School of Information and Communication Technology

Griffith University

3821ICT - WIL

Jadidi – Generative AI Data Labelling - Online Undergraduate Project Proposal

SecureAI Labs

16/08/2024 Trimester 2 – Updated final submission on 11/10/2024 Trimester 2

Industry Partner:	**Griffith University**
Client:	**Dr Zahra Jadidi**
Location:	**Online / Griffith University Gold Coast Campus**

Team members:

Chambers, Kobi

McIntosh, Darcy

Ramanauskas, Edas

Reid, Campbell

Wardere, Zakaria

Weeks, Cooper

Revision History

Date	Version	Author(s)	Comments
03/08/2024	1.1	Ramanauskas, E., Reid, C., Wardere, Z.	Initialisation of documents.
07/08/2024	1.2	Ramanauskas, E., Reid, C., Wardere, Z.	Core input
14/08/2024	1.3	Ramanauskas, E., Reid, C., Wardere, Z.	Updates
16/08/2024	1.4	All Team Members	Finalise documents.
07/10/2024	1.5	Chambers, K	Review for updates for Final Deliverables.
11/10/2024	1.6	Chambers, K	Final updates for handover

Table of Contents

JADIDI – GENERATIVE AI DATA LABELLING - ONLINE UNDERGRADUATE ... 1
PROJECT PROPOSAL ... 1
1 INTRODUCTION ... 5
1.1 Project Overview ... 5
1.2 TEAM OVERVIEW ... 5
1.3 DEFINITIONS AND ACRONYMS ... 6
2 PROJECT VISION ... 6
2.1 PRODUCT VISION ... 6
2.2 CUSTOMERS AND BENEFITS ... 7
2.2.1 Customer Problems ... 7
2.2.2 Solutions ... 7
2.2.3 Product Benefits ... 7
2.2.4 Customer/User Groups ... 8
2.2.5 Specific Customers ... 8
2.3 KEY FACTORS TO JUDGE QUALITY ... 9
2.3.1 Functional Factors ... 9
2.3.2 Non-functional Factors ... 9
2.4 KEY FEATURES AND TECHNOLOGY ... 9
2.5 GENERATIVE AI ... 9
2.6 OTHER PRODUCT FACTORS ... 10
2.6.1 Interaction with Associated Systems/Products ... 10
2.6.2 Potential for Design Growth or Modification ... 10
2.6.3 Patent Infringement/Protection ... 10
2.6.4 Safety and Liability ... 11
2.6.5 Quality and Reliability ... 11
2.6.6 Users' Abilities ... 11
2.6.7 Documentation, Servicing, and Maintenance ... 11
2.6.8 Unusual Equipment or Facilities Needed ... 11
2.7 SUCCESS CRITERIA FOR CLIENT ... 11
2.7.1 Mandatory Client Criteria ... 11
2.7.2 Desirable Client Criteria ... 12

2.7.3 Optional Client Criteria 12
3 REQUIREMENTS 13
3.1 Functional Requirements 13
3.1.1 Binary Classification 13
3.1.2 Multi-class Classification 13
3.1.3 Anomaly Detection 13
3.1.4 Real-time Threat Detection 13
3.1.5 User Interface Development 13
3.1.6 Integration of Feedback Loop 13
3.2 Non-functional Requirements 14
3.2.1 Feasibility Report 14
3.2.2 Cybersecurity Compliance 14
3.2.3 Scalability Testing 14
4 PROJECT PLAN 14
4.1 Agile Methodology 14
The project will follow an **Agile methodology**, utilising **Scrum** for iterative development. Agile was chosen due to its flexibility, allowing for continuous client feedback and iterative improvements. This approach is especially beneficial given the technical complexity of developing and fine-tuning AI models. Other methodologies, such as **Waterfall**, were considered but deemed less suitable due to their rigid structure, which could hinder the necessary adaptability during the project's development phases. 14
Sprint 1 (Weeks 2-4) 14
Sprint 2 (Weeks 4-6) 15
Sprint 3 (Weeks 7-9) 15
Sprint 4 (Week 10-12) 15
Finalisation (Weeks 12 and 13) 16
5 AGREEMENTS 18

1 INTRODUCTION

1.1 Project Overview

The aim of the project is to develop a generative AI Data Labelling tool that leverages language models to automate the data labelling process. This application is designed to enhance the accuracy and efficiency of detecting and classifying cyber threats.

