GPTMB 2025

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

ISBN: 978-1-68558-287-6

July 6th– 10th, 2025

Venice, Italy

GPTMB 2025 Editors

Júlio Monteiro Teixeira, Federal University of Santa Catarina (UFSC), Brazil

Clement Leung, Chinese University of Hong Kong Shenzhen, China

GPTMB 2025

Forward

The Second International Conference on Generative Pre-trained Transformer Models and Beyond (GPTMB 2025), held on July 6th- 10th, 2025 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 2025 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 2025. 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 2025 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that GPTMB 2025 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 Venice provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

GPTMB 2025 Chairs

GPTMB 2025 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

GPTMB 2025

Committee

GPTMB 2025 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

GPTMB 2025 Technical Program Committee

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
Maksim Eremeev, Genentech, USA
Gerhard Heyer, Universität Leipzig, Germany
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
Prajna Upadhyay, BITS Pilani Hyderabad Campus, India
Pierre Vilar Dantas, Federal University of Amazonas (UFAM), Manaus, Brazil
Hao Yan, George Mason University, USA
Abdou Youssef, The George Washington University, USA

Copyright Information

Table of Contents

Designing A New Graduate Course on Artificial Intelligence for Cybersecurity

Ping Wang
Department of Computer and Information Systems
Robert Morris University
Pittsburgh, PA, USA
wangp@rmu.edu

***Abstract* — Artificial intelligence (AI) technologies and solutions are increasingly integrated into various applications and domains of studies. Generative AI (Gen AI) also has significant impacts and implications for the fast-growing field of Cybersecurity and cybersecurity education for workforce development. This research proposes the design of a new graduate master's level credit course to integrate AI into cybersecurity education. This new course explores the evolving impacts of artificial intelligence on the cybersecurity ecosystem. The course is intended for students to learn to identify and evaluate AI-powered cyber threats and attacks and their implications as well as to utilize AI-powered systems for enhancing cyber threat detection, incident response, security automation, vulnerability analytics, and security risk assessment. The proposed course design will summarize initial suggestions of main topics, outcomes, activities, and assessment criteria for implementation.**

Keywords* – *AI; cybersecurity; vulnerability; learning outcomes; assessment.

I. Introduction

Artificial intelligence (AI) is a fast-growing, promising, inter-disciplinary and comprehensive technology solution supported by advanced computing, machine learning, data and knowledge representation, robotics, and optimization. With the strong potential to increase automation, efficiency and productivity, Generative AI (Gen AI) is increasingly adopted and used in various industries and fields of studies including Cybersecurity. Cybersecurity is also an increasingly critical area for national security and economic prosperity in the digital age due to rising and evolving cyber threats and risks. As a double-edged sword, Gen AI powered tools and solutions present opportunities for more efficient and effective cybersecurity measures such as in network traffic analysis and in cyber threat detection, risk assessment, and incident response, along with risks and challenges for cybersecurity in the case of malicious use of AI for more devasting cyber-attacks [1]-[3].

There are strong short-term and long-term demands for skilled workers in Cybersecurity around the world and especially in the United States that are projected to far outpace the average national job growth in the next decade [4][5]. Higher education is the main avenue expected for providing the pipeline of qualified professionals to meet the growing cybersecurity workforce demand. The U.S. National Centers of Academic Excellence in Cybersecurity (NCAE-C) designation program jointly sponsored by the National Security Agency (NSA) and Department of Homeland Security (DHS) is a national standard for reviewing, certifying, and maintaining high quality of cybersecurity education programs with rigorous and consistent requirements for program evaluation as well as up-to-date knowledge units (KUs) aligned to cybersecurity knowledge, skills, and abilities (KSAs) [6][7]. A recent global cybersecurity workforce study report shows that cybersecurity organizations and professionals need to keep up with AI as a major technology innovation in order to maintain and improve their efficiency and agility [8]. Therefore, cybersecurity education programs need to incorporate AI in the curriculum and course design. This study will briefly review relevant background and summarize the initial design of the proposed graduate AI for Cybersecurity course.

II. Background

Gen AI solutions have the capacity to help cybersecurity professionals to detect, analyze, and defend against cyber threats and attacks. Specific to cybersecurity, large language models (LLMs) and generative security models of Gen AI bring the major benefits of early threat detection, efficiency and accuracy in vulnerability and threat analysis and risk assessment, automated incident response, preventive and secure software development, as well as efficient training of cybersecurity professionals [9][10]. Recent research on AI for Cybersecurity shows that Gen AI applications have the capacity and strengths to automate repetitive security tasks, speed up cyber threat detection, penetration testing and response, and improve the accuracy of countermeasures to address cyber vulnerabilities and risks [11]-[13]. Therefore, a new course on AI for Cybersecurity should cover the security benefits of Gen AI and its applications and models.

Gen AI can be a double-edged sword to Cybersecurity, which also brings risks, challenges, and limitations for cybersecurity solutions. Unauthorized and malicious users could use AI tools to generate code and launch more powerful and devastating attacks and exploitations targeting known vulnerabilities [1][14]. For legitimate users, Gen AI applications, such as ChatGPT, may provide misleading results or "hallucinations", which is a substantial limitation [14][15]. In addition, there are concerns with the security and privacy risks of Gen AI applications that may disclose private and confidential user data on public domains [3][14][16]. Therefore, a new course on AI for Cybersecurity should also reveal and address the risks and limitations of AI models and applications.

