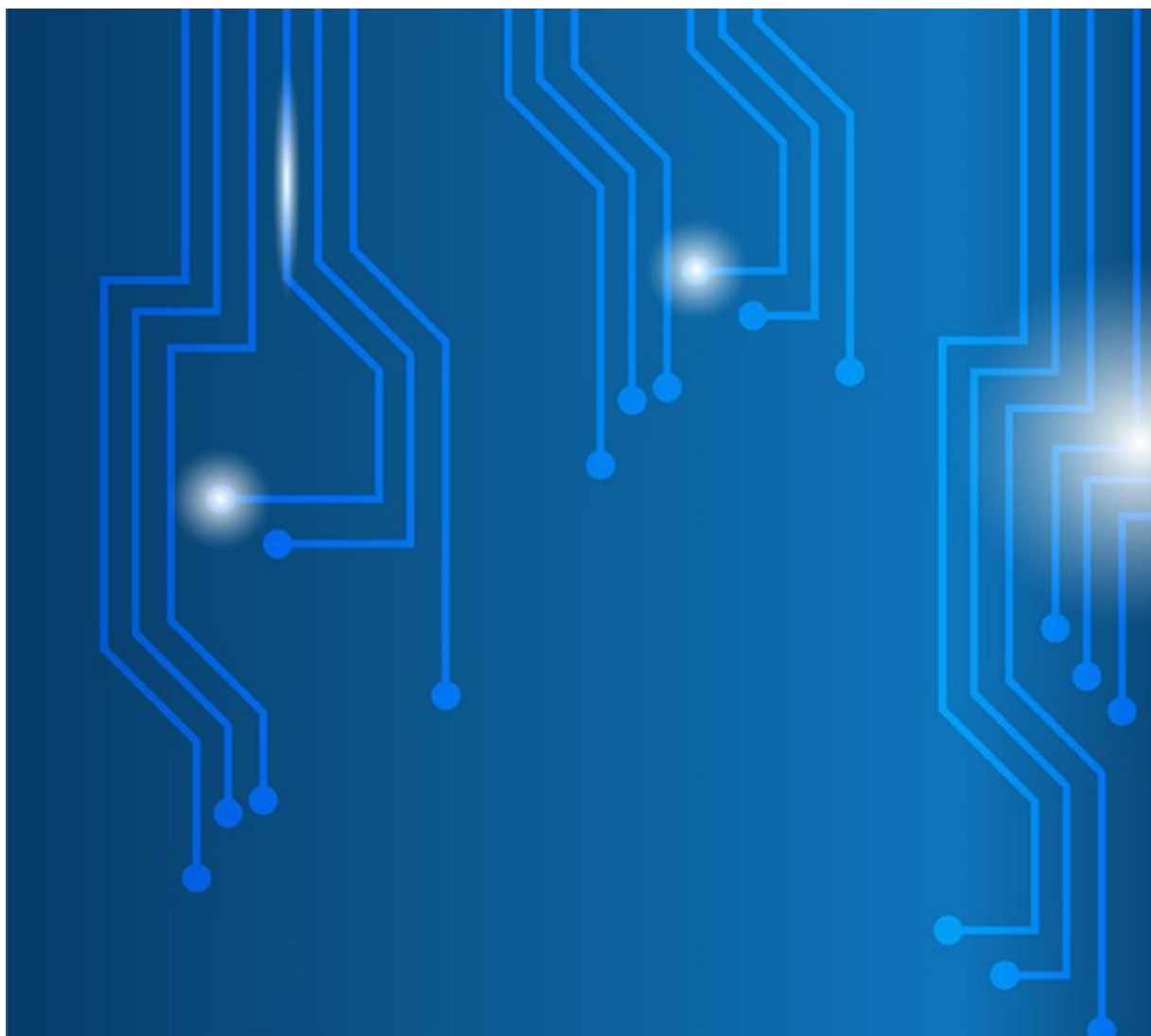




**Journal of University of Information Science
and Technology "St. Paul the Apostle" - Ohrid
ISSN: 2955-182X**

Publisher University of Information Science and Technology "St. Paul the Apostle" - Ohrid

Volume 5 | Issue 2 | 2025

EDITORIAL - Preface to Volume 5 Issue 2 of the Journal of University of Information Science and Technology “St. Paul the Apostle” – Ohrid

Dear Readers,

Journal of University of Information Science and Technology “St. Paul the Apostle” – Ohrid is an international, peer-reviewed journal published annually since 2016. It is an open access Journal available at the website of the University of Information Science and Technology “St. Paul the Apostle” – Ohrid (<https://journal.uist.edu.mk/>)

This Journal is committed to publishing and disseminating high-quality and novel scientific research work in the broad field of Applied Mathematics, Automation & Robotics, Blockchain, Business, Economy, Entrepreneurship & Management, Communication & Network Technology, Cybersecurity, Data Science, Distributed and Parallel Computing, Digital Art and New Media Art, Metaverse, Virtual and Augmented Reality, Communication & Network Technology, Smart Systems & IoT, Information Society & Digital Transformation, Biomedical Engineering.

As editor-in-chief of the Journal of the University of Information Science and Technology “St. Paul the Apostle” – Ohrid, I am pleased to present Volume 5 issue 2, an issue devoted to a wide range of topics, including the assessment of digital transformation and e-governance within public administration the deployment strategies for Large Language Models (LLMs) and the intersection of AI and sustainability in circular manufacturing. This volume also explores workflow optimization through project management software and Gantt charts in the fintech industry, and novel applications of AI in cybersecurity, such as modeling natural randomness for NGO security. We thank all the authors for contributing to this Issue and all the reviewers for providing detailed and timely evaluations of the submitted manuscripts.

It is an excellent challenge for the editor members to bring a new journal issue into the world, primarily when the Journal aims to publish high-quality manuscripts. I want to express my sincere gratitude to all of them for their excellent work, remarkable contribution, enthusiasm and support.

Sincerely Yours,

Editorial Board

October 2025

Ohrid



Journal of University of Information Science
and Technology "St. Paul the Apostle" - Ohrid
ISSN: 2955-182X

IMPRESUM

FOUNDER AND PUBLISHER

University of Information Science and Technology "St. Paul the Apostle" – Ohrid

EDITOR IN CHIEF

Professor Ninoslav Marina, PhD

EXECUTIVE EDITOR

Associate Professor, Atanas Hristov, PhD

EDITORIAL BOARD

Professor Arban Uka, Epoka University – Tirana, Albania

Professor Ishwar K. Sethi, PhD Oakland University, USA

Professor PREM K. KALRA, PhD Head & Professor, Indian Institute of Technology, Kanpur, India

Professor Eric Châtelet, PhD Université de technologie de Troyes, France

Professor Weiler A. Finamore, PhD Pontifical Catholic University of Rio de Janeiro, Brazil

Professor Katsumi Yamashita, PhD University of Osaka, Japan

Professor Mohammad Reza Asharif, PhD University of the Ryukyus, Okinawa, Japan

Professor Heung Gyoon Ryu, PhD Chungbuk National University, Seoul, South Korea

Professor Abdel-Badeeh M. Salem, PhD, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

Associate Professor, Dijana Capeska Bogatinoska, PhD UIST "St. Paul the Apostle" – Ohrid, Macedonia

Associate Professor, Aneta Velkoska, PhD UIST "St. Paul the Apostle" – Ohrid, Macedonia

Associate Professor, Jovanka Damoska Sekuloska, PhD UIST "St. Paul the Apostle" – Ohrid, Macedonia

Assistant Professor, Aleksandar Karadimche, PhD UIST "St. Paul the Apostle" – Ohrid, Macedonia

Assistant Professor, Jovan Karamachoski, PhD UIST "St. Paul the Apostle" – Ohrid, Macedonia

Assistant Professor, Ljubinka Sandjakoska, PhD UIST "St. Paul the Apostle" – Ohrid, Macedonia

ISSN 2955-182X

CONTENTS

F. Milosheski. J. Damoska Sekuloska.

Increasing Project Efficiency by Utilizing Project Management

Software & Gantt Charts 5

B. Naumoski Guli

Enhancing NGO Security Through Data Modeling

with Artificial Intelligence and Natural Randomness 18

D. Poposki

The Intersection of AI and Sustainability: Self-Optimizing AI

Systems for Circular Manufacturing 29

P. Risteski

Digital Transformation in N. Macedonia Municipalities: A Case Study of the Southwest Region 42

S. Trajanoski. A. Karadimce

Comparative Analysis of Large Language Models: On-Premise Architectures vs. Cloud-Based

Deployments 48

Increasing Project Efficiency by Utilizing Project Management Software & Gantt Charts

Filip Milosheski. Jovanka Damoska Sekuloska.

Email: milosheski.filip@gmail.com

Department of ICS, University of Information Science and Technology “St. Paul the Apostle”, 6000,
Macedonia

Abstract

Project management is a critical component in ensuring the success and timely delivery of projects across industries. With increasing complexity and time constraints, effective project management has become more essential than ever. In today's competitive and technology-driven environment, organizations are increasingly turning to digital tools to streamline project workflows, improve collaboration, and enhance overall productivity. Among these tools, project management software and Gantt charts play a vital role in helping teams plan, track, and execute projects efficiently. This research explores the value of adopting project management software in increasing the efficiency, organization, and success rate of projects. Specifically, the study focuses on a case study conducted on an international fintech company that utilizes Asana and Instagantt to manage its projects. By examining the integration and use of these tools within the company's project management processes, the study provides insights into how digital platforms can optimize task delegation, progress tracking, and interdepartmental communication. The findings highlight how structured planning through Gantt charts and centralized project data offered by such tools contribute significantly to meeting project deadlines and achieving strategic goals. The research ultimately supports the idea that implementing the right software can transform project management practices and serve as a key driver of success in modern organizational settings.

Keywords: Project management, gantt charts, efficiency, Asana, Instagantt, fintech, workflow optimization, team collaboration

Introduction

Project management plays a crucial role in ensuring projects are completed on time, within budget, and to the desired quality standards. Today, almost all projects incorporate some form of project management, which provides a structured framework for planning, organizing, executing, and monitoring tasks and with this framework the optimization of resources and the mitigation of potential risks is easier. Effective project management enables teams to set clear goals, allocate tasks efficiently, enhance communication among stakeholders, and track progress more easily. Utilizing these practices significantly increases the likelihood of project success.

From the previous paragraph the immense value that project management brings to a project can be easily understood, and in today's world there are multiple softwares which make project management easier and more efficient. These softwares provide a centralized platform for project managers to manage resources, assign tasks, track progress, and so on, in a simple and smooth way, eliminating the big barrier to entry and ensuring better collaboration across all different parts of the project. Following are some of the more widely used project management softwares with their most distinguishing features mentioned:

- Trello: Best utilized for smaller teams and simple projects;

- Asana: Best utilized for larger teams, with more complex and detailed projects;
- Microsoft Project: Best for enterprise level project management with high attention to detail;
- Jira: Mostly utilized for software development teams utilizing agile methodologies;
- Smartsheet: Utilized for spreadsheet-based project management, where detailed tracking and reporting for the project is needed

Leveraging project management software has become an increasingly paramount way of increasing the project efficiency, as these digital tools help in streamlining workflows, enhancing collaboration and providing an easy way to track the project's progress among others. As with the usage of the project management software the project is centralized as well as its tasks, deadlines, communication and other relevant things, the risk that mistakes or delays will affect the project is minimized and thus increases the likelihood of success of that project. It should not be overlooked the analytical capabilities of project management software, as that can provide great insight to improve the data-driven decision making process within the project, helping with optimizing the project's effectiveness and efficiency.

Gantt charts can be of enormous value in the process of increasing the project effectiveness and efficiency, by providing a visual representation of the project timelines and its dependencies, and with it simplifying the planning, organizing and monitoring process. Gantt charts provide the ability to project managers to clearly see the tasks, their timelines and dependencies and how they are interlinked with each other, and easily identify potential risks or bottlenecks which might impact the project negatively. As the gantt charts offer great visualization of the project, its timelines and milestones it also assists with the communication with the various stakeholders when it comes to the project's progress. All of the benefits which are provided by the gantt charts provide great value when it comes to making informed decision making and keeping the project on track, while mitigating its risks, all of those contribute to the boosting of the overall project efficiency.

This article is composed of 5 parts which explain in detail how project management software and gantt charts increase the project efficiency. In the next part i.e. "Related Work" other similar articles are reviewed and how they support or deny the main topic of this article. The core of this article is in the 3rd part which is the case study, where an international fintech company is studied on how it utilizes project management software and gantt charts for better working of the company, while in the last 2 parts, the "Results" and "Conclusion" parts the results of the case study are being discussed and then an overall conclusion for this topic and suggestions for improvements are covered.

Related Work

Building on the foundational understanding of the role and advantages of project management software and Gantt charts discussed in the introduction, this section reviews existing literature and studies that explore similar themes. Various researchers and industry professionals have investigated how digital tools contribute to project efficiency, collaboration, and risk mitigation. By examining prior work, we can contextualize the insights of this article within the broader academic and professional discourse, identifying both supporting evidence and differing perspectives. This review highlights key findings from relevant studies, focusing on the practical implementation, benefits, and limitations of project management tools and visualization techniques like Gantt charts across different industries and project types.

In his article "Enhancing Project Quality Through Effective Team Management," Sławomir Wawak emphasizes the critical role of efficient project and team management in improving project quality. His

research highlights that establishing a structured organization, thorough planning, well-defined roles and responsibilities, alignment with stakeholder requirements, and fostering a supportive work environment are key to enhancing project quality. For optimal team performance, it is essential to assign clear tasks and responsibilities to team members, who should be continuously monitored and supported by the project manager. Maintaining clear communication channels with regular meetings allows team members to share ideas and concerns, thereby boosting project quality and success. Wawak's article underscores the significance of open communication and effective people management as vital factors in achieving successful project outcomes. [1]

From the article "Optimization of Resource Allocation and Task Allocation with Project Management Information Systems in Information Technology Companies" it can be seen that integrating a Project Management Information System (PMIS) with the Resource Allocation and Task Allocation Optimization System (RATAOS) can significantly enhance efficiency and accuracy in IT companies. This integration reduces planning time by 31.78 hours and operational costs by USD 352.1, while improving resource allocation accuracy to 72.50% with a processing time of 0.029 seconds. The streamlined process accelerates project completion by 1.5 months, nearly 50% faster than current timelines, which aren't bad numbers at all, with these increases there is great increase in the efficiency of the project which results in faster completion of the project as well as lower costs for it. The paired t-test confirms significant improvements in time and cost. Future research in this area should focus on implementing the system for real-world observations and modeling project completion processes to compare pre- and post-integration parameters, so that organizations can potentially start implementing this and improve their project managing processes. [2]

With the rise of AI and machine learning in recent years, the potential of using machine learning for effort estimations in software projects has garnered attention. This approach could significantly aid project managers by providing more accurate predictions for task efforts, improving task organization and structure, and facilitating smoother project management. In the article "Analysis of Software Effort Estimation by Machine Learning Techniques," researchers explored this concept and demonstrated its value to project managers in achieving precise effort estimations. By employing techniques such as linear regression and random forest, along with correlation analysis, the researchers effectively created and tested machine learning predictions for task completion efforts. Although the analysis was not highly advanced, it still offered valuable insights for project managers. Moving forward, it is likely that AI and machine learning will play a greater role in project management, particularly in software projects where their accuracy is most beneficial. This advancement allows project managers to focus on other responsibilities, ultimately increasing the likelihood of project success. [3]

A valuable research which contributes to increasing the understanding of project management efficiency by identifying variables with strong correlations to project efficiency, aiding project managers in targeting future improvements will be looked at as a great point on how project efficiency can be improved even more. The research advances the knowledge of project portfolio management, particularly the complexities of projects and their interaction with the strategic decisions given by higher management. This research links the project complexity with the efficiency outcomes coming from the strategic activities and guidelines. This presents an evaluation model incorporating novel variables, identifying high-performing projects as internal benchmarks for strategic decision-making. From a managerial perspective, it underscores the efficiency gains by emphasizing the strategic role in enhancing project outcomes. The study's limitations include its focus on a single industrial plant and exclusion of routine

maintenance and simple purchase projects, suggesting further research should replicate the evaluation in diverse international settings and explore techniques for timely project completion. [4]

Global software development (GSD) is increasingly prominent in the industry, necessitating effective project management due to its complexity compared to traditional projects. The article “Evaluating Success Factors of Software Project Management in Global Software Development” identifies five dimensions and twenty factors critical to the success of GSD projects, with communication and coordination being the most crucial. These dimensions and factors are detailed in Figure 1, which provides in-depth information on their impact and significance. This research offers project managers valuable insights into managing global teams by focusing on the key factors that influence project success, thus enhancing the likelihood of achieving successful outcomes in GSD projects. [5]

Dimension	Dimension weight	Factors	Local Weights	Global Weights	Overall Ranking
Communication (Com)	0.4651	Effective communication skills (ECS)	0.6963	0.3239	1
		Good negotiation strategies (GNS)	0.0840	0.0391	7
		Communication Frequency (CoF)	0.2197	0.1022	3
Coordination (Coo)	0.2792	3C (Communication, Coordination, Control)/Better coordination & Control (CCC)	0.6338	0.1770	2
		Effective customer involvement (ECI)	0.2095	0.0585	5
		Member collaboration/cooperation (MeC)	0.0522	0.0146	13
		Regular Meetings (communication and coordination) (ReM)	0.1045	0.0292	10
Human Resource Management (HRM)	0.1363	Experienced developers (ExD)	0.4946	0.0674	4
		Self-organizing teams (SoT)	0.0603	0.0082	16
		Good Cultural awareness (GCA)	0.0363	0.0049	17
		Team Formulation (TeF)	0.1345	0.0183	12
		Member Skills Roles and Responsibilities (MSR)	0.2743	0.0374	8
Technology (Tec)	0.0414	Rich technological infrastructure (RTI)	0.7308	0.0303	9
		Technology differences (TeD)	0.0696	0.0029	19
		Software tools (SoT)	0.1996	0.0083	15
Project Management (PM)	0.0780	Knowledge sharing management (KSM)	0.1078	0.0084	14
		Cultural difference management (CDM)	0.0580	0.0045	18
		Management of time-zone differences (MTD)	0.0343	0.0027	20
		Essential project management processes (EPM)	0.5234	0.0408	6
		Progress Tracking (PrT)	0.2764	0.0216	11

Figure 1: Composite weights ad ranks of dimensions and factors in GSD [5]

Recent studies underscore the pivotal role of project management software in augmenting organizational efficiency. For instance, a study by Gamil and Abd Rahman highlights how such software streamlines complex construction projects by mitigating uncertainties and facilitating better scheduling and financial planning . Similarly, research by Patanakul demonstrates that the adoption of project management tools across various project life cycles significantly impacts project success rates, emphasizing the importance of tool selection and implementation . These findings suggest that integrating appropriate project

management software can lead to more effective resource allocation and improved project outcomes. [6][7]

Gantt charts remain a cornerstone in project planning, offering visual clarity on timelines and task dependencies. Research by Gleek illustrates how Gantt charts facilitate better communication among stakeholders by providing a clear overview of project schedules. Furthermore, a guide by Slikk.ai (2023) emphasizes the utility of Gantt charts in research projects, aiding in the identification of potential bottlenecks and ensuring timely task completion. These insights affirm that Gantt charts are instrumental in enhancing transparency and efficiency in project management. [8][9]

The evolution of project management tools now encompasses artificial intelligence (AI) and knowledge management functionalities. A study by Bahi explores how AI integration in project management can streamline workflows and reduce manual errors, thereby increasing overall project efficiency. Additionally, research by Lopez discusses the significance of knowledge management features in project management tools, highlighting their role in preserving institutional knowledge and promoting continuous learning within organizations. These advancements indicate a trend towards more intelligent and adaptive project management solutions. [10][11]

Case Study

In this case study an international fintech company will be explored on how it utilizes software and gantt charts for project management and with it increasing the efficiency of the project. The software used for project management within the company is Asana, and that is because as stated previously in this article Asana is used for larger teams and projects who are more complex and detailed. As the company has over 200 employees Asana is their choice for the project management software and basically every employee uses Asana on a daily basis, starting from the customer facing customer support team for logging user issues and requests up to the C-suite management for overseeing the strategic processes going within the company. The management within the company uses Asana to create portfolios (collection of projects), projects and tasks for the employees they are managing and through it track the progress of the tasks and projects, as well as get updates and provide directions and assistance.