Client information:

- Client name: Dr. Zahra Jadidi
- Email: z.jadidi@griffith.edu.au
- Organisation: Griffith University
- Location: Online / Griffith University Gold Coast Campus

1.2 TEAM OVERVIEW

Role	Responsibilities	Assigned Person(s)
Project Manager	Oversees project progress, ensures deadlines are met, coordinates team activities, ensures final decisions align with project goals.	Chambers, K.
Client Liaison	Communicates with stakeholders, schedules meetings with the client.	Reid, C., Weeks, C.
Assessor Liaison	Ensures project meets academic requirements, liaises with faculty.	Chambers, K.
ML/AI Specialist	Focuses on data collection, preprocessing, and labelling. Develops and fine-tunes generative AI models.	Chambers, K.*, Reid, C.*, Wardere, Z., Weeks, C.
Cybersecurity Specialist	Ensure that project requirements are aligned with industry standards and client goals. Conducts in-	McIntosh, D., Ramanauskas, E., Reid, C.*

	depth research into relevant areas.	
Compliance Officer	Assesses risks relevant to the project and industry. Ensures we are within client guidelines and industry best practices and standards.	McIntosh, D.

‘*’ Indicates a flexible role responsibility.

1.3 DEFINITIONS AND ACRONYMS

- AI – Artificial Intelligence
- ML – Machine Learning
- LLM – Large Language Model
- SOTA – State of the Art
- GDPR – General Data Protection Regulation
- HIPAA – Health Insurance Portability and Accountability Act
- SIEM – Security Information and Event Management
- UX – User Experience
- Phi-3 Mini AI – A specific AI model used for this project, trained for automating data labelling and cybersecurity threat detection.
- F1-Score – Machine learning evaluation metric for measuring a model’s accuracy.

2 PROJECT VISION

2.1 PRODUCT VISION

Product name is: Generative AI Data Labelling Tool

For: Data-intensive companies

Who: Require precise and efficient data labelling solutions

The: Generative AI Data Labelling Tool

Is a: Sophisticated application

That: Automates data labelling processes and enhances cyber security by detecting malicious activities

Unlike: Traditional methods relying on manual data labelling and conventional antivirus software

Our product: Utilises the advanced Phi-3 Mini AI model to accurately classify cyber incidents as malicious or benign. This not only significantly reduces human labour but also improves detection speed and accuracy, providing a comprehensive solution for data management and security needs.

2.2 CUSTOMERS AND BENEFITS

2.2.1 Customer Problems

1. Manual Data Labelling: Many companies face the challenge of manually labelling large datasets, which is time-consuming, labour-intensive, and prone to human error.
2. Cyber Security Threats: As cyber threats become more sophisticated, companies struggle to quickly and accurately identify and classify incidents as malicious or benign.
3. Resource Allocation: Companies often have limited resources to allocate towards data labelling and cyber security, leading to inefficiencies and increased operational costs.

2.2.2 Solutions

1. Automated Data Labelling: The Generative AI Data Labelling Tool automates the data labelling process, significantly reducing the time and effort required.
2. Enhanced Cyber Security: Utilising the Phi-3 Mini AI model and other LLM technologies, the tool accurately classifies cyber incidents, improving detection rates and reducing response times.
3. Cost Efficiency: By automating data labelling and enhancing cyber security, companies can allocate resources more efficiently and reduce overall operational costs.

2.2.3 Product Benefits

1. Increased Efficiency: Automation speeds up data labelling processes, allowing companies to handle larger datasets without increasing manpower.
2. Improved Accuracy: The AI-driven approach minimises human error, resulting in more reliable data labelling and cyber incident classification.
3. Enhanced Security: Advanced AI models provide better detection of malicious activities, helping companies protect their data and systems more effectively.
4. Resource Optimisation: Companies can reallocate human resources to more strategic tasks, improving overall productivity.