For pedagogical and educational effectiveness, a new course design should reflect the cognitive development process of different levels or stages of learning objectives in the updated

ISBN: 978-1-68558-287-6

Bloom’s taxonomy, which lays out the following 6 levels of progressive learning objectives and achievements [17]:

- Recall information, facts, terms, and basic concepts
- Describe and interpret facts and ideas to demonstrate comprehension
- Apply knowledge and techniques learned to solve problems in new situations
- Analyze information to identify causes, motives, and relationships
- Evaluate information or ideas based on certain criteria to make judgements
- Develop and propose new or alternative solutions

III. PROPOSAL

The proposed new course is a 3-credit course for a master’s degree program in applied AI at an NCAE-C designated university in the United States. The new course focuses on the evolving impacts of AI on the cybersecurity landscape and teaches students to identify and evaluate AI-powered cyber risks and solutions. The specific learning outcomes are:

- Identify and describe AI-powered cyber threats and attacks
- Evaluate AI-powered cyber threats and attacks and security implications and solutions
- Identify and describe positive impacts of AI in cybersecurity
- Identify and apply AI-driven solutions, techniques, and tools for cybersecurity
- Evaluate secure development practices for protecting applications in the age of AI
- Assess and evaluate AI-powered cybersecurity risks and solutions.

For a graduate level course, it is important include more advanced level learning objectives of analysis, evaluation, and solution development in Bloom’s taxonomy.

A variety of teaching and learning activities are suggested for this new course, including presentations, hands-on demos, discussions, and a comprehensive project assignment for problem solving. The project assignment includes progressive development of an initial project plan involving identification of AI-related cyber threats and risks, a midterm progress report, and a final report and presentation that are submitted for grading and assessment. The main assessment criteria for the project include problem description, analysis, and evaluation and discussion of proposed solutions, tools, and methods. Student presentations demonstrate their problem solving skills.

IV. CONCLUSION

This abstract presents preliminary research on proposing a new graduate course on AI for Cybersecurity. Future research will report the actual implementation, empirical data, and areas of improvements identified for the course design.

ACKNOWLEDGEMENT

This research is supported by a grant from the U.S. National Science Foundation (NSF) – NSF Grant ID 2234554.

REFERENCES

[1] S. Wilson, “Cybersecurity and artificial intelligence: Threats and opportunities,” Contrast Security, 2023.

[2] B. Schneier, “The coming AI hackers,” in The Cyber Projecct: Council for the Responsible Use of AI, Harvard Kennedy School, 2021.

[3] NIST/US Department of Commerce, “NIST Trustworthy and Responsible AI, NIST AI 600-1,” July 2024. Available: https://doi.org/10.6028/NIST.AI.600-1

[4] U.S. BLS, “Occupational Outlook Handbook – Information Security Analysts,” 2025. Available: https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

[5] M. Hogan et al., “Cybersecurity Workforce Data Initiative: Cybersecurity Workforce Supply and Demand Report,” National Science Foundation, 2024. Available: https://ncses.nsf.gov/about/cybersecurity-workforce-data-initiative

[6] P. Wang, M. Dawson, K.L. Williams, “Improving cyber defense education through national standard alignment: Case studies,” *International Journal of Hyperconnectivity and Internet of Things.* 2018, *2*(1), pp. 12-28.

[7] U.S. National Security Agency, “National Centers of Academic Excellence in Cybersecurity,” Available: https://www.nsa.gov/Academics/Centers-of-Academic-Excellence/

[8] (ISC)2, “Global Cybersecurity Workforce Prepares for an AI-Driven World,” 2024. Available: https://www.isc2.org/research

[9] A. Mamgai, “Generative AI with cybersecurity: friend or foe of digital transformation?”. Available: https://www.isaca.org/resources/news-and-trends/industry-news/2023

[10] P. Wang and H. D’Cruze, "AI-Assisted Pentesting Using ChatGPT-4" In Advances in Intelligent Systems and Computing, vol 1456. Springer, 2024. Available: https://doi.org/10.1007/978-3-031-56599-1_9

[11] R. Kaur, D. Gabrijelcic, and T. Klobucar, (2023). “Artificial intelligence for cybersecurity: Literature review and future research directions, Information Fusion 97 (2023) 101804, pp. 1-29

[12] P. Wang and H. D’Cruze, "AI-Assisted Pentesting Using ChatGPT-4" In Advances in Intelligent Systems and Computing, vol 1456. Springer, 2024. Available: https://doi.org/10.1007/978-3-031-56599-1_9

[13] S. Temara, “Maximizing penetration testing success with effective reconnaissance techniques using ChatGPT,” Research Square, 2023, pp. 1-10, DOI: https://doi.org/10.21203/rs.3.rs-2707376/v1

[14] M. Gupta, K. Aryal, and L. Praharaj, “From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy,” in IEEE Access, vol. 11, 2023, pp. 80218-80245

[15] X. Zhan, Y. Xu, and S. Sarkadi, “Deceptive AI ecosystems: The case of ChatGPT,” In ACM conference on Conversational User Interfaces (CUI ’23), July 19–21, 2023, Eindhoven, Netherlands.

[16] World Economic Forum, “Artificial Intelligence and Cybersecurity: Balancing Risks and Rewards”, White Paper, January 2025. Available: https://reports.weforum.org/

[17] L.W. Anderson, and D.R. Krathwohl, A taxonomy for learning, teaching, and assessing, Boston: MA: Allyn and Bacon, 2001.

ISBN: 978-1-68558-287-6

Retrieval Performance in RAG Systems: A Component-Level Evaluation Framework

Alexander Kreß*, Alexander Lawall*, Thomas Zöller*
*IU International University of Applied Science
Erfurt, Germany
alexander.kress@iu-study.org, alexander.lawall@iu.org, thomas.zoeller@iu.org

***Abstract*—Retrieval-Augmented Generation (RAG) systems are relevant for improving factuality in Large Language Model (LLM) outputs, yet their evaluation remains challenging due to their multi-component architecture. This paper introduces plot-RAG (pRAG), a novel evaluation framework that visualizes component-level performance in RAG systems, providing granular insights into retrieval and re-ranking processes, without requiring resource-intensive LLM-based evaluation. The effectiveness of pRAG is demonstrated by analyzing a real-world technical documentation question-answering system. Additionally, the methodology for generating and validating synthetic evaluation datasets is presented, showing they can match or exceed manually prepared datasets for RAG assessment. The experiments confirm that the retrieval component represents the most critical performance bottleneck in RAG systems, and a formula is provided to determine the optimal retrieval size based on response time requirements. These contributions enable a more efficient and targeted evaluation of RAG systems, particularly in specialized domains where the creation of ground truth data typically requires substantial expert involvement.**

Index Terms*—*retrieval-augmented generation; evaluation framework; synthetic datasets; component-level analysis.