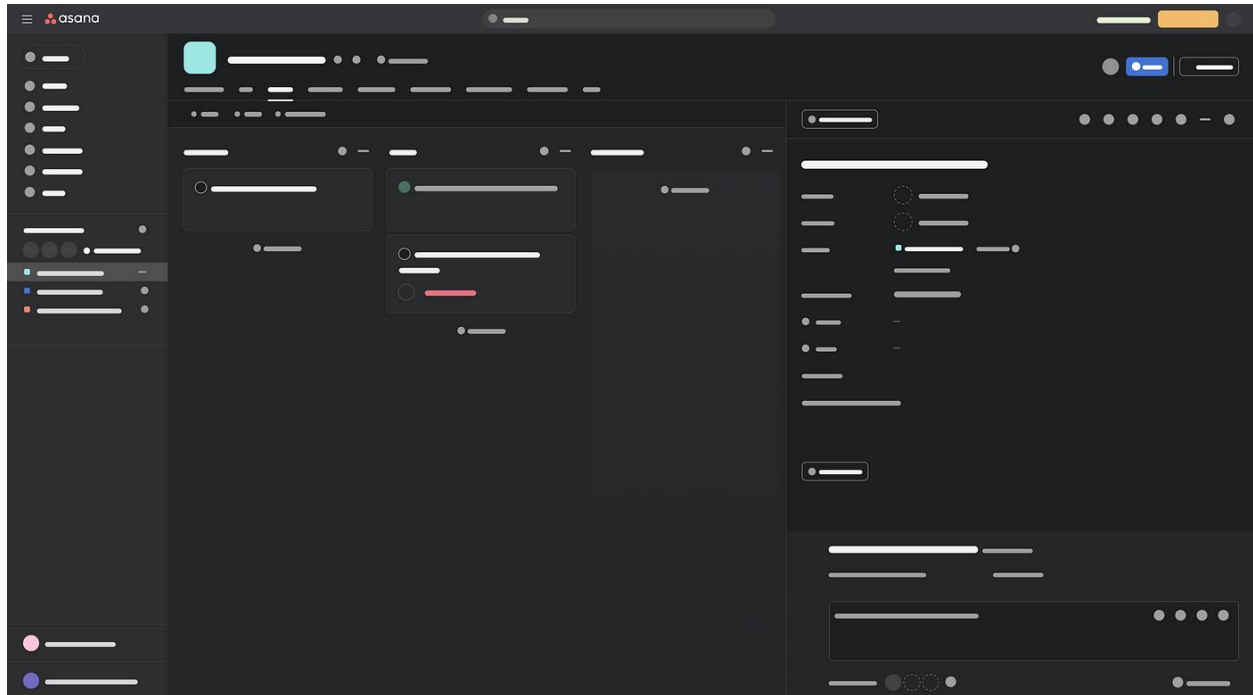


Figure 2: Asana Board View

In addition to Asana, management uses Instagantt as well as additional software used for project management. Asana offers multiple various views of projects such as the list, board, gantt and calendar view, which all are useful in their own way and depending on the preferences of the manager, however when it comes to the gantt charts within Asana they are lacking for the needs of management and that is why Instagantt is also used. Instagantt as the name states is mainly used for gantt charts which are charts who show the dependencies, timelines and responsible people across projects and the tasks in them. Instagantt was chosen because of the easy integration with Asana i.e. it is quite simple to import and automatically update all projects and tasks present in Asana to Instagantt and manipulate them even through Instagantt. The gantt charts on Instagantt as compared to Asana are better, as they offer more analysis options, easier interface to work within for the gantt charts and easier tracking and visualization for the gantt charts.

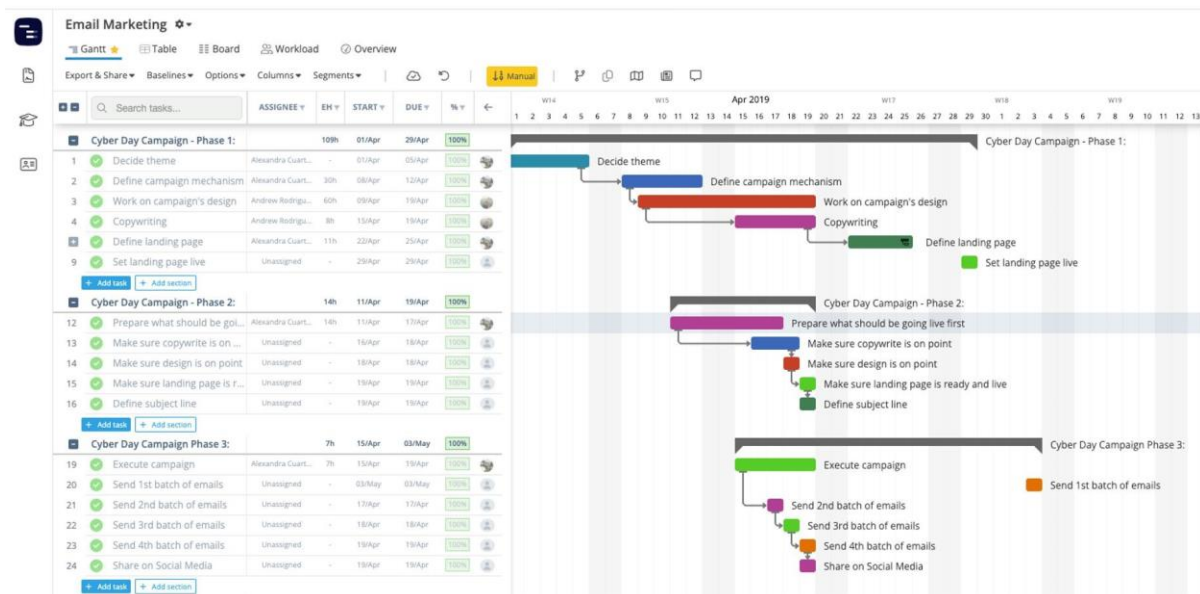


Figure 3: Gantt Chart in Instagantt

With the basic information for the softwares used provided, now it will be explained more in-depth how project and product managers are utilizing these softwares in their daily work. Each project manager creates a portfolio in Asana which is the central place where all of their projects should be located, with the purpose of easy navigation across all projects and for other stakeholders and senior management to easily understand which projects are under which project manager. As new projects are assigned to the project manager they would be creating a new project in Asana and adding it to their portfolio, as well as projects are completed, the completed project would also be completed in Asana and removed from the portfolio in order to keep a clean and organized structure within Asana.

Name	Status	Task progress	Date	Priority	Owner	Save layout as default
Editorial Calendar	On track	17%	Today – 31 Dec	Med...	AW	65h 00m
On Tour: London Event	On track	8%	13 Sep – 31 Dec	High	AW	61h 45m
On Tour: Tokyo Event	On track	18%	22 Oct, 2023 – 22 Sep, 2024	Med...	AW	261h 00m
Product Launch Campaign Tracker	At risk	34%	2 Sep – 10 Nov	Med...		54h 00m
Request for Creative Production	On track	21%	23 Sep – 2 Dec	High	AW	167h 45m

Figure 4: Portfolio in Asana

Apart from creating the project and adding it to the relevant portfolio the first step when a new project is created is to state the relevant sections and create the initial tasks for that project. Usually the initial tasks are related to the research and scoping of the project as a new project always needs to be scoped out and a scoping document should be created explaining what does the project do, who are the relevant stakeholders, what are the requirements, how it will be created, what are the different phases and anything else relevant for the project. The scope documents are written by the project managers in consultation with the relevant stakeholders and once they are done, all of the stakeholders need to sign off on the phase of the scope which is to go into implementation, in order to ensure that everyone is happy and agrees with the scope of the project and avoid rescoping while in implementation.

With the scope being finished the rest of the tasks needed for the implementation of the project such as design, development, marketing and so on are being created and timelines for each task is stated and tasks are assigned to the responsible people with clear descriptions on each tasks so that everyone know what is their part in the project and what they need to do. All of the communication for a task, such as updates, changes in due date, changes in direction or everything regarding the task is going through the comment section on the task, in that way communication through other platforms is removed in order to have all of the information about the task within Asana and having it as a single source of truth, so that everyone who needs to get information about the given task can find everything within a single place.

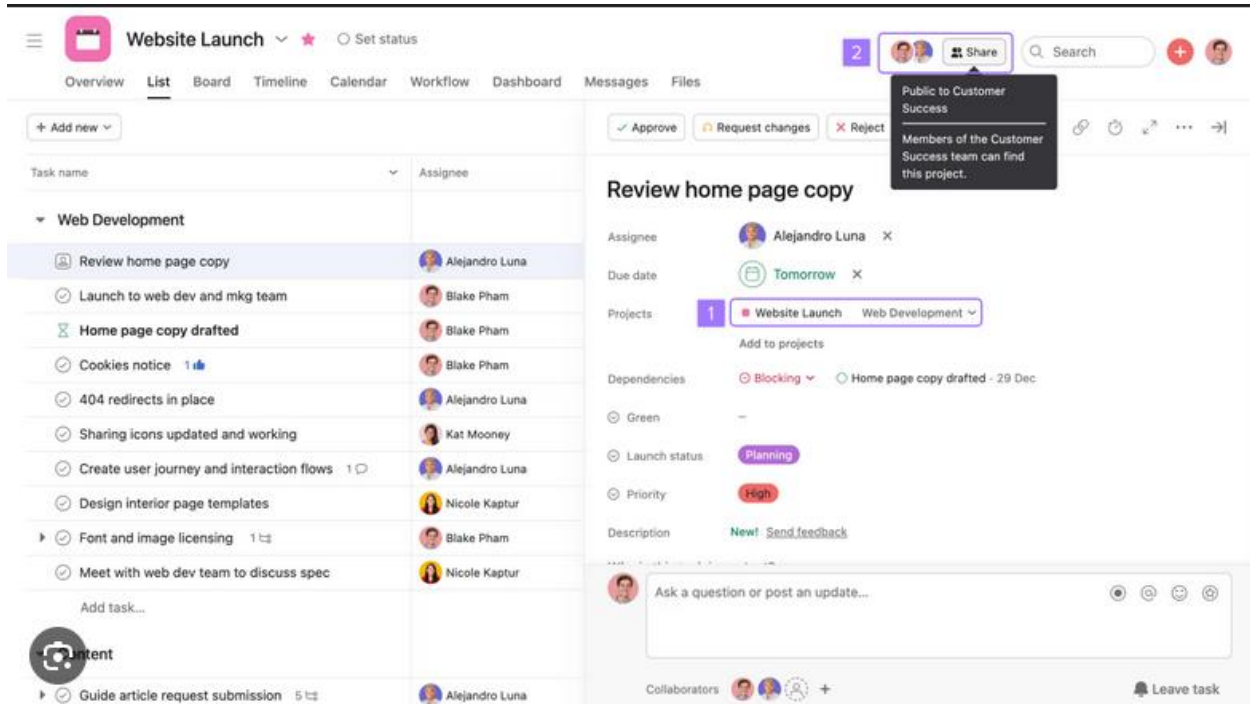


Figure 5: Task in Asana

Once the project is underway the tracking of the progress and monitoring what is happening with it, is being done mainly through Instagantt and in some cases through Asana as well. On a daily basis the project managers are looking through the gantt chart of the project to ensure that all tasks are progressing as they are planned to and understanding the dependencies between tasks. At least once a week more senior management and other relevant stakeholders have a meeting where the project manager goes through the gantt chart of the project to show the progress that is being made with the project and notify them about any changes, potential issues or escalations which are ongoing with the project. As the gantt chart is easy to understand visually it is simple and easy for all relevant stakeholders to understand how the project is going.

When the project manager is going through the gantt chart of the project if they see a task where they need to intervene such as new directions, clarifications or similar the project manager can easily through Instagantt comment on the task, reassign it or perform any other relevant action upon in, in order to ensure the process continues smoothly. In the cases where a task is at risk of not being completed within the timeline or it is already past it's due date, the project manager is checking whether there is an update on the task on why is that, and if there is no update they would be writing in the comment section, tagging the responsible person and ask for an update why the task is at risk and what would be the new due date for it, so that other dependent tasks can be adjusted accordingly.

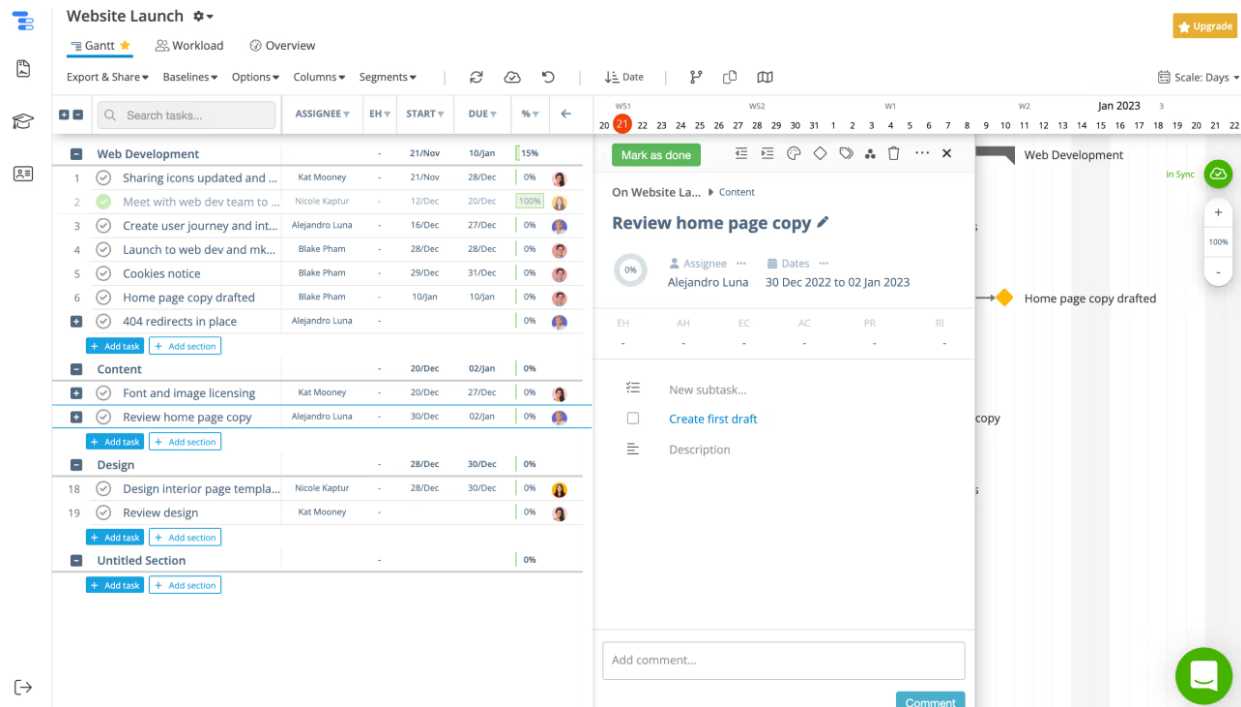


Figure 6: Task Manipulation in Instagantt

The baselines and workload features within Instagantt can also be quite valuable for the project managers as they can provide valuable insight. Baselines are in essence screenshots of the gantt charts at a given moment, and they are quite useful for seeing previous states of the gantt charts and how timelines have changed from what was initially planned, which can provide valuable insight in where most often the timelines have been predicted wrongly which can be a valuable lesson for future projects. The workload feature within Instagantt shows a version of the gantt chart view however instead of by sections as structured in Asana it is showing the tasks that each contributor to the project has, their timelines and in essence it can be easily observed if someone has a bigger workload than others, or if someone has a smaller workload and can be assigned additional tasks.

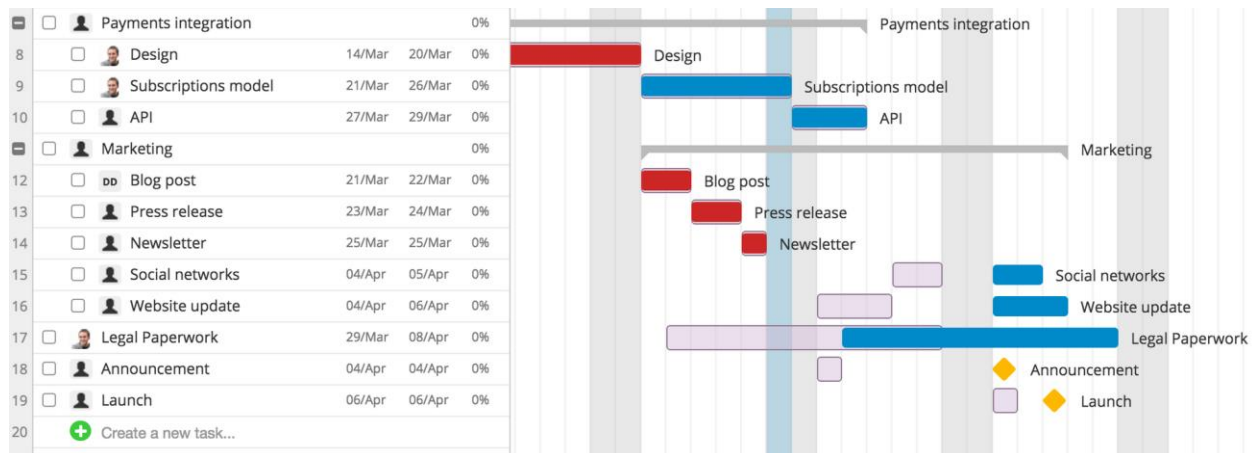


Figure 7: Baselines in Instagantt

As each major section of the project is being finished it is always required for it to go through the quality assurance stage. For example, in the case when the design of the project is done, that design needs to be reviewed by the project manager in line with the requirements stated in the scope document. As the company in question is a fintech company, almost all of the projects include the development team which in most cases is divided into the frontend and backend team, and whenever one of these teams finished their work on the project, a designated quality assurance specialists tests that the work done is done properly with the utmost attention to detail and that any relevant requirements are being taken into account and implemented. Without this step a project cannot be completed as without going through it, there is no way to be completely sure that the project which is completed is of good quality and fulfills the requirements stated in the scope document.