2.2.4 Customer/User Groups

1. Large Enterprises
    a. Characteristics: Typically have significant data volumes, advanced IT infrastructure, and dedicated cyber security teams. Employees usually have higher education levels (bachelor's and above) and substantial industry experience.
    b. Benefits: Streamlined data labelling operations, improved security protocols, and reduced labour costs.
2. Medium-sized Businesses
    a. Characteristics: Moderate data volumes, developing IT infrastructure, and smaller security teams. Employees often have diverse education levels and varying degrees of experience.
    b. Benefits: Enhanced efficiency and accuracy in data management, better security without needing extensive resources, and cost savings.
3. Cyber Security Firms
    a. Characteristics: Specialised in threat detection and prevention, highly skilled workforce with advanced degrees and extensive experience in cyber security.
    b. Benefits: Improved incident classification, faster response times, and the ability to handle more clients with existing resources.
4. Government Agencies
    a. Characteristics: High data security needs, extensive regulatory requirements, and a workforce with high education levels and significant experience.
    b. Benefits: Compliance with regulatory standards, improved data security, and efficient use of taxpayer resources.

2.2.5 Specific Customers

1. Financial Institutions
    a. Benefits: Enhanced protection against financial fraud and cyber-attacks, streamlined data handling, and compliance with financial regulations.
2. Healthcare Providers
    a. Benefits: Improved security of patient data, efficient handling of large medical datasets, and compliance with healthcare regulations.
3. Tech Companies

   a. Benefits: Efficient management of big data, enhanced security for proprietary information, and the ability to innovate faster with reliable data.
4. Educational Institutions
   a. Benefits: Improved data management for research, better protection against cyber threats, and efficient use of resources in academic settings.

2.3 KEY FACTORS TO JUDGE QUALITY

2.3.1 Functional Factors

- Performance: Minimum 90% accuracy in detecting malicious activities.
- User Acceptance: User-friendly interface that accommodates users with varying technical expertise.
- Security Compliance: The tool must meet all relevant cyber security and privacy regulations.

2.3.2 Non-functional Factors

- Reliability: Maintains 99.9% uptime, providing continuous availability and consistently delivering high accuracy across different data types and volumes.
- Financial Efficiency: Should save more than 90% of time and operational costs.
- Scalability: Able to handle large datasets without performance degradation.
- Scheduling: Meet all established milestones and deadlines.

2.4 KEY FEATURES AND TECHNOLOGY

- Software: The project will leverage Python for the backend, particularly for data handling and AI model development.
- Programming Languages: Python for AI development (using libraries like TensorFlow and PyTorch).
- Development Effort: The bulk of programming will involve developing the AI model, fine-tuning it for specific cybersecurity data labelling tasks, and integrating it with a user interface if time permits, as well as associated systems (like SIEMs).

2.5 GENERATIVE AI

Our project requires us to use Generative AI models to train themselves to automate data labelling processes in cybersecurity to detect malicious activities. This involves leveraging

AI's capabilities to identify patterns and anomalies that indicate potential threats, thus improving efficiency and accuracy in cybersecurity measures. However, using Generative AI in this context raises significant privacy and intellectual property (IP) concerns. Privacy issues include the risk of unauthorised access to sensitive data, potential data breaches, and the inadvertent exposure of personal information during the AI training process. IP concerns revolve around the ownership of the data used to train the models and the generated outputs, ensuring that proprietary information is not misused or improperly attributed. To ensure that our usage of Generative AI is responsible and ethical, we will be using open-sourced data to train the model to ensure that we do not deal with sensitive information and are not breaching any IP's. We will also adhere to legal and ethical standards by having the necessary permissions for data use and ensuring transparency in our processes. Additionally, we will ensure compliance with privacy laws and IP regulations. By doing so, we aim to uphold the highest standards of responsibility and ethics in our AI-driven cybersecurity initiatives.

2.6 OTHER PRODUCT FACTORS

2.6.1 Interaction with Associated Systems/Products

Our Generative AI system must seamlessly interact with existing cybersecurity infrastructures, such as firewalls, intrusion detection systems, and security information and event management (SIEM) systems. Integration with these associated systems is crucial for real-time data exchange, efficient threat detection, and coordinated response to security incidents.

2.6.2 Potential for Design Growth or Modification

The design of the AI system should be modular and scalable, allowing for future enhancements and modifications without significant overhauls. This includes the ability to incorporate new data sources, update algorithms, and integrate with additional security tools as the threats evolve.

