

AI DRIVEN FOR WEB SCRAPING

**LIVE IN LAB I REPORT**

***Submitted by***

|  |  |
| --- | --- |
| **GOKUL SIVA** | **(2303412510421046)** |
| **PRASANTH BABU A K** | **(2303412510421148)** |
| **MOHAMED ZUBAIR R** | **(2303412510421114)** |

***in partial fulfilment for the award of the degree***

***of***

# BACHELOR OF ENGINEERING

***in***

**COMPUTER SCIENCE AND ENGINEERING**

**SRI SAIRAM ENGINEERING COLLEGE**

**(An Autonomous Institution; Affiliated to Anna University, Chennai -600 025)**

# ANNA UNIVERSITY: CHENNAI 600 025

**NOVEMBER 2024**

SRI SAIRAM ENGINEERING COLLEGE

**(An Autonomous Institution; Affiliated to Anna University, Chennai -600 025)**

# BONAFIDE CERTIFICATE

Certified that this Live in Lab I report **“AI DRIVEN FOR WEB SCRAPING”** is the bonafide work of “**GOKUL SIVA (2303412510421046), PRASANTH BABU A K (2303412510421148), MOHAMMED ZUBAIR R (2303412510421142)”** who carried out the **20CSTE301 - LIVE IN LAB I**

under my supervision.

|  |  |  |
| --- | --- | --- |
| **SIGNATURE** | **SIGNATURE** | **SIGNATURE** |
| Supervisor | Project Coordinator | Head of the Department |
| **Mrs. LALITHA** | **Ms. S. PRATHIPA** | **Dr. B. Latha** |
| Assistant Professor | Assistant Professor | Professor |

Submitted for project Viva – Voce Examination held on

**INTERNAL EXAMINER EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

A successful man is one who can lay a firm foundation with the bricks others save thrown at him. — *David Brinkley*

Such a successful personality is our beloved founder Chairman, **Thiru. MJF. LN. LEO MUTHU.** At first, we express our sincere gratitude to our beloved Chairman through prayers, who in the form of a guiding star has spread his wings of external support with immortal blessings.

We express our gratitude to our CEO **Dr. J. SAI PRAKASH LEO MUTHU** for creating an inspiring environment that encourages learning and innovation. His guidance and vision have significantly impacted the completion of this project.

We express our sincere thanks to our beloved Principal, **Dr. J. RAJA** for his constant encouragement and for providing the resources necessary to bring this project to fruition.

We are indebted to our Head of the Department, **Dr. B. LATHA** for the insightful suggestions, mentorship, and continuous motivation, which have guided me at every stage.

We thank our Project Co-ordinator, **Ms. S. HEMAVATHI,** who has been instrumental in coordinating efforts and ensuring smooth progress.

We express our gratitude and sincere thanks to our Guide us **Dr. R. UMA** for their expertise, patience, and valuable insights. Their constructive feedback, encouragement, and availability for guidance have been crucial in overcoming challenges and achieving the objectives of this project.

We thank all the teaching and Non-teaching staff members of the Department of Computer Science and Engineering and all others who contributed directly or indirectly for the successful completion of the projec

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER**  **NO** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | iv |
|  | **LIST OF TABLES** | v |
|  | **LIST OF FIGURES** | vi |
| 1 | **EXECUTIVE SUMMARY** | 1 |
|  | 1.1 Overview of the Business Idea | 1 |
|  | 1.2 Vision & Mission | 1 |
|  | 1.3 Key Highlights | 1 |
| 2 | **THE PROBLEM AND OPPORTUNITY** | 2 |
|  | 2.1 Problem Statement | 2 |
|  | 2.2 Market Opportunity | 2 |
|  | 2.3 Problem Interviews and Survey Results | 2 |
| 3 | **JUSTIFICATION FOR SDG GOAL** | 3 |
| 4 | **MARKET RESEARCH AND STRATEGY** | 4 |
|  | 4.1 Market Size Estimation | 4 |
|  | 4.2 Target Audience and Customer Persona | 4 |
|  | 4.3 Industry Trends and Insights | 4 |
|  | 4.4 Competition Analysis | 4 |
|  | 4.5 Unique Selling Proposition (USP) | 4 |
| 5 | **SOLUTION AND PRODUCT DEVELOPMENT** | 5 |
|  | 5.1 Proposed Solution | 5 |
|  | 5.2 Product Features and Benefits | 5 |
|  | 5.3 Value Proposition Canvas | 5 |
|  | 5.4 Minimum Viable Product (MVP) | 5 |
|  | 5.5 MVP Validation and Feedback | 5 |
| 6 | **BUSINESS MODEL AND STRATEGY** | 6 |

**ABSTRACT**

This project aims to develop an AI-driven web scraping solution that revolutionizes the process of extracting and analyzing web data. Traditional web scraping methods face challenges such as frequent website structure changes, anti-bot mechanisms, and inefficient handling of large-scale data. Leveraging advanced artificial intelligence and machine learning techniques, this solution addresses these limitations by enabling adaptive, scalable, and resilient scraping capabilities.