I. Introduction

Retrieval-Augmented Generation (RAG) systems are important for improving the factuality and reliability of Large Language Model (LLM) outputs, especially in domain-specific applications. Despite their adoption, evaluating these systems remains challenging, particularly when considering their multi-component nature and varying performance across different use cases [1].

A. Motivation and Problem Statement

The evaluation of RAG systems faces several challenges that current LLM approaches fail to adequately address. While LLMs alone can be evaluated using established benchmarks, RAG systems introduce additional complexities due to their multi-stage architecture spanning document processing, retrieval, re-ranking, and generation components [2][3]. As noted by [2], dynamic data environments further complicate evaluation, as the underlying knowledge sources often change over time.

Current evaluation frameworks typically produce aggregate metrics that mask the performance of individual components, making it difficult to identify specific bottlenecks or optimization opportunities [2][4]. Manual evaluation methods are becoming increasingly inefficient, necessitating automated approaches that can scale with system complexity. Additionally, temporal aspects of RAG performance — such as latency variations across different technical configurations — are rarely incorporated into evaluation methodologies despite their critical importance in real-world applications [2][5].

Established benchmark datasets like HotpotQA [6] and MS MARCO [7] have proven inadequate for evaluating modern RAG systems [8], as they fail to capture the nuanced retrieval and generation scenarios encountered in specialized domains. The availability of ground truth data will become rare in the future [9]. While synthetic dataset generation offers promising alternatives [10], systematic approaches for validating these datasets and incorporating them into holistic evaluation frameworks remain underdeveloped.

B. Research Gap

Despite the proliferation of evaluation methods for RAG systems, major gaps persist in current approaches. Existing frameworks like RAGAS [11] rarely provide granular insights into component-level performance, instead focusing on end-to-end evaluation that obscures the contribution of individual technical elements [2][3]. As [12] observes, the various technical alternatives available at each stage of the RAG pipeline create a complex evaluation space that remains largely unexplored. [13] mentioned that this gap cannot be closed by asking LLMs for reasoning.

The role of re-ranking models [14][15] and hybrid retrieval techniques like the combination of embeddings and BM25 [16] in RAG performance is inadequately addressed by current evaluation approaches. Furthermore, while the importance of synthetic datasets for evaluation is increasingly recognized [17], methodologies for generating and validating these datasets remain ad-hoc and not standardized. These gaps demand a comprehensive evaluation framework that addresses both the technical complexity of RAG systems and the practical challenges of meaningful assessment [10].

C. Research Questions

This paper addresses the primary research question: "Which technical concepts are necessary to successfully evaluate RAG systems?". Secondary research questions are investigated to explore this question more comprehensively:

1) "How can we effectively evaluate the retrieval component in RAG systems?"
2) "How can synthetic datasets be efficiently generated and validated for RAG evaluation?"
3) "What approaches show promise for evaluating the entire RAG pipeline?"

D. Contributions

This paper makes several contributions to the field of RAG system evaluation:

- The introduction of a methodology for generating and validating synthetic evaluation datasets that can scale efficiently across domains and use cases.
- The development of a visualization approach (pRAG) for assessing retrieval component performance that incorporates both quality metrics and temporal analysis.
- The provision of empirical findings from applying the framework to a real-world RAG system designed for technical documentation question answering.

The framework addresses critical gaps in existing evaluation approaches by offering a more granular, component-specific assessment methodology that can adapt to the evolving landscape of RAG system design.

E. Paper Structure

The remainder of this paper is organized as follows: Section II describes the methodology, including the architecture of the evaluation framework, synthetic dataset generation approach, and component-specific assessment techniques. Section III presents the results of applying the framework to a case study RAG system and discusses key findings and implications. Finally, Section IV concludes the paper and outlines directions for future work.

II. Methodology

A. RAG System Architecture

The RAG system employs a microservice-based architecture designed for scalability and modular development. The system processes user queries through these key components: When a user submits a query via the frontend, the middleware API coordinates the workflow. First, the pre-processing API generates keywords and embeddings from the query for semantic comparison. These are passed to a vector handling API that performs hybrid retrieval, combining BM25 [18], keyword matching, and embedding-based semantic search through paradeDB, a PostgreSQL extension supporting vector operations.

Retrieved contexts and metadata flow back to the middleware API, which forwards them to the pre-processing API where a cross-encoder re-ranker prioritizes the most semantically relevant documents. Finally, these re-ranked contexts together with an initial prompt are provided to a LLM that generates a comprehensive response based on the available information and returns it to the user via the frontend.

B. System Implementation

The system is deployed on a Kubernetes cluster with the frontend developed in React and backend services in Python. For the knowledge base, 50 technical documents from HORSCH machinery manuals using Azure Document Intelligence are processed to convert PDF content into processable text.

For embedding generation, the Hugging Face multi-qa-MiniLM-L6-cos-v1 Sentence Transformer model [19] was implemented, selected for its balance of English language capabilities and computational efficiency. Documents were chunked to match the model's maximum token length and stored with machine-specific metadata.

We evaluated two cross-encoder models for re-ranking: msmarco-MiniLM-L6-en-dev1 [20] and ms-marco-MiniLM-L-6-v2 [21], which reorder retrieved contexts based on query relevance. For response generation, we utilized ChatGPT-3.5-Turbo-0125 with crafted prompts to ensure responses were relevant, accurate, and focused on HORSCH machinery documentation.

Performance timing was implemented using Python's time module, capturing execution duration for each component to enable system optimization.