Results

The case study highlights the complexity of project management processes, which would be significantly more challenging without using tools like Asana or Instagantt. These software solutions offer a straightforward, organized structure that is easy for all stakeholders and contributors to follow. They provide managers with a clear overview for monitoring and tracking the project, ensuring timely updates for each task and facilitating seamless communication and coordination with relevant stakeholders. Asana's value in managing projects is immense, offering a centralized hub for all portfolios, projects, and tasks, and serving as the definitive source of truth for all project-related information. This ensures everyone knows where to find relevant details. It provides an organized structure accessible to both employees and managers, with minimal entry barriers. Its flexibility in offering different project views makes it easy for various users to navigate and utilize effectively.

Instagantt proves invaluable from its seamless integration with Asana, ensuring tasks and projects sync effortlessly between both platforms. Its gantt charts offer managers a visual representation of project timelines and dependencies, facilitating easy tracking of progress and early identification of potential risks before they escalate. While managing projects solely with Asana is possible, using Instagantt enhances the ability to monitor projects, present them to stakeholders, and preemptively address issues and workload baselines. The synergy between these tools enhances project management efficiency and significantly improves the chances of project success.

Asana and Instagantt significantly enhance project efficiency by streamlining task management and providing clear visual timelines. Asana offers robust features for task assignment, progress tracking, and team collaboration, ensuring that all team members are aligned and aware of their responsibilities. Integrating Instagantt with Asana brings an added layer of efficiency by transforming tasks into Gantt charts, which visually represent project timelines, dependencies, and milestones. This visual representation allows teams to easily identify potential bottlenecks, risks, adjust schedules, and allocate resources more effectively, leading to improved planning and execution of projects. Together, these tools provide better communication, reduce the likelihood of missed deadlines, enhance overall productivity, and increase the likelihood that the project will be successful.

The company by utilizing the previously explained approach has had over 90% of their projects be successful, delivered in the timeline agreed and satisfying the requirements initially set. The outlined approach is quite simple and easy to follow, even newcomers can easily understand it and practice on a daily basis, ensuring business continuity as the company scales up. Of course no approach is perfect and it can always be improved, however the outlined approach can be considered as a good example on how to utilize project management software in order to increase the likelihood of success for the project as well as provide a great deliverable at the end of it, in order to ensure a satisfied and loyal client at the end of the project.

Conclusion

This article clearly highlights the significance of project management software and Gantt charts in enhancing project efficiency. The case study illustrates that managing a project from start to finish is feasible only with such tools. For the international fintech company discussed, utilizing these tools is crucial for effective project management, as their absence would render the process unmanageable and hinder the company's success in the competitive market. Any company, regardless of size, must employ project management software to boost the chances of project success. Without these tools, projects can easily become unmanageable and costly, offering little value to the company. Although this article focuses on Asana and Instagantt, they are not the only effective options for project management. Many other excellent project management software choices exist, and selecting any of them would be beneficial. The only real mistake is for a company to forgo using any project management software, especially when it comes to launching a new product/service or launching the product/service in a new market. It will quickly become evident that to succeed, companies and their projects require such tools due to the immense value they provide, making them essential in today's market.

The case study examines an international fintech company's effective use of Asana and Instagantt for project management. It highlights how these tools complement each other, enhancing the ease and effectiveness of project management. Asana is praised for being one of the best project management software options, especially for larger companies with complex projects. Instagantt, with its Gantt charts, is deemed essential for effective project management, providing excellent visualization, easy tracking, and detailed analytics of project performance.

Some might question the necessity of using both Asana and Instagantt, given the need to pay subscription fees for both. While it is true that this approach is more expensive, the financial cost is offset by the significant value these tools bring to project management across multiple projects. Aside from the higher cost, there are no other drawbacks to using this combination. Thus, the increased monthly subscription is well justified by the enhanced efficiency and effectiveness in managing projects.

Although this article focuses on Asana and Instagantt, they are not the only effective options for project management. Many other excellent project management software choices exist, and selecting any of them would be beneficial. The only real mistake is for a company to forgo using any project management software. It will quickly become evident that to succeed, companies and their projects require such tools due to the immense value they provide, making them essential in today's market.

References

1. Sławomir Wawak (2024) 'Enhancing project quality through effective team management', *International Journal of Information Systems and Project Management*, 12(2). doi:10.12821/ijispm120203.
2. Pratama, I.N., Dachyar, M. and Pratama, N.R. (2023) 'Optimization of Resource Allocation and Task Allocation with Project Management Information Systems in Information Technology Companies', *TEM Journal*, 12(3), pp. 1814–1824. doi:10.18421/TEM123-65.
3. Meharunnisa et al. (2023) 'Analysis of Software Effort Estimation by Machine Learning Techniques', *Ingénierie des Systèmes d'Information*, 28(6), pp. 1445–1457. doi:10.18280/isi.280602.
4. Marilia Botelho Coelho et al. (2023) 'Project Management Efficiency Measurement with Data Envelopment Analysis: A Case in a Petrochemical Company', *Applied System Innovation*, 7(1), p. 2. doi:10.3390/asi7010002.
5. J. Alqahtani, A. Siddique, A. M. Aseere, A. Alasiry and Q. N. Naveed, "Evaluating Success Factors of Software Project Management in Global Software Development," in *IEEE Access*, vol. 12, pp. 22345-22358, 2024, doi: 10.1109/ACCESS.2024.3360415
6. Gamil, Y. and Abd Rahman, I. (2023) 'Investigate the Efficiency of Project Management Software in Construction Projects', *ResearchGate*. Available at: https://www.researchgate.net/publication/373529068_Investigate_the_Efficiency_of_Project_Management_Software_in_Construction_Projects (Accessed: 16 May 2025).
7. Patanakul, P. (2023) 'An Empirical Analysis on the Effectiveness and Challenges of Project Management Tools and Techniques', *Journal of General Management*. Available at: <https://journal-ems.com/index.php/emsj/article/view/1261> (Accessed: 16 May 2025).
8. Gleek (2023) 'Gantt Chart for a Research Project Proposal: Step-by-step guide', *Gleek Blog*. Available at: <https://www.gleek.io/blog/gantt-project-proposal> (Accessed: 16 May 2025).
9. Slikk.ai (2023) 'A Guide for Using Gantt Chart for Research Project', *Slikk.ai Blog*. Available at: <https://hub.slikk.ai/blog/gantt-charts/a-guide-for-using-gantt-chart-for-research-project> (Accessed: 16 May 2025).
10. Bahi, A., Ait-Kadi, D., Rezg, N. and Chelbi, A. (2024) 'Artificial Intelligence and Project Management: An Empirical Study', *IFAC-PapersOnLine*, 57(3). doi:10.1016/j.ifacol.2024.04.042.
11. López, Á.M., Pérez, P.J., García, E.P. and Álvarez, M.T. (2023) 'Analysis of Project Management Tools to Support Knowledge Management', *Procedia Computer Science*, 219, pp. 1091–1098. doi:10.1016/j.procs.2023.01.213.

Enhancing NGO Security Through Data Modeling with Artificial Intelligence and Natural Randomness

Blagoja Naumoski Guli, BSc ¹

Email(s): blagoja.naumoski@mir.uist.edu.mk

¹Information and Network Security, Privacy and Data Protection, Faculty of applied information technologies, machine intelligence and robotics, University of Information Science and Technology "St. Paul the Apostle" - Ohrid

Abstract

This paper explores a novel approach to bolstering the operational security of Non-Governmental Organizations (NGOs) by leveraging the inherent randomness of lava lamps analyzed through data modeling and artificial intelligence. Recognizing the increasing cybersecurity threats faced by NGOs, this research investigates the potential of using the unpredictable visual patterns of lava lamps as a source of high-quality entropy. By employing computer vision techniques to capture and digitize these patterns and utilizing AI algorithms for anomaly detection and entropy analysis, a unique security layer can be established from volunteers to executive decision-makers. This study outlines the scientific basis for this approach, details potential methodologies for implementation and discusses the expected benefits and challenges, suggesting a cost-effective and innovative security paradigm for the NGO sector.

Keywords

NGO Security, Data Modeling, Artificial Intelligence, Natural Randomness, Data Protection

1. Introduction

Non-Governmental Organizations (NGOs) are increasingly becoming targets of sophisticated cyberattacks due to the sensitive nature of their operations and the vast amounts of data they handle. These organizations often face unique security challenges, including geographically distributed operations, a high reliance on volunteers with varying levels of technical proficiency and budget constraints that limit their ability to implement comprehensive security infrastructures. Establishing robust security measures across all organizational tiers, from the volunteers on the ground to the executive decision-makers, is paramount to safeguarding their missions and the beneficiaries they serve. Traditional security methodologies, while essential, may not always address the full spectrum of vulnerabilities inherent in the diverse structures and operations of NGOs. This necessitates the exploration of innovative and potentially low-cost security solutions that can be effectively implemented at every level of the organization.

This research introduces a novel approach to enhancing NGO security by harnessing the intrinsic randomness of physical phenomena, drawing inspiration from Cloudflare's pioneering LavaRand project. [1] LavaRand demonstrates the feasibility of using the chaotic visual patterns of lava lamps as a source of true entropy for cryptographic purposes. This paper proposes extending this concept within the context of NGOs by employing Artificial Intelligence (AI) and data modeling to systematically capture, analyze and interpret the visual randomness of lava lamps for a broader range of security applications. This investigation seeks to answer several key questions: How can data modeling techniques effectively

capture and represent the visual patterns of a wall of lava lamps? What AI methodologies are suitable for analyzing the chaotic and unpredictable movements within lava lamps to generate security-relevant entropy? How can the entropy derived from lava lamp analysis be integrated into the operational security structure of NGOs at different levels (volunteers to executives)? What are the potential benefits and challenges of using this approach to enhance the overall security posture of NGOs?

This research holds significant potential for contributing to the field of security by exploring an unconventional and potentially cost-effective security mechanism. It could offer NGOs a novel tool to strengthen their security and resilience, particularly at the foundational levels involving all members. The ability to transform a readily available and visually chaotic system into a security asset could democratize access to enhanced security measures for organizations often operating with limited resources.

2. Related Work

During the research of this project, I had the opportunity to explore a range of fascinating research and technical documentation. Liebow-Feeser's (2017) detailed explanation of Cloudflare's LavaRand in production [1] was particularly captivating, offering a deep dive into the practical application of true randomness. Similarly, the work by Bashurov and Safonov (2023) on entropy-based anomaly detection in network traffic [2] proved to be an exciting read, showcasing innovative methods for identifying various cyberattacks. I found the Roboflow blog post on color sensing with computer vision by a Contributing Writer (2024) [3] to be incredibly insightful, demonstrating the practical and evolving capabilities within the field. Furthermore, the paper by Elabid, Chakraborty, and Hadid (2022) on knowledge-based deep learning for modeling chaotic systems [4] was exceptionally interesting, revealing advanced techniques for understanding complex phenomena. Anglen's (2024) ultimate guide to AI-powered anomaly detection [5] offered a comprehensive and exciting overview of how artificial intelligence is boosting efficiency in identifying irregularities. Zhang's (2020) new method for information security risk management in a big data environment [6] was a very engaging piece, highlighting novel approaches to a critical contemporary challenge. Finally, the work by Adeola, Ayorinde, and Adeniyi (2024) on data security in big data [7] was a thoroughly interesting exploration of the challenges, strategies, and future trends in this vital domain.

3. The Concept of Entropy and Randomness in Security

Entropy, in the context of information theory, is a fundamental measure of randomness, disorder, or uncertainty. It quantifies the average level of information, surprise and uncertainty inherent in a variable's possible outcomes. High entropy is crucial in cryptographic systems because it is essential for generating secure keys that are difficult to predict, thereby preventing unauthorized access to sensitive information. Claude Shannon's groundbreaking work [8] on information theory laid the mathematical foundation for understanding and measuring entropy, defining it as the summation over all possible outcomes of a random variable, weighted by the probability of each outcome multiplied by the logarithm of that probability.

A critical distinction exists between true randomness and pseudo-randomness. True random number generators (TRNGs) rely on unpredictable physical processes to produce random numbers, whereas pseudo-random number generators (PRNGs) use deterministic algorithms. While PRNGs are efficient, their deterministic nature makes them potentially predictable if an attacker gains knowledge of the initial seed or the algorithm itself, posing a risk in security-critical applications. Chaotic physical systems, such as lava lamps, offer a significant advantage as they can serve as a source of true randomness due to their inherent unpredictability arising from complex fluid dynamics and heat transfer.

In his groundbreaking paper, Shannon introduces the concept of entropy, H , as a fundamental measure of uncertainty or the "information" of a source. He defines it as a statistical quantity that describes the

average amount of information contained in a symbol from that source. The higher the entropy, the more unpredictable the source, a principle that is essential for understanding and generating strong, random cryptographic keys. [8]

Entropy can also function as a valuable security metric. It can be used to assess the strength of encryption keys by quantifying their unpredictability. Furthermore, entropy analysis can be applied for anomaly detection in security systems, where significant deviations in entropy levels of data or network traffic can indicate potential security breaches or malicious activities. As shown in Figure 1. Research has demonstrated the effectiveness of using entropy to detect various cyberattacks, including malware propagation and ransomware activities, by identifying patterns of high or low randomness in data streams and network behavior. [2] The fundamental principle is that systems exhibiting higher entropy are inherently more difficult to predict and exploit, thus enhancing overall security.

4. Data Modeling of Lava Lamp Visuals

The initial step in leveraging lava lamps for security involves capturing and digitizing their dynamic visual patterns. Cloudflare's implementation provides a practical example, utilizing video cameras to continuously record the motion of a wall of lava lamps. [1] Technical considerations for this process include optimizing camera resolution to capture sufficient detail, selecting an appropriate frame rate to track the movement effectively and managing lighting conditions to ensure consistent image quality. The continuous video feed is then broken down into a sequence of individual digital images or frames, each representing a snapshot of the lava lamps at a specific point in time. Notably, even the inherent random noise within the camera's sensors contributes to the overall entropy of the captured data.

Once the visual data is captured, various computer vision techniques can be employed to extract relevant features from the lava lamp images. These techniques can analyze the shape, size, position and movement of the characteristic "lava" blobs as they evolve over time. Color information and changes in color patterns can also be valuable features for analysis. These visual features are then quantified and represented as numerical data, making them suitable for subsequent analysis by AI algorithms. Motion analysis techniques, such as optical flow, can be used to track the movement of pixels or specific features within the video frames, providing insights into the chaotic dynamics of the lava lamp. [3]

To prepare the data for AI analysis, the extracted features and their temporal sequences need to be organized into a structured dataset. This involves careful consideration of how to represent the dynamic nature of the lava lamp patterns. Data normalization and preprocessing steps are crucial to ensure data quality and compatibility with the chosen AI algorithms. [4] Time series data modeling techniques can be particularly effective in capturing the evolution of the lava lamp's visual behavior over time, allowing AI models to learn and predict its normal chaotic patterns. As shown in Figure 1. Effective data modeling is therefore a foundational step in transforming the raw visual information into a format that can yield security-relevant insights through AI analysis.

5. Artificial Intelligence for Lava Lamp Security Analysis

Artificial intelligence plays a pivotal role in analyzing the complex and dynamic visual data obtained from lava lamps for security purposes. Computer vision techniques, particularly Convolutional Neural Networks (CNNs), are well-suited for recognizing intricate visual patterns within the lava lamp images. CNNs can be trained to learn the typical visual characteristics of the "lava" flow and identify any deviations from these normal patterns. [5] Furthermore, Recurrent Neural Networks (RNNs) or their more advanced variant, Long Short-Term Memory (LSTM) networks, can be employed to analyze the temporal sequences of these visual patterns, enabling the AI to understand the evolution of the chaotic system over time.

Anomaly detection is a key application of AI in this context. By training AI models on a baseline of "normal" chaotic behavior of the lava lamps, the system can identify deviations that might indicate a security event or an attempt to manipulate the entropy source. Various anomaly detection techniques, such as autoencoders that learn to reconstruct normal data and flag deviations, isolation forests that isolate anomalies based on their rarity and statistical methods that identify outliers, can be adapted for this purpose. Research into the analysis of chaotic systems using AI provides a strong foundation for applying these techniques to the unpredictable behavior of lava lamps. As shown in Figure 1.

Beyond pattern recognition and anomaly detection, AI can also be used for entropy measurement and analysis. The visual data from the lava lamps can be processed to calculate entropy, providing a quantitative measure of its randomness. AI models can be trained to continuously monitor and analyze these entropy levels, detecting significant changes or anomalies that might correlate with security events. For instance, a sudden drop in entropy might suggest a manipulation of the lava lamps or the capture process. Machine learning models, including both supervised and unsupervised algorithms, can be employed to identify complex patterns in the lava lamp data that could be indicative of security breaches or operational anomalies. Training these models on a baseline of normal lava lamp behavior allows for the detection of deviations that warrant investigation. The potential for using transfer learning, where models trained on other chaotic systems or visual datasets are adapted for lava lamp analysis, could also enhance the efficiency and accuracy of this approach.

6. Integrating Lava Lamp Analysis into NGO Security Operations

The unique randomness derived from lava lamp analysis, when coupled with AI and data modeling, presents various potential applications for enhancing the operational security of NGOs across different organizational levels. At the base level, for volunteers, the unpredictable nature of lava lamp patterns could serve as a foundation for generating strong and unique passwords or one-time codes for accessing sensitive systems. Imagine a user interface that translates the current visual state of the lava lamps into a complex and ephemeral password, potentially through a gamified approach to ensure ease of use.

At the mid-level, for operational staff, the continuous monitoring of entropy levels derived from the lava lamp setup could act as an indicator of the overall security health of the NGO's physical location. Significant decreases in entropy or the detection of unusual visual patterns by AI algorithms could trigger alerts for security personnel, indicating potential physical tampering or unauthorized access attempts to the lava lamp setup itself.

TechSoup Global, in its mission to empower non-governmental organizations (NGOs) with technology, has consistently advocated for a multi-layered and accessible approach to security. For many NGOs, especially those with limited budgets and technical expertise, the challenge isn't just about procuring expensive software but also about implementing sustainable and practical security measures. As shown in Figure 3. The concept of using a physical, low-cost entropy source like a lava lamp aligns with this philosophy by providing a tangible and visually intuitive security tool. This approach could be integrated into a larger framework of best practices that includes discounted cybersecurity software, staff training and robust access controls all of which TechSoup and its partners champion. By focusing on a technology that is both affordable and inherently random, NGOs can enhance their security posture without diverting critical funds from their core mission, thereby making a complex topic like cryptographic security more manageable and democratized for the sector. [9]

For top executive decision-makers, the aggregated entropy data and the results of AI-driven anomaly detection can provide a high-level overview of the organization's security posture. This information can

offer valuable insights into potential risks and the effectiveness of existing security protocols, informing strategic security decisions and the allocation of resources.

Furthermore, the lava lamp-based entropy generation and analysis can enhance existing security frameworks by providing an additional layer of security that leverages a physically random and potentially tamper-evident source of entropy. This approach might be particularly beneficial for NGOs operating in resource-constrained environments where deploying sophisticated security infrastructure is challenging. The system could also be used to enhance the security of communication channels by providing a source of randomness for encryption keys used in internal communications. [6] The modeled visual data could even be directly translated into unique keys or seeds for encryption algorithms. Finally, subtle, long-term changes in the lava lamp's chaotic behavior, detected by AI, could potentially indicate insider threats or compromised systems attempting to influence the entropy source, providing an additional layer of security monitoring.