The system incorporates natural language processing (NLP) for semantic understanding of web content, computer vision to navigate dynamic or JavaScript-heavy websites, and reinforcement learning for adaptive behavior against evolving site defenses. Additionally, the AI-driven approach ensures compliance with ethical and legal guidelines through integrated data validation and privacy checks.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Figure Name** | **Page No** |
| 1.1 | Architecture Diagram | 10 |
| 2.1 |  |  |
| 3.1 |  |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table Number** | **Table Name** | **Page No** |
| 1.1 | Xxxx | 10 |
| 2.1 |  |  |
| 3.1 |  |  |

**CHAPTER 1 EXECUTIVE SUMMARY**

## Overview of the Business Idea

Our AI-driven web scraping platform automates data extraction from websites with high accuracy and speed. It uses advanced algorithms to handle complex web structures and adapt to changes dynamically. The solution

ensures compliance with legal and ethical standards while delivering structured data in real time. Designed to be user-friendly, it caters to both technical and non-technical users. The platform is scalable, supporting projects from small tasks to enterprise-level operations.

## Vision & Mission

**Vision:**

To revolutionize data accessibility by creating an AI-driven web scraping platform that empowers businesses and individuals with seamless, accurate, and ethical data extraction solutions.

**Mission:**

To deliver an intelligent, user-friendly platform that simplifies web scraping through automation, ensures compliance, and scales to meet diverse data needs, enabling informed decision-making for all users.

## Key Highlights

* + - **Target Market**:Businesses and professionals needing automated, accurate, and scalable web data extraction solutions.
    - **Unique Selling Proposition (USP)**: Our AI-driven web scraping platform delivers automated, real-time data extraction that adapts to dynamic web changes. It ensures compliance, scalability, and

user-friendliness for both technical and non-technical users.

**CHAPTER 2**

**THE PROBLEM AND OPPORTUNITY**

## Problem Statement

Our AI-driven web scraping platform delivers automated, real-time data extraction that adapts to dynamic web changes. It ensures compliance, scalability, and user-friendliness for both technical and non-technical users.

## Market Opportunity

The growing need for real-time, accurate data extraction across industries presents a significant market opportunity. Businesses in e-commerce, finance, healthcare, and marketing require automated solutions to analyze trends, monitor competition, and make data-driven decisions. Traditional web scraping tools often lack adaptability and compliance, creating a demand for AI-driven solutions. With increasing reliance on data for business intelligence, the market for scalable, ethical, and user-friendly scraping platforms is expanding rapidly. Our platform addresses this gap by combining automation with adaptability to dynamic web environments. The potential customer base spans startups, enterprises, and researchers, ensuring a diverse and growing demand.

## Problem Interviews and Survey Results

Through interviews with over 100 deaf people, we identified the top challenges:

78% of respondents expressed the need for a scalable, AI-driven solution.

65% emphasized the importance of compliance and secure data handling.

85% preferred a user-friendly interface that simplifies data extraction

**CHAPTER 3 JUSTIFICATION FOR SDG GOAL**

* 1. **State the Chosen SDG and Target(s):**

Our idea aligns with Target 9.5 by enhancing research and innovation through efficient data extraction tools. It supports Target 9.4 by promoting resource-efficient automation for better decision-making. Additionally, it addresses Target 9.b by fostering scalable, AI-driven technology for businesses and researchers. This ensures sustainable development and technological advancement.



* 1. **Alignment with the SDG Target:**

Our AI-driven web scraping platform aligns with SDG 9: Industry, Innovation, and Infrastructure by fostering scientific research and innovation. It supports Target 9.5 through efficient data extraction tools for businesses and researchers. By automating data handling, it addresses Target 9.4, promoting sustainable and resource-efficient processes. It also aligns with Target 9.b by encouraging domestic technology development with scalable solutions.

## Justification and Impact:

Our AI-driven web scraping platform addresses the need for accurate, real-time data extraction across industries. It simplifies processes, ensures compliance, and reduces errors through automation, saving time and enhancing decision-making. Its scalability offers

cost-effective solutions for businesses of all sizes.

