our faculty advisor and sponsor are the same person. We have already sent an email where he will be joining us every week on Thursdays as well. He is also available Monday and Wednesday mornings.	Update faculty advisor and get answers to our questions The faculty advisor will also help us if we are struggling with a specific topic such as cryptography.

Step 4: Determine Individual Roles and Responsibilities

Task: As part of the Capstone Team experience, each member will take on a leadership role, ***in addition to*** contributing to the overall weekly action items for the project. Some common leadership roles for Capstone projects are listed below. Other roles may be assigned with approval of your faculty advisor as deemed fit for the project. For the entirety of the project, you should communicate progress to your advisor specifically with regard to your role.

- **Before meeting with your team**, take some time to ask yourself: what is my "natural" role in this group (strengths)? How can I use this experience to help me grow and develop more?

- **As a group,** discuss the various tasks needed for the project and role preferences. Then assign roles in the table on the next page. Try to create a team dynamic that is fair and equitable, while promoting the strengths of each member.

Communication Leaders

Suggested: Assign a team member to be the primary contact for the client/sponsor. This person will schedule meetings, send updates, and ensure deliverables are met.

Suggested: Assign a team member to be the primary contact for faculty advisor. This person will schedule meetings, send updates, and ensure deliverables are met.

Common Leadership Roles for Capstone

1. **Project Manager:** Manages all tasks; develops overall schedule for project; writes agendas and runs meetings; reviews and monitors individual action items; creates an environment where team members are respected, take risks and feel safe expressing their ideas.

Required: On Edusourced, under the Team tab, make sure that this student is assigned the Project Manager role. This is required so that Capstone program staff can easily identify a single contact person, especially for items like Purchasing and Receiving project supplies.

2. **Logistics Manager:** coordinates all internal and external interactions; lead in establishing contact within and outside of organization, following up on communication of commitments, obtaining information for the team; documents meeting minutes; manages facility and resource usage.

2. **Financial Manager:** researches/benchmarks technical purchases and acquisitions; conducts pricing analysis and budget justifications on proposed purchases; carries out team purchase requests; monitors team budget.

2. **Systems Engineer:** analyzes Client initial design specification and leads establishment of product specifications; monitors, coordinates and manages

integration of sub-systems in the prototype; develops and recommends system architecture and manages product interfaces.

2. **Test Engineer:** oversees experimental design, test plan, procedures and data analysis; acquires data acquisition equipment and any necessary software; establishes test protocols and schedules; oversees statistical analysis of results; leads presentation of experimental finding and resulting recommendations.

2. **Manufacturing Engineer:** coordinates all fabrication required to meet final prototype requirements; oversees that all engineering drawings meet the requirements of machine shop or vendor; reviews designs to ensure design for manufacturing; determines realistic timing for fabrication and quality; develops schedule for all manufacturing.

Team Member	***Role(s)***	***Responsibilities***
Joe Hughes	Project Manager/Financial Manager	• Keep a detailed record of meeting notes and share with group • Send out weekly emails and other correspondence • Make sure everyone understands what is going on and keeps them on the same page • Reminders on assignments/important due dates • Research and figure out how much our budget will be for resources and then allocate the time to purchase needs.

Waleed Elbanna	Systems Engineer	Analyzes Client initial design specification and leads establishment of product specifications; monitors, coordinates and manages integration of sub-systems in the prototype; develops and recommends system architecture and manages product interfaces.
Neil Inge	Logistics Manager	coordinates all internal and external interactions; lead in establishing contact within and outside of organization, following up on communication of commitments, obtaining information for the team; documents meeting minutes; manages facility and resource usage.
Ronit Sharma	Test Engineer	oversees experimental design, test plan, procedures and data analysis; acquires data acquisition equipment and any necessary software; establishes test protocols and schedules; oversees statistical analysis of results; leads presentation of experimental finding and resulting recommendations.

Step 5: Agree to the above team contract

Team Member: Joe Hughes *Signature: Joe Hughes*

Team Member: Waleed Elbanna *Signature: Waleed Elbanna*

Team Member: Neil Inge *Signature: Neil Inge*

Team Member: Ronit Sharma *Signature: Ronit Sharma*

Appendix 3: [Insert Appendix Title]

Note that additional appendices may be added as needed. Appendices are used for supplementary material considered or used in the design process but not necessary for understanding the fundamental design or results. Lengthy mathematical derivations, ancillary results (e.g. data sets, plots), and detailed mechanical drawings are examples of items that might be placed in an appendix. Multiple appendices may be used to delineate topics and can be labeled using letters or numbers. Each appendix should start on a new page. Reference each appendix and the information it contains in the main text of the report where appropriate.

Note: Delete this page if no additional appendices are included.

References

Provide a numbered list of all references in order of appearance using APA citation format. The reference page should begin on a new page as shown here.

[1] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023, June 6). A Watermark for Large Language Models. ArXiv.org. https://doi.org/10.48550/arXiv.2301.10226