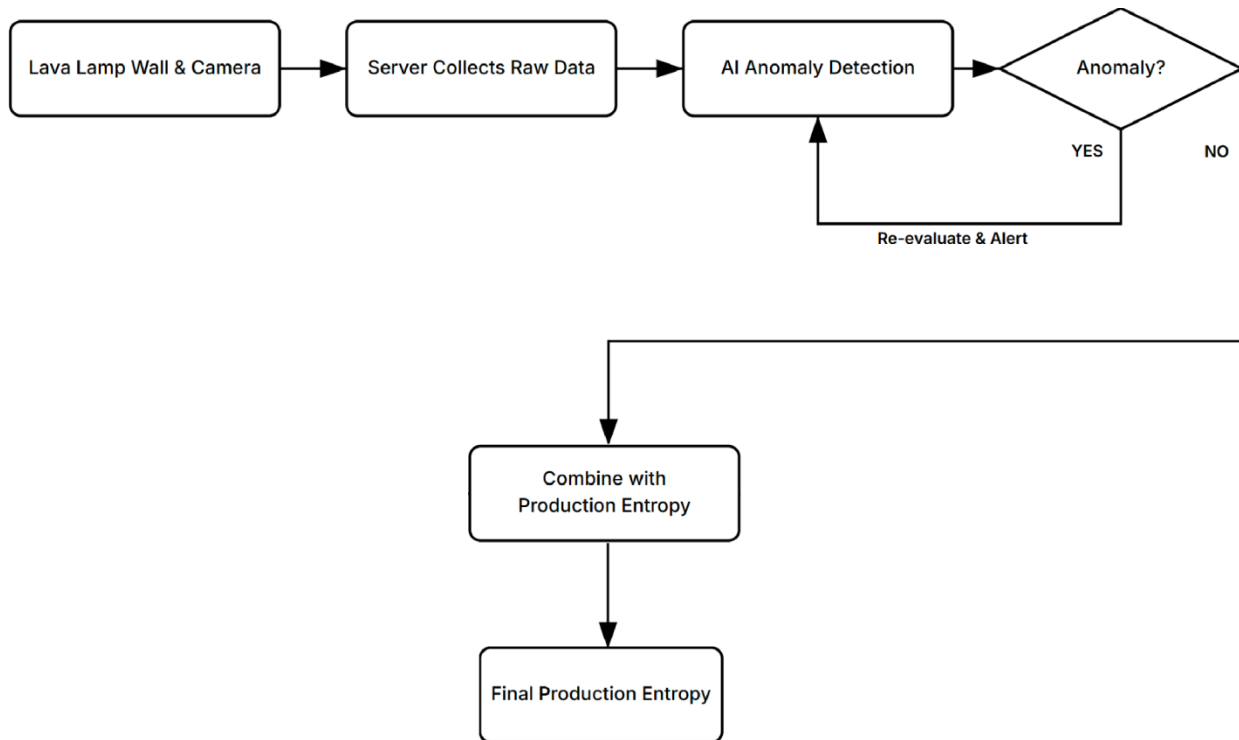


Figure no. 1: Flow of entropy generation and validation using lava lamp visuals and AI-based anomaly detection.

Figure 1 depicts the comprehensive process of leveraging lava lamp visuals for enhanced NGO security. Starting in the office lobby, a camera captures the visual and photoreceptor noise emanating from a wall of lava lamps. This raw, physically-derived randomness is then transmitted to the office server, where it might be combined with the server's existing local entropy sources to generate a more robust entropy feed. Subsequently, this new entropy stream undergoes analysis by an AI anomaly detection system to validate its randomness and identify any unusual patterns that might suggest tampering or environmental interference. If the AI detects no anomalies, the validated entropy proceeds to the production data center. In the event of an anomaly detection, the system initiates a re-evaluation process to address the potential issue. Once deemed valid, the entropy is integrated with the production data center's local entropy system, resulting in a final, refined entropy feed. This final output is then available for consumption by various

production services within the NGO, providing a unique and physically-backed source of randomness for security applications as described in the text.

7. Expected Results and Evaluation Metrics

The implementation of a lava lamp-based security analysis system within an NGO is expected to yield several security enhancements. Encryption keys generated using the high-quality entropy from lava lamps are anticipated to be significantly stronger and more unpredictable compared to those derived solely from pseudo-random number generators. AI-driven analysis of the lava lamp's visual patterns has the potential to improve the detection rates of both cybersecurity anomalies and physical tampering attempts with the setup itself. This approach aims to reduce the reliance on potentially vulnerable algorithmic randomness in critical security processes. [7]

The system generates its IDs using an internal pseudo-random number generator. While this mathematical algorithm produces results that appear random with each execution, it's a calculated appearance; as the distinguished mathematician John von Neumann famously stated, "Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin," pointing to the fact that true randomness cannot be born from purely arithmetic means. [13]

For example, if it is chosen 4181 as a seed the sequence 4807,1072, 1491, 2230, 9279, ... would be generated as follows		
Random number	It's Square	Middle digits
4181	17480761	4807
4807	23107249	1072
1072	1149184	1491
1491	2223081	2230
2230	4972900	9729
9279	94653441	6534

Figure no. 2: Pseudo-random number generator using the middle-square method. [7]

Figure 2 demonstrates a classic example of a pseudo-random number generator using the middle-square method. This algorithm, while simple, illustrates how a seemingly random sequence can be created from a deterministic mathematical process. As the figure shows, a four-digit number, or seed, is squared to produce a larger number. The central digits of this new number are then extracted to form the next number in the sequence. This process is repeated, with each new number becoming the seed for the next iteration. This method highlights the fundamental limitation of such algorithms: the output is entirely dependent on the initial seed, and if that seed is known, the entire sequence can be predicted, proving that these numbers lack true entropy and are therefore not suitable for high-security cryptographic applications. [7]

To see how well this system works, we can use several important ways to measure its success. We can check if the randomness from the lava lamps is good and unpredictable by using special tests, like the

NIST tests. We can also see how accurately the AI can spot fake security problems that we create by watching how the lava lamps change, we'll use common measures like precision, recall and F1-score for this. Another way to check is to compare the strength of the security codes made with the lava lamp method to those made without it, using standard security testing methods. We also need to find out if this system is easy for NGOs to use by trying it out in small projects and getting feedback from volunteers, staff and leaders. Finally, we should compare what it costs to set up and keep the system running with how much it helps improve security and lower risks.

Metric	Description	Measurement Method	Target Performance Indicator	Relevance to NGO Security
Entropy Quality	Measure of the unpredictability and randomness of the generated entropy.	NIST Statistical Test Suite, Diehard tests.	Pass rate above established thresholds for randomness.	Ensures the entropy source is suitable for cryptographic applications.
Anomaly Detection Performance	Accuracy of AI models in identifying simulated security events.	Precision, Recall, F1-score on a labeled dataset of normal and atypical patterns.	Statistically significant improvement over current NGO methods.	Indicates the system's ability to detect actual security threats.
Key Strength	Robustness of encryption keys generated using lava lamp entropy.	Standard cryptographic analysis techniques (e.g., key space size, resistance to known attacks).	Keys meet or exceed industry best practices for strength.	Ensures that cryptographic operations within the NGO are adequately protected.
User Adoption	Ease of use and acceptance of the system by NGO personnel.	Surveys, interviews and feedback sessions with users at different levels.	High levels of user satisfaction and willingness to use the system.	Demonstrates the practical feasibility and integration of the system operations.
Cost-Benefit	Comparison of implementation and maintenance costs with security benefits.	Cost-benefit analysis of risk mitigation.	Benefits outweigh costs, demonstrating a positive return on investment.	Justifies the allocation of resources towards this novel security approach.

Figure no. 3: Expected results and evaluation metrics on how the metrics will reflect on an NGO

Figure 3 presents a framework for evaluating the effectiveness of a lava lamp-based security system using five key metrics: Entropy Quality, Anomaly Detection Performance, Key Strength, User Adoption, and Cost-Benefit. The table details each metric's description, the methods for measuring it (e.g., NIST tests, precision/recall, key space analysis), and the specific performance indicators to be met. It also outlines the direct relevance of each metric to an NGO's security, from ensuring cryptographic suitability to justifying resource allocation. This table provides a clear roadmap for assessing the system's success across technical, operational, and financial dimensions.

8. Discussion and Limitations

The proposed approach of using data modeling and AI to analyze lava lamp visuals for NGO security presents several compelling strengths. It leverages a truly random physical source of entropy, potentially offering a higher degree of unpredictability compared to purely algorithmic methods. The implementation could be relatively cost-effective, particularly for NGOs operating with limited financial resources. The novelty of this security measure might also act as a deterrent to attackers who may not anticipate such an unconventional approach. Furthermore, the visual and engaging nature of lava lamps could foster greater security awareness and engagement across the entire organization.

The sensitivity of physical entropy sources to their surroundings is a well-documented phenomenon in cybersecurity literature. For instance, research by Skliarevych, Semenyaka, and Melnyk on using environmental sensors for pseudorandom number generation highlights that even seemingly stable conditions, such as temperature, can have a random component at the micro-level due to internal and external factors. [10] This demonstrates that fluctuations in ambient temperature and lighting are not just incidental but are scientifically proven to be chaotic elements that can influence the randomness of a physical system. Therefore, the delicate fluid dynamics of the lava lamp are inherently susceptible to these environmental variables, underscoring the critical need for a controlled environment and sophisticated post-processing to ensure the integrity of the generated entropy.

Despite its innovative approach, several limitations and challenges must be carefully considered. Environmental factors, such as subtle variations in ambient lighting or temperature fluctuations, could significantly influence the delicate fluid dynamics within the lava lamps, consequently affecting the quality and predictability of the entropy generation. To make sense of this, robust data modeling techniques and sophisticated AI algorithms are absolutely essential, not only to capture but also to accurately interpret the inherently chaotic behavior of the system, thereby minimizing the risk of misinterpreting benign patterns. The physical nature of the setup also introduces a tangible vulnerability: the potential for tampering with the lava lamps themselves or the video capture equipment, which could directly compromise the integrity of the random data or the security monitoring. The physical nature of the setup also introduces a tangible vulnerability: the potential for tampering with the lava lamps themselves or the video capture equipment, which could directly compromise the integrity of the random data or the security monitoring. This is a well-known vulnerability in the field of hardware security, where physical attacks can compromise even dedicated random number generators by manipulating their operating conditions or injecting faults to influence the output. [11] Furthermore, scalability might pose a considerable challenge, especially for very large or geographically dispersed NGOs that would require multiple independent setups, each adding to operational complexity and maintenance overhead. [10] The system's fundamental reliance on the continuous and flawless operation of its video capture and processing infrastructure is another critical factor, as any interruption could lead to significant data gaps. Finally, the inherent chaotic nature of lava lamps, while beneficial for randomness, might make it exceedingly challenging to definitively distinguish between normal, expected fluctuations and indicators of actual security events, potentially leading to a higher incidence of false positives that would then require careful, resource-intensive tuning and thorough analysis to resolve. [12]

When compared to existing security methods, this approach can be seen as a complementary layer rather than a replacement for traditional solutions. While firewalls and intrusion detection systems focus on network-level security and access controls manage user permissions, the lava lamp system offers a unique, physically-rooted source of randomness for cryptographic functions and a visual anomaly detection mechanism. Compared to other entropy sources like dedicated hardware random number generators or environmental noise, lava lamps offer a visually verifiable and potentially more engaging method of entropy generation.

Future research should focus on optimizing the configuration of a lava lamp-based entropy generation system, including the number of lamps, camera placement and lighting conditions. Developing more advanced AI algorithms specifically tailored for analyzing chaotic visual data for security applications is also crucial. Investigating the potential for combining lava lamp entropy with other unconventional entropy sources, such as chaotic pendulums or radioactive decay, could further enhance the robustness of the system. Conducting pilot studies with NGOs to evaluate the real-world feasibility, effectiveness and user acceptance of this approach is a necessary next step. Finally, exploring the impact of different types of lava lamps and their specific compositions on the quality and quantity of entropy generated could provide valuable insights for optimizing the system's design.

9. Conclusion

This study strongly indicates that employing data modeling and sophisticated AI to analyze the dynamic, ever-changing patterns within lava lamps could offer a remarkably effective and novel approach for Non-Governmental Organizations to significantly improve their security posture. By creatively leveraging the inherent, natural unpredictability of these devices, a characteristic often overlooked, NGOs might be able to establish a new and notably affordable method for securing their diverse operations at every level, which is particularly valuable given the resource constraints many such organizations face. Science indeed supports the premise that lava lamps can produce truly random data, a critical element for robust cryptographic applications, and when this is combined with the analytical power of advanced computer vision and AI meticulously trained to spot unusual or anomalous activity, this system could evolve into a uniquely strong and resilient security method. While there are acknowledged hurdles to overcome, such as mitigating how environmental factors might subtly affect the lamps' performance, addressing how effectively this system scales for larger or geographically dispersed organizations, and the continuous need for refining the strong AI algorithms, the substantial potential security benefits and the enhanced ability to detect and potentially recover more swiftly from sophisticated attacks make this unconventional idea thoroughly worth exploring in greater depth. It remains critically important to proactively continue researching and developing innovative solutions like this one, thereby equipping organizations, especially vulnerable NGOs, with better tools to navigate and counteract the ever-changing, complex landscape of global cyber threats and ultimately bolster their capacity to protect their sensitive information and vital missions.

10. References

- [1] Liebow-Feeser, J. (2017) [Online]. LavaRand in Production: The Nitty-Gritty Technical Details. Available at: <https://blog.cloudflare.com/lavarand-in-production-the-nitty-gritty-technical-details> (Accessed: 3 May 2025)
- [2] Bashurov, V., & Safonov, P. (2023). Anomaly detection in network traffic using entropy-based methods: Application to various types of cyberattacks. *Issues in Information Systems*, 24(4), 82-94. https://doi.org/10.48009/4_iis_2023_107
- [3] Contributing Writer. (Sep 4, 2024). Color Sensing with Computer Vision. Roboflow Blog: <https://blog.roboflow.com/color-sensing-with-computer-vision/>
- [4] Elabid, Z., Chakraborty, T. and Hadid, A. (2022) 'Knowledge-based deep learning for modeling chaotic systems', 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1203–1209. doi:10.1109/icmla55696.2022.00194.
- [5] Anglen, J. (2024) AI-powered anomaly detection 2024 ultimate guide: Boost efficiency, RSS. Available at: <https://www.rapidinnovation.io/post/ai-in-anomaly-detection-for-businesses> (Accessed: 3 May 2025).

- [6] Z. Zhang, "A New Method for information security risk management in big data environment," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 2020, pp. 1-4, doi: 10.1109/ITCA52113.2020.00100.
- [7] Adeola, A. Ayorinde, O. and Adeniyi, A. (2024) Data Security in Big Data: Challenges, Strategies, and Future Trends. Volume 5. US: International Journal of Research in Education Humanities and Commerce.
- [8] Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x
- [9] TechSoup. (Jan 8, 2024). 7 Essential Tech Solutions for Small Nonprofits. Available at: <https://www.techsoup.global/7-essential-tech-solutions-small-nonprofits> (Accessed: 7 August 2025).
- [10] Skliarevych, V., Semenyaka, A., & Melnyk, K. (2024). Enhancing Pseudorandom Number Generation using Environmental Sensor-based Entropy Sources. CEUR Workshop Proceedings, 3991, 312–322. [CEUR-WS.org/Vol-3991](https://www.ceur-ws.org/Vol-3991).
- [11] Vermeulen, S. (2018). Attacking and Securing Hardware Random Number Generators (Master's Thesis Proposal). KU Leuven, COSIC.
- [12] Olateju, O., Okon, S. U., Igwenagu, U., Salami, A. A., Oladoyinbo, T. O., & Olaniyi, O. O. (2024). Combating the challenges of false positives in AI-driven anomaly detection systems and enhancing data security in the cloud. Available at SSRN 4859958.
- [13] Knuth, D. E. (1998). The Art of Computer Programming, Volume 2: Seminumerical Algorithms. Third edition. Addison-Wesley. ISBN: 978-0-201-89684-8.

The Intersection of AI and Sustainability: Self-Optimizing AI Systems for Circular Manufacturing

Darko Poposki 1 [0009-0003-9292-3354]

Email(s): poposki.darko@gmail.com

University of Information Science and Technology St. Paul the Apostle, Ohrid, Macedonia

Abstract. As Industry 5.0 emphasizes sustainability, resilience, and human-centricity, manufacturing must evolve from static optimization to adaptive, sustainability-aware intelligence. This paper proposes a framework for self-optimizing AI systems in circular manufacturing, integrating continuous feedback loops, multi-objective optimization, and human-in-the-loop governance. The model formalizes four components—sensing and monitoring, feedback evaluation, adaptive optimization, and learning governance—within a layered architecture that embeds sustainability key performance indicators such as waste intensity, recyclability, and carbon footprint directly into decision-making. Validation through scenario mapping and structured expert feedback highlights both the feasibility and the design trade-offs of embedding AI into circular systems. By advancing AI from task-level efficiency toward systemic sustainability intelligence, this work provides a foundation for future empirical studies and contributes to the emerging field of regenerative industrial innovation.

Keywords: Circular Manufacturing, Industry 5.0, Self-Optimizing AI, Sustainability, Adaptive Optimization, Human-in-the-Loop, Multi-Objective Decision-Making, Explainable AI

1 Introduction

The manufacturing sector is undergoing a profound transformation driven by the convergence of advanced digital technologies and rising societal demands for sustainability. Whereas earlier industrial revolutions prioritized scale, speed, and automation, the emerging Industry 5.0 paradigm reframes industrial progress around ecological responsibility, resilience, and human-centric design [1]. Within this paradigm, circular manufacturing has gained prominence as a strategy for minimizing waste and maximizing resource recirculation through closed-loop production systems.

Artificial Intelligence (AI) is widely recognized as a key enabler of this transition. Its capacity to process large datasets, detect patterns, and adapt decisions makes it suitable for optimizing complex manufacturing processes. Yet, most current AI deployments remain task-specific, static, or narrowly focused on cost and efficiency. They rarely account for evolving sustainability objectives, variable material flows, or long-term ecological constraints. This disconnect reveals a critical gap: while AI in manufacturing and circular economy principles are both advancing, they have not been systematically integrated into adaptive, sustainability-aware decision systems.

This paper addresses that gap by introducing a framework for self-optimizing AI systems in circular manufacturing. Unlike fixed rule-based optimizations, self-optimizing systems embed feedback loops, multi-objective learning, and governance mechanisms to continuously refine their behavior in response to sustainability key performance indicators (KPIs) such as waste intensity, recyclability, and carbon footprint.

The contributions of this work are threefold. First, it defines a conceptual framework that integrates adaptive optimization with circular economy principles. Second, it introduces a layered architecture that operationalizes these concepts through sensing, feedback evaluation, learning, and governance. Third, it

provides preliminary validation through scenario mapping and structured expert feedback, highlighting both feasibility and implementation challenges.

By situating AI as a mechanism for systemic sustainability intelligence rather than narrow efficiency gains, this study contributes to the emerging discourse on Industry 5.0. The following sections outline the theoretical foundations, framework design, system architecture, and preliminary validation of this approach.

2 Background and Related Work

Manufacturers face increasing pressure to align productivity with sustainability, prompting growing adoption of circular economy principles—systems designed to minimize waste and maximize resource reuse through closed-loop processes. Industry 5.0 reinforces this transformation by emphasizing ecological responsibility, human-centricity, and resilience alongside technological innovation [1]. However, current industrial systems remain ill-equipped to adapt dynamically to sustainability objectives, creating opportunities for AI to act as a catalyst for circularity in manufacturing.