C. plot-RAG (pRAG): A Novel Evaluation Framework

1) Motivation and Design: A key contribution of this work is pRAG, a novel visualization and evaluation framework specifically designed to address the lack of quantitative, interpretable evaluation methods for RAG systems. pRAG provides granular insights into the performance of individual RAG components, particularly the critical retrieval and re-ranking stages. This contrasts with current evaluation approaches, which often focus on end-to-end performance or rely on limited metrics like recall and precision, which are susceptible to outliers [22].

2) Visualization Components: The pRAG visualization (see Figure 1) displays multiple dimensions of system performance simultaneously:

- **Context position tracking:** Visualizes where relevant contexts from ground truth appear in both retrieval and re-ranked results (blue numbers).
- **Retrieval method comparison:** Distinguishes between embedding-based and BM25 keyword-based retrievals (y-axis).
- **Ground truth distribution:** Shows distances between relevant contexts in the document corpus (green numbers).
- **Quantitative metrics overlay:** Presents calculated performance metrics alongside visual representations (top right corner).
- **Right contexts quantity:** Number of relevant contexts from ground truth at this position based on the entire evaluated data set (numbers in parentheses).

3) Metrics Integration: pRAG calculates and visualizes several critical metrics:

- **Specialized recall metrics:**
  - Recall Emb: Effectiveness of embedding-based retrieval
  - Recall BM25: Performance of keyword-based retrieval
  - Recall Full Retrieval: Combined unique contexts retrieval rate
  - Recall Reranking from Retrieval: Preservation of relevant contexts after re-ranking
- **Ranking quality:** Normalized Discounted Cumulative Gain (NDCG) calculation highlighting the importance of positioning relevant information earlier in results
- **Retrieval optimization:** Recommended retrieval sizes for both embedding and BM25 components

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-287-6

4) Actionable Insights: The pRAG framework provides actionable insights by visually exposing:

- Which retrieval method (BM25 or embeddings) more effectively captures relevant contexts
- How effectively the re-ranker prioritizes relevant contexts
- Optimal retrieval configuration parameters
- Performance bottlenecks in specific components

This visualization approach enables the identification of system weaknesses without requiring extensive manual analysis, making it particularly valuable for ongoing RAG system development and optimization.

Figure 1 shows the unitization of the pRAG approach in the analysis of retriever performance. The particular results are further discussed in Section III-A

D. Synthetic Dataset Generation

For evaluation, both manually curated and synthetically generated question-answer pairs based on three technical manuals for products from the HORSCH portfolio: Avatar 12/40 SD, Joker RX, and Tiger MT were created. These documents were selected based on machine sales volume analysis, indicating likely user query subjects.

For each document, we prepared 50 question-answer pairs with relevant contexts as ground truth. From each set, five pairs were randomly selected as examples for synthetic generation. Using these examples iteratively with different ground truth contexts, we generated 45 synthetic question-answer pairs per document using three different language models: GPT-4o-Mini, Gemini-1.5-Flash, and Nemotron-4-340b-Instruct. Also, two comparison methodologies were implemented:

1) **Absolute comparison:** evaluating curated vs. synthetic datasets based on different contexts
2) **Relative comparison:** generating synthetic data using ground truth from the curated dataset

Quality assessment employed a GAN-like approach, using language models (GPT-4o-Mini, Llama-3-Patronus-Lynx-8B-Instruct [23], and Prometheus-7b-v2.0 [24]) as discriminators to evaluate response quality with Pass/Fail determinations and comparative quality judgments.

E. Experimental Setup

Multiple experimental configurations are conceptualized to evaluate different aspects of the RAG system and demonstrate the utility of the pRAG framework. For enhanced retrieval configurations, we used a basic setup and changed specific technical components for enhanced setups:

a) Basic Synthetic Data Evaluation (Setup A):

- Generator: ChatGPT-3.5-Turbo-0125
- Retrieval: paradeDB (BM25+Embeddings)
- Re-ranker: Cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Retrieval size: 8 contexts each for BM25 and embeddings
- Re-ranking size: 4 contexts

b) Enhanced Retrieval Configuration (Setup B):

- Generator: ChatGPT-3.5-Turbo-0125
- Metadata: Machine Name
- Re-ranker: Cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Retrieval size: 60 contexts (BM25+Embeddings)
- Re-ranking size: 20 contexts

c) Enhanced Retrieval Configuration (Setup B-1):

- New Re-ranker: Cross-encoder/ms-marco-MiniLM-L-6-v2

d) Enhanced Retrieval Configuration (Setup B-2):

- New Method: HyDE Integration

For additional quantitative evaluation, we implemented the RAGAS framework to assess context precision, answer credibility, relevance, and accuracy. We compared RAGAS with the pRAG framework to substantiate the validity of the pRAG approach. We supplemented the pRAG approach with timing analysis of each system component, capturing minimum, maximum, and median execution times.

The expert evaluation was conducted with domain specialists who assessed question-answer pair quality in a blinded format, comparing synthesized and manually curated responses without knowledge of their origin to eliminate bias. For this experiment, we used the Basic Setup B with three different datasets.

In this comprehensive methodology using the novel pRAG evaluation framework, we aimed to evaluate not only the overall RAG system performance but also the viability of synthetic data for ongoing system improvement. We aimed to address the issue of available ground truth datasets by generating synthetic datasets automatically based on contexts from the database.

III. Results and Discussion

A. Performance Analysis of RAG Components Using pRAG

1) Retrieval Component Performance: The analysis demonstrates that each component contributes differently to the overall performance and can be individually assessed through visualization with pRAG. Figure 1 illustrates the pRAG visualization, where the positions of contexts in the ground truth collection are mapped against their retrieval positions. It shows that several relevant contexts (positions 6, 7, and 8) were missed by BM25 but captured by embedding-based retrieval. “Large” gaps in the diagram can support decision-making on whether increasing the retrieval size at the cost of performance should be implemented to identify only a few additional relevant contexts. The automated evaluation of RAG systems with pRAG does not require an LLM as a judge. This makes the evaluation more resource-efficient, considering the substantial computational power required by LLMs.

