GPTMB 2026

The Third International Conference on Generative Pre-trained Transformer Models and Beyond

ISBN: 978-1-68558-402-3

July 5 - 9, 2026

Nice, France

GPTMB 2026 Editors

Konstantinos (Constantine) Kotropoulos, Aristotle University of Thessaloniki, Greece

GPTMB 2026

Forward

The Third International Conference on Generative Pre-trained Transformer Models and Beyond (GPTMB 2026), held on July 5th- 9th, 2026 focused on advanced topics on GPTM and AI/Deep Learning and target the challenges of using at large scale of GPTM-based tools. The event considers the research works and the current challenges including input data, process truthfulness, impact on existing human perception, and lessons learned from experiments.

The advances on Machine Learning (ML) and Deep Learning (DL) change the nature of summarization and text generation. GPTM (Generative Pre-trained Transformer Models) are ML models that use DL techniques to generate natural language text. As for any model, the accuracy of the output is driven by the quality of input data (sensitivity, specificity) and the processing mechanisms.

The current achievements were warmly received by industrial media corporations and scientist communities. At the same time several aspects related to trust, bias, liability, and regulations because of the high probability of spreading untrue and difficultly to be cross-checked output.

We take here the opportunity to warmly thank all the members of the GPTMB 2026 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to GPTMB 2026. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also thank the members of the GPTMB 2026 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that GPTMB 2026 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of large language models. We also hope that Nice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

GPTMB 2026 Chairs

GPTMB 2026 Steering Committee

Petre Dini, IARIA USA/EU
Isaac Caicedo-Castro, University of Córdoba, Colombia
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Stephan Böhm, RheinMain University of Applied Sciences – Wiesbaden, Germany
Zhixiong Chen, Mercy College, USA
Joni Salminen, University of Vaasa, Finland
Christelle Scharff, Pace University, USA
Gerald Penn, University of Toronto, Canada
Konstantinos (Constantine) Kotropoulos, Aristotle University of Thessalonik, Greece
Thomas Zöller, IU International University of Applied Sciences, Germany

GPTMB 2026

Committee

GPTMB 2026 Steering Committee
Petre Dini, IARIA USA/EU
Isaac Caicedo-Castro, University of Córdoba, Colombia
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Stephan Böhm, RheinMain University of Applied Sciences – Wiesbaden, Germany
Zhixiong Chen, Mercy College, USA
Joni Salminen, University of Vaasa, Finland
Christelle Scharff, Pace University, USA
Gerald Penn, University of Toronto, Canada
Konstantinos (Constantine) Kotropoulos, Aristotle University of Thessalonik, Greece
Thomas Zöller, IU International University of Applied Sciences, Germany

GPTMB 2026 Technical Program Committee

Sergi Alvarez-Vidal, Universitat Autònoma de Barcelona, Spain
Thales Bertaglia, Maastricht University, Netherlands
Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany
Marietjie Botes, University of Luxembourg, Luxembourg / University of KwaZulu-Natal, South Africa
Isaac Caicedo-Castro, University of Córdoba, Colombia
Steve Chan, Decision Engineering Analysis Laboratory, USA
Zhixiong Chen, Mercy College, USA
Qiang (Shawn) Cheng, University of Kentucky, USA
Liang Ding, Alibaba Group, China
Maksim Eremeev, Genentech, USA
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Roshni Iyer, University of California, Los Angeles (UCLA), USA
Palak Jain, Google Research, India
John Kos, Design Intelligence Lab - Georgia Institute of Technology, USA
Konstantinos (Constantine) Kotropoulos, Aristotle University of Thessaloniki, Greece
Matt Kretchmar, Denison University, USA
Gerald Penn, University of Toronto, Canada
Ariadna Quattoni, UPC Barcelona, Spain
Kunal Rao, NEC Laboratories America, Inc., USA
Lina Rojas, Orange, France
Joni Salminen, University of Vaasa, Finland
Christelle Scharff, Pace University, USA
Kazim Sekeroglu, Southeastern Louisiana University, USA
Sumit Shekhar, Adobe Systems, India
Swati Tyagi, JP Morgan Chase & Co., Wilmington, DE, USA
Elena Umili, Sapienza University of Rome, Italy
Prajna Upadhyay, BITS Pilani Hyderabad Campus, India
Pierre Vilar Dantas, Federal University of Amazonas (UFAM), Manaus, Brazil

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

PosePrompt: Soft-Prompted Large Language Models for Natural Language Description of Human Motion from 2D Pose

Liyun Gong
School of Agri-food Technology and Manufacturing
University of Lincoln
Lincoln, UK
Email: lgong@lincoln.ac.uk

Oluwapelumi Alagbe
School of Engineering and Physical Sciences
University of Lincoln
Lincoln, UK
Email: 29404843@students.lincoln.ac.uk

Sheldon McCall
School of Engineering and Physical Sciences
University of Lincoln
Lincoln, UK
Email: ShMccall@lincoln.ac.uk

Philp Williams
Great Northern Physiotherapy
Lincoln, UK
Email: phil@greatnorthernphysiotherapy.co.uk

Huizhi Liang
School of Computing
Newcastle University
Newcastle upon Tyne, UK
Email: Huizhi.Liang@newcastle.ac.uk

Miao Yu
School of Engineering and Physical Sciences
University of Lincoln
Lincoln, UK
Email: myu@lincoln.ac.uu

***Abstract*—Natural language description of human motion enables interpretable motion understanding beyond categorical action labels. While motion–language models predominantly rely on 3D motion capture or parametric body models, such as A Skinned Multi-Person Linear Model (SMPL), these representations require specialised hardware or complex reconstruction pipelines. In contrast, 2D human pose estimation is widely accessible, yet remains underexplored for motion description. We propose PosePrompt, a parameter-efficient framework that generates natural language descriptions directly from 2D pose sequences using large language models. Spatiotemporal pose dynamics are encoded into compact motion features and injected into a pre-trained language model via soft prompt learning through multi-head attention layers, avoiding full model fine-tuning. This approach leverages strong linguistic priors while remaining effective in low-data settings. Experiments on a curated pose–text dataset demonstrate that PosePrompt produces accurate and coherent motion descriptions from 2D poses extracted from video sequences, with lightweight training requirements.**

Keywords- 2D poses; LLM; softprompt; multi-head attention.

I. INTRODUCTION

Understanding and describing human motion is a fundamental problem in computer vision, with applications spanning activity recognition, healthcare monitoring, human–computer interaction, and sports analysis. Natural language descriptions of motion provide semantic, human-interpretable explanations of movement patterns, enabling richer interaction and more advanced downstream reasoning beyond discrete action labels.

Recent advances in vision–language models and large language models (LLMs) have demonstrated strong capabilities in image and video understanding, including general video captioning and human motion description [1,2]. However, most existing LLM-based human description approaches rely on parametric 3D body models, such as SMPL, which require specialised motion-capture systems as well as heavy fine-tuning of LLMs, which incur substantial computational overhead, limiting their scalability and practical deployment.

To address these challenges, we propose **PosePrompt**, a parameter-efficient framework that adapts pre-trained LLMs to generate natural-language descriptions directly from 2D human pose sequences via soft prompt learning. Rather than relying on 3D motion representations such as SMPL, PosePrompt exploits 2D pose sequences, which capture the essential kinematic structure of human motion and have been widely adopted in action-recognition research. A lightweight pose encoder transforms the 2D skeletal sequences into compact spatiotemporal representations and produces a set of learnable continuous soft-prompt embeddings. These embeddings are injected into a pre-trained LLM to guide text generation. Crucially, this design avoids fine-tuning the entire language model, enabling efficient adaptation while preserving the strong linguistic priors of the pre-trained LLM.

We evaluate PosePrompt on a curated dataset of short human motion clips paired with textual descriptions, constructed from multiple public action datasets. Experimental results demonstrate that the proposed lightweight PosePrompt framework can generate accurate and semantically meaningful text descriptions using purely 2D

 ISBN: 978-1-68558-402-3

pose inputs, highlighting the potential of soft-prompted LLMs as an effective and scalable solution for interpretable human motion understanding from 2D skeletal data.

In Section 2, related work on pose-based action recognition and motion description is reviewed. In Section 3, the proposed PosePrompt methodology is described, including pose extraction, preprocessing, the pose encoder, and soft prompt generation. In Section 4, the experimental setup and results are presented. In Section 5, the conclusions and future work are provided.

II. RELATED WORKS

There exists a substantial body of work exploiting 2D and 3D human pose representations for motion understanding, particularly in the context of action recognition and, more recently, motion description.

Pose-Based Action Recognition:

Skeleton- and pose-based action recognition has been extensively studied as an alternative to RGB video, owing to its robustness to appearance variations, background clutter, and illumination changes. Early deep-learning approaches primarily employed recurrent architectures, such as long short-term memory (LSTM) networks, to model temporal dynamics in pose sequences for action classification [3]. These methods treat pose sequences as time series and capture motion patterns through sequential modeling of joint coordinates. Subsequent work introduced spatiotemporal graph convolutional networks (ST-GCN) [4] and their variants, which represent the human skeleton as a graph with joints as nodes and bones as edges. By explicitly modeling both spatial relationships between joints and their temporal evolution, these approaches more effectively capture structured motion dependencies than earlier time-series-based methods as in [3].

Beyond graph-based formulations, alternative representations have also been explored. Some approaches convert pose sequences into heatmap volumes or pseudo-images and apply 2D or 3D convolutional neural networks [5], achieving competitive performance while avoiding explicit graph construction. More recently, transformer-based architectures have been introduced for pose-based action recognition [5], leveraging self-attention mechanisms to model long-range temporal dependencies and complex inter-joint interactions across entire motion sequences. Despite these advances, most existing pose-based methods focus on predicting discrete action labels, offering limited semantic interpretability.

From Motion Labels to Motion Description:

Compared with action recognition and text-to-motion generation, motion-to-text description remains relatively underexplored. Early studies in this area predominantly adopted sequence-to-sequence learning paradigms. For instance, [6] employed an LSTM-based seq2seq framework to map 3D human motion sequences to natural language descriptions, while subsequent work incorporated generative adversarial learning into seq2seq models to better capture the diversity and realism of motion–language mappings [7].

More recently, driven by advances in LLMs, increasing attention has been paid to exploiting LLMs for motion interpretation using 3D motion representations, particularly those derived from the Archive of Motion Capture As Surface Shapes (AMASS) dataset and parameterized by SMPL or SMPL-X body models [8]. In particular, several recent works [1][2][9] explicitly treat human motion as a foreign language by discretizing continuous motion sequences into motion tokens and extending the vocabulary of language models. This formulation enables motion and text to be represented within a unified token space, allowing bidirectional translation between motion and natural language through end-to-end fine-tuning of LLMs.

While these LLM-based approaches have substantially advanced motion–language understanding, they typically rely on accurate 3D motion capture data or reliable estimation of SMPL parameters. Such requirements often involve specialised hardware, controlled capture environments, or computationally expensive reconstruction and training pipelines. Consequently, their applicability in unconstrained real-world settings—where only monocular RGB video is available—remains limited. This limitation motivates the exploration of motion-to-text description directly from more accessible representations, such as 2D pose sequences, without the need for heavy fine-tuning of large language models.

III. METHODOLOGY

Given a sequence of 2D human poses extracted and pre-processed from a video clip, the goal of this work is to generate a natural-language description that provides a semantic summary of the observed human motion. We address this task by leveraging a large language model enhanced through soft prompt learning, enabling effective adaptation to pose-based motion understanding, as illustrated in Figure 1. The proposed methodology is described below:

i. 2D pose extraction and pre-processing:

We employ the Detectron2 library [10] to extract 2D human poses from the original video sequences. Utilizing this library, we obtain the 2D skeleton, from which we extract the (x, y) coordinates of 12 key body points (including left/right shoulders, elbows, wrists, hips, knees, and ankles) for each video frame. After key point extraction by Detectron2, several preprocessing tasks are executed. Initially, we designate the midpoint of the shoulders as the origin and adjust the position of each key point accordingly to derive relative key point positions. Following this, we standardize the key point positions using the length of the trunk, defined as the distance between the midpoint of the shoulders and that of the hips, as a normalization factor. Upon completion of the preprocessing steps, an individual key point $\boldsymbol{p}_i$ is transformed according to the following formula:

$$\widehat{\boldsymbol{p}}_i = \frac{\boldsymbol{p}_i - \boldsymbol{p}_{mid_{shoulder}}}{L} \quad (1)$$

where $\boldsymbol{p}_{mid\_shoulder}$ represents the position of the middle of the left and right shoulders, and L represents the trunk length.

Through the utilization of the aforementioned preprocessing methods, we ascertain that the extracted 2D poses maintain invariance concerning scale and position.

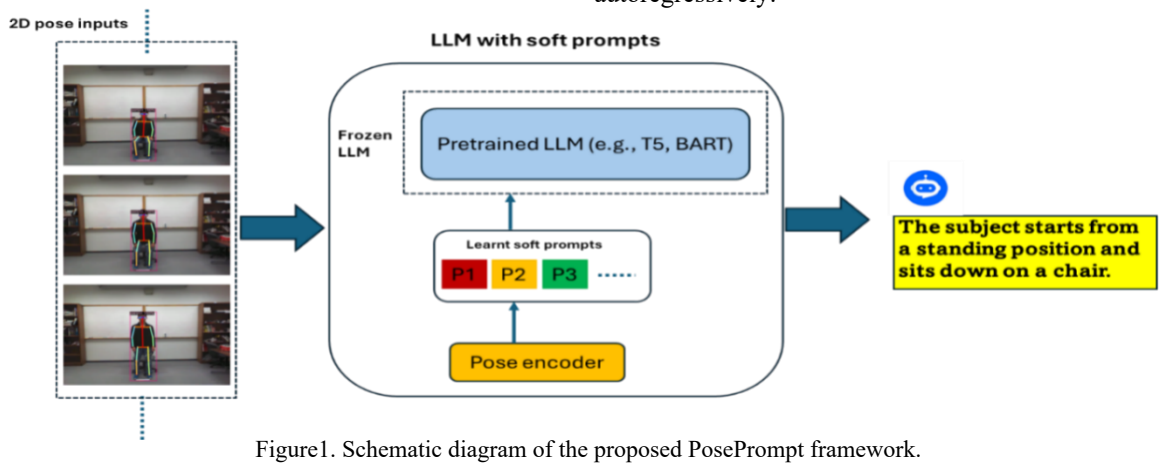


Figure1. Schematic diagram of the proposed PosePrompt framework.

Each extracted and processed 2D pose in a video sequence is concatenated and represented as

$$\mathbf{P} = \{\mathbf{p}_t\}_{t=1}^{T}, \mathbf{p}_t \in \mathbb{R}^{J\times 2}, \quad (2)$$

where $\boldsymbol{p}_t$ denotes the pre-processed 2D pose at time step t, consisting of $J = 12$ body joints. T represents the total number of frames in the sequence. The resulting pose sequence $\boldsymbol{P}$ serves as the input to the proposed PosePrompt framework as below.

ii. PosePrompt for motion to text

The proposed **PosePrompt** framework consists of three main components: i. Pose Encoder ii. Soft Prompt Generator and iii, Pre-trained Language Model, as shown in Figure 1.

Instead of training a motion-to-text model from scratch, we leverage a pre-trained large language model (LLM) and adapt it to pose-based motion interpretation using soft prompt learning. To capture motion dynamics from 2D poses, we employ a lightweight pose encoder f_{pose} that maps the input pose sequence $\mathbf{P}$ to a fixed-length feature representation:

$$\mathbf{z} = f_{\text{pose}}(\mathbf{P}), \mathbf{z} \in \mathbb{R}^{d}. \quad (3)$$

The pose encoder operates on normalized joint coordinates to model motion dynamics. In practice, f_{pose} can be implemented using a transformer-based encoder with positional embeddings [11], which can capture both short-term joint movements and longer-range temporal dependencies.

Given the pose feature $\mathbf{z}$, we generate a sequence of K soft prompt vectors:

$$\mathbf{S} = g(\mathbf{z}) \in \mathbb{R}^{K\times d_{\text{LM}}}, \quad (4)$$

where $g(\cdot)$ is a prompt projection network (e.g., a linear or MLP layer), and d_{LM} is the embedding dimension of the LLM. These soft prompts could be prepended to the input token embeddings of the language model, together with the textual tokens:

$$[\mathbf{S}; \mathbf{E}(x_1), \ldots, \mathbf{E}(x_n)], \quad (5)$$

where $\mathbf{E}(\cdot)$ denotes the token embedding function and x_i are textual tokens. Conditioned on the soft prompts, the pre-trained LLM could generate a natural language description autoregressively:

$$p(\mathbf{Y} \mid \mathbf{P}) = \prod_{l=1}^{L} p(y_l \mid y_{<l}, \mathbf{S}). \quad (6)$$

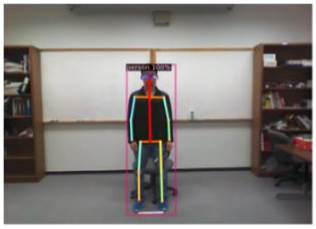
where $\mathbf{Y}$ represents the output texts.

Both the pose encoder and the prompt projection network are trained using a standard cross-entropy loss between the generated tokens and the corresponding ground-truth motion descriptions. This training objective enables the model to learn appropriate soft prompts from input 2D pose sequences for text generation, while keeping the large language model frozen. As a result, the proposed approach achieves efficient adaptation while preserving the rich linguistic knowledge of the pre-trained model.

IV. Experimental Results

We evaluated the proposed PosePrompt framework for motion-to-text generation using a dataset composed of samples from three public datasets: Weizmann, Intelligent Sensing Lab Dataset (ISLD), and Unbiasing through Textual Descriptions (UTD) [12]. In total, the dataset contains 530 video samples, each paired with a corresponding natural-language description. The data are split into 70% for training and 30% for testing.

The pose encoder adopts a multi-head attention architecture consisting of three attention layers, each with four heads and an output feature dimension of 256. The number of soft prompt vectors (denoted as K in Equation (4)) is set to 40 for each input sequence. Two pre-trained (LLM) backbones are investigated in this work: a lightweight T5-small model and a relatively larger Bidirectional and Auto-Regressive Transformer (BART) -base model [13]. The pose



Output text: The subject starts in a seated position and stands up to a full upright posture.

Output text: A person is bending down and reaching with one hand to pick up an object.

Output text: The subject stands upright, starts with their hands close to the chest, and pushes them forward.