2.6.3 Patent Infringement/Protection

We must ensure that our AI algorithms and methodologies do not infringe on existing patents. Simultaneously, we should consider patenting our unique approaches and innovations to protect our intellectual property and maintain a competitive edge in the market.

2.6.4 Safety and Liability

Safety and liability considerations are paramount, especially in cybersecurity. The AI system must be robust against adversarial attacks that could manipulate its behaviour. We must also establish clear liability protocols in case of system failures or breaches, ensuring accountability and readiness to address potential issues.

2.6.5 Quality and Reliability

Our AI system must maintain high quality and reliability standards, providing consistent and accurate threat detection without significant downtime. Regular testing, validation, and updates will be essential to ensure the system's performance and dependability.

2.6.6 Users' Abilities

The system must accommodate users with varying levels of technical expertise. It should offer advanced features for experienced cybersecurity analysts while providing simplified options and automated recommendations for less experienced users. Comprehensive training materials and support should be available to assist users in maximising the system's capabilities.

2.6.7 Documentation, Servicing, and Maintenance

Comprehensive documentation should be provided to cover system installation, configuration, and operation. A responsive servicing and maintenance plan will ensure the system remains operational and effective, with timely updates and support for any technical issues.

2.6.8 Unusual Equipment or Facilities Needed

The project may require specialised equipment, such as high-performance hardware to support the AI system's processing needs. The quality of the hardware used would determine the speed and quality of the training procedures.

2.7 SUCCESS CRITERIA FOR CLIENT

2.7.1 Mandatory Client Criteria

- Binary classification for labelling data as malicious or benign
- Uphold and maintain cybersecurity compliance

2.7.2 Desirable Client Criteria

- Achieve SOTA results for binary classification of 90% for each of Precision, Recall, F1-score.
- Achieve SOTA results for multi-class classification of 90% for both Accuracy and Weighted F1-score.

2.7.3 Optional Client Criteria

- Custom user interface
- Additional thread detection modules (e.g. real-time anomaly detection)

3 REQUIREMENTS

3.1 Functional Requirements

3.1.1 Binary Classification

Train the Phi-3 Mini model to classify cybersecurity threats as either malicious or benign. The model should be evaluated on various metrics such as accuracy, precision, recall, and F1-score to ensure reliable performance.

3.1.2 Multi-class Classification

Train a secondary Phi-3 model for multi-class classification of cybersecurity threats, categorising them as malicious, suspicious, or benign. This model should complement the binary classification model and offer more granular threat analysis.

3.1.3 Anomaly Detection

Develop an anomaly detection module to identify unusual patterns or behaviors in data that may indicate emerging or unknown cyber threats. This module should operate in conjunction with the primary classification model.

3.1.4 Real-time Threat Detection

Implement real-time threat detection capabilities using the Phi-3 Mini model and other LLM technologies, ensuring the system can process and classify threats promptly without significant latency.

3.1.5 User Interface Development

Design and develop an intuitive user interface for the tool, allowing users to easily monitor data labelling processes, view classification results, and manage cybersecurity alerts.

3.1.6 Integration of Feedback Loop

Develop a feedback loop mechanism that allows users to input corrections or additional data to improve the model's accuracy over time. This feature should help in continuously refining the model’s performance based on real-world usage.

3.2 Non-functional Requirements

3.2.1 Feasibility Report

Submit a comprehensive report analysing the feasibility of Large Language Models (LLMs) as cybersecurity threat detection mechanisms. The report should include an assessment of potential benefits, limitations, and integration challenges.

3.2.2 Cybersecurity Compliance

Ensure that the tool complies with relevant cybersecurity regulations and standards (e.g., GDPR, HIPAA) across different industries, particularly for financial institutions and healthcare providers.

3.2.3 Scalability Testing

Conduct scalability testing to ensure the tool can handle large datasets typical of large enterprises and government agencies without compromising performance. This should include stress testing under high data throughput scenarios.

4 PROJECT PLAN

4.1 Agile Methodology

The project will follow an **Agile methodology**, utilising **Scrum** for iterative development. Agile was chosen due to its flexibility, allowing for continuous client feedback and iterative improvements. This approach is especially beneficial given the technical complexity of developing and fine-tuning AI models. Other methodologies, such as **Waterfall**, were considered but deemed less suitable due to their rigid structure, which could hinder the necessary adaptability during the project's development phases.