Since pRAG visualizes the full set of retrieved contexts, the precision metric can be omitted. However, integrating the recall value into the diagram is beneficial to complement the visualization with a quantitative metric. The average values from setups in Section II-E b), c), and d) are presented in Table I.

The pRAG visualization enabled a dedicated evaluation of retrieval techniques, revealing that relevant contexts were retrieved either through BM25 or embedding-based methods

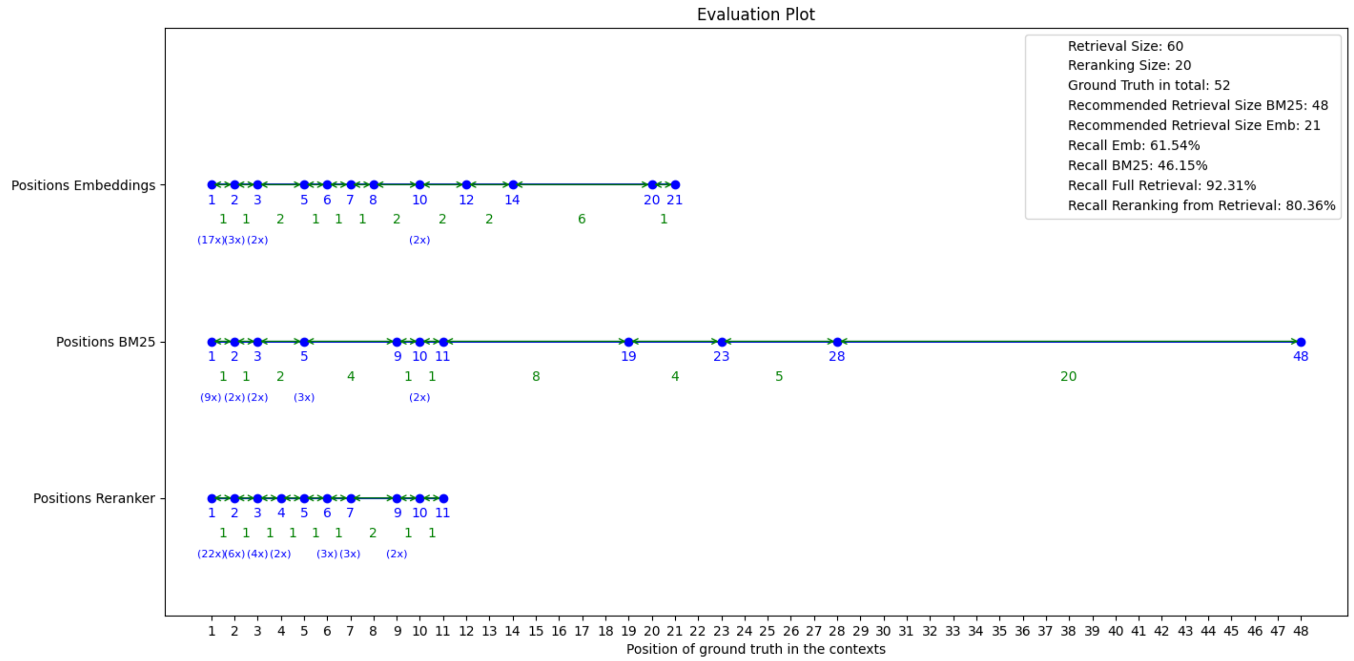

Figure 1. pRAG Visualization Showing Position of Retrieved Contexts Relative to Ground Truth based on Setup B-1.

TABLE I
RECALL METRICS ACROSS ENHANCED RETRIEVAL CONFIGURATION

Metric	Setup B	Setup B-1	Setup B-2
Recall BM25	0.4314	0.4615	0.3654
Recall Embeddings	0.6863	0.6154	0.3846
Recall Full Retrieval	0.9412	0.9231	0.6923

and also in both. This points out the importance of evaluating different retrieval strategies individually, as relevant contexts may be identified in one approach but not in another. Further analysis demonstrated that embedding-based retrieval significantly outperformed lexical methods for datasets containing technical terminology. This underlines the necessity of hybrid retrieval approaches, where the combination of strategies ensures a more comprehensive retrieval process and improves overall performance.

2) Optimal Retrieval Size Determination: The experiments demonstrate a relationship between retrieval size and answer quality. With higher retrieval size RAG systems have to handle more irrelevant contexts. Therefore, the optimal retrieval size can be determined using:

$$\text{Retrieval Size} = \frac{\text{Current Retrieval Size} \times \text{Average Response Time}}{\text{Acceptable Response Time}}$$

This formula provides a practical guideline for balancing response time against completeness. As shown in Figure 1, there is no need to put the retrieval size to 60 because the latest relevant contexts were found in positions 21 by embeddings and position 48 by BM25.

The pRAG analysis revealed diminishing returns beyond certain retrieval sizes. For example, in setup B-2 (cf. Figure 1), increasing BM25 retrieval size from 28 to 48 yielded only one additional relevant context, suggesting a practical cut-off point based on efficiency considerations.

B. Synthetic Dataset Evaluation Results

1) Comparative Quality Assessment: To assess the effectiveness of synthetic versus manually prepared datasets, we evaluated both using specialized discriminator models. The results of this evaluation indicate that synthetically generated data achieves comparable or superior performance. Specifically, manually prepared datasets did not offer a notable advantage, and Lynx even performed better on the synthetic data. This confirms that synthetic datasets can provide a similar level of performance to manually prepared ones. Detailed performance results are presented in Figure 2.