Output text: The person walks forward at a steady pace.

Figure 2. Output text results from video clip samples

encoder and prompt projection networks are trained using the AdamW optimizer [14], with a learning rate of $2 \times 10^{-4}$and a weight decay of 0.01 with 1000 epochs.

After training, the PosePrompt framework is evaluated on the different test datasets. Qualitative results are illustrated in Figure 2, which shows example motion descriptions generated from the extracted 2D pose sequences. Furthermore, we conduct both subjective and objective evaluations across the entire test set. Subjective evaluation is performed by human comparison between the ground-truth descriptions to evaluate the text generation accuracy, while objective evaluation is conducted using the Bilingual Evaluation Understudy (BLEU) metric [15], as reported in Table I. The results indicate that the T5-based PosePrompt model achieves higher motion-to-text generation accuracy than its BERT-based counterpart, attaining over 90% accuracy of the generated text descriptions.

TABLE I. POSEPROMPT RESULTS BASED ON DIFFERENT LLM MODELS

	T5-small	***BART-base***
No. of parameters	~60M	~139M
Subjective Predict Accuracy	91.19%	88.68%
BLEU	92.01	88.19

V. CONCLUSIONS AND FUTURE WORK

We proposed PosePrompt, an efficient framework for generating natural language descriptions of human motion directly from 2D pose sequences using soft-prompted large language models. By conditioning a frozen LLM on pose-derived motion features, our approach avoids full model fine-tuning while effectively bridging skeletal representations and language generation. Experiments on a curated pose–text dataset demonstrate that PosePrompt could achieve accurate and coherent motion descriptions based on 2D poses. These results highlight the potential of soft-prompted LLMs for interpretable human motion understanding from widely available and privacy-preserving 2D pose data. Future work will extend PosePrompt to a broader range of large language model architectures and investigate its application to domain-specific tasks, such as gait analysis and healthcare assessment.

ACKNOWLEDGMENT

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 778602 ULTRACEPT.

REFERENCES

[1] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "MotionGPT: human motion as a foreign language*", NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, no. 880, Pages 20067 – 20079.

[2] Y. Wang et al., "MotionGPT-2: A General-Purpose Motion-Language Model for Motion Generation and Understanding," Online. Available from: https://arxiv.org/html/2410.21747v1

[3] J. Liu, G. Wang, P. Hu, L. Duan, and A. Kot, "Global Context-Aware Attention LSTM Networks for 3D Action Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[4] S. Yan, Y. Xiong, and D. Lin," Spatial-temporal graph convolutional networks for skeleton-based action recognition," *AAAI'18*, no. 912, pp. 7444-7452, 2018.

[5] M. Liu et al., "3D Skeleton-Based Action Recognition: A Review," [Online]. Available from: https://arxiv.org/abs/2506.00915

[6] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *Robotics and Autonomous Systems,* vol. 109, pp. 13-26,2018.

[7] Y. Goutsu, and T. Inamura, "Linguistic Descriptions of Human Motion with Generative Adversarial Seq2Seq Learning," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021.

[8] N. Mahmood, N. Ghorbani, N. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of Motion Capture As Surface Shapes," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019.

[9] C. Guo, S. Wang, and L. Cheng, "TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts," *European Conference on Computer Vision,* Tel Aviv, Israel, 2022.

[10] Y. Wu, A. Kirillov, F. Massa, WY. Lo, and R. Girshick, Detectron2, [Online]. Available from: https://github.com/facebookresearch/detectron2

[11] A. Vaswani, et al., Attention is All you Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[12] MPOSE2021, [Online]. Available from: https://github.com/PIC4SeR/MPOSE2021_Dataset
[13] Pretrained models, [Online]. Available from: https://huggingface.co/transformers/v4.6.0/pretrained_models.html
[14] I. Loshcgilov and F. Hutter, "Decoupled Weight Decay Regularization," *International Conference on Learning Representations* (ICLR 2019), LA, USA, 2019.
[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu , BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL), Philadelphia, 2002.

Mitigating Hallucinations in Retrieval-Augmented Generation through Stateful Agentic Workflows

Zahra Fakoor Harehdasht
Department of IT and Technology
IU International University of Applied Sciences
Erfurt, Germany
za.fakoor@gmail.com

Marius Schönberger
Department of IT and Technology
IU International University of Applied Sciences
Erfurt, Germany
marius.schoenberger@iu.org

Silke Vaas
Department of IT and Technology
IU International University of Applied Sciences
Erfurt, Germany
silke.vaas@iu.org

***Abstract*—Applications of Large Language Models often produce incorrect information, commonly referred to as hallucinations, which poses a significant challenge for enterprise knowledge management systems that rely on reliable automated responses. This paper investigates whether stateful, multi-agent workflows can mitigate such errors more effectively than conventional linear retrieval-augmented generation pipelines. The study contributes to the growing body of research on generative artificial intelligence and large language model applications by focusing on reliability, verification mechanisms, and the architectural design of retrieval-augmented generation workflows. The main contributions of this paper are a specialized multi-agent RAG architecture with dedicated agents for query rewriting, routing, hybrid retrieval, and verification, as well as a state-machine-based orchestration framework enabling iterative reasoning and conditional execution beyond linear pipelines. The architecture follows a two-stage mitigation strategy combining proactive retrieval refinement and reactive verification, including the ability to abstain when evidence is insufficient. It was evaluated through controlled experiments using synthetic enterprise datasets and compared it with a baseline linear retrieval-augmented generation pipeline. Performance was measured using an established evaluation framework for retrieval-augmented generation systems. The results demonstrate significant improvements across key evaluation metrics, including a 42.5% relative increase in faithfulness and substantial gains in answer accuracy and context recall. These improvements primarily result from separating error sources across specialized agents and the integration of automated verification steps. The findings highlight the importance of structured orchestration and verification mechanisms for enhancing the reliability of large language model applications. The results indicate that stateful, multi-agent architectures can increase the trustworthiness of enterprise generative artificial intelligence systems without requiring fine-tuning of the underlying language model.**

***Keywords*—Retrieval-Augmented Generation; Hallucination Mitigation; Large Language Models; Multi-Agent Systems**

I. INTRODUCTION

Enterprises increasingly rely on large collections of heterogeneous digital information, including technical documentation, operational runbooks, knowledge bases, and API documentation. Unlike traditional structured databases, this information often exists in semi-structured or unstructured formats, making effective retrieval challenging [1]. Conventional enterprise search systems typically rely on keyword matching, which performs poorly when user queries differ semantically from the terminology used in source documents. Such vocabulary mismatches can prevent relevant information from being retrieved and reduce the overall accessibility of organizational knowledge [2]. As a result, knowledge discovery remains inefficient despite the availability of large information repositories [3].

Recent advances in large language models (LLMs) have enabled new approaches for interacting with enterprise knowledge systems. Retrieval-Augmented Generation (RAG) combines LLMs with external document retrieval, allowing generated responses to be grounded in specific knowledge sources rather than relying solely on model training data [4]. While RAG improves semantic search and question answering capabilities, it remains vulnerable to hallucinations [5]. In enterprise environments, such errors can undermine trust and may lead to incorrect technical decisions or misleading information [6].

Conventional RAG systems typically follow a linear retrieve–then–generate pipeline, limiting iterative reasoning and structured verification [7]. Consequently, hallucinations are mainly addressed through retrieval improvements or prompt engineering, while architectural mitigation approaches remain underexplored [8]. To address this limitation, this paper proposes a stateful agentic workflow that decomposes the RAG process into specialized agents for query refinement, retrieval strategy selection, response generation, and output verification. These agents operate within a state-machine-based orchestration framework that enables iterative reasoning and validation before responses are returned to the user [9]. This architecture introduces structured verification and multi-agent coordination to mitigate hallucinations and improve the reliability of RAG-based enterprise knowledge systems [10], [11].

The remainder of this paper is structured as follows. Section 2 reviews related work on RAG, hallucination taxonomies, and agentic AI workflows. Section 3 introduces the proposed system architecture, including the data ingestion pipeline, the agentic graph design, and the hallucination mitigation strategy. Section 4 describes the evaluation methodology, outlining the experimental setup, baselines, metrics, and quantitative findings. Section 5 discusses the results and their technical and practical implications. Section 6 concludes the paper and highlights directions for future research.

II. Related Work

To provide a foundation for the present study and position it within the existing body of knowledge, the following section reviews relevant related work.

A. Retrieval-Augmented Generation

RAG, first introduced by Lewis et al. [4], combines LLMs with external knowledge retrieval to ground generated responses in updatable document sources. The architecture links a pre-trained Large Language Model (parametric memory) with an external document store (non-parametric memory), enabling access to domain-specific knowledge without retraining the model [4]. Before query time, documents are segmented into smaller chunks, which are converted into embedding vectors and stored in a vector index to enable similarity-based retrieval. As discussed by Gao et al. [10], the chosen chunking strategy significantly affects retrieval performance because segment size influences both semantic coherence and retrieval precision.

During inference, RAG typically follows a two-stage pipeline [4]. First, a retriever performs a similarity-based retrieval to identify the most relevant document chunks. These passages are then incorporated into the prompt presented to the language model, which generates an answer based on the context provided. To improve retrieval quality, several extensions have been proposed. Query transformation methods reformulate user queries to improve retrieval effectiveness [13], [14]. Hybrid search combines dense vector retrieval with lexical approaches, such as Best Matching 25 (BM25) [15], often merging rankings using techniques like Reciprocal Rank Fusion [16]. In addition, reranking methods refine initially retrieved candidates using more precise relevance models, including cross-encoders or LLM-based scoring approaches [17]. These techniques aim to improve the quality of retrieved evidence, which is critical for reducing hallucinations in RAG systems.

B. Hallucination Taxonomy

Hallucinations remain a persistent challenge in LLMs, even when RAG is used to ground responses in external knowledge sources. As noted by Ji et al. [5], hallucinations occur when generated outputs contain statements that are unsupported by the retrieved evidence or contradict the underlying sources.

Within RAG pipelines, Zhang & Zhang [18] distinguish between two primary categories of hallucination. Retrieval-induced hallucinations arise when the retrieval component fails to provide relevant or sufficient context. This can occur when irrelevant documents are retrieved, when passages lack critical information, or when the retrieved sources contain conflicting evidence. In such cases, the language model may rely on its internal parameters to infer missing information.

By contrast, generation-induced hallucinations originate during the generation stage. Even when relevant context is available, the model may misinterpret source material, or introduce unsupported details while producing fluent responses, making these errors difficult to detect [18].

These failure modes indicate that improving retrieval alone is insufficient for reliable RAG systems. Effective mitigation therefore requires architectural mechanisms that verify generated statements against retrieved evidence and enforce stronger grounding during the generation process.

C. Agentic AI Workflows

Managing the complexity of advanced RAG pipelines with a single monolithic architecture can be difficult. Multi-agent systems offer an alternative approach by decomposing complex tasks into smaller components handled by specialized agents. According to Russell & Norvig [19], an agent is an autonomous entity capable of perceiving its environment, making decisions, and acting toward defined goals. Wooldridge [20] further characterizes agents through properties such as autonomy, reactivity, pro-activeness, and the ability to interact with other agents.

Recent research has applied this paradigm to LLM–driven systems, often referred to as LLM-Agents [21]. In such architectures, complex workflows are divided into modular components that perform distinct tasks. Within RAG pipelines, these agents can assume roles such as query refinement, document retrieval, response generation, and verification, enabling clearer separation of responsibilities and more flexible system design.

To coordinate these interactions, frameworks such as LangGraph provide orchestration mechanisms for stateful agent workflows. Unlike linear pipelines, the graph-based architecture allows iterative and conditional execution paths, enabling agents to exchange information, trigger additional steps, or perform verification loops during runtime [9].

However, while agentic workflows have been proposed to structure complex LLM-driven systems, their application as architectural mechanisms for mitigating hallucinations in retrieval-augmented generation remains comparatively underexplored. This multi-agent perspective provides a flexible foundation for improving reliability in RAG systems, as it enables targeted interventions, such as verification steps, to mitigate hallucinations while maintaining modular and extensible system architectures.

III. System Architecture

The architecture is divided into an offline data ingestion path and an online multi-agent query pipeline (see Figure 1). Both components use a shared knowledge store implemented with PostgreSQL and the pgvector extension [22].

A. Data Ingestion

The system utilizes a modular Extract, Transform, Load (ETL) pipeline to convert heterogeneous inputs (i.e., PDFs and other documents) into a queryable format. Text is extracted from PDFs while preserving their original layout and hierarchical structure to maintain semantic coherence in complex documents. PostgreSQL with the pgvector extension is used as the unified knowledge store, enabling the storage of transactional metadata alongside 3072-dimensional embeddings generated using OpenAI's text-embedding-3-large model [22], [23].

Content is segmented using recursive character splitting [12], which partitions text along logical boundaries rather than limited arbitrary character counts, thereby preserving

semantic context. Each chunk is enriched with an LLM-generated headline to improve vector search accuracy, representing a form of proactive query transformation [13].

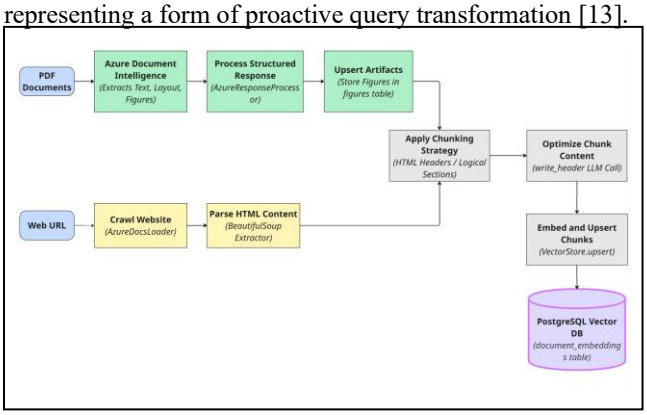


Figure 1. The dual-source data ingestion pipeline (adapted from [10])

To support multimodal queries, the ingestion pipeline, presented in Figure 1, extracts figures and tables and sorts their metadata and captions in a relational table linked to the corresponding text chunks via unique identifiers.

B. The Agentic Graph

The query pipeline is modeled as a stateful graph using LangGraph rather than a brittle linear RAG [9]. This architecture, illustrated in Figure 2, centers on a shared state object that functions as the system's short-term memory, which stores the rewritten query, the retrieved document sets, and the verification scores across all nodes. The shared state acts as a centralized blackboard where each agent writes its output (e.g., the reformulated query or retrieved chunks). This allows the Verification Agent to compare the final answer against the original query and the retrieved context stored earlier in the state, ensuring end-to-end consistency.

By leveraging this shared state, the system moves through nodes that make local decisions, which allows conditional branching and iterative loops instead of a fixed sequence. This allows the system to dynamically reason about and refine information quality before delivering the response. The pipeline consists of the following key nodes:

- Query Rewriter Agent: User prompts are often fuzzy or ambiguous. This agent reformulates user inputs into optimized search queries, improving semantic retrieval precision compared to standard keyword-based approaches [13].
- Router Agent: To optimize performance and cost, this agent analyzes the rewritten query and determines whether to query the internal vector store or external sources based on the detected intent.
- Hybrid Search: Semantic retrieval alone may miss exact technical terms such as error codes or product versions. The system therefore combines keyword-based search (for exact lexical matching and acronyms) with semantic search (for conceptual similarity and intent) executing both in parallel and merging the results to improve coverage and recall.
- Batch Reranker Agent: Initial retrieval may produce noisy or irrelevant results. This agent uses an LLM to score the relevance of retrieved candidates in a single

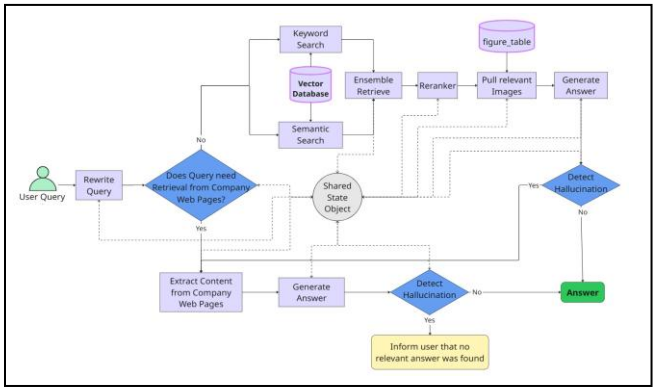


Figure 2. Proposed architecture for RAG hallucination mitigation via agentic workflow.

batch call [17], reducing noise and ensuring that the Generation Agent receives high-quality contextual input.

To ensure consistent behavior, each agent utilizes structured system prompts with strict output constraints. For example, the Verification Agent is prompted: 'Given the context [C], does the answer [A] contain any statements not explicitly supported? Answer only with a JSON object: {is_hallucination: bool, confidence: float}.' Similar constraints are applied to the Query Rewriter and Reranker to ensure the state object receives parseable data for programmatic decision-making.

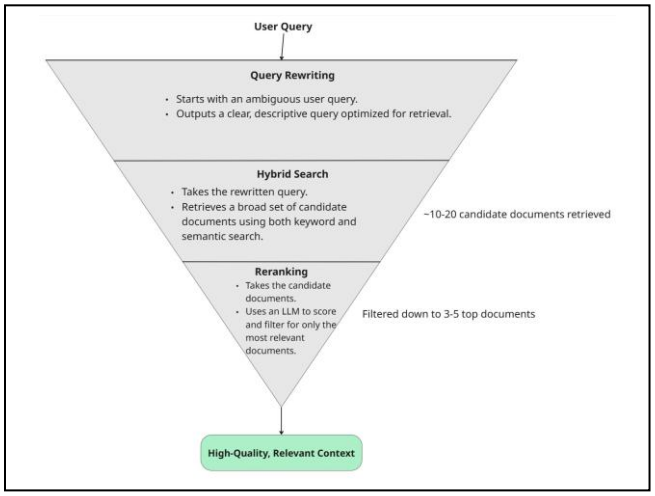


Figure 3. Proposed architecture for RAG hallucination mitigation via agentic workflow (adapted from [16])

C. Hallucination Mitigation Strategy

The core contribution of this architecture is a two-layered hallucination mitigation framework:

The first layer, *proactive mitigation,* targets retrieval-induced hallucinations [18] by reducing noise that may mislead the LLM. Through query rewriting and document

reranking, the system ensures that the generator receives a compact, high-quality set of relevant passages. Figure 3 illustrates this process as a filtering funnel, where each stage progressively refines the context provided to the generator.