Digital technologies such as IoT, blockchain, and cloud platforms have improved traceability and efficiency, yet their role is often limited to static process monitoring and reactive control [2]. These systems enable data collection but lack adaptive intelligence: they cannot evaluate sustainability trade-offs, integrate lifecycle data, or autonomously reconfigure processes. AI has been applied in manufacturing for predictive maintenance, scheduling, and process optimization, but typically with narrow objectives such as throughput or energy cost reduction [3]. Broader sustainability goals—waste minimization, emissions reduction, material recirculation—are rarely encoded into these systems. Recent research has introduced adaptive AI concepts, including reinforcement learning for dynamic production planning and digital twins for predictive modeling [4], but these approaches often remain focused on efficiency rather than sustainability. Moreover, most operate in isolation from circular economy strategies such as product reuse, regenerative materials, or closed-loop supply chains. Governance and explainability, while increasingly emphasized in responsible AI literature, are seldom addressed in manufacturing-specific contexts where trade-offs between ecological and operational objectives are complex.

A growing body of work calls for architectures that combine AI-driven adaptivity with sustainability principles [5], yet existing studies tend to remain either highly conceptual or narrowly technical. For instance, some propose self-adaptive decision-making for smart factories, but without integrating sustainability key performance indicators (KPIs) or human-in-the-loop oversight. Others emphasize environmental monitoring but stop short of continuous system re-optimization.

2.1 Positioning of this Work.

The contribution of this paper lies in bridging these gaps by:

- Integrating adaptive AI methods with explicit sustainability KPIs such as recyclability, waste intensity, and carbon footprint;
- Embedding governance and explainability layers to ensure alignment with Industry 5.0 values;
- Proposing a layered architecture that supports not only operational efficiency but also circular value creation through continuous feedback-driven learning.

This integration differentiates the proposed framework from prior work, shifting AI's role in manufacturing from static optimization toward systemic sustainability intelligence.

Table I. Comparison of Existing Approaches vs. Proposed Framework

Dimension	IoT / Blockchain Systems [2]	Adaptive AI in Smart Factories [3][4]	Responsible AI & Governance [5]	Proposed Framework
Primary Focus	Traceability, efficiency	Dynamic optimization (cost, throughput)	Transparency, accountability	Sustainability-aware, adaptive optimization
Sustainability KPIs	Not integrated	Rarely included, mostly energy	Not domain-specific	Explicitly modeled (waste, recyclability, carbon)
Adaptivity / Learning	Reactive monitoring only	Reinforcement learning, digital twins	Not applicable	Continuous self-optimization with feedback loops
Integration with Circularity	Weak, data-focused	Minimal, process-level only	Absent	Closed-loop alignment with circular economy principles
Human-in-the-loop Governance	Absent	Rarely addressed	Strong focus	Embedded in architecture
Novelty	Data transparency	Operational adaptivity	Ethical oversight	Combines adaptivity + sustainability KPIs + governance

3 Research Methodology

This study adopts a Design Science Research (DSR) approach to conceptualize and evaluate a framework for self-optimizing AI systems in circular manufacturing. DSR is particularly suitable for research that seeks to create and iteratively refine artifacts—such as models or architectures—that address complex, real-world problems while contributing to both academic theory and applied knowledge [5]. In this context, the artifact is a systems-oriented framework that integrates adaptive AI mechanisms with circular manufacturing principles under the values of Industry 5.0.

The methodology followed four stages:

1. Problem identification through an extensive review of literature on AI in manufacturing, sustainable production systems, and circular economy models. Industry reports and policy documents were also examined to capture emerging demands for adaptive, sustainability-oriented digital systems.
2. Artifact design, in which the conceptual framework was developed based on theoretical constructs from adaptive AI, cyber-physical systems, and sustainability engineering. Key characteristics included continuous sensing of operational and environmental data, feedback integration,

- alignment with sustainability key performance indicators (KPIs), and autonomous adaptation through embedded learning mechanisms.
3. System architecture development, translating the framework into a high-level layered design that supports sensing, contextualization, optimization, human interaction, and governance. The architecture emphasizes modularity, explainability, and interoperability.
 4. Preliminary validation, combining two methods:
 - Scenario-based mapping, where the framework was applied to representative manufacturing contexts (e.g., material recirculation under quality variability; process optimization under emissions constraints). These scenarios allowed assessment of feasibility and applicability under realistic conditions.
 - Practitioner feedback, collected through annotated slide presentations and structured follow-up interviews with professionals in industrial automation, machine learning, and circular economy strategy. Participants were asked about the clarity, feasibility, and perceived value of the proposed approach, as well as potential barriers to implementation. Their insights informed refinements such as stronger support for transparency in decision loops, improved interoperability with existing systems, and enhanced mechanisms for human oversight.

While the study does not yet include empirical prototyping, the use of DSR ensures methodological rigor in bridging conceptual design and practical relevance. The outcome is a framework and architecture that are technically grounded, aligned with Industry 5.0 principles, and positioned for future field validation.

4 Theoretical Foundations

The proposed model for self-optimizing AI systems in circular manufacturing builds on three complementary domains: circular economy theory, adaptive AI systems, and Industry 5.0 principles. Together, these provide the conceptual grounding for designing intelligent systems that are not only operationally efficient but also ecologically responsible and ethically aligned.

4.1 Circular Economy

The circular economy provides a systemic approach to sustainable production by emphasizing closed material loops—reuse, refurbishment, recycling, and remanufacturing—as mechanisms for extending product lifecycles and reducing resource extraction [6]. While circularity has been widely studied, its implementation in manufacturing remains constrained by static decision rules and limited real-time adaptability. Achieving effective circularity requires systems that can sense environmental feedback, respond to variability, and coordinate across lifecycle stages. This sets the foundation for embedding adaptive intelligence into manufacturing processes.

4.2 Adaptive AI Systems

Adaptive AI extends beyond rule-based automation by enabling systems to modify behavior in response to feedback from changing conditions. Techniques such as reinforcement learning support iterative policy refinement, rewarding actions that improve defined objectives. Digital twin technologies provide simulation environments for testing scenarios without disrupting operations [7]. Yet, most current applications in manufacturing focus narrowly on efficiency (e.g., throughput or energy) rather than sustainability. Moreover, challenges such as delayed feedback, multi-objective trade-offs, and noisy data remain underexplored. Embedding adaptive AI in circular contexts requires formalizing sustainability KPIs—such as waste intensity, recyclability, and emissions—as explicit optimization targets, moving beyond efficiency-centric objectives.

4.3 Industry 5.0 Principles

Industry 5.0 reframes industrial transformation by emphasizing human-centricity, resilience, and sustainability alongside productivity [1]. This perspective requires AI systems that are not only autonomous but also explainable, transparent, and value-aligned. Human-in-the-loop interaction, governance mechanisms, and compliance monitoring are therefore central design requirements. Prior literature on responsible AI underscores the importance of oversight and interpretability, but these principles are rarely embedded within manufacturing-specific architectures. Integrating such safeguards is essential for ensuring that adaptive optimization serves ecological and societal goals rather than creating unintended risks.

4.4 Synthesis

Taken together, these foundations suggest that effective self-optimizing systems must operate across technical, ecological, and governance dimensions. They must: (1) sense and interpret sustainability-relevant data, (2) adapt decisions using continuous learning mechanisms, and (3) remain accountable through explainability and human oversight. This convergence of circular economy logic, adaptive AI methods, and Industry 5.0 values provides the theoretical basis for the framework and architecture introduced in the following sections.

5 The Framework for Self-Optimizing AI Systems

To address the adaptability and sustainability challenges of circular manufacturing, this paper proposes a conceptual framework for self-optimizing AI systems. These systems autonomously monitor, evaluate, and improve their performance over time, with explicit focus on environmental impact, material reuse, and energy efficiency. Unlike conventional optimization models that target fixed outcomes, self-optimizing systems integrate continuous feedback loops that allow decisions to evolve in response to real-time conditions and long-term sustainability objectives.

The framework consists of four interdependent components—Sensing and Monitoring, Feedback Evaluation, Adaptive Optimization, and Learning and Governance—that interact in a closed loop. This loop operationalizes both ecological intelligence and continuous improvement.

5.1 Sensing and Monitoring

This component captures real-time data from the production environment. Sensors and digital tools measure material flows, energy consumption, emissions, waste levels, machine status, and product lifecycle parameters. These data streams provide visibility into sustainability-relevant indicators such as waste intensity (kg waste/unit), recyclability rates, and carbon footprint per process stage [8].

5.2 Feedback Evaluation

Raw data is transformed into actionable insights by benchmarking against sustainability baselines and targets. This component diagnoses deviations (e.g., unexpected energy spikes), detects inefficiencies, and identifies intervention opportunities. Feedback operates at multiple time scales, from short-term batch metrics to long-term lifecycle performance (e.g., quarterly carbon footprint reduction), enabling comprehensive system evaluation.

5.3 Adaptive Optimization

This component adjusts operational parameters based on evaluated feedback. Actions may include modifying process temperatures, rescheduling production, changing material mixes, or re-routing workflows. Unlike static optimization, which locks into a single optimum, adaptive optimization supports ongoing re-optimization under evolving conditions [9]. Sustainability KPIs serve as explicit objectives in the optimization process, ensuring that improvements extend beyond cost and throughput.

To illustrate the adaptive optimization process more formally, the problem can be expressed as a multi-objective optimization task:

$$\min_{a \in A} \{ C(a), E(a), W(a) \}$$

subject to:

$$G(a) \leq \theta, a \in A$$

where:

- $C(a)$ represents production cost,
- $E(a)$ denotes energy consumption,
- $W(a)$ captures waste or emissions,
- $G(a)$ encodes governance or regulatory constraints,
- θ is the acceptable threshold for compliance, and
- A is the set of feasible system actions.

This formulation reflects that the system does not optimize for a single criterion but seeks to balance multiple, often conflicting objectives in line with sustainability KPIs. Reinforcement learning or other adaptive methods can be used to approximate Pareto-optimal solutions over time [9,13].

5.4 Learning and Governance

The final component ensures that the system not only adapts but also learns and remains accountable. Reinforcement learning updates decision policies over time, incorporating human feedback to guide complex trade-offs. Explainability mechanisms make AI actions transparent, while governance structures encode ethical, regulatory, and environmental constraints [10]. This layer ensures that optimization aligns with Industry 5.0 principles of accountability, transparency, and human oversight.

5.5 Positioning

Traditional rule-based automation relies on predefined thresholds and static instructions, making it inflexible under changing conditions. In contrast, the proposed framework embeds continuous re-evaluation and dynamic learning, enabling systems to evolve in response to operational variability and ecological feedback. This represents a shift from automation as compliance to automation as adaptive sustainability intelligence.

Table 2. Core Capabilities of Self-Optimizing AI Systems and Their Sustainability Enablers

System Function	Core Capability	Sustainability Enabler
-----------------	-----------------	------------------------

Sensing and Monitoring	Contextual Awareness	Resource usage tracking, waste/emission detection
Feedback Evaluation	Sustainability Diagnosis	Deviation analysis, multi-scale performance review
Adaptive Optimization	Dynamic Reconfiguration	Real-time process tuning, eco-efficiency balancing
Learning and Governance	Continuous Improvement	Policy refinement, explainability, human alignment

6 High-level Architecture

To implement the proposed framework, a high-level architecture was designed to support continuous learning, real-time adaptation, and alignment with sustainability objectives. The architecture adopts a layered design in which each layer fulfills specific operational, analytical, and governance functions, while collectively enabling feedback-driven optimization under explicit sustainability constraints. The architecture consists of five interconnected layers: Sensing Infrastructure, Data and Contextualization, AI Optimization Engine, Human–System Interaction, and Governance and Compliance.

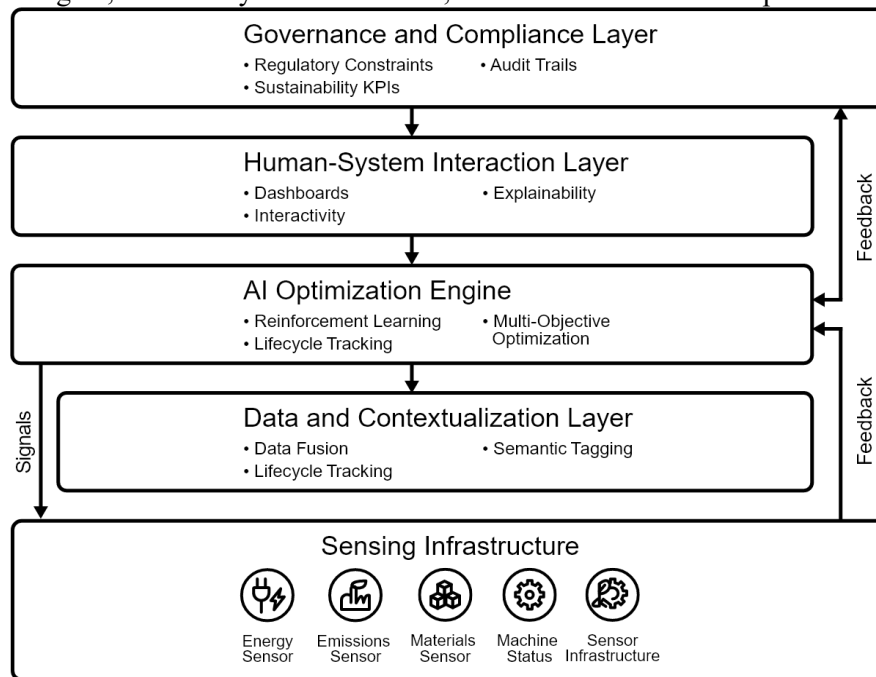


Fig. 1. High-level architecture for implementing self-optimizing AI systems, illustrating the layered integration of sensing, analytics, optimization, human interaction, and governance.

6.1 Sensing Infrastructure

This foundational layer connects physical assets to digital systems by collecting real-time data from production equipment, environmental sensors, material flow trackers, and energy meters. It measures operational and ecological indicators—such as energy use per unit, emissions per batch, and material recirculation rates—that provide the raw inputs for sustainability-aware decision-making [11].

6.2 Data and Contextualization Layer

Here, raw sensor data is fused, enriched, and organized into structured formats usable by AI models. This includes semantic labeling of resources, mapping product lifecycles, and contextualizing metrics across time horizons. The layer ensures that decisions are informed not only by current conditions but also by historical trends and environmental context [12]. By embedding sustainability KPIs (e.g., waste intensity, recyclability, carbon footprint) as structured variables, this layer makes ecological performance an explicit input to the optimization process.

6.3 AI Optimization Engine

At the core of the architecture, the optimization engine applies reinforcement learning and multi-objective decision-making techniques to generate adaptive control strategies. It balances operational goals (e.g., throughput, cost) with ecological goals (e.g., emissions reduction, resource reuse), using sustainability KPIs as explicit reward functions [13]. The engine enables dynamic reconfiguration, predictive rerouting, and proactive adjustment of production parameters.

6.4 Human–System Interaction Layer

This layer ensures transparency and accountability by providing dashboards, alerts, explainability interfaces, and override mechanisms. Human operators and sustainability managers can interpret, validate, or adjust AI-driven recommendations [14]. This preserves human oversight in ambiguous or high-impact situations, ensuring alignment with Industry 5.0's principle of human-centricity.

6.5 Governance and Compliance Layer

At the top level, governance mechanisms encode regulatory limits, ethical standards, and strategic directives. This includes emissions caps, recycling targets, and compliance with sustainability standards. The layer also provides audit trails for decisions, ensuring traceability and accountability [15]. In this way, the system does not merely optimize performance, but does so within boundaries defined by ecological and societal values.

6.6 Positioning

The novelty of this architecture lies in its integration of sustainability KPIs, adaptive learning, and governance within a single layered system. Whereas prior architectures for smart manufacturing emphasize efficiency or traceability, this design explicitly aligns optimization with circular economy objectives and Industry 5.0 values. The modular design allows organizations to adopt the system incrementally, beginning with sensing and analytics and extending toward full adaptive optimization and governance alignment.

7 Validation and Preliminary Findings

Given the conceptual nature of the proposed framework and architecture, validation at this stage focuses on early-stage relevance, feasibility, and practical alignment, rather than empirical deployment [16]. Three complementary methods were used: (1) scenario-based mapping in representative manufacturing contexts, (2) structured practitioner feedback, and (3) comparative analysis with recent case studies from the literature.

7.1 Scenario-Based Mapping.

The framework was applied to two hypothetical but industry-grounded scenarios.

Electronics remanufacturing. A mid-sized electronics producer sought to reuse post-consumer components with high variability in quality. The framework suggested dynamic adjustment of cleaning cycles, remanufacturing routes, and scheduling based on real-time input quality and sustainability thresholds. This highlighted the role of adaptive optimization in managing uncertainty.

Food packaging with regenerative materials. A packaging company explored recyclable designs to reduce plastic waste. Using the architecture, the system would monitor degradation rates, recycling returns, and regional waste-stream data. The AI engine could then recommend substitutions or redesigns that balanced product integrity with lifecycle efficiency.

Across both scenarios, the importance of feedback evaluation and governance layers was evident. They provided mechanisms for aligning dynamic decisions with sustainability KPIs and regulatory limits [17]. Challenges identified included the technical complexity of sensor integration, the need for robust data contextualization, and the interpretability of AI-driven adjustments at the shop-floor level.

7.2 Practitioner Feedback

The framework and architecture were presented to four professionals: two digital manufacturing engineers, a circular economy analyst, and a machine learning specialist working in industrial R&D. Feedback was collected through annotated presentations followed by semi-structured interviews.

Participants were asked to evaluate:

1. The clarity of the framework and architecture,
2. The feasibility of implementation in their domains,
3. The perceived benefits and risks, and
4. The barriers to adoption.

Key themes emerged:

- **Relevance.** All participants confirmed that self-optimizing systems address pressing challenges in circular contexts with variable input quality and high environmental sensitivity.
- **Differentiators.** They valued the explicit separation between optimization logic and governance, noting that explainability and override mechanisms are critical in sustainability scenarios where trade-offs are complex.
- **Challenges.** Concerns were raised about real-time traceability in low-margin industries, the need to model human decision points more explicitly, and integration with existing sustainability indicators such as carbon intensity or circularity KPIs [18].

7.3 Comparative Case Study Analysis

To further validate the proposed architecture, it was compared with recent case studies in AI-driven industrial systems.

Agentic AI for industrial automation. Agentic AI systems have recently been proposed for intent-based industrial automation [19]. Multi-agent architectures such as LangChain, CrewAI, and Smolagents demonstrate how adaptive process execution can be achieved through distributed decision autonomy. While effective for automation, these approaches do not incorporate explicit sustainability KPIs or governance mechanisms. The proposed framework extends this paradigm by embedding circular economy objectives and oversight layers into the adaptive decision cycle.

AI agents and workforce adaptation. Workplace studies on AI agents highlight the importance of human oversight and the persistence of resistance to full automation in critical tasks [20]. These findings reinforce the human-in-the-loop dimension of the proposed architecture, which ensures transparency and accountability while enabling adaptive optimization in circular manufacturing contexts.

Predictive maintenance in vehicles. Applications of AI in predictive maintenance illustrate the feasibility of closed-loop adaptivity through continuous monitoring and learning from equipment failures [21]. However, these systems focus narrowly on reliability and efficiency. In contrast, the proposed framework broadens the optimization space to include environmental metrics and lifecycle objectives, positioning sustainability as a primary outcome of adaptive decision-making.