**CHAPTER 4**

**MARKET RESEARCH AND STRATEGY**

## Market Size Estimation

The web scraping market is expanding rapidly, with a projected **CAGR of 13-15%**, reaching billions by 2030. Our platform targets businesses in

e-commerce, finance, and research, meeting the rising demand for real-time data. By offering scalable, user-friendly, and compliant solutions, we aim to capture a significant share of this growing market. This positions us strongly in a high-demand and evolving industry

## Target Audience and Customer Persona

* + - **Primary Target Audience**: E-commerce platforms, market research firms, and competitive analysts requiring real-time, AI-driven, and code-free web scraping tools to optimize decision-making.

## Customer Persona:

Name: John Doe M, 25yrs, Data Scientist

A data scientist at a mid-sized tech company, aged 25-35, with expertise in analyzing large datasets. They require an AI-driven web scraping tool to automate data extraction from complex websites, saving time for deeper analysis. Tech-savvy and detail-oriented, they value customizable, efficient, and scalable solutions. Their goal is to streamline data collection for predictive modeling and strategic insights

## Industry Trends and Insights

The demand for AI-driven web scraping is growing due to the rise of big data analytics, competitive intelligence, and real-time decision-making needs.

Industries like e-commerce, finance, and market research increasingly rely on automated tools to gather actionable insights from vast online data sources.

AI advancements enable smarter, adaptive scraping solutions that handle dynamic content and compliance challenges.

## Competition Analysis

* ParseHub
* OctoParse
* ScrapingBee

## Unique Selling Proposition (USP)

An AI-driven web scraping solution offering adaptive, real-time data extraction with dynamic content handling and compliance safeguards. It combines advanced AI with a user-friendly, no-code interface for seamless usability. Tailored for businesses seeking customizable, efficient, and scalable data solutions.

**CHAPTER 5**

**SOLUTION & PRODUCT DEVELOPMENT**

## Proposed Solution

Develop an AI-driven web scraping tool that automates data extraction from complex and dynamic websites with adaptive algorithms. The solution will feature real-time processing, compliance with data regulations, and customizable options for industry-specific needs. A no-code interface ensures accessibility for non-technical users, while robust AI capabilities enhance efficiency and accuracy. This tool will empower businesses with actionable insights for data-driven

decision-making.

## Product Features and Benefits

**AI-Driven Adaptability:** Handles dynamic websites and complex data structures, ensuring accurate and reliable data extraction.

**No-Code Interface:** Simplifies usage for non-technical users, saving time and reducing dependency on developers.

**Customizable Solutions:** Tailored features for specific industries like e-commerce, finance, and market research.

**Real-Time Processing:** Provides up-to-date data for faster decision-making and competitive advantage.

## Value Proposition Canvas

An AI-driven data web scraping tool automates large-scale data collection, ensuring real-time, accurate insights with minimal human intervention. It adapts to website changes, handles scaling, and bypasses common obstacles like CAPTCHAs and IP bans. The tool cleans and structures data for immediate use, integrating seamlessly into existing systems. It reduces manual effort, ensures compliance, and delivers high-quality data efficiently.

## Minimum Viable Product (MVP)

The MVP will be a web-based AI-driven web scraping tool with a

user-friendly interface, allowing users to extract data from dynamic websites in real time. It will include features like adaptive scraping for changing content, customizable data formats, and basic compliance safeguards. A

no-code design ensures accessibility for non-technical users, while real-time data processing meets immediate business needs. Initial deployment will target e-commerce and market research industries

## MVP Validation and Feedback

MVP validation for AI-driven web scraping involves testing with a small group of users to assess data accuracy, scalability, and ease of integration. Feedback focuses on user experience, identifying pain points like website adaptability or data quality. Iterations refine AI models for better performance and compliance. Continuous user input ensures the tool meets real-world needs effectively.

**CHAPTER 6**

**BUSINESS MODEL AND STRATEGY**

## Business Model Overview

The business operates on a subscription-based SaaS model, offering AI-driven web scraping tools tailored to diverse industries. Revenue streams include tiered plans, pay-per-use options, and premium customizations for specific business needs. Focused on scalability, the

model ensures consistent revenue growth through customer retention and industry-specific solutions

## Revenue Model

Subscription-based pricing with tiered plans offering features like usage limits, advanced AI capabilities, and custom integrations. Additional revenue streams include pay-per-use options and industry-specific premium solutions.