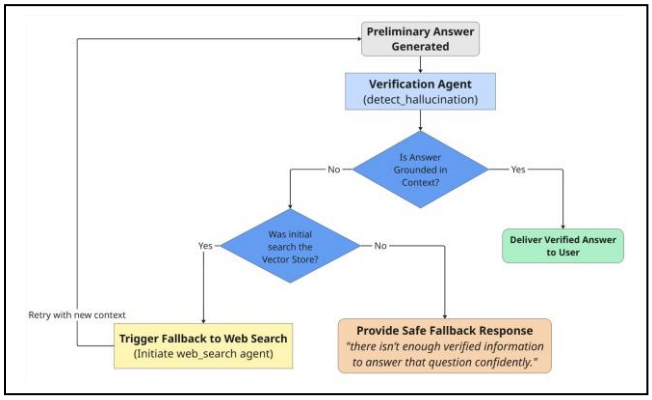


Figure 4. The Reactive Verification and Fallback Loop for hallucination mitigation (adapted from [18])

The second layer, reactive mitigation, addresses generation-induced hallucinations [18]. After a draft answer is generated, it is passed to a Verification Agent that evaluates the response against the reranked source documents. The agent acts as an automated fact-checker and emits a boolean flag (*is_hallucination*) along with a confidence score. If a hallucination is detected, the system withholds the response and triggers a fallback mechanism, either retrieving additional evidence from external sources or returning a safe "insufficient evidence" response. This process is illustrated in Figure 4.

IV. EVALUATION

The evaluation was designed to assess whether the proposed stateful agentic workflow reduces hallucinations and improves answer accuracy in a technical enterprise setting.

A. Experimental Setup

To ensure a controlled environment, a fresh PostgreSQL database was populated with synthetic, enterprise-style documents. These documents covered four categories: IT security, human resource policy, technical training, and employee onboarding. The synthetic dataset consisted of 150 documents totaling approximately 600 pages of text. The 'Technical Training' category specifically included 15 multimodal documents containing images, such as server architecture diagrams and network flowcharts. The content was designed to mimic internal wikis, including semi-structured tables for HR policies and unstructured technical logs for IT security.

Two models from OpenAI were used, *GPT-4o* for agentic reasoning, and *text-embedding-3-large* for vector generation [23]. A curated evaluation dataset of questions was used to test different retrieval paths and capabilities, both straightforward fact retrieval and more complex synthesis tasks. The 50 evaluation queries were classified as follows: Text-based (35 queries, e.g., 'What are the steps for employee onboarding?') and Multimodal/Tabular (15 queries, e.g., 'According to the IT diagram, which port is used for the API gateway?'). The query distribution across document categories was as follows: IT Security (15), HR Policy (12), Technical Training (13), and Onboarding (10).

B. Baselines

Two configurations were compared:

- Baseline RAG: A simplified, linear pipeline with a basic semantic retriever and a generator LLM.
- Agentic RAG (the proposed system): The full multi-agent system, which includes the Query Rewriter, Batch Reranker, and the reactive verification loop.

C. Metrics

Performance was evaluated using the Retrieval-Augmented Generation Assessment (RAGAS) framework [11], using GPT-4 as the evaluator LLM. RAGAS provides a set of metrics designed to measure RAG pipelines [11]:

- *Faithfulness*: Measures whether the answer is grounded solely in the retrieved context and serves as the primary indicator for hallucination.
- *Answer Correctness*: The factual accuracy of the response relative to the ground truth.
- *Context Recall*: Evaluates whether the retriever successfully identified all necessary information required to answer the query.
- *Answer Relevancy*: Measures how well the generated response addresses the intent of the user's prompt.

D. Quantitative Findings

The performance of both systems was measured across a set of 50 different queries. The overall results, shown in Table I., represent the average score across all evaluations, where 1.00 indicates a perfect score.

TABLE I. Comparative RAGAS Results

Metric	Baseline RAG	Agentic RAG (Artifact)
Faithfulness	0.59	0.84
Answer Correctness	0.66	0.89
Context Recall	0.65	0.86
Answer Relevancy	0.74	0.93

These quantitative results indicate a consistent performance advantage for the agentic workflow. A detailed analysis of how these specific architectural components contributed to these increases in scores is presented in the following section.

V. RESULTS AND DISCUSSION

Building on the conducted evaluation, the following section presents the obtained results and discusses their implications regarding hallucination reduction and answer accuracy.

A. Retrieval-Augmented Generation Assessment scores

The quantitative evaluation in Table 1 demonstrates consistent improvements across all RAGAS dimensions when comparing the linear baseline RAG pipeline to the proposed stateful agentic architecture.

Most notably, faithfulness increased from 0.59 to 0.84, representing a relative improvement of 42.5%. This substantial gain indicates that the proposed architecture effectively reduces both retrieval- and generation-induced hallucinations.

The improvement in answer correctness (0.66 → 0.89) shows that stronger contextual grounding leads to higher factual accuracy. Importantly, correctness did not increase at the expense of relevancy, which improved from 0.74 to 0.93. This suggests that query rewriting and reranking stages do not merely constrain outputs conservatively but align retrieval more closely with user intent.

The rise in context recall (0.65 → 0.86) further supports the claim that the combination of hybrid retrieval and reranking effectively reduces hallucinations caused by insufficient or noisy context. In baseline RAG systems, such conditions often force the generator to interpolate from parametric memory. By increasing retrieval coverage and reducing the noise prior to generation, the agentic system shifts the process from generative inference to evidence-based synthesis.

A qualitative analysis of execution paths revealed that straightforward factual queries (e.g., 'What is the standard employee notice period?') typically followed a linear path with high initial retrieval confidence. In contrast, ambiguous or complex queries (e.g., 'Cross-reference recent security guidelines with internal protocols to identify compliance gaps') frequently triggered the Query Rewriter to iterate and the Router to pull from external web sources when the internal vector store returned low-confidence scores. However, even with these adaptive paths, the system still faced challenges with multi-hop reasoning and queries requiring information spread across multiple disparate documents. In these cases, the Batch Reranker occasionally filtered out a necessary 'linking' document if its individual relevance score appeared low, contributing to the gap between Context Recall (0.86) and the higher-performing metrics like Answer Relevancy (0.93).

Beyond the primary evaluation using GPT-4o, supplementary tests were conducted using other large language models, specifically Claude 3.5 Sonnet and the open-source Llama 3.1 (70B). These preliminary runs showed that the relative performance gains remained consistent across different model architectures. While baseline scores varied, the agentic workflow consistently improved the faithfulness and accuracy of the output compared to linear pipelines, confirming that the benefits of stateful orchestration are consistent across models.

Overall, the improvements across all metrics indicate that effective hallucination mitigation emerges primarily from architectural orchestration rather than from improvements in the underlying foundation model.

B. Why the agentic system outperformed the baseline

The performance improvements observed can be attributed to three structural characteristics of the system.

1) Architectural Separation of Failure Modes

Linear RAG pipelines implicitly assume that retrieval quality alone determines output reliability [4]. In contrast, the proposed system models hallucination as a two-class problem: retrieval-induced and generation-induced. The proactive mitigation layer addresses the former by improving context quality prior to generation, while the reactive verification layer targets the latter through post-generation validation. This operationalizes the hallucination taxonomy [18] within an executable workflow.

2) Graph-Based Orchestration

Unlike linear pipelines, the system maintains a shared workflow state that propagates rewritten queries and intermediate results. This enables conditional branching, iterative refinement, and fallback strategies. As a result, RAG is transformed from a feedforward pipeline into a controlled, stateful optimization process.

3) Verification as Confidence Calibration

The verification agent acts as a control mechanism by evaluating the consistency between generated claims and source documents. This shifts the optimization objective from plausibility to evidence.

Importantly, the verification loop allows the system to abstain (“insufficient evidence”) rather than generate unsupported content, which contributes significantly to the observed improvements in accuracy.

Baseline RAG systems lack such a mechanism and therefore tend to produce overly confident responses under uncertainty.

C. Technical and Practical Implications

The results provide several insights for the deployment of retrieval-augmented generation systems in enterprise environments.

First, they suggest that hallucination mitigation should be understood primarily as an architectural challenge rather than a purely model-level problem. Simply improving the underlying LLM is not sufficient to ensure reliable results [5]. Instead, improvements in reliability result from structured orchestration of retrieval and generation components [12]. In particular, the coordinated use of query optimization, retrieval refinement, reranking, and verification significantly strengthens the factual grounding of the generated responses.

From a practical perspective, this architectural approach offers an additional advantage: reliability improvements can be achieved without modifying or fine-tuning the underlying base model. This enables a model-agnostic design, allowing organizations to switch or upgrade LLM providers without redesigning the surrounding system. Such flexibility is particularly valuable in enterprise settings where model capabilities, costs, and vendor strategies evolve rapidly.

Second, the introduction of a verification loop substantially increases the system’s trustworthiness. In compliance-sensitive areas such as human resources or IT security, incorrect or unsupported statements can lead to

operational or legal risks. A verification layer that validates generated responses against retrieved evidence therefore serves as an important safeguard, preventing unfounded or fabricated statements from being returned to the user.

A potential risk in this architecture is 'error propagation,' where an initial misjudgment by the Query Rewriter or Reranker biases all subsequent agents. Without human-in-the-loop validation, the system relies on the LLM's internal 'self-correction' capability. Future iterations should include a human-flagging mechanism for low-confidence outputs.

Overall, these findings indicate that the reliability of enterprise RAG systems depends less on increasingly powerful language models and more on carefully designed retrieval, orchestration and verification processes.

VI. Conclusion and Future Work

Building upon the presented results and discussion, the final section draws conclusions, addresses limitations, and identifies opportunities for future work.

A. Architectural Contributions to Reliable RAG Systems

This study demonstrates that hallucination reduction in retrieval-augmented generation systems can be significantly improved through architectural design. In particular, separating error modes via a two-stage mitigation strategy, combined with a graph-based orchestration, and a verification agent capable of abstaining when evidence is insufficient, substantially enhances system reliability.

These improvements are achieved without fine-tuning the underlying foundation model, highlighting that central role of architectural design in ensuring reliable system behavior.

From a design science research perspective, the resulting artifact offers both a functional prototype for use in companies and a conceptual model for structuring hallucination-aware RAG systems.

B. Trade-off: Reliability vs. Latency

The improved reliability of the proposed architecture comes at the cost of additional computational overhead. Additional agents, reranking steps, and the verification loop introduce the number of model calls and, consequently, higher response latency compared to a minimal linear RAG pipeline. This reflects an inherent trade-off between response speed and output reliability.

C. Directions for Future Research

In enterprise settings where incorrect information may lead to operational, financial, or compliance risks, prioritizing reliability over minimal latency is often justified. Nevertheless, future implementations may benefit from adaptive strategies that dynamically adjust verification depth based on query criticality, enabling a balance between efficiency and robustness. Several directions for future research emerge from this work. One promising avenue is the integration of knowledge graphs into the architecture. While the current pipeline primarily operates on unstructured text, extending it to incorporate structured representations - such as entities (i.e., services, products, or infrastructure components) - could enable more advanced multi-hop reasoning across documents. However, this approach introduces additional challenges, including schema design and entity disambiguation, which must be addressed carefully to maintain system reliability [24].

Another direction involves domain-specific fine-tuning of core models or reranking components, which may enhance grounding in specialized enterprise corpora such as technical documentation or regulatory guidelines.

Finally, future studies could evaluate agentic RAG architectures on larger and more diverse real-world datasets to better assess scalability and robustness under production conditions. Such investigations would provide deeper insights into the performance of architectural hallucination mitigation strategies in dynamic enterprise knowledge environments.

References

[1] A, McAfee and E. Brynjolfsson, “Big data: The management revolution,” Harvard Business Review, vol. 90, no. 10, pp. 61–68, Oct. 2012. [Online]. Available from https://hbr.org/2012/09/big-datas-management-revolutio, 2026.03.27.

[2] G. W. Furnas, T. K. Landauer, L. M. Gomez and S. T. Dumais, “The vocabulary problem in human-system communication,” Communications of the ACM, vol. 30, no. 11, pp. 964-971, 1987, doi: 10.1145/32206.3221.

[3] M. Alavi and D. E. Leidner, “Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues,” MIS Quarterly, vol. 25, no. 1, pp. 107-136, 2001, doi: 10.2307/3250961.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M.Lewis, W-T.Yih, T. Rocktäschel, S. Riedel and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020, doi: 10.48550/arXiv.2005.11401.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H.S. Chan, A. Madotto and P. Fung, “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023, doi: 10.1145/3571730.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82-115, 2020, doi: 10.48550/arXiv.1910.10045.

[7] S. Mishra, S. Niroula, U. Yadav, D. Thakur, S. Gyawali, and Shiva Gaire, “SoK: Agentic Retrieval-Augmented Generation (RAG): Taxonomy, Architectures, Evaluation, and Research Directions, arXiv preprint, 07 Mar. 2026, doi: 10.48550/arXiv. 2603.07379 [Online]. Available from https://arxiv.org/html/2603.07379v1. 2026.03.27.

[8] L.M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, F. Seidel, “Retrieval augmented generation for large language models in healthcare: A systematic review”. PLOS Digit Health. 2025 Jun 11;4(6):e0000877. doi: 10.1371/journal.pdig.0000877. PMID: 40498738; PMCID: PMC12157099. [Online]. Available from https://pmc.ncbi.nlm.nih.gov/articles/PMC12157099/, 2026.03.27.

[9] LangChain, “LangGraph,” [Online]. Available from https://python.langchain.com, 2026.03.27.

[10] Confident AI Inc., *DeepEval. The LLM Evaluation Framework*. [Online]. Available from https://deepeval.com, 2026.03.27.

[11] S. Es, J. James, L. E. Anke, and S. Schockaert, “RAGAs: Automated evaluation of retrieval augmented generation,” Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics System Demonstrations (EACL 2024), March 2024, pp. 150-158, doi: 10.18653/v1/2024.eacl-demo.

[12] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint, pp. 1-24, 2024, doi: 10.48550/arXiv.2312.10997.

[13] X. Ma, Y. Gong, P. He, H. Zhao and N. Duan, “Query Rewriting for Retrieval-Augmented Large Language Models,” Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), Dec. 2023, pp. 5303-5315, doi: 10.18653/v1/2023.emnlp-main.322.

[14] H. S. Zheng, S. Mishra, X. Chen, H. T. Cheng, E. H. Chi, Q. V. Le and D. Zhou, “Take a step back: Evoking reasoning via abstraction in large language models,” arXiv preprint, pp. 1-38, Mar. 2024, doi: 10.48550/arXiv.2310.06117.

[15] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009, doi: 10.1561/1500000019.

[16] G. V. Cormack, C. L. Clarke and S. Buettcher, “Reciprocal rank fusion outperforms condorcet and indi-vidual rank learning methods,” Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 758-759. [Online]. Available from https://cormack.uwaterloo.ca/cormacksigir09-rrf, 2026.03.27.

[17] N. Thakur, N. Reimers, A. Rucklé, A. Srivastava and I. Gurevych, “BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models,” arXiv preprint, pp. 1-24, Oct. 2021, doi: 10.48550/arXiv.2104.08663.

[18] W. Zhang and J. Zhang, “Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review,” Mathematics, vol. 13, no. 5, 856, 2025. doi: 10.3390/math13050856.

[19] S. J. Russell and P. Norvig, Artificial intelligence: A modern approach. 4th ed., Harlow, UK: Pearson Education Limited, 2022.

[20] M. Wooldridge, An introduction to multiagent systems. 2nd ed., West Sussex, UK: John Wiley & Sons.

[21] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, ... and T. Gui, “The rise and potential of large language model based agents: A survey,” Science China Information Sciences, vol. 68, no. 2, 121101, 2025, doi: 0.1007/s11432-024-4222-0.

[22] A. Lamba, “pgvector (v0.7.2),”, GitHub. [Online]. Available from https://github.com/pgvector/pgvector, 2026.03.27.

[23] OpenAI, “New embedding models and API updates,” 2024. [Online]. Available from https://openai.com/index/new-embedding-models-and-api-updates, 2026.03.27.

[24] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, “Unifying Large Language Models and Knowledge Graphs: A Roadmap,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 7, pp. 3580-3599, July 2024, doi: 10.1109/TKDE.2024.3352100.

Evaluating Policies to Improve Relational-Level Learning in Engineering Mathematics: An Approach based on Machine Learning and the Monte Carlo Method

Isaac Caicedo-Castro*†‡ Rubby Castro-Púche†§ Oswaldo Vélez-Langs*‡
*Socrates Research Team
†Research Team: Development, Education, and Healthcare
‡Department of Systems and Telecommunications Engineering
§ Department of Social Sciences
University of Córdoba
Carrera 6 No. 76-305, 230002, Montería, Colombia
e-mail: {isacaic | rubycastro | oswaldovelez}@correo.unicordoba.edu.co

***Abstract*—In this research, we combine machine learning and Monte Carlo simulation to estimate the probability that Systems Engineering students at the University of Córdoba fail to attain the relational level of the Structure of the Observed Learning Outcomes (SOLO) taxonomy in mathematics courses. Attainment of this level indicates the ability to integrate mathematical concepts and techniques in order to solve engineering problems. Logistic Regression is employed to model this probability using measures of student self-efficacy. Because the available dataset provides only limited coverage of the input space, Monte Carlo simulation is used to generate additional student profiles and evaluate counterfactual intervention scenarios aimed at improving self-efficacy. The results indicate that the baseline probability of failing to attain the relational level is 56.94%. Among the individual interventions considered, reducing mathematics anxiety yields the largest reduction in this probability, followed by strengthening mastery-learning beliefs and improving perceived teaching support. The simultaneous implementation of all interventions produces the greatest effect, reducing the estimated probability to 47.69%. These findings suggest that the combination of machine learning and Monte Carlo simulation provides a computational framework for evaluating educational interventions prior to their implementation and supports evidence-based decision-making in engineering education.**

Keywords-machine learning; logistic regression; support vector machines; Gaussian processes for classification; educational data mining; Monte Carlo method; learning outcomes.

I. Introduction

Learning applied mathematics is challenging across educational levels, from elementary education to postgraduate study. This difficulty stems from both the formal nature of mathematics and the complexity involved in translating real-world situations into mathematical representations. Consequently, engineering students often experience difficulties in integrating concepts, techniques, and methods when addressing problems within their respective fields, such as industrial, electrical, and mechanical engineering.