TABLE II
GENERATOR-DISCRIMINATOR COMBINATIONS

No.	Generator - Discriminator Combination
1	GPT-4o Mini - GPT-4o Mini
2	Gemini-1.5-Flash - GPT-4o Mini
3	Neomotron-4-340b-Inst. - GPT-4o Mini
4	GPT-4o Mini - Llama-3-Patronus-Lynx-8B-Inst.
5	Gemini-1.5-Flash - Llama-3-Patronus-Lynx-8B-Inst.
6	Neomotron-4-340b-Inst. - Llama-3-Patronus-Lynx-8B-Inst.
7	GPT-4o Mini - Prometheus-7b-v2.0
8	Gemini-1.5-Flash - Prometheus-7b-v2.0
9	Neomotron-4-340b-Inst. - Prometheus-7b-v2.0

2) Human Expert Validation: Human evaluators assessed pairs of question-answer examples from both dataset types. In

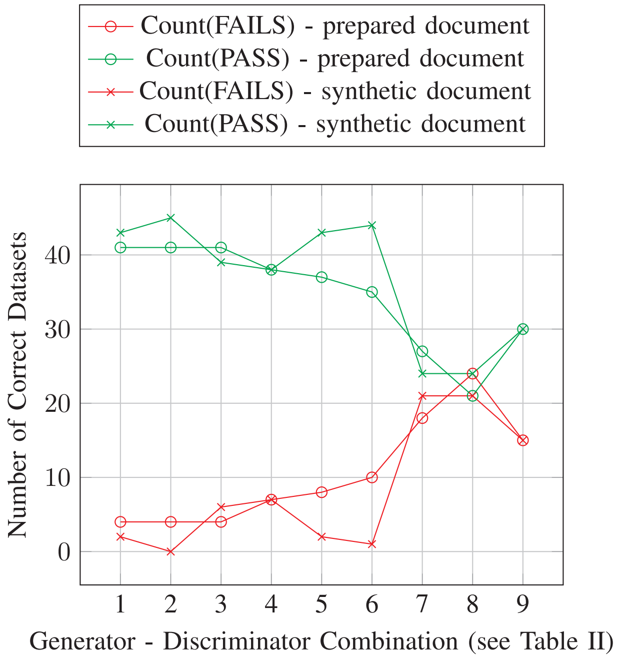

Figure 2. Comparison of Prepared and Synthetic Document for Different Generator-Discriminator Combinations.

68% of comparable cases, experts preferred synthetically generated data, with the remaining 32% showing no clear preference. Table III summarizes these findings. Expert evaluators noted that synthetic datasets showed stronger logical coherence and clearer question formulation. However, they identified a consequential limitation: synthetic datasets generated from tabular data frequently contained factual errors or misinterpretations of numerical relationships, suggesting a specific weakness in current LLM approaches to tabular content. For the evaluation, we used setup B with the three different datasets. The model Nemotron was chosen for its ability to generate better synthetic data [25].

TABLE III
HUMAN EXPERT PREFERENCES IN DATASET EVALUATION

Dataset Pair	Prefer Prepared	Prefer Synthetic	No Preference
Avatar/Nemotron	14	20	16
JokerRX/Nemotron	8	9	31
TigerMT/Nemotron	4	24	22

C. Technical Component Performance Insights

1) Comparative Analysis of Retrieval Enhancements: We evaluated technical enhancements to the base RAG architecture, including re-ranking models and the integration of HyDE (Hypothetical Document Embeddings) [26]. This method decomposes dense retrieval into two distinct tasks: First, it uses an instruction-following language model (like InstructGPT) to generate a hypothetical document in response to a user query. In the second step, an unsupervised contrastively-trained encoder (like Contriever) encodes this hypothetical document into an embedding vector. This vector identifies a neighborhood in the corpus embedding space, from which similar real documents are retrieved based on vector similarity. Table IV summarizes these findings.

TABLE IV
PERFORMANCE COMPARISON OF RETRIEVAL ENHANCEMENT TECHNIQUES

Technique	Recall Re-rank from Retrieval	NDCG	Mean Resp. Time (s)
msmarco-MiniLM-L6-en-de-v1	0.7544	0.54	3.41
ms-marco-MiniLM-L-6-v2	0.8036	0.63	4.08
HyDE Integration	0.7179	0.35	5.43

Contrary to [26], the HyDE approach showed reduced performance despite increased processing time. The pRAG analysis revealed that HyDE's theoretical advantage in generating better query representations did not improve the retrieval of relevant contexts in our test datasets.

Among re-ranking models, ms-marco-MiniLM-L-6-v2 demonstrated the best performance with 80% recall from retrieval but required 20% more processing time than msmarco-MiniLM-L6-en-de-v1. The time-performance analysis shows this tradeoff across various system components.

IV. CONCLUSION AND FUTURE WORK

This paper contributes to the evaluation methodology of RAG systems. Our primary findings are the critical role of ground truth data in conducting valid evaluations of domain-specific RAG applications.

A. Key Contributions

Our research has validated three key advances in RAG evaluation:

1) **The pRAG Visualization:** pRAG provides granular insights into component-level performance that conventional aggregated metrics cannot reveal. This approach allows precise identification of retrieval bottlenecks and optimization opportunities within complex RAG architectures. The visualization-based approach of pRAG offers insights into system performance beyond what metrics-only frameworks provide. pRAG is a resource-saving evaluation technology for RAG systems without any usage of LLM-powered evaluation.
2) **Viability of Synthetic Datasets:** The results confirm comparable or superior evaluation quality of synthetically generated question-answer pairs compared to manually prepared datasets. This significantly reduces the resource burden for domain-specific RAG applications while maintaining evaluation rigor.
3) **Retrieval Optimization Guidelines:** The retrieval component represents the most critical performance bottleneck in RAG systems and we provide a practical formula for determining optimal retrieval size based on response time requirements.

The integration of these approaches enables more efficient and targeted evaluation of RAG systems, particularly in specialized domains where ground truth data creation conventionally requires substantial expert involvement.

B. Future Research Directions

Several promising research directions emerge from this work:

1) **Dynamic Evaluation of Evolving RAG Systems:** Future research should explore automated evaluation approaches for continuously changing RAG systems, potentially integrating pRAG with streaming metrics.
2) **Multi-modal Data Analysis:** Our work focused exclusively on textual data. Extending these evaluation methods to incorporate images, tables, and other data modalities represents an important next step.
3) **Enhanced Synthetic Data Generation:** While our synthetic datasets performed well, specific weaknesses were identified with tabular data. Future work should address these limitations and explore character-based generation approaches to increase dataset heterogeneity.
4) **Generator Component Analysis:** The relationship between retrieval metrics and generation quality is of interest. Future work should explore how retrieved contexts influence the generation process and final answer quality.