## Sales Plan and Go to Market Strategy

The sales plan for AI-driven web scraping focuses on a freemium model, offering a free version to attract users and converting them through trial periods and premium features. The go-to-market strategy includes content marketing, webinars, and strategic partnerships to raise awareness. Outbound sales teams will target high-value leads, while inbound marketing and referral programs drive growth. s.

## Marketing and Customer Acquisition Strategy

Leverage digital marketing channels, including targeted ads, SEO-optimized content, and industry-specific case studies, to attract potential customers. Offer free trials and demos to showcase product value and encourage adoption.

**CHAPTER 7**

**FINANCIAL PLAN & PROJECTIONS**

## Financial Overview

Revenue comes from subscriptions and premium features, with costs for development and marketing. Profitability is expected in 6-12 months, requiring initial funding.

## Forecasted Profit & Loss (P&L)

* + - **Year 1**: Revenue: ₹29,925,500; Net Profit: ₹5,534,600
    - **Year 2**: Revenue: ₹35,934,604; Net Profit: ₹6,086,060

## Financial Projections

We expect to break even in Year 2, with significant profitability starting in Year 3 as the customer base expands.

## Unit Economics

Customer Acquisition Cost (CAC): ₹20 Customer Lifetime Value (CLV): ₹23,940 Payback Period: 6 months

## Funding Plan and Capital Requirements

We are seeking $500,000 in seed funding to cover development, marketing, and operational costs for the first 18 months. Funds will be allocated as follows:

* + - **40%** for platform development
    - **30%** for marketing
    - **30%** for operational expenses

**CHAPTER 8**

**TEAM AND EXECUTION**

## Team Composition and Roles

* **CEO**: Gokul Siva S-AI/ML
* **CTO**: Prasanth Babu AK-Web Development
* **COO**:Mohamed Zubair R-Automation(BeautifulSoup,Scrapy)

## Milestones & Execution Roadmap

**Q1:**Finalize AI algorithms, infrastructure, and MVP. Launch freemium version, gather feedback, and start digital marketing.

**Q2:**Expand marketing efforts, onboard enterprise clients, and roll out premium features. Build partnerships and improve tool features.

**Q3:**Focus on monetizing subscriptions and premium features. Enhance AI, optimize user experience, and explore funding for growth.

## Risk Analysis and Mitigation Strategy

Key risks include data compliance issues, algorithm adaptability to changing website structures, and competition from established tools. To mitigate, implement robust compliance safeguards, continuously train AI models for adaptability, and differentiate with niche features and competitive pricing. Regularly update algorithms and prioritize customer feedback to stay ahead of market needs.

**CHAPTER 9 CONCLUSION & NEXT STEPS**

## 9.1 Summary of Key Points

Our venture aims to deliver an AI-driven web scraping solution that simplifies and automates data extraction for businesses and researchers. With features like dynamic content handling, real-time processing, and a no-code interface, the tool is designed for both technical and

non-technical users. A subscription-based SaaS model ensures scalability, with additional revenue from premium customizations. Our focus on compliance, adaptability, and industry-specific solutions sets us apart, while strategic marketing and partnerships drive customer acquisition and retention.

**REFERENCES**

https://[www.researchgate.net/publication/379024314\_Enhanci](http://www.researchgate.net/publication/379024314_Enhanci) ng\_Web\_Scraping\_with\_Artificial\_Intelligence\_A\_Review