This research was conducted within the Bachelor's programme in Systems Engineering at the University of Córdoba, Colombia, where an Outcome-Based Education (OBE) approach is integrated with the Structure of the Observed Learning Outcomes (SOLO) taxonomy. Within this framework, learning achievement is classified into five levels according to the grade scale [0, 5], as follows:

i) Prestructural (0.0–2.0): The student has not yet demonstrated an understanding of the key concepts.
ii) Unistructural (2.1–2.9): The student demonstrates an understanding of a single aspect of the task.
iii) Multistructural (3.0–3.7): The student demonstrates an understanding of several aspects of the task, although these are not yet integrated.
iv) Relational (3.8–4.5): The student is able to integrate multiple aspects of the task into a coherent whole.
v) Extended Abstract (4.6–5.0): The student demonstrates a deep understanding of the underlying concepts and is able to apply them in novel contexts.

A student is deemed to have passed an assessment upon achieving at least the multistructural level, corresponding to a grade of 3.0 or higher. Similarly, progression within a course requires a final grade of at least 3.0. However, the development of effective engineering problem-solving skills is generally associated with the relational and extended abstract levels, particularly in mathematics-related courses.

Accordingly, the first objective of this research is to estimate the probability that a student fails to attain at least the relational level in mathematics as a function of self-efficacy, defined as an individual's belief in their capability to perform actions required to achieve a particular outcome [1]. Self-efficacy is shaped by several sources, including mastery experiences, social modelling through the observation of peers, social persuasion such as instructor feedback, and emotional and motivational states.

Regarding the mathematical notation, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$ be a dataset comprising observations of students' self-efficacy perceptions and their performance in engineering mathematics. The ith student is represented by a real-valued D-dimensional feature vector $x_i \in \mathcal{X} \subseteq \mathbb{R}^D$, whose jth component corresponds to a particular self-efficacy attribute. For example, x_{ij} may represent the extent to which the ith student believes they can successfully solve most mathematics problems during an assessment.

The corresponding target variable is defined as $y_i = 1$ if the ith student failed to attain the relational level, as determined by the mean grade obtained in previous mathematics courses, and

$y_i = 0$ otherwise. Here, M denotes the number of observations in the dataset.

Given a new student represented by the feature vector x_k, the objective is to estimate the probability $P(\hat{y}_k = 1 \mid x_k)$, where $\hat{y}_k = 1$ indicates that the model predicts the student will fail to attain the relational level. Here, x_k denotes the input vector, while y_k and $\hat{y}_k$ represent the observed and predicted outcomes, respectively.

A further objective of this research is to estimate the probability of failing to attain the relational level under a range of hypothetical interventions designed to reduce this risk. To this end, we pondered the following research questions:

i) How would this probability change if students perceived greater teaching support?
ii) How would this probability change if mathematics anxiety were reduced?
iii) How would this probability change if mastery confidence were increased?
iv) How would this probability change if greater teaching support, reduced mathematics anxiety, and increased mastery confidence were considered simultaneously?

To address the two objectives outlined above, a dataset comprising 93 observations and 26 variables was collected (i.e., $M = 93$ and $D = 26$).

Several machine learning methods were evaluated for estimating the probability that a student x_k fails to attain the relational level, that is, $P(\hat{y}_k = 1 \mid x_k)$. Across the 10-fold cross-validation experiments, Logistic Regression achieved the highest accuracy among the methods considered. A similar pattern was observed for precision.

Finally, the Monte Carlo method is employed to simulate the proposed intervention scenarios and thereby address the research questions outlined above. According to the simulation outcomes, the baseline probability of failing to attain the relational level in mathematics is estimated at 57.94%. The simulation results suggest that this estimate decreases to 54.54% under improved teaching support, 53.64% under reduced mathematics anxiety, and 53.99% under increased mastery confidence. The largest reduction is observed when all three interventions are considered simultaneously, yielding an estimated probability of 47.69%.

The remainder of the paper is organised as follows. Section II reviews the relevant literature, Section III describes the methodology, and Sections IV and V present the results and discussion. Section VI concludes the paper.

II. Prior Research

Educational data mining methods have been widely employed to predict student dropout, delayed graduation [2][3][4], and the risk of course failure or withdrawal [5][6][7][8][9][10][11][12][13][14]. In contrast, we examine the relationship between self-efficacy and engineering students' attainment of the relational level in mathematics, which reflects their ability to integrate mathematical concepts and apply them to problem-solving tasks within the SOLO framework.

Self-efficacy, introduced by Albert Bandura in 1986, refers to an individual's belief in their capability to successfully achieve a particular level of performance [1].

The relationship between self-efficacy and mathematics has been extensively investigated in the literature [15][16][17][18][19][20]. Similar associations have also been reported for reading skills [21].

Several of these studies have focused on mathematics at the undergraduate level [15][16][20], whereas others have examined self-efficacy in middle and secondary school settings [17][18][19].

Our research differs from prior one in three principal respects.

First, the questionnaire employed in our research captures broader beliefs regarding mathematical self-efficacy, rather than focusing on task-specific skills such as histogram construction [20].

Second, the relationship between self-efficacy and the outcome variable is formulated in our research as a classification problem, whereas previous studies have typically adopted a regression framework. In addition, our analysis is based on an educational process extending over at least two and a half years, reflecting the period during which students might complete the core mathematics curriculum. Our research further combines empirical observations with simulation to investigate intervention scenarios that would be difficult to evaluate experimentally because of the time, resources, and ethical considerations involved.

Third, our analysis focuses specifically on engineering mathematics, thereby providing a more clearly defined application context than that considered in related recent work [20].

A similar combination of machine learning and Monte Carlo methods has previously been employed to analyse the risk of failing physics courses [22][23] and student demotivation in scientific computing [24][25].

Logistic Regression and the Monte Carlo method have been applied to estimate the risk of failing physics courses based on students' academic progress during a semester [23]. This work extends an earlier study in which Monte Carlo simulation alone was used for the same purpose on a smaller dataset [22]. In that context, student progress is measured using at most two grades, making visualisation feasible. By contrast, our research considers a multivariate input space comprising more than three input variables. Consequently, both studies combine predictive modelling and Monte Carlo simulation, although we additionally evaluate Gaussian Processes and Support Vector Machines alongside Logistic Regression.

Prior research has also combined regression models with Monte Carlo simulation to analyse the risk of student demotivation within a multivariate setting [24][25]. While both lines of research employ supervised learning methods, our research addresses a classification problem, whereas the aforementioned studies focus on regression.

The Monte Carlo method has also been applied to predict student performance in engineering, medical, and mixed student cohorts, where it was found to outperform linear regression,

particularly in capturing non-linear grade transitions and stochastic variability [26].

To our knowledge, the combination of methods adopted in our research has not previously been applied to investigate the relationship between self-efficacy and the probability of attaining the relational level in engineering mathematics.

III. Research Methodology

In this research, supervised machine learning methods were used to estimate the probability that the kth student fails to attain the relational level in mathematics, which is denoted as $P(\hat{y}_k = 1 \mid x_k)$. The analysis considered probabilistic classifiers, including Logistic Regression and Gaussian Processes [27], together with Support Vector Machines. Although Support Vector Machines do not provide probabilistic outputs inherently, their predictions might be calibrated to yield probability estimates [28]. Further details of these methods may be found in [29][30].

A dataset was collected at the beginning of 2026 to train the models described above. The corresponding questionnaire was designed such that its items represented the input variables included in the dataset. Responses were recorded on a Likert scale ranging from 0 (*strongly disagree*) to 5 (*strongly agree*).

The questionnaire was administered to 93 students enrolled between the fifth and eighth semesters of the programme, which is designed to be completed over 10 semesters. Participant identities were anonymised prior to analysis. Table I presents the questionnaire and the correspondence between its items and the 26 input variables of the models, represented by the components of the vector x_i, for $i = 1, \ldots, 93$.

The target variable is derived from the mean grade obtained across the core mathematics courses of the programme, namely Calculus I (Differential Calculus), Calculus II (Integral Calculus), Calculus III (Vector Calculus), Differential Equations, and Linear Algebra. The distribution of the average grades used to define the target variable is presented in Figure 1, while Figure 2 shows the resulting class distribution.

As the dataset is not perfectly balanced, model performance is evaluated using stratified 10-fold cross-validation. This procedure partitions the data into 10 folds, enabling model training and evaluation across 10 iterations. The distribution of records within each test fold is reported in Table II. Prior to model fitting, all input variables are standardised by centring them to zero mean and scaling them to unit variance.

Within the collected dataset, the proportion of students who failed to attain the relational level is 56.99%, corresponding to $P(y = 1) = 56.99\%$. However, Figure 3 shows that the input variables do not uniformly cover the input space, with some values unobserved in the available data. For example, the variables $x_{i,1}$, $x_{i,2}$, $x_{i,3}$, and $x_{i,15}$ never take the value 2. To explore regions of the input space that are sparsely represented or absent from the observed dataset, the Monte Carlo method is employed [31], assuming a multivariate normal distribution for the input variables.

The second objective of the research is addressed by estimating the probability that students fail to attain the relational level under a range of intervention scenarios. To this end, the Monte Carlo method is employed to simulate the implementation of these interventions.

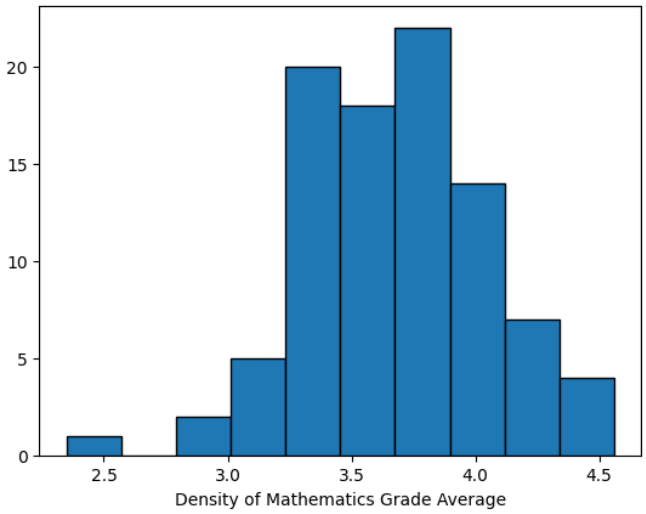


Figure 1. Histogram of the students' Average Grade obtained in mathematics courses

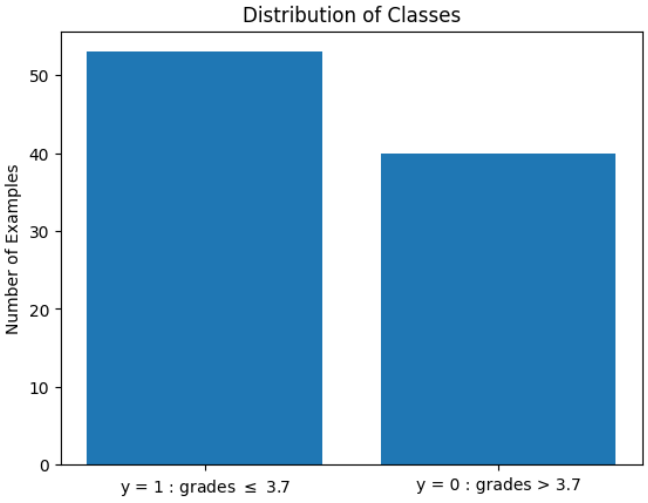


Figure 2. 53 out of 93 students (57%) did not reach the relational level in mathematics

Formally, the estimation of $P(\hat{y} = 1)$ requires the evaluation of the following integral:

$$P(\hat{y} = 1) = \int_{\mathcal{X}} P(\hat{y} = 1 \mid x) P(x) dx. \quad (1)$$

As the integral cannot be evaluated analytically, the Monte Carlo method is employed to approximate it. To this end, a matrix $X \in \mathbb{R}^{N \times D}$ is generated, whose rows correspond to samples drawn from a multivariate normal distribution with zero mean and covariance matrix S. The zero-mean assumption reflects the prior standardisation of the input variables. Here, N denotes the number of generated samples and D the number of input variables.

TABLE I. DESCRIPTION OF THE DATASET COLLECTED IN THIS RESEARCH

Input Variable	Question	Mean	Standard Deviation	Minimum Value	Maximum Value	Mode	Median
$x_{i,1}$	I can successfully solve most of the mathematics problems assigned to me.	3.65	0.71	2.00	5.00	4.00	4.00
$x_{i,2}$	If I try hard enough, I can complete even difficult mathematics tasks.	4.23	0.69	2.00	5.00	4.00	4.00
$x_{i,3}$	I can learn new mathematics topics without needing constant help.	3.62	0.83	2.00	5.00	4.00	4.00
$x_{i,4}$	When I make mistakes in mathematics, I can identify them and improve.	3.96	0.70	3.00	5.00	4.00	4.00
$x_{i,5}$	I am confident in my ability to understand complex mathematical concepts.	3.63	0.88	1.00	5.00	4.00	4.00
$x_{i,6}$	I often have difficulty successfully completing mathematics problems.	3.09	0.95	1.00	5.00	3.00	3.00
$x_{i,7}$	Compared with most of my classmates, I am able to achieve good results in mathematics.	3.43	0.74	2.00	5.00	3.00	3.00
$x_{i,8}$	I learn mathematics as quickly as other students in my class.	3.60	0.88	2.00	5.00	4.00	4.00
$x_{i,9}$	Seeing other students succeed in mathematics increases my belief that I can succeed as well.	4.19	0.95	1.00	5.00	5.00	4.00
$x_{i,10}$	When other students understand mathematics faster than I do, I begin to doubt my own ability.	2.54	1.17	1.00	5.00	3.00	3.00
$x_{i,11}$	I feel that my mathematical ability is inferior to that of most of my classmates.	2.58	1.08	1.00	5.00	3.00	3.00
$x_{i,12}$	My teachers believe that I can succeed in mathematics.	3.60	0.87	1.00	5.00	3.00	4.00
$x_{i,13}$	When my teacher encourages me, my confidence in mathematics increases.	4.11	0.90	1.00	5.00	5.00	4.00
$x_{i,14}$	Positive feedback about my work in mathematics strengthens my belief in my ability.	4.22	0.81	1.00	5.00	4.00	4.00
$x_{i,15}$	When I receive negative feedback in mathematics, I begin to doubt my competence.	2.70	1.20	1.00	5.00	3.00	3.00
$x_{i,16}$	I feel supported by my teacher in developing my mathematical skills.	3.65	1.01	1.00	5.00	3.00	4.00
$x_{i,17}$	I feel calm when solving mathematics problems.	3.78	0.96	1.00	5.00	3.00	4.00
$x_{i,18}$	I become anxious when mathematics problems are very challenging.	3.44	1.09	1.00	5.00	3.00	3.00
$x_{i,19}$	I feel overwhelmed when I face difficult mathematics topics.	3.29	1.19	1.00	5.00	3.00	3.00
$x_{i,20}$	I feel confident even when working with unfamiliar mathematical material.	2.94	0.97	1.00	5.00	3.00	3.00
$x_{i,21}$	Even relatively simple mathematics tasks can make me feel stressed.	2.38	1.06	1.00	5.00	2.00	2.00
$x_{i,22}$	I enjoy solving challenging mathematics problems.	3.14	1.10	1.00	5.00	3.00	3.00
$x_{i,23}$	I believe mathematics is important for becoming a successful systems engineer.	4.22	0.81	2.00	5.00	5.00	4.00
$x_{i,24}$	I study mathematics only to obtain good grades.	2.98	1.11	1.00	5.00	3.00	3.00
$x_{i,25}$	I study mathematics only to fulfil the requirements for obtaining the Systems Engineering degree.	2.94	1.06	1.00	5.00	3.00	3.00
$x_{i,26}$	I feel curious about learning mathematics.	3.63	1.03	1.00	5.00	4.00	4.00

The multivariate normal distribution is assumed in order to preserve the dependence structure among the input variables. For example, variables $x_{i,18}$ and $x_{i,21}$ are expected to be statistically dependent because both relate to the anxiety experienced by the ith student.

This assumption is motivated by the mathematical tractability of the multivariate Gaussian distribution and its ability to represent relationships among variables through the covariance matrix S. Consequently, the dependence structure of the input variables is approximated using a latent multivariate Gaussian model parametrised by the empirical covariance matrix. Although this approximation might not possibly capture every aspect of the observed dependence, it provides a computationally efficient and statistically reasonable basis for Monte Carlo simulation.

The integral introduced above is approximated using Monte Carlo integration:

$$P(\hat{y} = 1) \approx \frac{1}{N} \sum_{i=1}^{N} P(\hat{y}_i = 1 \mid X_{i,:}), \tag{2}$$

where $X_{i,:}$ denotes the ith row of the matrix X.

The evaluation of a policy π leads to the conditional probability

$$P(\hat{y} = 1 \mid \pi) = \int_{\mathcal{X}} P(\hat{y} = 1 \mid x, \pi) P(x \mid \pi) dx. \tag{3}$$

This quantity is estimated by Monte Carlo simulation after fixing selected variables to values consistent with the policy under consideration. The resulting approximation is given by

$$P(\hat{y} = 1 \mid \pi) \approx \frac{1}{N} \sum_{i=1}^{N} P(\hat{y}_i = 1 \mid X_{i,:}, \pi). \tag{4}$$

To simulate improved teaching support, variables $X_{:,12}$, $X_{:,13}$, $X_{:,14}$, $X_{:,15}$, and $X_{:,16}$ are fixed at their respective 75th percentiles, namely 0.44, 0.97, 0.96, 0.33, and 0.33.

To simulate reduced mathematics anxiety, variables $X_{:,17}$ and $X_{:,20}$ are fixed at their 75th percentiles (1.31 and 0.01, respectively), whereas variables $X_{:,18}$, $X_{:,19}$, and $X_{:,21}$ are fixed at their 25th percentiles (-0.37, -0.24, and -0.38, respectively).

To simulate increased mastery confidence, variables $X_{:,1}$, $X_{:,2}$, $X_{:,3}$, $X_{:,4}$, and $X_{:,5}$ are fixed at their 75th percentiles (0.55, 1.22, 0.39, 0.8, and 0.44, respectively), while $X_{:,6}$ is fixed at its 25th percentile (-1.19).

Finally, the simultaneous implementation of all interventions is simulated by fixing the corresponding variables according to the specifications described above.