In conclusion, the combination of pRAG visualization and synthetic dataset generation represents an advancement in RAG system evaluation methodology. These approaches provide practical tools for researchers and practitioners seeking to optimize RAG implementations for specialized knowledge domains in a more efficient and targeted assessment of individual components.

REFERENCES

[1] X. Wang *et al.*, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 716–17 736.

[2] H. Yu *et al.*, "Evaluation of retrieval-augmented generation: A survey," in *CCF Conference on Big Data*. Springer, 2024, pp. 102–120.

[3] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 194–199.

[4] A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2395–2400.

[5] T. Kenneweg, P. Kenneweg, and B. Hammer, "Retrieval augmented generation systems: Automatic dataset creation, evaluation and boolean agent setup," *arXiv preprint arXiv:2403.00820*, 2024, [retrieved: May, 2025].

[6] Z. Yang *et al.*, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018, [retrieved: May, 2025].

[7] D. F. Campos *et al.*, "Ms marco: A human generated machine reading comprehension dataset," *ArXiv*, vol. abs/1611.09268, 2016, [retrieved: May, 2025]. [Online]. Available: https://api.semanticscholar.org/CorpusID:1289517

[8] K. Zhu *et al.*, "Rageval: Scenario specific rag evaluation dataset generation framework," *arXiv preprint arXiv:2408.01262*, 2024, [retrieved: May, 2025].

[9] R. Liu *et al.*, "Best practices and lessons learned on synthetic data," *arXiv preprint arXiv:2404.07503*, 2024, [retrieved: May, 2025].

[10] S. Kim *et al.*, "Evaluating language models as synthetic data generators," *arXiv preprint arXiv:2412.03679*, 2024, [retrieved: May, 2025].

[11] S. Es, J. James, L. E. Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.

[12] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, 2023, [retrieved: May, 2025].

[13] I. Mirzadeh *et al.*, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv:2410.05229*, 2024, [retrieved: May, 2025].

[14] Y. Yu *et al.*, "Rankrag: Unifying context ranking with retrieval-augmented generation in llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121 156–121 184, 2024.

[15] G. d. S. P. Moreira *et al.*, "Enhancing q&a text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for rag," *arXiv preprint arXiv:2409.07691*, 2024, [retrieved: May, 2025].

[16] P. Mandikal and R. Mooney, "Sparse meets dense: A hybrid approach to enhance scientific document retrieval," *arXiv preprint arXiv:2401.04055*, 2024, [retrieved: May, 2025].

[17] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, "Machine learning for synthetic data generation: a review," *arXiv preprint arXiv:2302.04062*, 2023, [retrieved: May, 2025].

[18] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[19] Hugging Face, "Model card for multi-qa-minilm-l6-cos-v1," [Online]. Available: https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1, 2021, [retrieved: May, 2025].

[20] ——, "Model card for msmacro-minilm-l6-en-de-v1," [Online]. Available: https://huggingface.co/cross-encoder/msmarco-MiniLM-L6-en-de-v1, 2021, [retrieved: May, 2025].

[21] ——, "Model card for ms-macro-minilm-l-6-v2," [Online]. Available: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2, 2021, [retrieved: May, 2025].

[22] D. Park and S. Kim, "Probabilistic precision and recall towards reliable evaluation of generative models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 20 099–20 109.

[23] S. S. Ravi, B. Mielczarek, A. Kannappan, D. Kiela, and R. Qian, "Lynx: An open source hallucination evaluation model," *arXiv preprint arXiv:2407.08488*, 2024, [retrieved: May, 2025].

[24] S. Kim *et al.*, "Prometheus 2: An open source language model specialized in evaluating other language models," *arXiv preprint arXiv:2405.01535*, 2024, [retrieved: May, 2025].

[25] B. Adler *et al.*, "Nemotron-4 340b technical report," *arXiv preprint arXiv:2406.11704*, 2024, [retrieved: May, 2025].

[26] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels, 2022," *URL https://arxiv. org/abs/2212.10496*, 2022.

Assessing the Capabilities of Large Language Models in Translating American Sign Language Gloss to English

Jalal Al-Afandi[*], Péter Pócsi[†], Gábor Borbély[†],
Helga M. Szabó[†], Ádám Rák[*][†], Zsolt Robotka[*][†], András Horváth[*]

alafandi.mohammad.jalal@hallgato.ppke.hu, { rak.adam, horvath.andras}@itk.ppke.hu
[*]Peter Pazmany Catholic University, Faculty of Information Technology and Bionics

{peter.pocsi, gabor, helga, zsolt}@deepsign.ai
[†] DeepSign Technologies Ltd.

***Abstract*—In this paper, we investigate the ability of large language models (LLMs) to translate American Sign Language with GLOSS annoation into English without fine-tuning or architectural modifications. Our findings show that pretrained transformers achieve translation quality comparable to human experts. While prompt engineering enhances accuracy for simpler models, it has minimal impact on more advanced ones. Additionally, when generating multiple translation variants, the first response is typically the most accurate, with subsequent outputs declining in quality. These results underscore the strong zero-shot translation capabilities of LLMs and highlight their potential for scalable ASL-GLOSS translation applications.**

***Keywords*-ASL-GLOSS translation; Generative pretrained transformers, large language models**

I. Introduction

Large Language Models (LLMs) have emerged as a transformative force in natural language processing, demonstrating remarkable versatility across various applications, including text generation, summarization, and machine translation. These models, often referred to as foundation models, are trained on vast corpora of text and possess extensive knowledge of human languages. Their ability to generalize across a wide range of tasks has enabled them to achieve impressive performance, even in low-resource language translation tasks. Recent studies have shown that LLMs excel in one-shot and few-shot learning scenarios [1], where only a limited number of examples are available. This makes them particularly suitable for translating languages with scarce training data.