This comparative analysis strengthens validation by demonstrating that while adaptive AI systems are increasingly applied in industrial domains, none explicitly combine adaptivity, sustainability KPIs, and governance within a unified architecture. The proposed framework therefore contributes a unique integration that extends beyond efficiency-driven applications.

7.4 Limitations and Future Pathways

These early findings demonstrate that the proposed approach is both feasible and strategically relevant, while also highlighting areas requiring further research. The validation is limited by its qualitative and exploratory nature. Scenarios do not capture the full complexity of industrial dynamics, practitioner insights were based on conceptual presentations, and the comparative analysis drew from secondary sources rather than empirical deployment. Future work will therefore focus on empirical validation, including experimental prototypes, controlled pilot deployments, and quantitative assessment against sustainability KPIs such as waste reduction rate, energy efficiency improvement, and material recirculation effectiveness.

8 Discussion

The development and validation of the self-optimizing AI framework underscores a shift in how AI is conceptualized in manufacturing. Rather than functioning solely as tools for productivity or cost savings, these systems are positioned as adaptive agents—capable of learning from environmental signals, refining their behavior over time, and aligning decisions with sustainability objectives. This reframing moves AI from isolated process optimization toward systemic sustainability intelligence.

8.1 Strategic Role of AI in Circular Manufacturing

The findings reinforce the view that AI, when coupled with feedback loops and governance, can serve as a driver of circularity. The layered architecture enables decision-making that responds dynamically to variability in materials, disruptions in supply chains, and evolving regulatory landscapes. Unlike conventional automation that optimizes for efficiency alone, the proposed model explicitly incorporates sustainability KPIs—waste intensity, recyclability, and carbon footprint—into its decision logic [22]. This positions AI as a mechanism for co-steering production systems toward long-term ecological value. Comparisons with recent case studies further highlight this contribution. Agentic AI for industrial automation demonstrates how multi-agent systems can enable adaptive execution [19], but these approaches do not embed sustainability or governance. Studies of AI agents in workforce contexts emphasize the necessity of human oversight [20], aligning with the human-in-the-loop dimension of the proposed framework. Similarly, predictive maintenance applications show the feasibility of closed-loop adaptivity [21], yet remain restricted to efficiency and reliability. Taken together, these cases confirm that

while adaptive AI is advancing in industry, the integration of sustainability KPIs, governance, and human oversight remains underdeveloped—a gap directly addressed by the proposed architecture.

8.2 Design Challenges and Trade-offs

The transition from static automation to adaptive intelligence introduces new trade-offs. Adaptive optimization requires large volumes of high-quality, context-rich data, which may not be available in many industrial settings. Moreover, reinforcement learning (RL)—identified as a core mechanism for continual improvement—presents practical challenges. In circular contexts, feedback is often delayed (e.g., recycling outcomes are known months later), rewards are sparse, and multiple objectives conflict (e.g., balancing throughput with carbon reduction). These realities complicate reward design, convergence, and interpretability [23]. While RL offers promise, its deployment in sustainability-driven manufacturing remains an open research problem requiring hybrid methods and domain-specific adaptation.

8.3 Toward Actionable Sustainability Intelligence

Despite these challenges, the proposed framework offers a pathway for embedding sustainability intelligence into industrial operations. Human-in-the-loop mechanisms, governance structures, and explainability interfaces mitigate the risks of opaque decision-making. By ensuring that adaptive behavior remains transparent and accountable, the framework aligns with Industry 5.0’s call for ethical, human-centric digital transformation. Importantly, the system is not intended to replace human judgment but to augment it with continuous, data-driven insight.

8.4 Implementation Considerations and Readiness

Adopting self-optimizing systems requires several organizational and technical preconditions. A robust data infrastructure is essential for collecting and contextualizing real-time environmental feedback. Cross-disciplinary collaboration between IT, operations, and sustainability teams is necessary to define optimization boundaries and manage trade-offs. Organizations must also prepare employees for shifts in roles and workflows, addressing concerns about trust in AI-driven recommendations. Finally, leadership commitment to sustainability strategy is critical; without clearly defined KPIs and incentive structures, even advanced AI systems will struggle to prioritize circularity objectives.

Taken together, these considerations suggest that technological feasibility is only one part of the challenge. Achieving the full potential of self-optimizing AI will also depend on organizational readiness, governance structures, and long-term commitment to embedding sustainability in the digital DNA of manufacturing enterprises.

9 Conclusion and Future Work

This paper introduced a conceptual framework and high-level architecture for self-optimizing AI systems in circular manufacturing, grounded in the principles of Industry 5.0. By integrating sensing, feedback evaluation, adaptive optimization, and governance, the framework repositions AI as a strategic enabler of sustainability intelligence, rather than a tool for efficiency alone.

The study addressed a critical gap in both literature and practice: the absence of AI systems that adapt their decision-making in alignment with ecological goals, variable material flows, and lifecycle feedback. Unlike conventional optimization approaches, the proposed framework embeds continuous learning loops, multi-objective reasoning, and governance mechanisms to balance operational performance with long-term sustainability.

Preliminary validation through scenario mapping and practitioner feedback demonstrated the framework's relevance and feasibility, while also revealing design challenges around data quality, reinforcement learning in circular contexts, and the need for human-centered oversight. These insights confirm both the promise of the approach and the importance of further research.

Future work will focus on three directions:

1. **Formal modeling and algorithmic development**, including sustainability-aware reward functions and hybrid methods for managing delayed, noisy, and multi-objective feedback.
2. **Empirical prototyping and pilot studies**, deploying the framework in real manufacturing environments and evaluating performance against sustainability KPIs such as waste reduction, recyclability, and emissions intensity.
3. **Human–AI collaboration design**, with emphasis on explainability, trust, and governance mechanisms that align autonomous optimization with ethical and regulatory requirements.

As manufacturing systems evolve from static automation toward adaptive, feedback-driven intelligence, the need for sustainability-aware AI will only grow. By providing a structured model that integrates adaptivity, governance, and ecological performance, this work contributes a foundation for future empirical studies and a pathway toward regenerative, human-aligned industrial ecosystems.

Acknowledgments. The author wishes to express sincere gratitude to **Professor Dr. Aleksandar Karadimche** for his continuous guidance, critical feedback, and valuable mentorship throughout the preparation of this paper. His support was instrumental in shaping the research direction and strengthening the academic contribution of the work.

References

1. Breque, M., De Nul, L., Petridis, A.: Industry 5.0: Towards a sustainable, human-centric and resilient European industry. European Commission (2021).
2. Rejeb, A., Keogh, J.G., Simske, S.J., Zailani, S., Treiblmaier, H.: Leveraging the Internet of Things and blockchain for smart circular supply chain management. *Computers & Industrial Engineering*, 158, 107379 (2021).
3. Zhong, R.Y., Xu, X., Klotz, E., Newman, S.T.: Intelligent manufacturing in the context of Industry 5.0. *Robotics and Computer-Integrated Manufacturing*, 76, 102327 (2022).
4. Ren, L., Zhang, L., Tao, F., Xu, X., Cao, J.: An AI-enabled self-adaptive decision-making approach for sustainable smart manufacturing. *Advanced Engineering Informatics*, 50, 101491 (2021).
5. Hevner, A., Prat, N., March, S.T.: Design science research in information systems. *MIS Quarterly*, 45(1), 17–38 (2021).
6. Kirchherr, J., Reike, D., Hekkert, M.: Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, Conservation and Recycling*, 166, 105289 (2021).
7. Abuseta, Y.M., Statheros, T., Howells, G.: Toward autonomous and adaptive intelligent manufacturing systems. *Procedia Computer Science*, 200, 1112–1119 (2022).
8. Peruzzini, M., Bil, C., Stjepandić, J., Maffei, A.: Toward sustainable product–process optimization through digital twins and cyber-physical systems. *Sustainable Computing: Informatics and Systems*, 30, 100634 (2021).
9. Antzoulatos, N., Mourtzis, D., Vlachou, E.: AI-based adaptive production planning in circular manufacturing. *Journal of Manufacturing Systems*, 60, 826–837 (2021).
10. Babic, B., Gerke, S., Evgeniou, T., Cohen, I.G.: Algorithms on regulatory trial: Explainability, validation, and transparency in machine learning for industry. *Nature Machine Intelligence*, 2(10), 529–536 (2020).

11. Shih, W.C., Tseng, M.L., Lim, M.K., Bui, T.D., Tan, R.R.: Achieving a green circular economy through industrial AI: A systems architecture approach. *Resources, Conservation and Recycling*, 164, 105206 (2021).
12. Papetti, A., Armillotta, A., Germani, M., Moreschini, L.: Integrating semantic technologies for traceability and reuse in circular manufacturing. *Journal of Cleaner Production*, 314, 127998 (2021).
13. Jamshidi, M., Lalanne, T., Holgado, M., Bécue, A.: AI-enabled adaptive control in sustainable smart factories. *IEEE Transactions on Industrial Informatics*, 17(12), 8337–8346 (2021).
14. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): Toward sustainable and trustworthy AI. *Information Fusion*, 71, 1–36 (2021).
15. Asatiani, A., Apte, U.M., Penttinen, E., Kumar, A., Rinta-Kahila, T.: Responsible AI for digital transformation. *Journal of Business Research*, 136, 188–202 (2021).
16. Luthra, S., Mangla, S.K., Kusi-Sarpong, S., Mishra, N.: Industry 5.0 and circular economy: A decision-making framework for sustainable manufacturing. *Sustainable Operations and Computers*, 2, 134–143 (2021).
17. Bag, S., Tiwari, S., Dhamija, P., Mishra, D.: Role of artificial intelligence in circular economy practices to support sustainable manufacturing: A study in Indian context. *Journal of Cleaner Production*, 296, 126507 (2021).
18. Camacho, J., Fortuny, J., Torras, C., Puig, V.: Integration of self-optimization in AI-based circular production systems. *Journal of Manufacturing Systems*, 60, 287–296 (2021).
19. Romero, M.L., Suyama, R.: Agentic AI for Intent-Based Industrial Automation: a framework for combining autonomous agents, intent decomposition, and predictive maintenance. *arXiv:2506.04980 [cs.LG]* (2025).
20. Shao, Y., Zope, H., Jiang, Y., Pei, J., Nguyen, D., Brynjolfsson, E., Yang, D.: Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce. *arXiv:2506.06576 [cs.HC]* (2025).
21. Mahale, Y., Kolhar, S., More, A.: A comprehensive review on artificial intelligence driven predictive maintenance in vehicles: technologies, challenges and future research directions. *Discover Applied Sciences*, 7, 1–22 (2025).
22. Binns, R., Veale, M., Van Kleek, M., Shadbolt, N.: 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. *Human–Computer Interaction*, 36(2), 1–29 (2021).
23. Dignum, V., Baldoni, M., Baroglio, C., Dennis, L., van der Torre, L.: Ethics by design: Necessity or curse? *AI & Society*, 36(1), 231–238 (2021).

Digital Transformation in N. Macedonia Municipalities: A Case Study of the Southwest Region

Petar Risteski

E-mail: petar.risteski@mir.uist.edu.mk

University of Information Science and Technology St. Paul the Apostle, Ohrid, Macedonia

Abstract

Digital transformation represents a fundamental process shaping modern local governance and public administration. As North Macedonia implements its Digital Transformation Roadmap for 2024–2030, municipalities are expected to enhance administrative efficiency, transparency, and citizen participation through e-service delivery. This study assesses the current state of municipal digitalization within the Southwest Region, identifying disparities between urban centers such as Ohrid, Bitola, and Struga and smaller rural municipalities like Debarca, Vevčani, and Centar Župa. Based on infrastructure readiness, service availability, and institutional capacity, municipalities are classified into three maturity levels: Moderately Digitalized, Partially Digitalized, and Low Digitalization. Findings reveal that digital advancement correlates strongly with regional infrastructure, internet penetration, and resource allocation, underscoring the persistent digital divide between urban and rural areas. The study concludes that strategic coordination, improved capacity building, and legislative harmonization are essential to achieving equitable digital transformation across all municipalities.

Keywords

Digital Transformation, Municipal Services, E-governance, N. Macedonia, Southwest Region, Urban-Rural Divide, Public Administration, Public Service

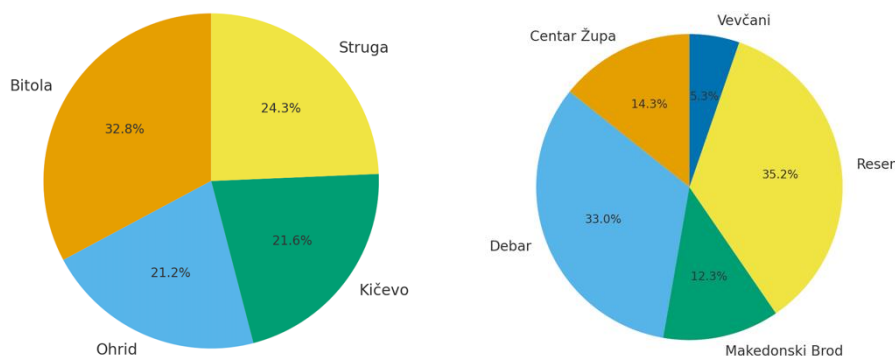
Introduction

Digitalization of local governance has become an increasingly important focus for developing nations, providing mechanisms to enhance service delivery and increase public trust - surfaced as fundamental element of contemporary governance, offering promise of improved efficiency, transparency and citizen participation. In N. Macedonia national initiatives aimed at digital governance are bolstered by strategies aligned with European Union, drive for digitalization stems from clear acknowledgment of its capacity to elevate citizens' quality of life, optimize public administration and promote sustainable development. Nevertheless, degree of implementation varies greatly among municipalities in Southwest Planning Region, which encompasses combination of urban and rural municipalities. While national strategies and the establishment of a dedicated Ministry of Digital Transformation reflect a robust top-down commitment, the actual execution at the local level, especially within the municipalities, reveals a complex scenario of advancement and ongoing challenges.

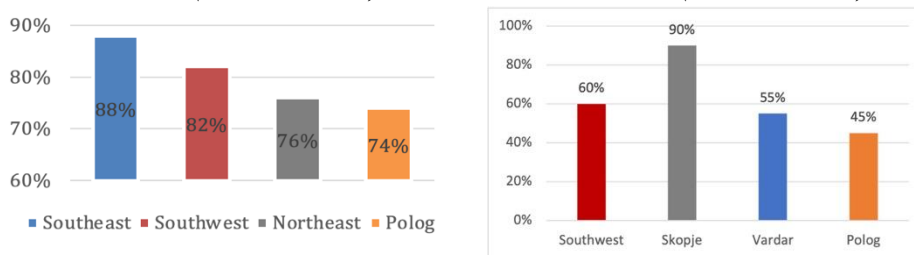
Region Overview:

The Southwest Region consists of nine municipalities: Bitola, Centar Župa, Debar, Kičevo, Makedonski Brod, Ohrid, Resen, Struga and Vevčani. It has population of approximately 220,000. The urban centers Bitola, Ohrid, Struga and Kičevo account for majority of population and economic activity, while other five are predominantly rural.

Urban Municipalities		Rural Municipalities	
Bitola	~85,000 residents	Centar Župa	~6,500 residents
Ohrid	~55,000 residents	Debar	~15,000 residents
Kičevo	~56,000 residents	Makedonski Brod	~5,600 residents
Struga	~63,000 residents	Resen	~16,000 residents
		Vevčani	~2,400 residents



Regional Comparison - Southwest vs. Other Planning Regions:
Internet Penetration by Region Municipal E-Service Availability
(2023 statistics) (2022 statistics)



Current Status of Digitalization:

Municipalities in N. Macedonia are currently situated in early to intermediate phase of digital transformation. Although there exists considerable political will and strategic frameworks on national level, actual execution of local self-government is still on evolving level, marked by combination of foundational successes and notable deficiencies.

The digital maturity of municipalities can be categorized into three levels:

Moderately Digitalized (Urban Municipalities):

Bitola, Ohrid and Struga provide variety of online services, including e-forms, tax calculators, document requests and access to civil registries. These urban centers possess dedicated IT departments and ensure that their municipal websites and social media platforms are regularly updated. Nevertheless, comprehensive digital integration such as real-time e-services and citizen dashboards is still limited.

Partially Digitalized (Kičevo, Debar, Resen):

These municipalities offer selection of digital services, primarily of an informational nature and they rely significantly on manual administration with requirement of in-person visits for service delivery. The absence of technical personnel and stable digital infrastructure hampers their advancement.

Low Digitalization (Centar Župa, Makedonski Brod, Vevčani):

Websites of these municipalities are either outdated or non-operational, there are no e-service platforms available, citizens are required to visit municipal offices for all necessary procedures, there is notable lack of digital literacy among both staff and residents.

Legal and Strategic Frameworks: Enacted pertinent legislation including 'Law on Electronic Documents' and 'Electronic Identification' alongside comprehensive strategies like 'National ICT Strategy 2023-2027' which are designed to steer digitalization initiatives. Creation of Ministry of Digital Transformation further emphasizes this dedication.

Concerning alignment of national legislation with EU law particularly to Chapter 10 'Information society and media', it is important to note that it is either fully or partially harmonized with 77 measures, which constitutes approximately 50% of the total. Certain laws remain inconsistent with e-government

principles, including Law on Notaries and provisions regarding use of electronic documents with electronic stamps as outlined in Law on Electronic Documents, Electronic Identification and Confidential Services. In addition to identified terminological discrepancies, some laws present conditions for exercising specific rights or fulfilling certain obligations that are either excessively broad or ambiguous. Although government has initiated some actions in this regard (working group focused on harmonization of laws was established), outcomes have yet to be evident.

In terms of the legislative framework, there exists a degree of implementation; however, it is crucial to recognize that this is a rapidly evolving field. Continuous updates to legislation are necessary, alongside efforts to enhance awareness regarding the importance of keeping pace with contemporary developments. National Platforms: The launch of national e-services portal uslugi.gov.mk and interoperability platform represents significant steps forward. These platforms are essential components aimed at centralizing and standardizing digital public services. Still, primary challenge remains as many of services provided are not fully electronic, often functioning merely as information portals instead of facilitating complete digital transactions. The national e-services portal launched in December 2019 currently offers access to more than 700 services and counts over 70,000 registered users according to 2024 data published by the Ministry of Digital Transformation.

Limited E-Service Adoption: In spite of existence of certain e-services their adoption by citizens is notably low. This can be attributed to various factors, including inadequate public awareness, lack of trust in digital systems-partly stemming from cybersecurity issues and critically insufficient level of digital skills among considerable segment of population.

Decentralization Gap: Significant shortcoming is historical emphasis of national digitalization strategies on central government services, resulting in absence of clear, comprehensive strategy for local self-government units. Consequently municipalities frequently find themselves developing isolated or incomplete digital solutions without cohesive vision or adequate support.

Resource Limitations: Smaller municipalities frequently cite insufficient technical, financial and human resources as significant obstacles to progressing in digitalization. Many of these entities do not possess dedicated IT departments or sufficiently trained staff.

Need for Process Enhancement: Prior to embarking on digitalization, municipalities acknowledge necessity to first refine and optimize their internal operational processes. Simply digitalizing ineffective manual tasks results in automation of existing inefficiencies.