IV. The Research Results

The evaluation results suggest that Logistic Regression achieves higher accuracy and precision than both Support Vector Machines (SVMs) and Gaussian Processes (GPs) (Table III). A plausible explanation is the relatively small sample size in relation to the dimensionality of the input space. With only 93 records and 26 input variables, more flexible non-linear models are likely to be more susceptible to overfitting, whereas Logistic Regression provides a simpler decision boundary that might generalise more effectively under limited-data conditions.

TABLE II. 10-Fold Cross-Validation Setting for the Dataset

Fold	Number of Examples for Fitting the Model			Number of Examples for Testing		
	Positive	Negative	Total	Positive	Negative	Total
1	47	36	83	6	4	10
2	47	36	83	6	4	10
3	47	36	83	6	4	10
4	48	36	84	5	4	9
5	48	36	84	5	4	9
6	48	36	84	5	4	9
7	48	36	84	5	4	9
8	48	36	84	5	4	9
9	48	36	84	5	4	9
10	48	36	84	5	4	9

Furthermore, Logistic Regression is particularly suitable when the relationship between the input variables and the log-odds of the target variable can be approximated linearly. Under such conditions, the model typically exhibits lower variance and greater interpretability than more flexible non-linear alternatives.

The results reported in Table III correspond to the hyperparameter configurations selected through stratified 10-fold cross-validation. All models were implemented using the scikit-learn library [32].

For SVMs, the best-performing polynomial kernel had degree one and was therefore equivalent to the linear kernel, which explains the identical results obtained with both kernels. The regularisation parameter C was 7.8×10^{-3} for the linear kernel and 0.25 for the polynomial kernel. For the sigmoid kernel, the optimal values of C and γ were 0.5 and 1.6×10^{-2}, respectively. When the radial basis function (RBF) kernel was employed, the corresponding values were $C = 2$ and $\gamma = 1.9 \times 10^{-3}$.

For Gaussian Processes, the best-performing configuration consisted of the sum of two kernels: an RBF kernel scaled by 3.1×10^{-3} with length scale equal to 1, and a Matérn kernel scaled by 1.9×10^{-3} using the same length scale. For the Rational Quadratic kernel, the length scale and shape parameter α were 1.5×10^{-2} and 1.9×10^{-3}, respectively.

The fold-wise results obtained for Logistic Regression during the 10-fold cross-validation procedure are reported in Table IV. The model achieved an average accuracy of 72.22% and an area under the receiver operating characteristic (ROC) curve of 0.72 (Figure 4), indicating performance substantially better than random classification. In addition, the recall of 91% corresponds to the correct identification of 48 out of 53 positive instances, as shown in the confusion matrix reported in Table V.

The regularisation parameter of the Logistic Regression model was selected using the elbow criterion illustrated in Figure 5. The comparison reported in Table III was conducted using $\lambda = 2 \times 10^{-2}$.

After fitting the Logistic Regression model to the complete dataset using the selected regularisation parameter, Monte Carlo

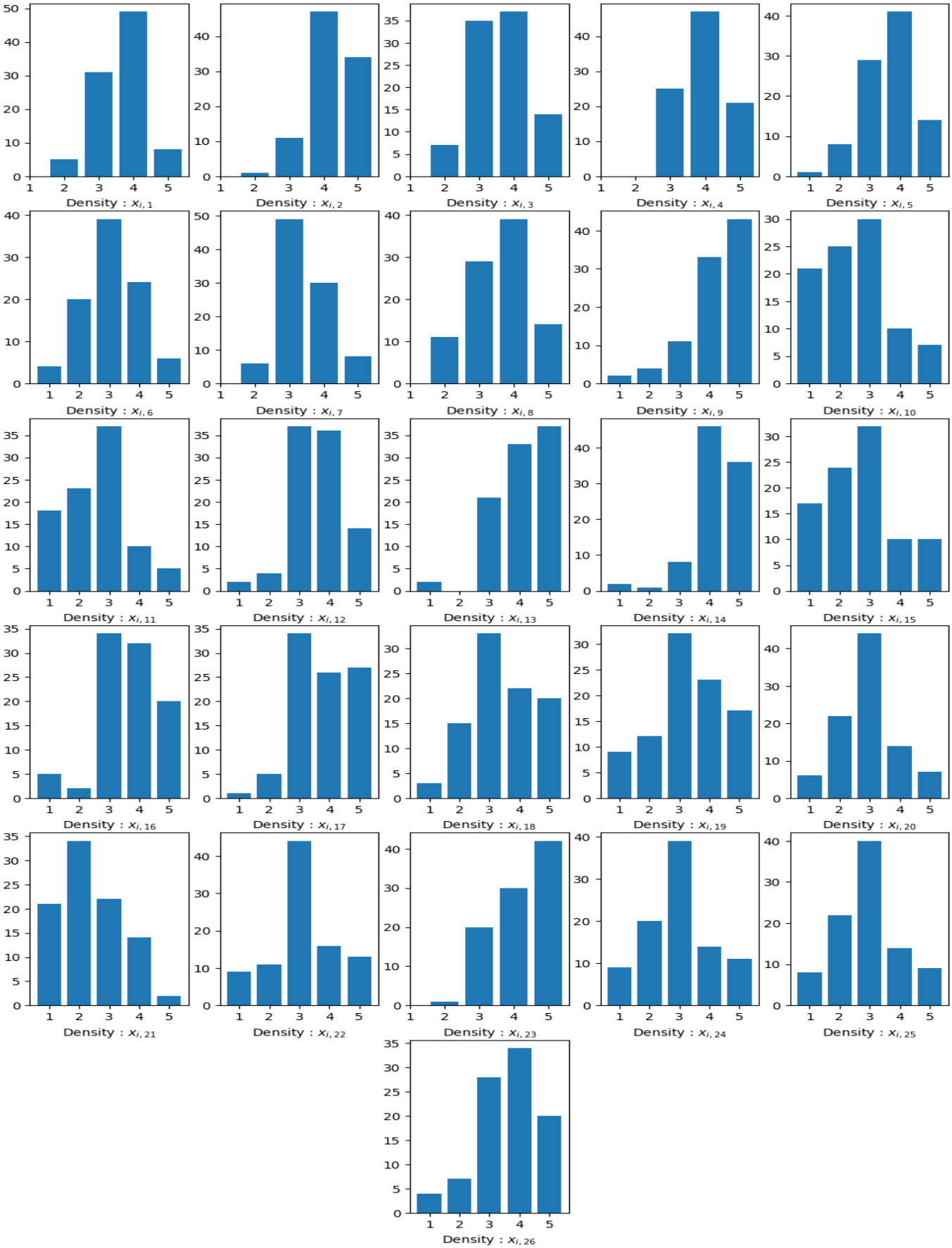


Figure 3. Histogram for each input variable

TABLE III. 10-FOLD CROSS-VALIDATION RESULTS

Machine Learning Model	Mean Precision (%)	Mean Recall (%)	Mean F_1 Score (%)	Mean Accuracy (%)
LR	**70.32**	91.00	78.94	**72.22**
SVM+Lin	67.33	**96.33**	79.00	70.11
SVM+Sig	68.00	**96.33**	**79.50**	71.11
SVM+Pol	67.33	**96.33**	79.00	70.11
SVM+RBF	67.33	**96.33**	79.00	70.11
GP+RQ	68.83	92.67	78.59	71.00
GP+DP	69.86	65.67	66.82	63.33
GP+M+RBF	69.04	83.00	74.80	68.78
GP+M	68.14	83.00	74.16	67.67
GP+RBF	69.39	84.67	75.68	69.78

In this research, we have adopted Logistic Regression (LR), Support Vector Machines (SVM) with Linear (SVM+Lin), Sigmoid (SVM+Sig), Polynomial (SVM+Pol), and Radial Basis Function (SVM+RBF) kernels. Additionally, we employed Gaussian Processes (GP) for classification with Rational Quadratic (GP+RQ), Dot Product (GP+DP), Matern (GP+M), and Radial Basis Function (GP+RBF) Kernels.

TABLE IV. 10-FOLD CROSS-VALIDATION RESULTS CORRESPONDING TO THE LOGISTIC REGRESSION MODEL

Fold	Precision (%)	Recall (%)	F_1 score (%)	Accuracy (%)
1	71.43	83.33	76.92	70.00
2	71.43	83.33	76.92	70.00
3	62.50	83.33	71.43	60.00
4	62.50	100.00	76.92	66.67
5	83.33	100.00	90.91	88.89
6	71.43	100.00	83.33	77.78
7	80.00	85.71	80.00	77.78
8	71.43	100.00	83.33	77.78
9	62.50	100.00	76.92	66.67
10	66.67	80.00	72.73	66.67
Mean	70.32	91.00	78.94	72.22

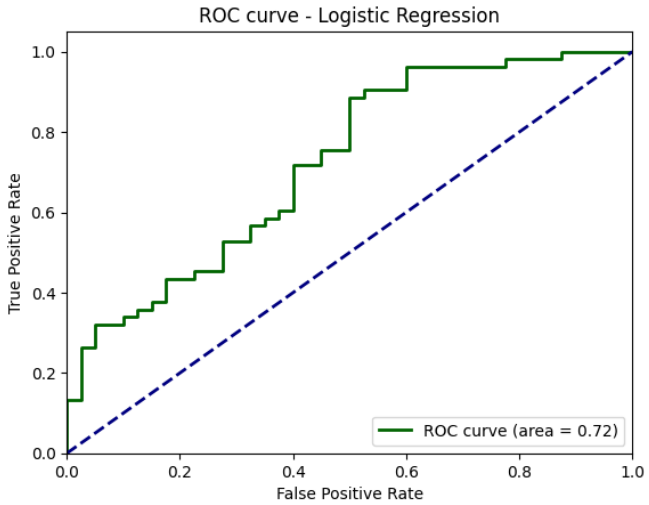


Figure 4. Receiver Operating Characteristic (ROC) curve for the Logistic Regression model

TABLE V. CONFUSION MATRIX ILLUSTRATING THE PERFORMANCE OF LOGISTIC REGRESSION MODEL

True Class	Forecasted Class $\hat{y}_i = 0$	Forecasted Class $\hat{y}_i = 1$	Total
$y_i = 0$	19	21	40
$y_i = 1$	5	48	53
Total	24	69	93

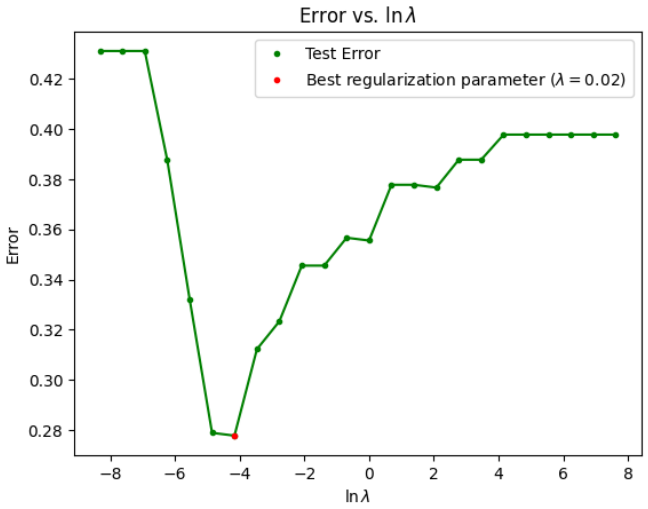


Figure 5. Elbow rule used to select the regularisation parameter for the logistic regression model

TABLE VI. MONTE CARLO SIMULATION OUTCOMES

Policy π	Probability of Failing to Reach the Relational Level (%)	95 % CI	Standard Error
Without Policy	$P(\hat{y} = 1) = 56.94$	[56.80, 57.06]	6.5×10^{-4}
Teaching Support	$P(\hat{y} = 1 \mid \pi) = 54.54$	[54.23, 54.56]	6×10^{-4}
Anxiety Reduction	$P(\hat{y} = 1 \mid \pi) = 53.64$	[53.52, 53.75]	8.1×10^{-4}
Mastery Learning	$P(\hat{y} = 1 \mid \pi) = 53.99$	[53.89, 54.10]	5.2×10^{-3}
All Previous Policies	$P(\hat{y} = 1 \mid \pi) = 47.69$	[47.55, 47.82]	7×10^{-3}

CI stands for Confidence Interval

simulation was employed to explore a broader region of the input space. The resulting estimates are reported in Table VI. All intervention scenarios reduce the probability of failing to attain the relational level. Among the individual interventions, anxiety reduction yields the largest decrease in this probability, whereas the simultaneous implementation of all interventions produces the greatest overall reduction.

The distributions shown in Figures 6–10 are consistent with the estimates reported in Table VI. In particular, Figure 6, corresponding to the baseline scenario, exhibits a distribution centred at higher predicted probabilities than the intervention scenarios. By contrast, Figure 10, which corresponds to the simultaneous implementation of all interventions, is centred at lower predicted probabilities, indicating the lowest estimated risk of failing to attain the relational level. Furthermore, the Monte Carlo estimates exhibit convergence in all scenarios, as illustrated in Figures 11–15.

V. ANALYSIS OF THE RESULTS

In this section, the performance of the classification models is analysed and the implications of the Monte Carlo simulation results are discussed.

The precision and accuracy of the classifiers are below 80%, suggesting that the predictive information contained in the dataset is limited, albeit still informative, as both metrics exceed

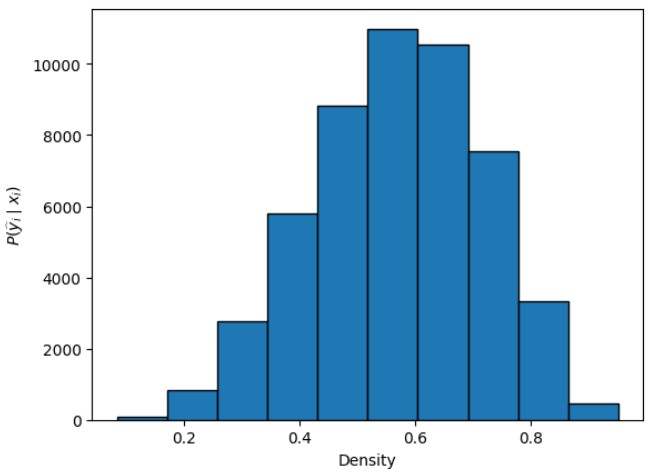


Figure 6. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation without a policy.

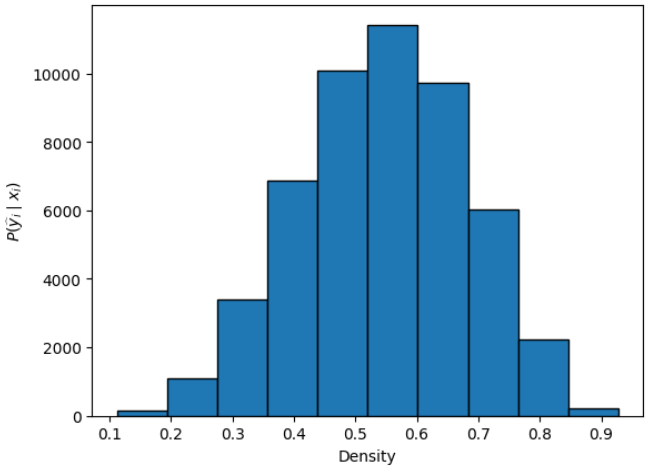


Figure 7. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the application of the teaching support policy.

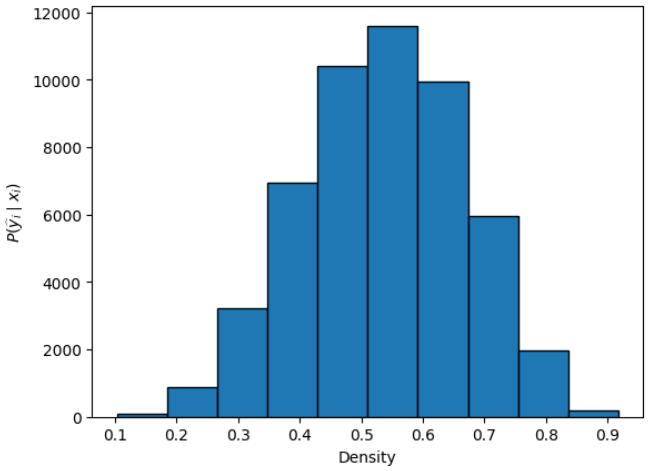


Figure 8. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the application of anxiety reduction policy.

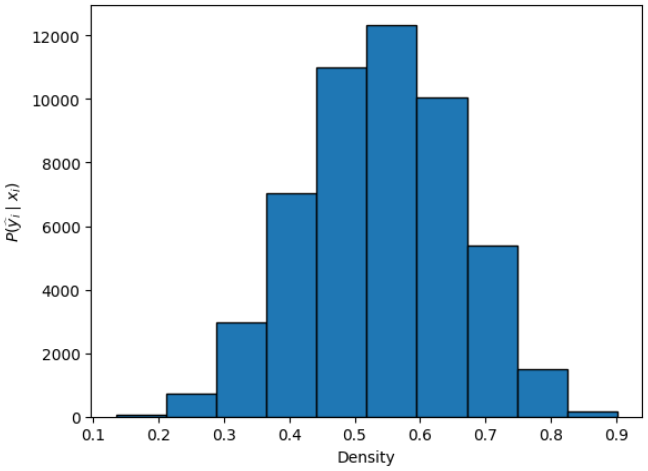


Figure 9. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the application of mastery learning policy.

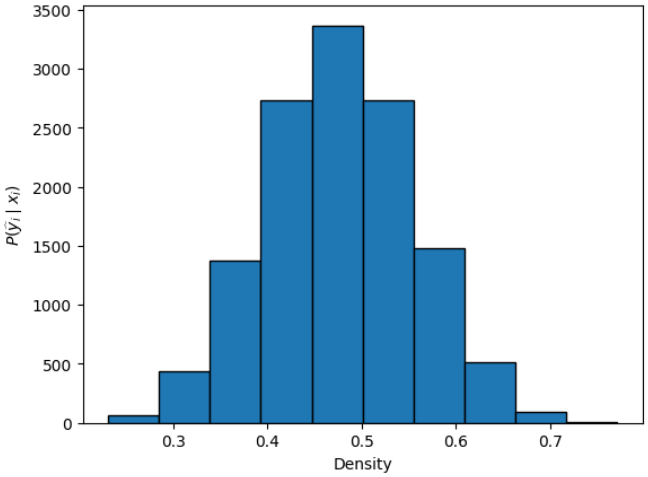


Figure 10. Distribution of probabilities $P(\hat{y}_i = 1 \mid x_i)$ obtained from the simulation for the application of all policies.