Sprint 1 (Weeks 2-4)

Focus: Initial planning, research, and feasibility

- Task 1: Project kick-off meeting and team alignment
- Task 2: Research and Literature Review on LLMs in Cybersecurity
- Task 3: Initial Feasibility Analysis Outline
- Task 4: Draft the structure of the Feasibility Report
- Task 5: Identify and gather necessary datasets for model training

- Task 6: Detailed Feasibility Analysis - Benefits and Limitations
- Task 7: Analyse Integration Challenges of LLMs in Cybersecurity
- Task 8: Prepare the first draft of the Feasibility Report
- Task 9: Review and refine the Feasibility Report with team input

Sprint 2 (Weeks 4-6)

Focus: Model training and compliance planning

- Task 10: Prepare for model training by refining dataset selection and preprocessing
- Task 11: Start Model Training for Binary Classification
    - Subtask: Data Preprocessing
    - Subtask: Model Selection and Initial Setup
- Task 12: Research and Plan for Cybersecurity Compliance Requirements
- Task 13: Begin Implementing Cybersecurity Compliance Measures
    - Subtask: Review GDPR, HIPAA, and other relevant standards
    - Subtask: Develop initial documentation for compliance

Sprint 3 (Weeks 7-9)

Focus: Model evaluation, compliance, and scalability testing preparation

- Task 14: Continue Model Training and begin Evaluation
    - Subtask: Evaluate the model on accuracy, precision, recall, and F1-score
    - Subtask: Iterative improvements based on evaluation results
- Task 15: Prepare for Scalability Testing
    - Subtask: Design Stress Testing Scenarios
    - Subtask: Set up Testing Environment
- Task 16: Finalise Cybersecurity Compliance Implementation
    - Subtask: Finalise compliance documentation
    - Subtask: Integrate compliance checks into the development pipeline

Sprint 4 (Week 10-12)

Focus: Scalability testing, final model refinements, optional features, and presentation