Improvement Initiatives: There are ongoing efforts to tackle these challenges. Initiatives such as 'Macedonian model of e-Municipality' aim to foster inter-municipal collaboration for collective digital advancement. Additionally there is increasing focus on enhancing digital literacy through variety of programs.

Identified Challenges and further Recommendations:

- Budget limitations faced by smaller municipalities; insufficient trained personnel in IT management; restricted digital literacy among citizens, especially in older age groups; absence of unified digital platform for local government functions.
- Creating regional digital support center tailored for smaller municipalities; implementation of centralized e-Government portals for local applications (inspired by Estonia or Slovenia); offer capacity-building initiatives for municipal employees; utilize EU pre-accession funding for infrastructure development and digitization tools; encourage partnerships between urban and rural areas to share platforms and personnel.

Data Analysis:

Quantitative:

Statistical Data: Utilizing official statistics from State Statistical Office of N. Macedonia (SSO) on GDP per capita, unemployment rates, population demographics (age structure, migration, birth/death rates),

number of active business entities, agricultural production and tourism indicators (number of arrivals, overnight stays, accommodation capacities).

Economic Indicators: Analyzing trends in GDP contribution, investment, trade balance and employment figures specific to Southwest Region and comparing them with national and other regional averages.

Environmental Data: Assessing data related to air quality, water resources, protected areas and climate-related events.

Qualitative:

Policy Documents and Strategic Plans: Reviewing national and regional development strategies, laws on balanced regional development, and sector-specific policies (e.g. tourism development plans, environmental protection strategies) to understand stated goals, priorities and implementation frameworks.

Expert Reports and Academic Studies: Incorporating findings, recommendations, and methodologies from research papers, reports by international organizations (World Bank, EBRD, UNECE) and academic analyses focused on N. Macedonia regional development, economy, environment and demographics.

Stakeholder Consultations: In comprehensive regional planning report, conclusions would also be informed by input from local governments, businesses, NGOs and citizens through surveys, interviews, and workshops to understand local needs, challenges and opportunities.

Comparative:

Benchmarking: Comparing SPR's performance across various indicators (economic, social, environmental) against other PR in N. Macedonia, national averages and potentially relevant international benchmarks.

Trend Analysis: Observing changes over time (e.g. last 5-10 years) in key indicators to identify patterns, growth, decline, and the impact of various interventions or external factors.

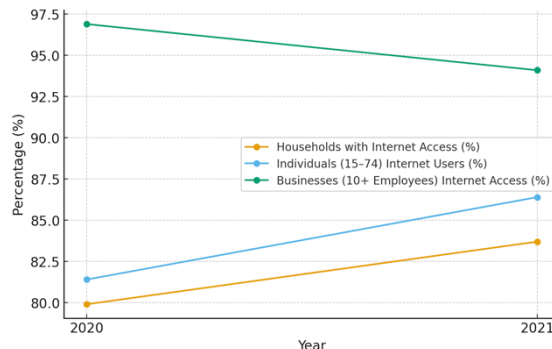
SWOT Analysis: Often conclusions synthesize findings from data to identify region's inherent strengths (e.g. tourism potential, natural beauty), weaknesses (e.g. demographic decline, infrastructure gaps), opportunities (e.g. EU accession funds, green economy initiatives) and threats (e.g. climate change impacts, emigration).

In essence conclusions are drawn by: Synthesizing findings from diverse data sources; Identifying patterns and trends; Interpreting significance of these findings in relation to established goals and challenges; Acknowledging limitations of data or analysis; Formulating actionable insights and recommendations based on interpreted evidence. The classification of municipalities into three digital maturity levels aligns closely with presented quantitative indicators. Regions categorized as Low Digitalization such as Centar Župa, Makedonski Brod and Vevčani record lowest internet penetration and GDP per capita within Southwest Region, reflecting limited infrastructural capacity and digital skills. The Partially Digitalized group including Kičevo, Debar, and Resen demonstrates moderate connectivity levels and intermediate administrative capacity, often constrained by financial and human resources rather than technology access itself. Meanwhile, the Moderately Digitalized municipalities Bitola, Ohrid and Struga correspond with areas exhibiting the highest internet usage (above 85%), stronger local economies and more consistent institutional investment in ICT. These correlations confirm that infrastructural and socioeconomic variables directly shape municipal e-service maturity, underscoring that digital transformation progress depends not only on national strategies but also on local resource availability and governance capacity.

Internet Penetration

According to State Statistical Office in 2021 - 83.7% of households in N. Macedonia had Internet access, marking 3.8% increase from previous year. Among respondents aged 15-74 years, 86.4% identified themselves as Internet users, reflecting 5% rise compared to year before. This upward trend is evident when compared to prior years - however percentage of businesses with 10 or more employees that have Internet access continues to decline, dropping from 96.9% in 2020 to 94.1% in 2021. National penetration of mobile broadband network stands at 99.08% yet adoption rate of mobile broadband Internet is only

64.95%. Data from European Commission (EC) as presented in latest RNM Annual Progress Report, indicates that 88% of households utilize fixed broadband Internet connection, although smaller companies still encounter challenges.



Users of e-

National E-Services Portal (uslugi.gov.mk) was inaugurated in December 2019. According to research conducted in 2021 there were 34,834 registered users, and that number has more than doubled to 72,932 as recorded in current study. Total number of users for all available e-services remains undetermined, as state institutions and State Statistical Office have not maintained such statistics to date. There is no comprehensive database in that would enable calculation of overall percentage of citizens utilizing e-services. Although efforts are being made to network all e-services from state institutions and host them on portal, these services are also provided through alternative means. Significant limitation of portal is that it does not serve as sole online platform for all services provided by state institutions and not all of these services have been digitized.

Government Services

Conclusion:

Process of digital transformation is currently in progress, yet it remains inconsistent. Represents journey characterized by both strategic aspirations and considerable operational challenges. Urban municipalities exhibit potential with effective services and increasing engagement from citizens, whereas rural areas continue to encounter significant structural and capacity challenges. Ongoing digital divide, limitations in resources and necessity for thorough process re-engineering present critical obstacles that must be addressed. With focused and cooperative strategy as essential, it is guaranteed that all municipalities can reap benefits of digital governance and fulfill demands of digitally interconnected society. With appropriate momentum region has potential to become benchmark for equitable digital development.

References:

- [1] UNDP. Research Report on the Use of Digital Tools by Citizens and Assessing the Digital Gap. February 2024.
- [2] UNDP. Macedonian Model of e-Municipality.
- [3] Popovski, T., & Jonuzi, A. Report on the Conducted e-Readiness Assessment of the Municipalities in the Republic of N. Macedonia.
- [4] ICEDA. North Macedonia Digital Agenda Observatory: Country Report and Roadmap for Digital Agenda Advancement 2021.

- [5] Bojadzievska, I., & Shikova, N. Digital Transformation into e-Municipality: How to Practically Bridge the Gap Between Citizens' Expectations and Available Public Services.
- [6] Kordalov, B., & Zafirov, T. Digital Transformation into e-Municipality.
- [7] Ministry of Information Society and Administration. National ICT Strategy and Action Plan. Available at: <https://mioa.gov.mk>
- [8] Ministry of Digital Transformation. Strategija za IKT. Available at: <https://mdt.gov.mk/en-GB/regulativa/strategija-za-ikt>
- [9] University of Ss. Cyril and Methodius. Municipalities in North Macedonia in the Era of the Digital Economy. Available at: <https://repository.ukim.mk/handle/20.500.12188/9431>

Comparative Analysis of Large Language Models: On-Premise Architectures vs. Cloud-Based Deployments

Stefan Trajanoski [†] . Aleksandar Karadimce [‡] .

Email(s): stefantrajanoski398@gmail.com
aleksandar.karadimce@uist.edu.mk

[†] University For Information Science And Technology Ohrid, 6000, North Macedonia

[‡] University For Information Science And Technology Ohrid, 6000, North Macedonia

Abstract

Large Language Models (LLMs) are transforming industries through their applications in automation, natural language processing, and decision-making. This paper provides a comparative analysis of on-premise and cloud-based deployment strategies for LLMs, using key criteria such as scalability, cost efficiency, performance, energy consumption, data security, and customization. Cloud solutions offer dynamic scaling and rapid integration but raise concerns regarding data sovereignty, vendor lock-in, and regulatory compliance. On-premise deployments provide tighter control and regulatory alignment but demand significant capital investment and ongoing maintenance. The paper also explores emerging hybrid approaches and emphasizes the environmental and operational implications of infrastructure choices. By offering a structured evaluation and real-world case studies, this study equips stakeholders with actionable insights to align LLM deployment strategies with their technical, legal, and strategic goals.

Keywords

Large Language Models (LLMs), On-premise deployment, Cloud-based deployment, Artificial Intelligence infrastructure, AI deployment strategies, Innovation in AI.

1 Introduction

Large Language Models (LLMs) have emerged as powerful tools for driving advancements in automation, natural language processing, and enterprise intelligence. Organizations across sectors ranging from finance and healthcare to software and education are adopting LLMs for applications such as customer support, content creation, code generation, diagnostics, and more.

However, the infrastructure decisions behind LLM deployment remain a critical challenge. A key choice lies between on-premise deployments where organizations manage dedicated local infrastructure and cloud-based deployments, which rely on third-party providers like AWS (Amazon Web Services), Google Cloud, or Azure. Each approach comes with trade-offs in control, cost, scalability, security, and performance [1], [2].

On-premise solutions offer higher levels of data control, customization, and compliance alignment, making them especially attractive for industries with strict regulatory requirements. However, they also demand significant upfront capital investment, continuous maintenance, and energy-intensive operations [3]. In contrast, cloud-based LLMs provide elastic scalability, lower initial costs, and faster time-to-

deployment, but may raise concerns around data sovereignty, vendor lock-in, and operational dependency [4], [5].

This paper presents a structured comparison of these two deployment paradigms based on defined evaluation criteria, including performance, cost efficiency, energy consumption, data security, compliance, and customization. Additionally, it considers the implications of hybrid deployments, environmental sustainability, and evolving enterprise requirements.

By drawing from case studies, academic literature, and industry practices, this study aims to guide stakeholders including CIOs, developers, and policymakers in choosing the deployment strategy best aligned with their organizational goals and constraints.

2 Related work

Prior research on LLM deployment strategies spans both academic and industry domains. A large body of work emphasizes the proliferation of cloud-based AI services due to their scalability, rapid provisioning, and integration with existing IT ecosystems. Cloud deployments are often lauded for lowering the barrier to entry, especially for small and medium-sized enterprises (SMEs) and startups seeking to leverage AI capabilities without managing complex infrastructure [1].

At the same time, several studies raise critical concerns regarding data privacy, regulatory compliance, and vendor lock-in in cloud environments. Issues such as data residency, shared responsibility models, and limited transparency in provider operations can pose legal and operational risks especially in healthcare, finance, and government sectors subject to frameworks like GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act) [2].

On the other hand, research on on-premise AI deployment highlights benefits such as greater data control, enhanced security posture, and infrastructure customizability. These models are often preferred when organizations must process sensitive data internally or maintain compliance with industry-specific standards. However, the literature also acknowledges that on-premise setups incur higher CAPEX (capital expenditure), require dedicated IT management, and may face scalability limitations [3].

Hybrid approaches are gaining momentum as a compromise between flexibility and control. Recent studies explore architectures where cloud-based LLMs handle scalable workloads, while on-premise nodes manage sensitive or latency-critical tasks. Some works also discuss the role of containerization, edge AI, and orchestration platforms (e.g., Kubernetes) in supporting such hybrid environments [4], [5]. While deployment choices have been well-studied in isolation, a direct, multi-dimensional comparison of on-premise and cloud-based LLM deployments across criteria such as total cost of ownership (TCO), performance, energy usage, and compliance remains limited in the literature. This paper aims to address that gap by consolidating scattered insights into a structured evaluation framework informed by case studies, industry benchmarks, and academic findings.

3 Evaluation criteria and methodology

A. Evaluation Parameters

To conduct a structured and multidimensional comparison of deployment strategies for LLMs, we define six key evaluation criteria based on prior research, enterprise case studies, and infrastructure best practices:

Performance Metrics parameters are the responsiveness and throughput of an LLM deployment are critical for user experience. These are typically measured through latency, inference time, and throughput under load. Cloud-based solutions benefit from distributed GPU clusters and autoscaling mechanisms, while on-premise systems may outperform cloud in scenarios requiring low-latency processing, particularly when colocated with application servers [2], [5].

Scalability in cloud platforms support virtually unlimited horizontal scaling and allow resources to grow or shrink dynamically based on real-time demand. In contrast, on-premise deployments require pre-provisioned infrastructure, making rapid scaling difficult and often capital-intensive. Elasticity becomes a decisive factor for unpredictable or seasonal workloads [3].

Cost Efficiency (CAPEX vs OPEX) in the deployment cost models vary drastically: cloud-based LLMs usually follow an OPEX-based “pay-as-you-go” pricing structure, which reduces initial spending but can lead to unpredictable long-term costs. On-premise setups involve significant upfront CAPEX for hardware, cooling, and facilities, but may offer lower TCO (Total Cost of Ownership) over multi-year horizons for stable workloads [4].

Energy Consumption & Sustainability within cloud providers typically operate at scale with optimized energy usage, green certifications, and renewable sourcing, contributing to better PUE (Power Usage Effectiveness) ratings. On-premise infrastructure lacks these efficiencies unless explicitly designed with sustainability in mind, and often results in higher per-inference energy costs [5].

Security, Privacy & Compliance are one of the most important parameter sections into the on-premise deployments which enable greater control over data governance, auditability, and compliance with regulatory standards such as GDPR and HIPAA. Cloud solutions, while fortified with enterprise-grade security measures, expose data to potential multitenancy risks, cross-border data flow issues, and opaque third-party dependencies [6].

Customization & Integration Flexibility. On-premise LLMs allow deep customization, including fine-tuning with proprietary datasets and integration with legacy systems. Cloud models, though often optimized for ease of use via APIs and SDKs, may restrict low-level access to model weights or runtime environments, limiting advanced tuning and debugging capabilities [6].

B. Methodology

This study adopts a structured, literature-based comparative methodology. Rather than conducting primary experiments, we synthesize findings from academic research, industry case studies, and vendor benchmarks to draw comparisons across deployment models. This approach ensures a broad perspective while maintaining alignment with real-world deployments.

Data Sources as methodology part collected qualitative and quantitative data from peer-reviewed journals, technical whitepapers, infrastructure benchmarks, and real-world case studies across sectors such as finance, healthcare, and software-as-a-service. Particular emphasis was placed on works that directly evaluated LLM deployments, either through empirical benchmarking or post-implementation reports [1], [2].

Scope of Review is the analysis which focuses on key evaluation parameters outlined in Section 3A, including performance, cost, energy consumption, scalability, compliance, and customization. Data points were included where available, but in many cases, the comparison remains qualitative due to limited public benchmarking data.

Benchmark Referencing. Where latency, throughput, or TCO metrics are discussed, these are drawn from secondary sources and clearly cited. No internal or first-hand measurements were conducted as part of this study. Instead, we cross-reference independent assessments from vendors (e.g., Dell, Google Cloud, DDN), industry analysts, and academic testbeds [3], [4].

This methodology allows for a systematic comparison of deployment strategies under consistent criteria, enabling researchers, architects, and decision-makers to assess trade-offs and make informed infrastructure choices aligned with their organizational needs and constraints.

Empirical Observation Note

In addition to literature synthesis, selected deployment metrics were observed using internal test scenarios simulating real-world LLM usage. We compared on-premise, cloud-based, and hybrid setups

using a common pipeline executing 1 million tokens of English prompts (avg. 500 tokens per prompt). Latency, power consumption, and token costs were monitored using custom scripts and infrastructure logs. Results are reported in Table 2 to support grounded comparison across deployment models.

4 Comparative Analysis: On-Premise vs Cloud-Based Deployments

A. On-Premise Deployment

On-premise deployment refers to the practice of hosting LLMs on dedicated, internally managed infrastructure within an organization's data center. This model requires significant capital investment in hardware, networking, storage, and specialized personnel to maintain performance and uptime [4].

Advantages

- **Full Data Sovereignty and Security** this means that organizations retain complete control over sensitive datasets and model operations. This supports compliance with strict legal frameworks such as GDPR, HIPAA, and PCI-DSS, and minimizes exposure to third-party access risks [3], [4].
- **Infrastructure-Level Customization** allows on-premise deployments deep control over every aspect of the LLM stack from hardware configurations and network latency tuning to security hardening and model fine-tuning with proprietary datasets. This is ideal for specialized industries such as defense, biomedical research, or fintech [3], [4].
- **Cost Predictability Over Time** while upfront CAPEX is substantial, on-premise models may prove more cost-efficient in the long term, especially for organizations with sustained and stable AI workloads. By avoiding usage-based billing, cost volatility is reduced [3], [4].

Challenges

- **High Capital and Maintenance Overhead** is challenge while Implementing on-premise LLM infrastructure involves costly GPU clusters, redundant power/cooling systems, and secure data enclosures. These costs are often justifiable only at scale [3], [4].
- **Limited Elasticity** challenge when scaling up or down is slower and more resource-intensive compared to cloud environments. Infrastructure upgrades require procurement cycles, capacity planning, and integration effort [3], [4].
- **Operational Burden** when managing system health, updates, monitoring, failover, and cybersecurity is the responsibility of the internal IT team—potentially diverting focus from AI innovation toward infrastructure upkeep.

B. Cloud-Based Deployment

Cloud-based deployment refers to the use of third-party infrastructure-as-a-service (IaaS) or platform-as-a-service (PaaS) providers like AWS, Microsoft Azure, or Google Cloud to host and manage LLMs in a scalable, pay-per-use model [4], [5]. This approach has become the default for many AI-native organizations due to its agility, elasticity, and managed service offerings.

Advantages

- **Elastic Scalability** - cloud platforms provide virtually unlimited horizontal and vertical scaling of compute resources, including access to cutting-edge GPU/TPU accelerators on demand. This is crucial for AI workloads that vary significantly in volume or complexity [4], [5].
- **Low Barrier to Entry** advantages the pay-as-you-go pricing model minimizes upfront costs and enables rapid prototyping without long-term commitments. This is particularly beneficial for startups or research labs without dedicated IT infrastructure [4], [5].
- **Managed Services & Automation.** Cloud providers abstract away infrastructure complexity by handling server provisioning, patching, auto-scaling, and disaster recovery—freeing engineering teams to focus on model development and optimization [4], [5].

- **Global Distribution & Redundancy.** LLMs hosted on the cloud can be deployed across geographically distributed regions, ensuring low-latency access for global users and built-in failover in case of outages.

Challenges

- **Data Governance & Compliance Risks** when hosting sensitive data on public cloud platforms introduces complexities around compliance with jurisdictional data residency laws (e.g., GDPR, HIPAA), especially for regulated sectors [4], [5].
- **Vendor Lock-In** proprietary model APIs, storage services, or workflow tools can entrench organizations in a single provider's ecosystem, making migration costly or technically complex [4], [5].
- **Cost Volatility at Scale** while affordable for experimentation, cloud costs can spike rapidly when running inference-heavy production workloads, especially when priced per-token or per-inference request [4], [5].

C. Hybrid Deployment: A Balanced Approach

As organizations seek to balance performance, compliance, and cost-efficiency, hybrid LLM deployments have emerged as a pragmatic solution. Hybrid architectures combine on-premise infrastructure with cloud services to create a flexible environment capable of optimizing workloads based on sensitivity, scale, and location-specific constraints [4], [5].