60%. Furthermore, the recall of Logistic Regression and the other models exceeds 90%, indicating that most students who are unlikely to attain the relational level are correctly identified. In this context, failing to identify a struggling student may have more serious consequences than providing additional support to a student who may not strictly require it.

The moderate predictive performance of the models is reflected in the distributions shown in Figures 6–10. Models with greater discriminative power might produce more polarised distributions, concentrating probability mass closer to the extremes. Nevertheless, such levels of uncertainty are not uncommon in educational psychology research, where human perceptions and behaviours are often characterised by substantial variability.

The combination of machine learning and Monte Carlo simulation provides a virtual environment for evaluating educational interventions without the time and resource requirements associated with large-scale experimental studies. Notably, the

 ISBN: 978-1-68558-402-3

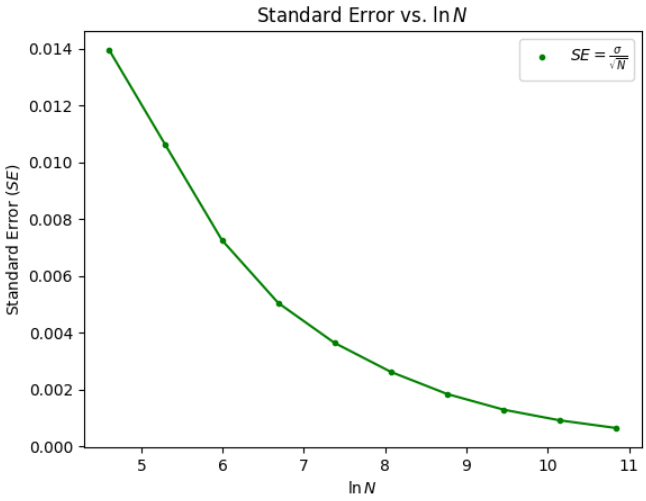


Figure 11. The simulation converges to the expected probability $P(\hat{y} = 1) = 56.94\%$ when no policy is applied as $N = 51,200$, with a standard error of 6.5×10^{-4}. The result lies within the 95% confidence interval of [56.80, 57.06].

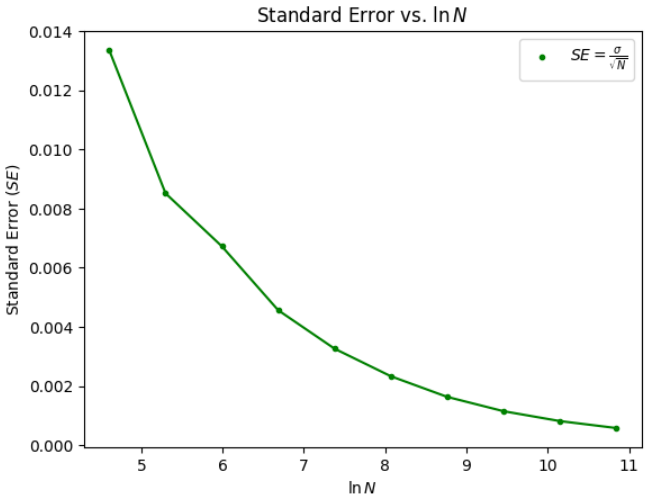


Figure 13. The simulation converges to the expected probability $P(\hat{y} = 1 \mid \pi) = 53.64\%$ when anxiety reduction policy is applied as $N = 51,200$, with a standard error of 8.1×10^{-4}. The result lies within the 95% confidence interval of [53.52, 53.75].

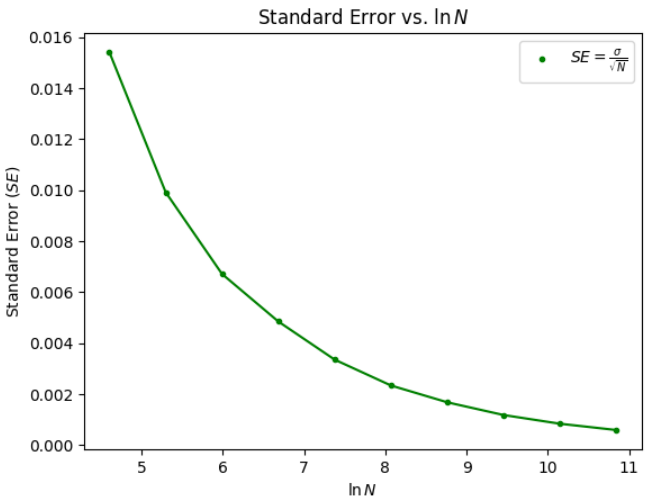


Figure 12. The simulation converges to the expected probability $P(\hat{y} = 1 \mid \pi) = 54.54\%$ when teaching support policy is applied as $N = 51,200$, with a standard error of 6×10^{-4}. The result lies within the 95% confidence interval of [54.23, 54.56].

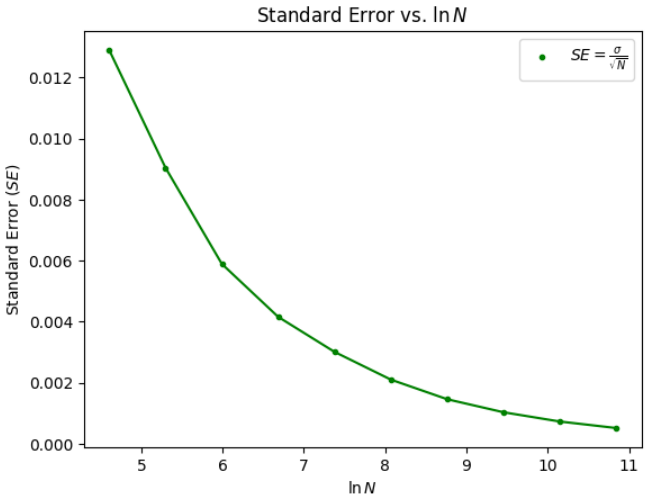


Figure 14. The simulation converges to the expected probability $P(\hat{y} = 1 \mid \pi) = 53.99\%$ when mastery learning policy is applied as $N = 51,200$, with a standard error of 5.2×10^{-3}. The result lies within the 95% confidence interval of [53.89, 54.10].

probability estimated under the baseline simulation is very close to the empirical proportion of students who failed to attain the relational level in the observed dataset. Specifically, $|P(\hat{y} = 1) - P(y = 1)| = |56.94 - 56.99| = 0.05$.

The small discrepancy between these values suggests that the multivariate Gaussian assumption provides a reasonable approximation for the purposes of simulation.

Despite these encouraging results, further research is required to improve predictive performance and strengthen the validity of the proposed framework. In particular, collecting a larger dataset and investigating lower-dimensional representations of the input space may improve model robustness and predictive accuracy.

The simulation outcomes provide evidence that might assist decision-makers in prioritising interventions aimed at helping students attain at least the relational level in engineering mathematics. Within the context of the present study, reducing mathematics anxiety appears to yield the largest individual reduction in risk, followed by increasing mastery confidence and improving teaching support. The greatest reduction is achieved when all interventions are implemented simultaneously.

VI. Conclusion and Future Work

In summary, the Monte Carlo simulations estimate that the probability of failing to attain the relational level in mathematics courses within the Bachelor of Systems Engineering programme

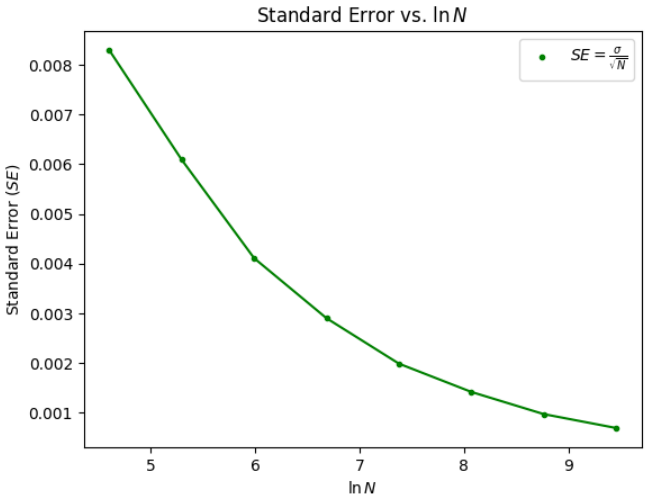


Figure 15. The simulation converges to the expected probability $P(\hat{y} = 1 \mid \pi) = 47.69\%$ when all policies are applied as $N = 12,800$, with a standard error of 7×10^{-3}. The result lies within the 95% confidence interval of [47.55, 47.82].

at the University of Córdoba is 56.94% (95% CI: [56.80, 57.06]).

The simulation of improved teaching support reduces this probability to 54.54% (95% CI: [54.23, 54.56]). This result suggests that students' perceptions of instructional support might play a relevant role in their ability to attain relational-level understanding. In particular, practices that strengthen students' confidence and provide constructive feedback may contribute to reducing the risk of underachievement.

Simulating reduced mathematics anxiety further decreases the estimated probability to 53.64% (95% CI: [53.52, 53.75]). This finding highlights the potential importance of emotional and affective factors in engineering mathematics education. Consequently, instructional strategies that minimise unnecessary stress and support students when confronting challenging material may help mitigate the risk of failing to attain the relational level.

Similarly, improving mastery-learning beliefs reduces the estimated probability to 53.99% (95% CI: [53.89, 54.10]). This result is consistent with the self-efficacy literature, which emphasises the importance of mastery experiences in shaping students' beliefs about their capabilities. Learning environments that promote incremental progress, sustained engagement, and repeated experiences of success might therefore contribute to improved outcomes.

The largest reduction is obtained when all interventions are implemented simultaneously, lowering the estimated probability to 47.69% (95% CI: [47.55, 47.82]). This result suggests that the factors considered in this research should not be viewed in isolation, as their combined influence appears to be greater than that of any individual intervention.

Overall, the results demonstrate that the combination of machine learning and Monte Carlo simulation provides a computational framework for evaluating educational interventions prior to their implementation. Such an approach enables decision-makers to explore the potential consequences of alternative strategies aimed at improving students' attainment of relational-level understanding in engineering mathematics.

Future work should focus on expanding the dataset in order to obtain more accurate and robust estimates. A larger sample would also facilitate the analysis of all levels of the SOLO taxonomy rather than restricting the problem to a binary classification task. In addition, the questionnaire employed in this study should be compared with alternative self-efficacy instruments reported in the literature.

Further research should also investigate dimensionality-reduction techniques, including Principal Component Analysis, Non-negative Matrix Factorisation, t-distributed Stochastic Neighbour Embedding, and Autoencoders, in order to identify latent representations of the input space and potentially improve predictive performance.

Finally, the intervention strategies examined through simulation should be implemented and evaluated empirically in educational settings to assess the extent to which the predicted effects are observed in practice.

ACKNOWLEDGEMENT

Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. The authors thank Universidad de Córdoba in Colombia for supporting this study. They also thanks all students who collaborated by answering the survey conducted for collecting the dataset used in this study. Finally, the author thanks the anonymous reviewers for their comments that contributed to improve the quality of this article.

REFERENCES

[1] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory* (Prentice-Hall series in social learning theory). Englewood Cliffs, NJ: Prentice-Hall, 1986, ISBN: 013815614X.

[2] D. E. M. da Silva, E. J. S. Pires, A. Reis, P. B. de Moura Oliveira and J. Barroso, 'Forecasting Students Dropout: A UTAD University Study', *Future Internet*, vol. 14, no. 3, pp. 1–14, 2022.

[3] I. Caicedo-Castro, O. Velez-Langs, M. Macea-Anaya, S. Castaño-Rivera and R. Catro-Púche, 'Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention', in *The 2022 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, International Academy, Research, and Industry Association, 2022, pp. 76–84.

[4] S. Zihan, S.-H. Sung, D.-M. Park and B.-K. Park, 'All-Year Dropout Prediction Modeling and Analysis for University Students', *Applied Sciences*, vol. 13, p. 1143, Jan. 2023.

[5] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Loumos, 'Dropout Prediction in E-Learning Courses through the Combination of Machine Learning Techniques', *Computers and Education*, vol. 53, no. 3, pp. 950–965, 2009, ISSN: 0360-1315.

[6] J. Kabathova and M. Drlik, 'Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques', *Applied Sciences*, vol. 11, p. 3130, Apr. 2021.

[7] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka and P. C. Nshimyumukiza, 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization', *Computers and Education: Artificial Intelligence*, vol. 3, p. 100 066, 2022, ISSN: 2666-920X.

[8] V. Čotić Poturić, I. Dražić and S. Čandrlić, 'Identification of Predictive Factors for Student Failure in STEM Oriented Course', in *ICERI2022 Proceedings*, ser. 15th annual International Conference of Education, Research and Innovation, Seville, Spain: IATED, 2022, pp. 5831–5837, ISBN: 978-84-09-45476-1.

[9] V. Čotić Poturić, A. Bašić-Šiško and I. Lulić, 'Artificial Neural Network Model for Forecasting Student Failure in Math Courses', in *ICERI2022 Proceedings*, ser. 15th annual International Conference of Education, Research and Innovation, Seville, Spain: IATED, 2022, pp. 5872–5878, ISBN: 978-84-09-45476-1.

[10] I. Caicedo-Castro, M. Macea-Anaya and S. Rivera-Castaño, 'Early Forecasting of At-Risk Students of Failing or Dropping Out of a Bachelor's Course Given Their Academic History - The Case Study of Numerical Methods', in *PATTERNS 2023: The Fifteenth International Conference on Pervasive Patterns and Applications*, ser. International Conferences on Pervasive Patterns and Applications, Nice, France: IARIA: International Academy, Research, and Industry Association, 2023, pp. 40–51, ISBN: 978-1-68558-049-0.

[11] I. Caicedo-Castro, M. Macea-Anaya and S. Castaño-Rivera, 'Early Forecasting of At-Risk Students of Failing or Dropping Out of a Bachelor's Course Given Their Academic History - The Case Study of Numerical Methods', in *PATTERNS 2023 : The Fifteenth International Conference on Pervasive Patterns and Applications*, 2023, pp. 40–51.

[12] I. Caicedo-Castro, 'Course Prophet: A System for Predicting Course Failures with Machine Learning: A Numerical Methods Case Study', *Sustainability*, vol. 15, no. 18, 2023, 13950.

[13] I. Caicedo-Castro, 'Quantum Course Prophet: Quantum Machine Learning for Predicting Course Failures: A Case Study on Numerical Methods', in *Learning and Collaboration Technologies*, P. Zaphiris and A. Ioannou, Eds., Cham: Springer Nature Switzerland, 2024, pp. 220–240, ISBN: 978-3-031-61691-4.

[14] I. Caicedo-Castro, 'An Empirical Study of Machine Learning for Course Failure Prediction: A Case Study in Numerical Methods', *International Journal on Advances in Intelligent Systems*, vol. 17, no. 1 and 2, pp. 25–37, 2024.

[15] N. E. Betz and G. Hackett, 'The Relationship of Mathematics Self-Efficacy Expectations to the Selection of Science-based College Majors', *Journal of Vocational Behavior*, vol. 23, no. 3, pp. 329–345, 1983.

[16] F. Pajares and D. M. Miller, 'Mathematics self-efficacy and mathematics performances: The need for specificity of assessment', *Journal of Counseling Psychology*, vol. 42, no. 2, pp. 190–198, 1995.

[17] F. Pajares and J. Kranzler, 'Self-Efficacy Beliefs and General Mental Ability in Mathematical Problem-Solving', *Contemporary Educational Psychology*, vol. 20, no. 4, pp. 426–443, 1995.

[18] F. Pajares and L. R. Graham, 'Self-Efficacy, Motivation Constructs, and Mathematics Performance of Entering Middle School Students', *Contemporary Educational Psychology*, vol. 24, no. 2, pp. 124–139, 1999.

[19] I. L. Nielsen and K. A. Moore, 'Psychometric Data on the Mathematics Self-Efficacy Scale', *Educational and Psychological Measurement*, vol. 63, no. 1, pp. 128–138, 2003.

[20] M. Coe, R. McFadden, B. Pratt-Sitaula and E. Baer, 'Geoscience Mathematics Self-Efficacy Scale (GeoMSES) for Majors-Level Undergraduates: Psychometric Data', *Education Sciences*, vol. 16, no. 2, p. 288, 2026.

[21] E. Rhew, J. Piro, P. Goolkasian and P. Rappold, 'The effects of a Growth Mindset on Self-Efficacy and Motivation', *Cogent Education*, vol. 5, no. 1, p. 1 492 337, 2018.

[22] I. Caicedo-Castro, R. Castro-Púche and S. Castaño-Rivera, 'Estimating the Risk of Failing Physics Courses through the Monte Carlo Simulation', in *GPTMB 2025: The Second International Conference on Generative Pre-trained Transformer Models and Beyond*, ser. International Conference on Generative Pre-trained Transformer Models and Beyond, Venice, Italy: IARIA: International Academy, Research, and Industry Association, 2025, pp. 15–22, ISBN: 978-1-68558-287-6.

[23] I. Caicedo-Castro, R. Castro-Púche and S. Castaño-Rivera, 'Estimating the Risk of Failing Physics Courses through Machine Learning and Monte Carlo Simulation', *International Journal on Advances in Systems and Measurements*, vol. 19, no. 1 and 2, 2026.

[24] I. Caicedo-Castro, O. Vélez-Langs and R. Castro-Púche, 'Using the Monte Carlo Method to Estimate Student Motivation in Scientific Computing', in *PATTERNS 2025: The Seventeenth International Conferences on Pervasive Patterns and Applications*, ser. International Conferences on Pervasive Patterns and Applications, Valencia, Spain: IARIA: International Academy, Research, and Industry Association, 2025, pp. 15–22, ISBN: 978-1-68558-263-0.

[25] I. Caicedo-Castro, R. Castro-Púche and O. Vélez-Langs, 'Identifying Factors that Increase the Risk of Demotivation in Scientific Computing Courses Using Monte Carlo Methods', *International Journal on Advances in Systems and Measurements*, vol. 18, no. 3 and 4, pp. 162–171, 2025.

[26] M. Hegazi, K. Chebil and S. A. Alqahtanib, 'A monte carlo-based model for predicting student performance', 2026.