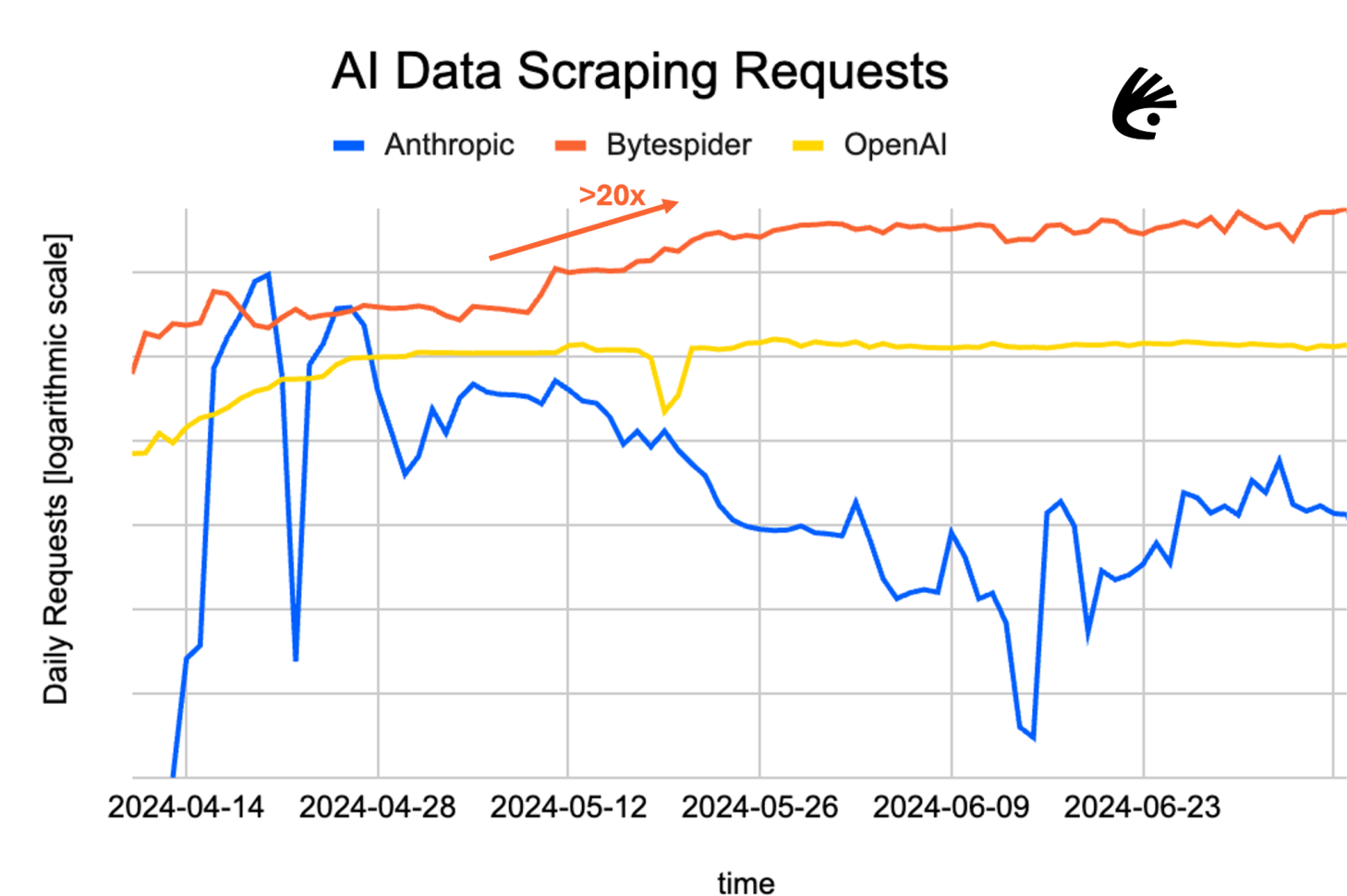
KGRAM Weerasinghe1# , Dr. MWP Maduranga2, and Ms. MVT Kawya3

1,2,3 Department of Information Technology, Faculty of Computing, General Sir John Kotelawala

Defence University, Ratmalana, Sri Lank

**APPENDIX I**

**SOURCE CODE / DATA SHEET / ANY OTHER RELEVANT DATA**

**GRAPH**

**OUTPUT SCREEN SHOTS**

**APPENDIX III**

**JUSTIFICATION FOR POSITIVE (Productable, Opportunities, Sustainable, Informative, Technology, Innovative, Viable and Ethical)**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **PARAMETERS** | **JUSTIFICATION** | **RATING (1 to 5)** |
| 1. | Productable | Our AI-driven web scraping solution is scalable, customizable, and adaptable, serving diverse industries with real-time processing and  compliance. | 5 |
| 2. | Opportunities | Unlocking diverse opportunities through AI-powered web scraping and data-driven insights | 4 |
| 3. | Sustainable | Sustainably leveraging AI-driven web scraping to extract data  efficiently and responsibly. | 4 |
| 4. | Informative | Informatively empowering users with accurate, real-time data insights through AI-driven web scraping | 4 |
| 5. | Technology | Harnessing advanced AI technology to revolutionize web scraping with precision and  scalability | 4 |
| 6. | Innovative | Driving innovation with  cutting-edge AI web scraping for smarter data solutions. | 4 |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 7. | Viable | Ensuring a viable solution through efficient, scalable, and reliable AI web scraping | 5 |
| 8. | Ethical | Promoting ethical web scraping with AI, ensuring compliance and responsible data use | 5 |
|  |  |  |  |