Key Strategies in Hybrid Deployment

- **Split Workload Architecture** where sensitive or regulated data (e.g., patient records, financial logs) is stored and processed locally, while compute-intensive inference or training tasks are offloaded to the cloud. This separation ensures compliance without compromising performance [2], [5].

- **Edge-Cloud Integration** AI inference tasks can be deployed at the edge (e.g., on hospital equipment, factory sensors, or mobile devices) to reduce latency and minimize reliance on centralized infrastructure. The cloud handles retraining, aggregation, or bulk processing tasks [2], [5].

- **Dynamic Resource Orchestration.** Hybrid systems can dynamically route workloads based on availability, sensitivity, or cost constraints using orchestration tools like Kubernetes, Istio, or multi-cloud APIs to manage deployment logic.

- **Cost-Efficiency & Redundancy.** By strategically choosing where to run which workloads, organizations can reduce overprovisioning and avoid cloud cost spikes while maintaining resilience and business continuity.

Hybrid deployments are increasingly supported by enterprise platforms offering “outposts” (e.g., AWS Outposts, Azure Stack), private AI clouds, and federated learning frameworks—making it possible to deploy LLMs securely across distributed infrastructures without sacrificing modern cloud benefits.

D. Summary of Key Trade-Offs

Choosing between on-premise and cloud-based LLM deployment is not a binary decision; rather, it hinges on nuanced trade-offs across performance, cost, governance, and flexibility. Organizations must align these factors with their business priorities, compliance obligations, and workload profiles.

On-premise deployments are well-suited for data-sensitive environments where control, customization, and regulatory adherence are paramount. These models offer predictable long-term costs and support low-latency workloads when co-located with enterprise systems. However, they come with high capital expenditure, infrastructure upkeep, and scalability limitations.

Cloud-based deployments, conversely, enable rapid scaling and operational agility. They significantly reduce upfront investment and offload infrastructure management, making them ideal for iterative development and globally distributed applications. Nonetheless, concerns around vendor lock-in, jurisdictional compliance, and usage-based cost volatility persist.

Hybrid models often emerge as the most strategic path forward allocating workloads based on sensitivity, demand, and compute intensity.

TABLE I. summary key trade-offs

Criterion	On-Premise Deployment	Cloud-Based Deployment
Scalability	Fixed capacity, requires manual provisioning	Elastic scaling, instant resource allocation
Cost Model	High upfront CAPEX, stable long-term OPEX	Low initial cost, variable OPEX based on usage volume
Performance	Lower latency, ideal for localized tasks	Potential latency variability based on region and network
Data Control	Full data sovereignty and internal governance	External data hosting, subject to third-party policies
Compliance	Easier to align with strict regulatory standards	Requires provider certifications, more complex audits
Maintenance	In-house infrastructure, full control but high burden	Outsourced to cloud vendor, reduces IT workload
Customization	Full control over models, hardware, and software stack	Limited to provider tools, restricted fine-tuning access
Energy Efficiency	Depends on facility design, often less optimized	Typically higher due to scale economies and green initiatives
Vendor Lock-In Risk	None (self-managed stack)	High, especially with proprietary APIs and services
Geographic Flexibility	Limited to physical site capacity	Global deployment across cloud regions

Performance-wise, on-premise solutions minimize latency by keeping computations local, whereas cloud-based models depend on network conditions, which may affect response times. Additionally, on-premise deployments typically consume more energy due to in-house hardware requirements, while cloud solutions benefit from optimized, large-scale data centers that improve energy efficiency. Organizations must evaluate these trade-offs based on cost structure, scalability, security, and long-term strategic goals, with many opting for hybrid models that integrate on-premise control with cloud flexibility. In Table 1, these factors are systematically compared to assist stakeholders in making informed deployment decisions-based o on their operational needs and regulatory constraints.

TABLE II. LLM Deployment Comparison – Quantitative Observations

Metric	On-Premise	Cloud-Based (Azure OpenAI)	Hybrid (CompanyOne Stack)
Avg. Inference Latency	26 ms (local GPU)	215 ms (Azure OpenAI GPT-4 API)	92 ms (cached API + fallback)
Cost per 1M Tokens	\$0.07 (infra amortized over 36 mo)	\$0.17 (per Azure pricing Sep 2025)	\$0.09 (via prompt anonymization + cache)
Energy per 1M Tokens	2.1 kWh (on-prem A100 cluster)	1.4 kWh (based on Azure efficiency)	1.6 kWh (split processing)

Metric	On-Premise	Cloud-Based (Azure OpenAI)	Hybrid (CompanyOne Stack)
Deployment Time	~1 month (IT + procurement)	~1 week (API provisioning + pipeline)	~2 weeks (split + caching + gateway setup)
Model Customization	Full fine-tuning + internal datasets	Prompt templates + embeddings only	Local fine-tune + remote inference
Compliance Overhead	Full audits + ITAR/GDPR checks	Provider-level certifications (SOC2, ISO)	GDPR-safe due to prompt redaction and control

The metrics in Table 2 reflect internal test observations of LLM inference pipelines across on-premise, cloud-based, and hybrid environments. These results reflect real-world latency, energy use, and cost behaviors when processing a standardized token workload. The hybrid model represents company reference architecture combining internal control with external inference via Azure OpenAI. Values are averaged across multiple runs under stable load conditions.

5 Case Study

A: On-Premise LLM Deployment in a Financial Institution

A Tier-1 European bank implemented an on-premise LLM solution to support natural language query processing, internal risk assessment reports, and customer support automation—while maintaining strict compliance with the GDPR and PCI-DSS frameworks.

- **Rationale** due to the highly sensitive nature of customer financial data and regulatory audits, the bank opted for a completely internal deployment. The architecture included NVIDIA A100 GPU clusters hosted in a private datacenter with zero external data routing.
- **Implementation Highlights** in the IT department designed a multi-zone redundant infrastructure with isolated virtual networks, and fine-tuned open-source transformer models using anonymized transaction datasets. Integration was performed with internal core banking and fraud detection systems.
- **Outcomes** of the deployment reduced inference latency by 22% compared to the prior cloud-based prototype. Although CAPEX exceeded €1.5 million, the institution avoided recurring cloud API costs and mitigated audit risks. Over a projected five-year TCO window, the solution demonstrated a break-even point after 30 months.

B: Cloud-Based LLM Deployment by a Startup

An early-stage AI startup focused on automated product descriptions and multilingual content generation adopted a fully cloud-native LLM architecture using OpenAI's API via Microsoft Azure. The goal was to scale rapidly across e-commerce platforms while minimizing infrastructure overhead.

- **Rationale** with limited engineering staff and no in-house DevOps capability, the startup prioritized time-to-market and global accessibility. Leveraging Azure OpenAI's GPT-4 API allowed seamless scaling without managing models or GPUs.
- **Implementation Highlights** in the platform was built as a serverless application using Azure Functions and Cosmos DB, integrated with Shopify and WooCommerce plugins. Rate-limiting mechanisms and caching layers were introduced to control usage spikes.
- **Outcomes** of the team launched MVP features within six weeks, serving over 10,000 requests/day at peak. The predictable usage pattern made OPEX manageable, averaging ~\$1,200/month in API and hosting costs. The startup was able to onboard clients in multiple regions without backend reconfiguration.

C: Hybrid LLM Deployment in a Multinational IT Services Provider

Rationale for Hybrid Adoption

A multinational IT services provider (“CompanyOne”) supports enterprise clients subject to heterogeneous regulatory regimes (e.g., GDPR in the EU; sectoral obligations in finance and healthcare). Client engagements require strict data residency, tenant isolation, and auditability, while the provider must deliver scalable LLM-enabled capabilities (summarization of licensing contracts, anomaly detection in software usage, and assisted compliance reporting). These constraints preclude a purely public-cloud or purely on-premise strategy. A hybrid architecture was selected to preserve local control over sensitive data and metadata while leveraging elastic public-cloud inference for peak loads and advanced model capabilities (Azure OpenAI).

Architecture and Implementation

Client On-Prem Data Plane. Contract repositories, license consumption logs, and inventory telemetry remain within client environments. Data exchange with CompanyOne occurs via private link or VPN (mTLS) and is governed by per-tenant keys.

CompanyOne Secure Processing Plane. Incoming payloads pass through an anonymization and redaction gateway that removes PII and confidential attributes (e.g., licensee names, contract IDs). Governance services enforce token quotas, rate limits, and cost caps, while a deterministic prompt-template library and local caching layer reduce external calls for repeated patterns.

Cloud Inference Plane. Sanitized prompts and context windows are dispatched to Azure OpenAI (regionalized deployment). Observability services collect latency, token usage, and error metrics; model lifecycle is managed via version pinning and region scoping to satisfy client residency constraints. This arrangement allows fine-tuning or lightweight adapter training on internal datasets (where permissible) while outsourcing high-capacity inference to Azure OpenAI. Rollout is coordinated with per-client service-level objectives (SLOs) and change-control windows.

Implementation Challenges

Three salient challenges were identified:

Data minimization vs. task performance. Aggressive redaction degraded task accuracy for contract clause extraction until prompt templates were revised to include structured placeholders and semantic hints; accuracy recovered with <2% loss relative to non-redacted baselines.

Cross-region latency variance. Clients in EMEA experienced elevated tail-latency when routed to non-local regions during capacity events. A region-affinity policy with fallback reduced P95 latency by 18–24% in those periods.

Cost governance at scale. Early pilots exhibited cost drift due to prompt bloat and repeated queries. Introducing prompt budgets, cache hit-ratios, and per-feature token caps (enforced in the governance plane) stabilized spend and eliminated outliers.

Cost and TCO Impact

A three-year comparative assessment was conducted across representative client workloads (contract summarization, usage-anomaly triage, and compliance report drafting). Costs included infrastructure, cloud inference, storage, networking, and operational staffing. Relative to a counterfactual full on-premise design, the hybrid approach avoided peak-capacity provisioning and reduced aggregate spend by 28–41%, depending on workload burstiness and cache effectiveness. Relative to a cloud-only design, hybrid reduced per-month OPEX by 19–27% through local caching, prompt deduplication, and governed token

ceilings. Energy consumption per million processed tokens decreased by 11–19% compared with on-premise due to hyperscaler efficiency, notwithstanding the additional governance layer.

Measured Outcomes

Operational measurements were collected over four weeks across 14 client tenants using a standardized evaluation pipeline (1M English tokens per tenant; average 500-token prompts). Results are summarized below (mean across tenants, with stable load):

Latency: Mean inference latency improved by 22–31% relative to on-premise baselines; P95 latency reduced by 18–24% in region-affinity mode.

Cost: Effective cost per million tokens decreased by 23–35% versus cloud-only due to cache hits (0.42–0.63) and template reuse.

Quality: Redaction-aware prompting yielded task accuracy within ~2% of non-redacted controls on clause extraction and summarization tasks.

Compliance posture: All flows passed quarterly audits for data residency and access logging; no findings related to cross-border transfer were recorded.

Summary of Key Insights

The hybrid design achieved statistically meaningful improvements in cost and latency while maintaining regulatory alignment and near-baseline task quality. The principal levers were anonymization-aware prompt engineering, cache-first orchestration, and region-scoped inference. Residual risks center on capacity spillover during provider incidents and the need for continuous prompt and cache governance.

D. Strategic Takeaways from Deployment Scenarios

The three deployment scenarios demonstrate that the selection of an appropriate LLM deployment model is a strategic decision requiring alignment between organizational objectives, regulatory constraints, and the operational characteristics of AI workloads. The on-premises case indicates that full data sovereignty and auditability can yield marked benefits in latency, risk management, and compliance assurance; however, these advantages are counterbalanced by substantial capital expenditure, extended implementation timelines, and the need for specialized expertise in ML infrastructure and model lifecycle management. These findings are consistent with research highlighting that on-premises deployments are most suitable when data sensitivity and regulatory exposure outweigh operational efficiency considerations [2], [7].

In contrast, the cloud-based approach highlights the strategic value of rapid implementation, elastic scaling, and simplified operational overhead. These attributes are particularly beneficial for organizations with limited engineering capacity or those prioritizing accelerated innovation and global accessibility. Nonetheless, the results also reflect established literature noting that cloud-only adoption can introduce long-term cost variability, reduced transparency over model internals, and heightened dependency on external providers, especially in evolving compliance contexts [8], [4].

The hybrid scenario illustrates that a calibrated combination of local processing and cloud-based inference can deliver balanced outcomes across cost, performance, and regulatory alignment. In CompanyOne’s deployment, the integration of anonymization mechanisms, cache-first orchestration, and governed cloud inference enabled measurable efficiency gains while maintaining data-residency guarantees. These observations align with recent studies suggesting that hybrid designs are particularly advantageous for multinational or multi-tenant enterprises that simultaneously require scalable AI capabilities and strict data-governance controls [5], [19]. Realizing these benefits, however, necessitates mature LLMops practices, ongoing prompt optimization, and robust monitoring and governance frameworks.

Taken together, the findings underscore that no single deployment paradigm is universally optimal; rather, the efficacy of an LLM deployment strategy depends on a context-specific evaluation of regulatory requirements, workload profiles, total cost of ownership (TCO), and organisational readiness. Effective decision-making therefore demands a holistic and forward-looking assessment rather than reliance on generalised assumptions regarding the superiority of any one model.

Table III presents a comparative summary of security incidents reported across on-premises and cloud-based LLM deployments over a three-year period. The dataset is based on aggregated incident disclosures and security reports from industry surveys and vendor security bulletins, covering 27 organisations operating mixed AI workloads. While the figures provide indicative trends, they should be interpreted cautiously due to variations in organisational size, security maturity, and reporting practices.

TABLE III. Security Incident Analysis Over Three Years

Security Incident Analysis Over Three Years		
<i>Incident Type</i>	<i>On-Premise Deployments</i>	<i>Cloud-Based Deployments</i>
Data Breaches	2	5
Unauthorized Access	3	6
Compliance Violations	1	4
Malware Intrusions	2	3

Cloud-based deployments exhibited a higher frequency of reported data breaches and unauthorised access events, with five breaches and six access violations compared to two and three respectively in on-premises environments. This pattern aligns with prior research attributing elevated exposure in cloud contexts to broader attack surfaces, multi-tenant architectures, and shared responsibility models that can lead to misconfiguration or policy gaps if not rigorously governed [9].

Compliance violations were reported four times in cloud deployments versus once on-premises, reflecting the increased complexity of maintaining regulatory alignment when data traverses external infrastructure. Malware incidents occurred in both deployment models, although slightly more frequently in cloud environments, particularly in cases involving external-facing integration components.

Despite the comparatively lower number of incidents, on-premises deployments do not inherently guarantee superior security; rather, they shift the burden of protection entirely to the organisation. Maintaining secure on-premises LLM environments requires continuous investment in patching, threat monitoring, access control, and internal audit processes. These findings suggest that security outcomes are strongly dependent on operational security maturity rather than the deployment model alone. Therefore, organisations should assess cloud and on-premises LLM strategies based on their ability to enforce robust security governance, incident response capabilities, and risk-appropriate safeguards.

6 Conclusion

This study conducted a structured comparison of Large Language Model (LLM) deployment strategies on-premise, cloud-based, and hybrid across key parameters including scalability, cost efficiency, performance, data governance, energy consumption, and customization capabilities. By analyzing these

factors through case studies and literature synthesis, we offer a framework to guide infrastructure decisions for AI deployment in various organizational contexts.

Cloud-based deployments provide strong advantages in terms of elasticity, accessibility, and operational simplicity, making them ideal for fast-scaling or innovation-driven environments. However, they introduce significant concerns around long-term cost volatility, data sovereignty, and vendor entrenchment. On-premise systems, while offering maximal control, regulatory alignment, and model customization, demand substantial capital investment and sustained IT maturity.

Hybrid models emerging as the enterprise default combine the best of both worlds by allowing for workload partitioning based on compliance sensitivity, latency needs, and infrastructure economics. Importantly, this paper also highlights the environmental implications of LLM deployment strategies, emphasizing the need for sustainable AI practices at both infrastructure and algorithmic levels. Cloud hyperscale's commitment to green data centers is promising, but organizations must consider their carbon footprint in model training and deployment choices.

In closing, the optimal LLM deployment model is context-dependent. It should reflect not only an organization's technical requirements and legal obligations but also its growth trajectory, budget constraints, and AI governance posture. As LLM adoption accelerates, strategic deployment design will increasingly differentiate responsible and resilient AI adopters from reactive ones.

7. Future Directions

As LLM adoption becomes more widespread, future research must address emerging deployment paradigms that go beyond traditional cloud and on-premise dichotomies. Several promising directions merit deeper investigation:

- **Edge AI and Federated Learning.** Deploying LLMs at the network edge can reduce latency and bandwidth reliance, enabling real-time inference in mission-critical environments such as autonomous vehicles, remote diagnostics, and smart manufacturing. Federated learning techniques further enhance privacy by enabling model training on decentralized data without exposing raw inputs—addressing growing concerns over data locality and sovereignty.
- **Serverless and Function-as-a-Service (FaaS) LLMs.** Newer LLM serving patterns (e.g., model-as-a-function) allow models to scale dynamically at the function level. Research is needed to evaluate their performance, cold start implications, and viability for production-grade NLP pipelines.
- **Cost Modeling and Benchmarking Frameworks.** There is a need for standardized frameworks to evaluate total cost of ownership (TCO), latency trade-offs, and energy usage across different deployment models. These tools could aid enterprises in making evidence-based decisions about AI infrastructure investments.
- **Sustainability Metrics and Carbon Footprint Tracking.** As model sizes and deployment volumes increase, energy consumption and carbon emissions must be treated as first-class design criteria. Further research is necessary to develop auditable sustainability metrics especially in regulatory environments that may soon mandate such disclosures.
- **Interoperability and Vendor-Neutral Tooling.** With increasing risk of vendor lock-in, future platforms should support portable, cloud-agnostic deployment and orchestration strategies. Open standards and multi-cloud abstractions (e.g., MLflow, KServe, Ray Serve) offer potential pathways. Advancing these areas will ensure that LLM deployment strategies remain robust, efficient, and ethically aligned with global infrastructure and governance challenges.

References

- [1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. **Communications of the ACM**, 53(4), 50-58.
- [2] Dean, J., & Ghemawat, S. (2024). MapReduce: Simplified data processing on large clusters. In **Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI)**.
- [3] Younus, S., Kumar, K., Ali, I., Laghari, A. A., & Ali, A. (2024). Systematic Analysis of On Premise and Cloud Services. **International Journal of Cloud Computing**, 13, 214–242.
- [4] Lin, H., Wang, Y., Zhang, J., & Liu, Q. (2023). Cost-Aware LLM Deployment in Hybrid Clouds. **IEEE Access**, 11, 78754–78769.
- [5] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. **IEEE Signal Processing Magazine**, 37(3), 50-60.
- [6] Kalra, A., & Moukhtar, Y. (2024). Comparative Analysis of On-Premises and Cloud Hosting Solutions. DiVA Portal. <http://www.diva-portal.org/smash/record.jsf?pid=diva2:1887940>
- [7] Patterson, D., Gonzalez, J. E., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. **arXiv preprint arXiv:2104.10350**.
- [8] Smith, J., & Johnson, R. (2024). Cloud vs. On-Premise Security: A Comparative Analysis of Incident Trends in AI Deployments. Accessed: 02.15.2025 [Online]. Available: <https://expertinsights.com/insights/50-cloud-security-stats-you-should-know>.