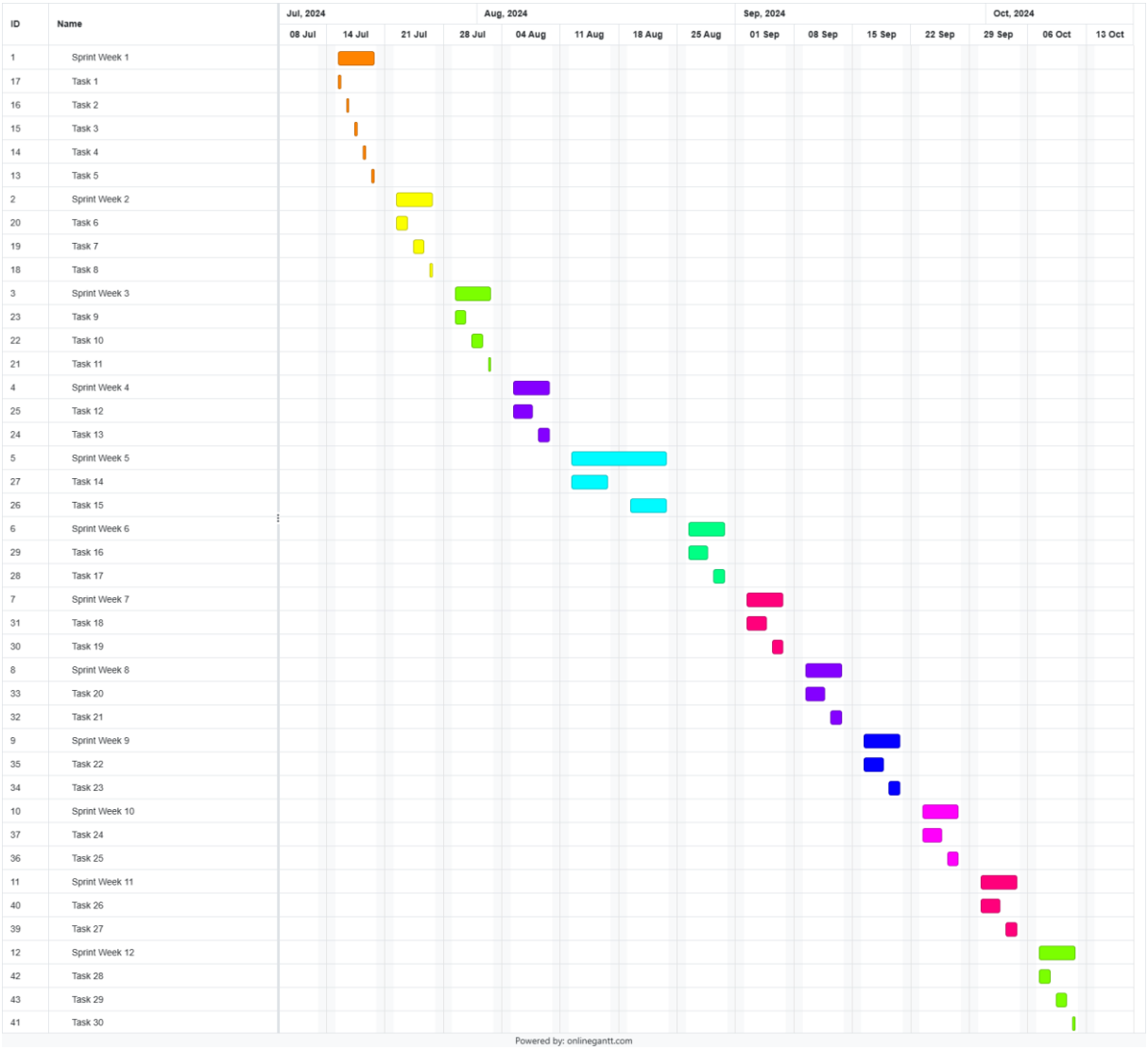
- Task 17: Conduct Scalability Testing

    - Subtask: Execute Stress Tests with Large Datasets
    - Subtask: Analyse Performance Metrics under High Data Throughput
- Task 18: Model Refinement based on Scalability Test Results
- Task 19: Optional - Begin Multi-class Classification Model Training
- Task 20: Develop Anomaly Detection Module
- Task 21: Begin Implementation of Real-time Threat Detection
    - Subtask: System Architecture for Real-time Processing
    - Subtask: Initial Integration with Phi-3 Mini Model
- Task 22: Finalise Work on Multi-class Classification Model
- Task 23: Present all finding to client

Finalisation (Weeks 12 and 13)

Focus: Finalisation and Review

- Task 24: Final Testing of All Components
    - Subtask: Functional Testing
    - Subtask: Security and Compliance Validation
- Task 25: Documentation and Final Report Compilation
    - Subtask: Compile Project Documentation
- Task 26: Project Closure and Retrospective
    - Subtask: Team Review and Lessons Learned


ID
Name
Jul, 2024
Aug, 2024
Sep, 2024
Oct, 2024
08 Jul
14 Jul
21 Jul
28 Jul
04 Aug
11 Aug
18 Aug
25 Aug
01 Sep
08 Sep
15 Sep
22 Sep
29 Sep
06 Oct
13 Oct
1 Sprint Week 1
17 Task 1
16 Task 2
15 Task 3
14 Task 4
13 Task 5
2 Sprint Week 2
20 Task 6
19 Task 7
18 Task 8
3 Sprint Week 3
23 Task 9
22 Task 10
21 Task 11
4 Sprint Week 4
25 Task 12
24 Task 13
5 Sprint Week 5
27 Task 14
26 Task 15
6 Sprint Week 6
29 Task 16
28 Task 17
7 Sprint Week 7
31 Task 18
30 Task 19
8 Sprint Week 8
33 Task 20
32 Task 21
9 Sprint Week 9
35 Task 22
34 Task 23
10 Sprint Week 10
37 Task 24
36 Task 25
11 Sprint Week 11
40 Task 26
39 Task 27
12 Sprint Week 12
42 Task 28
43 Task 29
41 Task 30
Powered by: onlinegantt.com

5 AGREEMENTS

All persons identified in this document sign the form below to indicate that they have read the Project Vision and Agreement and agree to the contents therein.

X Zahra Jadidi Zahra Jadidi 25/10/2024

Client

Edas Ramanauskas ER 11/10/2024

Zakaria Wardere  11/10/2024

Darcy McIntosh 

11/10/2024

Campbell Reid  11/10/2024

Cooper Weeks  11/10/2024

Kobi Chambers  11/10/2024