[27] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2006, ISBN: 026218253X.

[28] J. C. Platt, 'Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods', in *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.

[29] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2013, ISBN: 9780262018029.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[31] N. Metropolis and S. Ulam, 'The Monte Carlo Method', *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949, ISSN: 01621459, 1537274X.

[32] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A Narrative Exploration of the Effect of Artificial Intelligence on Managerial Functions in Africa

Edward Phele
Department of Information Systems, North-West University, South Africa. Email: roidqueipsy@gmail.com

Nathan Davids
Department of Information Systems, North-West University, South Africa. Email: natedogdavids12@gmail.com

Pako Ntchupetsang
Department of Information Systems, North-West University, South Africa. Email: pakonchupetsang13@gmail.com

Anita Ratlou
Department of Information Systems, North-West University, South Africa. Email: siyandaanita2@gmail.com

Mohammed Aqil Malek
Department of Information Systems, North-West University, South Africa. Email: malekaqil786@gmail.com

Joshua Ebere Chukwuere
Department of Information Systems, Technology Enhanced Learning and Innovative Education and Training in South Africa (TELIT-SA), North-West University, South Africa. Email: joshchukwuere@gmail.com

Aadil Malek
Department of Information Systems, North-West University, South Africa. Email: malekaadil493@gmail.com

Richard Abolade Akinkunmi
Department of Information Systems, North-West University, South Africa. Email: richardakinkunmi08@gmail.com

Abstract—**The role of Artificial Intelligence (AI) across sectors is massive, and its effects are felt in Africa. African businesses are beginning to adopt AI tools and applications for performing basic and complex real-time insights, predictive analytics, automation processes, enhancing competitiveness, operational efficiency, and strategic decision-making. This study aims to examine how AI affects management functions and the decision-making process in African organisations. The study employed a Narrative Literature Review (NLR) methodology in reviewing academic publications within the research topic to uncover gaps and achieve the objective. The findings revealed that AI affects managerial functions in Africa by enhancing consumer engagement and sustainable development, supporting managers to focus on innovation and value creation by streamlining the repetitive task process. In general, the study confirms that AI is becoming an effective enabler of management success in Africa's changing economic context, combined with ethical governance and digital literacy. However, the application of AI in managerial functions has a long way to go. In the future, academics can investigate how African organisations can formulate governance models (frameworks) to tackle ethical issues such as algorithmic bias, data privacy, plagiarism, and ethical decision-making. In addition, research can focus on finding the competencies and skills required for African managers to use AI in driving the decision-making process to improve performance and operational functions.**

Keywords- African organisations; Artificial intelligence (AI); Decision-making; Digital transformation; Managerial functions; Operational efficiency

I. INTRODUCTION

Artificial Intelligence (AI) is essentially reshaping humanity's lifestyles in the 21st century by changing the ways people live, educate, communicate, and work. The impact of Information and Communication Technology (ICT) in the world is felt in many sectors, such as agriculture, finance, education, and media, where it facilitates businesses in accessing broader markets, making better decisions, and offering better services [1]. The adoption of AI in Africa is not only seen as the continent's way of modernising, but at the same time as a source of solutions to age-old problems in the government, business, education, and development sectors. According to the United Nations Development Programme estimate for Africa, AI will potentially contribute $1.5 trillion economically and shape the socio-economic and technological sectors [2]. Companies on the continent are collectively getting used to the idea of employing AI as a means of decision-making, a tool for productivity, and an instrument of innovation, which leads to a gradual shift from traditional management to data-driven operations [1].

Artificial intelligence technologies are extensively implemented in data analysis in most African businesses, especially in the financial sector, agriculture, logistics, and creative industries, to achieve higher accuracy, eliminate the possibility of flesh-and-blood (human) errors, and furnish managers with up-to-the-minute information [20]. A good example is machine learning used by banks for the purpose of improving credit assessment and fraud detection; on the other hand, retail and logistics businesses use predictive analytics for demand forecasting and inventory optimisation. All of these allow managers to make decisions quickly and accurately, enabling them to strengthen their position in both local and global markets. The State of AI in Africa Report [3] also state that this trend towards technology is only a part of a bigger economic transformation that will be the primary driver of the digital economy in Africa, which is projected to be over USD 180 billion by 2030. The success of these projections lies in the available resources. The arguments presented in the African Union Digital Transformation Strategy (2024-2030) state that the success of the digital economy in Africa depends on the capacity to balance innovation and human values. To this point, it is necessary to comprehensively aggregate the role and impact of AI on managerial functions in Africa. Then, this narrative literature review aims to examine how artificial intelligence affects managerial functions and the decision-making process in African organisations. The rest of the paper covers the following sections 1) problem statement, 2) scope of the study and objectives, 3) Research method, 4) how AI enhances management capabilities, 5) organizational and systematic benefits and AI in business, 6) AI in both a positive and negative transformation catalyst for emerging African managers, 7) implications for higher education, and 8) conclusion.

II. PROBLEM STATEMENT

In addition to optimising efficiency, AI can also be a source of democratisation and the development of local businesses. The African youth are already engaged in creating AI-based solutions in various industries, including agriculture (precision farming), health (diagnostic imaging), and government (smart governance platforms). In this way, they are creating new markets that were not accessible in the past, in addition to spurring socioeconomic development. Additionally, the fact that AI will relieve the administrative side of jobs from mundane/stereotypical tasks will lead to a scenario where future leaders will be able to spend more time on the most innovative and strategic thinking, which are the most sought-after attributes of the modern business environment [4].

Nevertheless, the pace with which AI technology has been developed in Africa has been greater than the establishment of the various ethical and legal parameters needed to regulate practice, which imposes a significant burden on the continent. Although nations such as South Africa, Kenya, and Nigeria have already begun the AI strategy formulation process, the majority of African regions have yet to establish the regulatory frameworks upon which the responsible use of AI can be undertaken [5]. This situation has resulted in issues of privacy, algorithmic discrimination, and even misappropriation of personal data. According to the African Union Digital Transformation Strategy (2024-2030), the advantages of AI might be overshadowed by inequality and exploitation when there are no policy interventions. Surveillance, data privacy, and algorithmic bias are now issues of immediate concern [5].

The key issue then is not whether Africa can embrace AI, but rather whether it will be able to manage its usage. In the absence of regulation or control, AI is likely to exacerbate the existing disparities in social and economic life by placing the upper hand on those businesses that possess abundant data and more sophisticated technologies. Moreover, the lack of proper data protection frameworks suggests that user information can easily end up in the wrong hands, where it can be abused unethically, which may subsequently cause mistrust among users of digital platforms. This type of mismanagement can hurt employment opportunities, fairness in decision-making, and even democratic accountability in countries where governance is weak, as is the case in this country.

This is a two-edged sword-like situation with opportunities and challenges for the new managers. They must embrace AI as a major tool that can lead to higher efficiency, better decision-making, and better competition. At the same time, they should have the required knowledge and ethical sense, or rather "the right way of doing things," in order to manage AI in the correct way. Managers who are familiar with the social impacts of AI will have the ability to apply it in such a way that results in sustainability and fairness. This is especially vital for African managers operating in fast-developing but resource-limited environments. They play a key role in deciding how AI can be used to promote the integration of the less privileged in the development process, instead of further widening the gap between the rich and the poor.

III. SCOPE OF THE STUDY AND OBJECTIVES

A. Scope of the study

In this review of the literature, the specific context under consideration is in African contexts, which include Sub-Saharan Africa and North Africa. Although the review relies on the global AI management literature in which African-specific research is limited, the analysis focuses on studies, examples of cases, and empirical findings from African countries and regions. In cases where global research has been mentioned, it is explicitly taken into account that the study can be generalised to African business settings and can be transferred to them [6]. The African focus is explained by the fact that the African continent has its own socio-economic specifics, such as a young population, high rates of penetration of mobile devices, inconsistency in infrastructures, entrepreneurial mobility, and the presence of certain developmental challenges that result in unique conditions to address the managerial influence of AI [7].

The main aim of this literature review paper is to critically analyse and summarise available academic studies on the benefits of AI tools for future managers in Africa, given the growing penetration of AI in business processes. AI technologies can be enablers and amplifiers of managerial performance instead of threats to managerial fitness [8]. Due to the fast digitalisation of African economies and social aspects of society, it is urgently needed to comprehend how AI tools could be used strategically to increase managerial capabilities, organisational performance, and be part of the overall economic development goals [4]. The current literature remains mainly based on the experiences of developed economies with significantly different infrastructural, economic, and social circumstances [9].

B. Research main objective and subobjectives

To examine how artificial intelligence affects managerial functions and the decision-making process in African organisations.

This objective is subdivided into four sub-objectives for a better understanding of the topic of the study:

- **a.** Investigate how AI technologies enhance the managerial capabilities of African managers.
- **b.** To explore the benefits and negative perspectives of AI in business functions and operations.
- **c.** To examine the positive impact on managerial business operations in Africa.
- **d.** To assess the implications of AI for African higher education in preparing managers.

IV. RESEARCH METHOD

There are many types of research methodologies that can be used to carry out a literature research paper. One of them is a narrative literature review (NLR). NLR can be considered a traditional literature review methodology with the ability to synthesise existing research topics in a comprehensive way to discover gaps [10] [11]. NLR is flexible in allowing researchers to define inclusions and exclusions to deepen their understanding of the topic under discussion. Review what is in the literature on a certain topic, but the findings cannot be generalized [10]. To synthesise, the researcher can focus on a particular area of a study, and the inclusion criteria can be biased [10] [12].

This literature review methodology can be used in different disciplines, such as health, information systems, and many others, without a defined structure [10] [13] [14]. The role and implications of AI for emerging managers in Africa are still limited; however, there is a need to review the existing literature to identify the gaps in the literature. The identification of research gaps can lead to an understanding of the state of things in the research area and discipline, as well as providing a pathway for future research.

A. Search process and execution

The researcher applied this methodology in the study to better discover the research gap in the topic by looking at how artificial intelligence affects managerial functions and the decision-making process in African organisations. To achieve the research objective, there was a comprehensive literature review in established scientific databases, such as Web of Science, Scopus, IEEE Xplore, EBSCOhost, African Journals Online (AJOL), and ScienceDirect, as well as sources of grey literature such as Google Scholar. Keyword searches used included adoption AND decision-making of AI in Africa, AI AND Africa, AI AND African organisations, AI ethics AND governance in Africa, and Artificial Intelligence AND managerial functions.

The findings of the study will provide aggregated information and knowledge on the core effects of AI on managerial functions in African organisations. To achieve this, the study focuses on literature that covers AI managerial roles for managers (those within five years), or those who have moved from non-managerial positions to managerial positions, as well as mid-career management (under 40 years old and with less than 10 years of managerial experience). This population segment is tactical, as emerging managers are the future of African organisations and, at the same time, are digital natives who feel more comfortable with technology and are professionals whose career paths are most influenced by the implementation of AI in business.

B. Inclusion and exclusion criteria

Inclusion and exclusion criteria are used to increase transparency, strengthen credibility, and reduce bias [15, 16]. Inclusion criteria were used in the study to define the characteristics of the literature to be considered in accordance with the research objective.

The inclusion includes:

- **a.** Relevant literature on the topic, such as AI
- **b.** Peer-reviewed articles, book chapters, and conference proceedings.
- **c.** Current publications from the last five years and at most ten years.
- **d.** Publications in English.
- **e.** Literature studies on African emerging managers or managers.

Exclusion criteria are used to define the literature to be excluded from the study. The criteria were applied to ensure that the study is focused on the objective and to eliminate unnecessary literature [17]. The exclusion criteria are as follows:

- **a.** Publications outside the scope of the study topic and objectives.
- **b.** Non-academic literature.
- **c.** Publications not written in English.
- **d.** Redundant or duplicate studies.
- **e.** Complex AI studies without organisational or managerial links.

The review is clearly positioned to show the optimistic view of AI on the administrative effects; however, it also takes into account the critics and valid concerns regarding the issues with implementing AI.

C. Transparency and reproducibility

Transparency and reproducibility are two important aspects of the research process to ensure that adequate processes, rigor, and necessary considerations are in place and followed. However, traditionally, NLR allows authors to tell stories or provide detailed explanations about the topic under review based on the author's expertise. According to Paré and Kitsiou [32], NLR focuses less on extensive data extraction but more on thematic storytelling and qualitative synthesis. Nonetheless, given the different views of NLR, this study ensures that the following are applied to ensure transparency and reproducibility of its findings

a. **Search report and documentation:** A detailed literature search was done. These searches were conducted to select academic materials available on Scopus, Web of Science (WoS), and grey literature such as Google Scholar.
b. **Define eligibility criteria:** A clear set of eligibility criteria was established. This includes defining the inclusion and exclusion criteria as provided above.
c. **Acknowledge limitations:** Identify any biases in research interest and topic; rather, the focus should be on publication to improve quality.
d. **Reference software:** Apply reference management software to track and store the used citations. The researchers used Endnote software in tracking and exporting used articles into the library toward promoting reproductivity of its findings.

V. HOW AI ENHANCES MANAGEMENT CAPABILITIES

Across African companies and public institutions, AI is moving from pilots to targeted deployments that augment managerial work, especially in analytics-driven decision-making, operational productivity, strategic planning, competitive intelligence, and innovation management. Recent reviews and sector studies converge on the view that AI, when paired with data governance and skills development, improves performance outcomes and supports inclusive growth [1] [5]. The evidence focused on Africa is now visible in manufacturing, finance, supply chains, and public health logistics, illustrating measurable gains in quality and service reach [18].

a. **Enhanced decision-making:** Many managers make use of machine learning to make decisions based on complicated and dynamic data. Explainable machine learning and model risk management have improved South Africa's credit ecosystems [19]. This helps managers meet governance obligations and defend lending decisions. At the macro level, systematic reviews show that AI in supply chain management (SCM) strengthens managerial choices through demand forecasting, anomaly detection, and inventory optimisation, provided that data quality and integration are addressed [20]. Customised machine learning credit scoring helps managers better oversee small file customers in the banking sector in Africa and Small and Medium Enterprises (SMEs) finance. This type of support reduces risk and expedites access to more working capital.
b. **Improved operational efficiency and productivity:** Examples taken from South African manufacturing illustrate the successful application of AI to increase productivity and quality, as well as optimise the supply chain, in a way that efficiency gains are physically identified at the factory and network levels [18]. Moreover, in logistics and health, AI-enabled forecasting and allocation are currently increasing the service level: a 2025 review has unveiled AI's role in vaccine supply chain planning in African environments that propels demand prediction, routing, and cold chain risk management, which are the causes of wastage and stockout reduction [21].
   The same is true for bank operations in Africa, where efficiency improvements are also reported. For instance, it is expected that the Natural Language Processing (NLP) chatbot, along with predictive analytics, will lessen the number of queries that cause the accumulation of awaiting user requests and, at the same time, will prompt transaction handling, thus speeding up the process so the manager would be free to utilise staff-dedicated tasks of greater value instead [1]. The situation in South African SMEs is described as a positive turn: qualitative research finds that generative-AI tooling fastens content creation, customer engagement, and back-office automation, which, in this way, stimulates productivity in resource-constrained environments [22].
c. **Implications for managers:** Start with low-effort, high-impact, and more specifically defined tasks (forecasting, scheduling, triage, anomaly detection) that deal with data, and then move towards optimising complex processes in which you can measure throughput, error rates, and service times with defined KPIs [18] [21].
d. **Strategic planning and competitive intelligence:** Artificial Intelligence expands a manager's horizon scanning by interpreting and analysing unstructured data (news, filings, reviews) and structured data (for instance, prices, macro data, telemetry) into meaningful signals on market entry, pricing, and risk. An in-depth examination indicates that AI contributes to the implementation of dynamic capabilities of sensing, sense, seize, and reconfigure [1]. In the financial world, an increased focus on model risk and compliance is a significant strategic jump plan to implement AI through credit, fraud, and customer lifetime analytics [19]. At the national and sectoral levels, Africa-focused studies highlight that AI readiness and policy coherence within a country directly influence competitive outcomes. This provides important guidance to managers about ecosystem risk (skills, infrastructure) and collaboration [23].
e. **Innovation and product-development acceleration:** AI brings about an innovation acceleration primarily by significantly shortening the discovery and iteration cycle times; for example, rapid customer-journey analysis, idea screening, and simulation. Innovation in African economies that is driven by studies links digital/AI

transformation to productivity through new product and process innovation pathways [24]. In the case of inclusive finance, ML-based scoring and risk assessment tools facilitate the creation of new products for segments of the population that are not served by the traditional banking system, thus unlocking the potential of the market [25]. Meanwhile, sectoral work on renewable-energy supply chains shows that AI-assisted dynamic planning can simplify the installation process, which is an example of AI. Scale up the new solution architectures within African settings [26].

VI. ORGANISATIONAL AND SYSTEMIC BENEFITS OF AI IN BUSINESS

Artificial intelligence is transforming the manner in which businesses conduct business, handle information, and provide value in all fields in a complete transformation. AI acts as a driver and structure of organisational change, whether in financial management or customer interaction. The literature explains that AI not only leads to the promotion of automation, but also fosters the development of data-driven strategies, organisational responsiveness, and inclusive business processes. In this section, the systemic advantages of AI adoption are discussed in an organisational dimension, in the form of monetary optimisation, customer interaction, human resource creation, and cross-functional integration.

a. **Management optimisation:** AI has brought considerable improvements to financial management by improving the accuracy and efficiency of decision-making. With the help of Machine Learning algorithms, companies can predict their financial trends, identify fraud, and control their resources more effectively. Butson and Spronken-Smith [27] observed that real-time budgetary planning and cost savings, which can be achieved using AI-based predictive analytics, lead to less human error and increased accountability. In addition, robotic process automation (RPA) eliminates monotonous accounting processes and liberates financial departments so that they can focus on strategic projects by automating these processes. Increased transparency and real-time data reporting due to the integration of AI in enterprise resource planning (ERP) systems leads to better financial governance.

b. **Transformation of customer engagement:** The greatest advancement has been made in customer engagement due to AI. AI-driven customer relationship management (CRM) systems allow organisations to provide hyper-personalised interactions and predictive recommendations, enable round-the-clock responses, and enhance customer satisfaction. Predictive AI models are used to predict requirements and provide proactive delivery of the service by examining how a customer behaves. This has transformed marketing approaches so that firms can no longer be reactive in their marketing but must be anticipatory in order to foster loyalty and enduring value.

c. **Human resource management improvement:** AI has transformed human resource management through the simplification of recruitment, employee growth, and performance analysis. Automated systems eliminate bias. Enhance the quality of hires by helping with screening. AI-based applications, such as predictive analytics, can enable organisations to enhance their processes. Organisations must implement proactive retention initiatives by detecting the risks of employee turnover. In addition, AI-based learning systems will be able to customise training to suit each employee and create ongoing skill growth and interaction. Systematically, these tools will contribute to more data-driven and adaptive HR functions, matching workforce planning to organisational strategy. Functions, matching workforce planning to organisational strategy.

d. **Cross-functional coordination improvement:** AI enhances cross-functional coordination by means of organisational data integration across departments. Unified access is available as real-time analytics dashboards, providing advice to decision-makers in the fields of marketing, operations, and logistics, leading to improved collaboration. Research demonstrates that the speed of communication and the accuracy of information are enhanced with the assistance of data democratisation via AI. Teams quickly adapt to commercialise adjustments via AI-driven process management tools, thereby building organisational agility and innovation.

e. **Democratisation of business tools for SMEs:** AI has introduced sophisticated business applications to the masses, for instance, small and large businesses (SMBs). AI solutions based on the cloud, such as Cloud AI from Google and Azure AI from Microsoft, create predictive analytics, automation, and data visualisation that are more cost-effective for organisations. This accessibility has facilitated SMEs to compete on equal terms with larger organisations in AI-based marketing, logistics, and customer engagement. It is emphasised that reports indicate that barriers to entry can be reduced with the help of AI innovation, which helps organisations become more digital. Democratisation has been fundamental in promoting competition and sustainable growth in most industries.

f. **Inclusive growth and development:** Inclusivity also grows due to AI-driven transformation because it offers more opportunities to underserved groups and emerging markets. Automation, when combined with human upskilling programmes, leads to gains in productivity without disproportionate displacement of labour. AI fosters equal participation in the online economy by providing access to information, improving service delivery, and fostering entrepreneurship. In addition, AI-based educational technologies support workforce readiness and lifelong learning, which are vital for sustainable development in the Fourth Industrial Revolution.

VII. AI IN BOTH A POSITIVE AND NEGATIVE TRANSFORMATION CATALYST FOR EMERGING AFRICAN MANAGERS

The economy of Africa as a whole is made up of young people, making it easier to use AI responsibly. Emerging managers under 40 years of age are digital natives who see technology as a means to increase creativity, teamwork, and advancement in society, not as a threat. This advantage should be used to develop managers who are skilled in the ethical and strategic use of AI. African managerial environments are based on relationships and social principles such as Ubuntu, which focus on collaboration, empathy, and good outcomes for the community. When this is combined with AI's analytical power, these values can create a new kind of management that combines data-driven accuracy with human intuition. Due to this, our group believes that AI adoption should be done in a way that is specific to Africa: contextualised, ethical, and focused on development [6]. Towards a balanced and contextual AI narrative in Africa, we must consider the positive and negative transformative forces of AI that affect managerial functions in and outside African organisations.

A. AI as a positive transformation catalyst

The discussion of AI in the higher levels of business leadership has fluctuated constantly. Sometimes it is viewed as a great tool that can help people be efficient, but at times it is looked down upon. Some scholars have looked at the story of the AI revolution, and some of them see AI as a big help in the business world; others, on the other hand, see it as a threat to people's jobs and privacy, which managers have been in charge of until now. Our group uses the positive transformation paradigm as a logical way to explain that middle ground. The digital transformation is spreading rapidly across Africa, and the people who run African companies are not only knowledgeable about technology but also obsessed with it. That means AI is not only helping the continent get back on its feet, but it is also a source of technological advancement, which is the process of quickly overcoming traditional issues with infrastructure, so that innovation can happen right away [4].

Therefore, rather than causing harm, AI enhances capabilities that improve managerial skills, expand the boundaries of data-based research, and encourage social and economic involvement through participation in the global market [1]. A study conducted in Kenya, Nigeria, and South Africa, collecting data from the agriculture, banking, and logistics sectors, showed that the strategic adoption of AI positively impacts the outcomes of forecasting, planning, and resource allocation activities [18] [21].

B. AI as a negative transformation catalyst

Many people believe that the use of AI can cause more unemployment, reduce the appreciation of human labour, and make the gap between rich and poor countries even greater [28]. However, these problems are mostly caused by bad governance, not by problems with technology itself. The issues surrounding algorithms and the lack of creativity are true, but can be mitigated through clear regulations and ethical supervision. Africa's growing position has advantages because many countries are still working on their AI frameworks and can include accountability, fairness, and transparency from the start.

Regardless of negative views, new African research shows that AI increases rather than decreases managers' performance. Machine learning and predictive analytics are now very important for making data-driven decisions in finance, healthcare, and agriculture [19]. Managers who used to make decisions based on insufficient data can now make faster and more informed decisions. AI-powered credit and risk assessment systems in Nigeria's banking sector have lowered the number of bankruptcy claims and made it easier for small businesses to secure loans. This has given managers much more confidence [25]. South African manufacturers that use AI-based predictions have reported productivity gains of more than 20percent [18]. Eastern African logistics and transport companies also benefit from predictive maintenance, which reduces costs and makes their services more reliable [21]. The results show that AI improves managers' skills, encourages new ideas, and helps them remain flexible in their plans. This helps new African managers make smart, creative, and agile choices in changing situations [1] [26].

Although these are good things, it is important to remember that AI's success in Africa depends on strong governance frameworks. When there aren't any clear ethical rules, data protection laws, or ways for professionals to be held accountable, people might misuse the information or not trust it. However, this lack of governance is not a reason to say no to AI; it is a chance to come up with new ideas responsibly. Kenya's draft National AI Strategy and South Africa's AI ethics framework are examples of regulatory policies in Africa that show that regional organisations are already moving towards responsible governance.

VIII. IMPLICATIONS FOR HIGHER EDUCATION

The skills needed by aspiring managers are changing due to the rapid adoption of AI in African enterprises. Higher education institutions are essential to help students prepare for and benefit from this technological revolution [29]. AI can be introduced in universities to provide prospective managers with a competitive advantage. The benefits of AI include, but are not limited to, better decision-making, operational effectiveness, and strategy. Institutes can create digitally literate managers who can use AI technologies to enhance innovation, streamline operations, and anticipate market changes by introducing AI into management courses. In addition, correspondence can be improved by incorporating AI in the classroom, bridging the gap between theoretical knowledge and its practical application. Universities may provide students with hands-on practice in applying AI in the context of real business through case studies, simulations, and AI-based projects. In addition to technical expertise, this experiential training develops critical thinking, problem-solving, and adaptability skills, which

managers in dynamic African markets must possess [30] [31].

The proposed intervention will adhere to the principles of improving AI literacy in the curriculum by using the three dimensions of AI learning: knowledge, skills, and attitudes (KSA). Improving Curriculum through AI Literacy. The suggested intervention will be based on the principles of improving AI literacy in the curriculum through the three dimensions of AI learning: knowledge, skills, and attitudes (KSA). AI literacy is the foundational block of modern management education. To help students understand the technical and strategic elements of artificial intelligence, the courses must present them with machine learning, automation, predictive analytics, and what are commonly referred to as decision-support systems. New managers who have this kind of knowledge will be in a better position to critically assess AI-generated insights and make prudent decisions, as well as employ technology to create real company value.

Additionally, AI literacy reduces the dependence on intuition, as it empowers students to filter out relevant data and actions. Promote data-driven thinking, which is a much-needed skill for modern African managers who must spend resources in the most efficient manner when responding to market changes in the shortest time possible.

A. Addressing negative perspectives

Higher education institutions must train students to embrace technology by focusing on AI literacy, including its spread and positioning.

a. **Incorporating responsible AI practices and ethics:** Managers are expected to be capable of making morally and socially responsible decisions. The ethical and responsible use of AI in management should be part of the curriculum. Students are expected to study topics such as algorithmic bias, data privacy, transparency, and the overall effects of AI on society to develop a mindset that is both creative and responsible.
b. **Motivating experiential learning:** Practical experience is a critical part of the use of AI. Universities can also provide students with the experience of working with tools and data in real commercial contexts through internships, group projects, and simulations built using AI [3]. The experiential approach assists in the production of novel techniques and problem-solving skills in a non-threatening environment and enhances managerial competencies. In addition, experiential learning improves collaborative problem-solving and teamwork. To successfully use AI solutions in organisations, students gain an understanding of the challenges and opportunities associated with the introduction of AI [30] [31]. Universities are capable of producing managers who can fill the gap between theory and practice by modelling real-life business scenarios.
c. **Supporting contextual and inclusive AI education:** To ensure relevance and inclusiveness, education must be localized in the African business context. It becomes easier to understand how AI can be used to address context-specific issues for students. There are problems with local case studies, AI applications that are culturally sensitive, and regional economic facts. This plan encourages regional innovation and provides equitable access to information. Contextual education is also helpful to enable students to identify and mitigate the biases present in current AI technologies around the world. Through universities, managers can develop and adopt AI solutions that support different African communities through promotion. It will involve inclusive learning, which promotes social responsibility and technological innovation [30] [31].
d. **Enhancing university-industry cooperation:** Business partnerships are important in bridging the gap between theory and practice in the real world. African companies, AI groups, and universities can collaborate to provide internships, mentoring, and the use of state-of-the-art AI tools [3]. Students will understand the strategic value of AI in real-life business activities due to this interaction. It is also easier to work together and develop AI solutions that fit African scenarios. As students develop skills in the implementation of AI strategically and ethically, they also enable other industries in the area to be more innovative. This is because this academic-industrial partnership improves employability and prepares managers to lead AI-driven businesses successfully.
e. **Developing a sustainable AI ecosystem in higher education:** The development of an AI ecosystem should be based on effective governance structures, faculty education, and facility building. Universities will have to invest in technology laboratories, research projects, and professionalism to help students and teachers use AI tools effectively. Those programmes that are well supported ensure that AI is accepted in the long term and that its managerial and educational benefits are maximized. In addition, healthy ecosystems encourage cross-sector collaboration between government, business, and academic institutions, which contributes to continuous innovation and skill development. Institutions train managers with the ability to apply AI at the forefront through the integration of this technology with AI literacy, ethical, and applied learning in their curriculum. This drives innovation, competitiveness, and inclusive development in all of Africa.

IX. CONCLUSION AND FUTURE STUDIES

Artificial Intelligence is changing the African business environment and bringing new concepts to thinking, strategy, and leadership. AI can provide future managers with tools to improve the decision-making process, simplify work, and activate innovation. The automation of activities and the forecasting of information have made AI a fundamental element to remain competitive in the contemporary business environment in the African context. However, this massive technological progress comes at a very steep cost. The regulatory systems in Africa are not keeping up with the rapid diffusion of AI. This is the reason why there has been a risk-prone situation and there has been no resolution to

ethical and legal concerns. If systems are not in place to accommodate these risks, the same technology that can be progressive will, by default, breed new forms of inequality or discrimination.

The greatest priority that Africa should have is, therefore, AI governance based on a moral platform guaranteeing accountability, inclusiveness, and transparency. Businesses, schools, and governments should unite and provide simple rules that would protect citizens while stimulating innovation. The introduction of boards to normalize AI ethics, the promotion of digital literacy, and the increase in cooperation between policymakers and technology innovators are among the most crucial measures. For future research prospects, academics should investigate how managers navigate technostress when dealing with the high wave of AI evolution in Africa. Furthermore, academics can investigate in the future how organisations in Africa can formulate governance models (frameworks) to tackle ethical issues such as algorithmic bias, data privacy, plagiarism, and ethical decision-making. In addition, research can focus on finding the competencies and skills necessary for African managers to use AI in driving the decision-making process to improve performance and operational functions. Other future investigations can look at how to effectively use AI for inclusive growth and facilitate SMEs' growth while eliminating inequality in African management processes and economies.

REFERENCES

[1] J. Amankwah-Amoah and Y. Lu, "Harnessing AI for business development: a review of drivers and challenges in Africa," Production Planning & Control, vol. 35, no. 13, pp. 1551-1560, 2024.

[2] J. E. Ataguba, "Changes in socioeconomic inequality in self-assessed health in South Africa: The contributions of changes in inequalities between and within socioeconomic groups," SSM-Population Health, vol. 29, p. 101755, 2025.

[3] CIPIT. "The State of AI in Africa Report." The Centre for Intellectual Property and Information Technology Law (CIPIT). [retrieved: November, 2025] https://cipit.strathmore.edu/the-state-of-ai-in-africa-report/ (accessed.

[4] N. Ndung'u and L. Signé, "The Fourth Industrial Revolution and digitization will transform Africa into a global powerhouse," 2020.

[5] I. D. Mienye, Y. Sun, and E. Ileberi, "Artificial intelligence and sustainable development in Africa: A comprehensive review," Machine Learning with Applications, vol. 18, p. 100591, 2024.

[6] N. Adeleye, E. Osabuohien, and E. Bowale, "The role of institutions in the finance-inequality nexus in Sub-Saharan Africa," Journal of Contextual Economics–Schmollers Jahrbuch, no. 1–2, pp. 173-192, 2017.

[7] K. Banga and D. W. te Velde, "Digitalisation and the future of African manufacturing," Supporting Economic Transformation, 2018.

[8] H. J. Wilson and P. R. Daugherty, "Collaborative intelligence: Humans and AI are joining forces," Harvard business review, vol. 96, no. 4, pp. 114-123, 2018.

[9] A.-S. T. Olanrewaju, M. A. Hossain, N. Whiteside, and P. Mercieca, "Social media and entrepreneurship research: A literature review," International journal of information management, vol. 50, pp. 90-110, 2020.

[10] G. Paré and S. Kitsiou, "Methods for literature reviews," in Handbook of eHealth evaluation: An evidence-based approach [Internet]: University of Victoria, 2017.

[11] N. R. Haddaway et al., "Eight problems with literature reviews and how to fix them," Nature Ecology & Evolution, vol. 4, no. 12, pp. 1582-1589, 2020.

[12] H. Suri and D. Clarke, "Advancements in research synthesis methods: From a methodologically inclusive perspective," Review of Educational Research, vol. 79, no. 1, pp. 395-430, 2009.

[13] B. M. Silva, J. J. Rodrigues, I. de la Torre Díez, M. López-Coronado, and K. Saleem, "Mobile-health: A review of current state in 2015," Journal of biomedical informatics, vol. 56, pp. 265-272, 2015.

[14] M. Pautasso, "The structure and conduct of a narrative literature review," A guide to the scientific career: Virtues, communication, research and academic writing, pp. 299-310, 2019.

[15] D. Tranfield, D. Denyer, and P. Smart, "Towards a methodology for developing evidence‐informed management knowledge by means of systematic review," British journal of management, vol. 14, no. 3, pp. 207-222, 2003.

[16] R. F. Baumeister and M. R. Leary, "Writing narrative literature reviews," Review of general psychology, vol. 1, no. 3, pp. 311-320, 1997.

[17] M. J. Grant and A. Booth, "A typology of reviews: an analysis of 14 review types and associated methodologies," Health information & libraries journal, vol. 26, no. 2, pp. 91-108, 2009.

[18] M. L. Nzama, G. A. Epizitone, S. P. Moyane, N. Nkomo, and P. P. Mthalane, "The influence of artificial intelligence on the manufacturing industry in South Africa," South African Journal of Economic and Management Sciences, vol. 27, no. 1, pp. 1-10, 2024.

[19] H. A. du Toit, W. D. Schutte, and H. Raubenheimer, "Integrating traditional and non-traditional model risk frameworks in credit scoring," South African Journal of Economic and Management Sciences, vol. 27, no. 1, p. 5786, 2024.

[20] G. Culot, M. Podrecca, and G. Nassimbeni, "Artificial intelligence in supply chain management: A systematic literature review of empirical studies and research directions," Computers in industry, vol. 162, p. 104132, 2024.

[21] S. M. Musa, U. A. Haruna, L. J. Aliyu, M. Zubairu, and D. E. Lucero-Prisno III, "Leveraging AI to optimize vaccines supply chain and logistics in Africa: opportunities and challenges," Frontiers in pharmacology, vol. 16, p. 1531141, 2025.

[22] A. Gawe and A. Gudyanga, "The impact of artificial intelligence (AI) on development studies' learners at one State university campus in Zvishavane, Zimbabwe," Cogent Education, vol. 12, no. 1, p. 2539217, 2025.

[23] M. D. Dzandu, S. T. Asiedu, B. Pathak, and S. De Cesare, "Artificial Intelligence Harm and Accountability by Businesses-A Systematic Literature Review'," in Proceedings of the 27th International Conference on Enterprise Information Systems, 2025: SCITEPRESS, pp. 1012-1019.

[24] E. Egbeleo and K. Sodokin, "Digital transformation, institutional quality and productivity in Sub-Saharan Africa," Cogent Economics & Finance, vol. 13, no. 1, p. 2519924, 2025.

[25] D. Kazimoto and S. Baadel, "Empowering Africa's Disfranchised SMEs: Machine Learning-Based Credit Scoring for Informal African Merchants," Human-Centric Intelligent Systems, vol. 5, no. 3, pp. 376-384, 2025.

[26] S. Bawa, "Leveraging artificial intelligence and dynamic supply chains for renewable energy development in Africa's frontier markets," Energy & Environment, p. 0958305X251375931, 2025.

[27] R. Butson and R. Spronken-Smith, "AI and its implications for research in higher education: A critical dialogue," Higher Education Research & Development, vol. 43, no. 3, pp. 563-577, 2024.

[28] M. C.-T. Tai, "The impact of artificial intelligence on human society and bioethics," Tzu chi medical journal, vol. 32, no. 4, pp. 339-343, 2020.

[29] A. B. Samuels and U. Singh, "Education reimagined: South Africa's journey through the 4IR and beyond," Transformation in Higher Education, vol. 10, p. 482, 2025.

[30] E. O. Arakpogun, Z. Elsahn, F. Olan, and F. Elsahn, "Artificial intelligence in Africa: Challenges and opportunities," The fourth industrial revolution: Implementation of artificial intelligence for growing business success, pp. 375-388, 2021.

[31] F. Azaroual, "Artificial intelligence in Africa: challenges and opportunities," Policy Brief. PB, pp. 23-24, 2024.

[32] G. Paré, and S. Kitsiou, "Methods for literature reviews. In Handbook of eHealth evaluation: An evidence-based approach [Internet]". University of Victoria, 2021