Among the communities that could greatly benefit from these advancements are deaf and hard-of-hearing individuals. Sign languages serve as the primary mode of communication for these communities; however, the automatic translation of sign languages into spoken or written languages remains a significant challenge [2]. Developing effective translation solutions could substantially enhance accessibility and inclusivity, supporting social integration and improving communication opportunities for these individuals.

Automatic translation of sign languages typically follows a two-step pipeline [3][4], although end-to-end approaches have also been explored [5]. The first step involves recognizing and detecting visual symbols associated with sign language gestures. From these visual inputs, a structured intermediate representation can be derived, such as ASL-GLOSS. ASL-GLOSS serves as a symbolic transcription of American Sign Language (ASL) gestures, capturing the essential lexical components of signs while abstracting away certain nonmanual markers, including facial expressions, eye movements, and contextual cues. Although ASL-GLOSS simplifies the representation of sign language, it remains an incomplete encoding of meaning, as it lacks many elements necessary for full semantic understanding, although this limitation also applies to written text. Table I presents examples of English and ASL-GLOSS sentences.

TABLE I. Example sentences in English and ASL-GLOSS

English sentence	**ASL-GLOSS**
There are a lot of studies on speech disorders	STUDY ON SPEECH/ORAL fs-DISORDER A-LOT
While I was a graduate student, in a linguistics class, a professor gave a lecture about syntax.	DURING/WHILE IX-1p GRAD STUDENT IN CLASS_2 LONG-AGO LINGUISTICS TEACH+AGENT TEACH+AGENT DIRECT/EXPLAIN fs-SYNTAX
My mother taught my two brothers and me, so it was easier for us to move around.	part:indef MOTHER TEACH IX-1p+ AND TWO BROTHER EASY MOVE part:indef

Existing machine translation solutions for sign-to-English or ASL-GLOSS-to-English tasks typically rely on smaller, domain-specific models trained exclusively on sign and gloss-specific datasets [6]. However, these models often struggle with generalization due to their limited exposure to the target output language. We hypothesize that the broad linguistic knowledge embedded in LLMs can mitigate this issue by providing improved translations and using their comprehensive understanding of syntax, semantics, and common expressions in the target language.

In this work, we explore the capabilities of current LLMs in translating ASL-GLOSS into English. Our objective is to assess the direct translation quality of LLMs on ASL-GLOSS inputs and to establish a baseline accuracy for LLM-based gloss translation. Additionally, by analyzing potential translation errors and corrections, we aim to provide insights into the viability of LLMs as robust components in future sign language translation pipelines.

The remainder of this paper is organized as follows. Section II. provides an overview of the theoretical background and related work relevant to our study. In Section III., we introduce the proposed methodology, including dataset descriptions and evaluation metrics. The results and their analysis are discussed in Section IV., highlighting both quantitative and qualitative findings. Finally, Section V. concludes the paper by summarizing the main contributions and key findings along with the potential avenues for future research.

II. Exisiting solutions

Machine Translation (MT) of ASL encompasses various approaches, each leveraging different technologies to facilitate translation between ASL and spoken or written languages. Key methodologies include:

1) Rule-Based Systems: Early MT systems for ASL utilized rule-based approaches, where linguistic experts encoded grammatical and syntactic rules to map English text to ASL structures [7]. An example is the TEAM prototype, which analyzed English text's syntactic and morphological aspects before accessing a sign synthesizer to produce corresponding ASL signs via a computer-generated human avatar [8].
2) Statistical Machine Translation (SMT): SMT approaches rely on statistical models derived from bilingual corpora to predict translation probabilities. However, the scarcity of large-scale parallel ASL-English corpora has limited the effectiveness of SMT in ASL translation [9].
3) Neural Machine Translation (NMT): Recent advancements in NMT have shown promise in translating spoken languages. Applying NMT to ASL involves training deep learning models on annotated sign language datasets to capture the nuances of ASL grammar and expressions. Challenges include the need for extensive datasets and the complexity of modeling sign language's spatial and temporal aspects [6].
4) Vision-Based Recognition Systems: These systems employ computer vision techniques to interpret sign language from video input [10]. For instance, the Kinect Sign Language Translator utilizes Microsoft's Kinect sensor to capture signers' movements and translate them into spoken language using machine learning and pattern recognition [11].
5) Sensor-Based Recognition Systems: Some approaches use wearable sensors to detect hand movements and positions. For example, SignAloud incorporates gloves equipped with sensors that transliterate ASL into English by tracking hand movements and sending data to a computer system for analysis and translation [12].
6) Hybrid Systems: Combining multiple methodologies, hybrid systems aim to enhance translation accuracy. SignAll integrates computer vision and natural language processing to recognize hand shapes and movements, converting this data into simple English phrases to facilitate real-time ASL translation [13].

Despite these advancements, challenges persist, particularly in accurately interpreting the diverse and complex structures of ASL. Ongoing research aims to address these issues by developing more robust models and incorporating larger, more diverse datasets to improve the reliability and inclusivity of ASL machine translation systems.

III. Methodology

To thoroughly evaluate ASL-GLOSS to English translation, it is essential to carefully consider the data sources and models used in this study. The methodology section outlines our approach to selecting appropriate datasets, choosing relevant language models, and establishing a rigorous evaluation framework. These choices form the foundation for robust and reproducible experimental results.

A. Datasets

To evaluate the performance of LLMs in ASLGLOSS-to-English translation, we conducted an extensive review of available datasets. Our primary objective was to select a dataset that meets several critical criteria. The ideal dataset would be large-scale, contain video recordings of the signing person, provide gloss annotations of the signed sentences, and include high-quality English translations. Video recordings are particularly important as they serve as the most accurate reference for human translations, capturing the full range of visual cues necessary for understanding sign language, including hand movements, facial expressions, and other nonmanual markers. Additionally, we prioritized datasets that feature complex sentence structures and a broad spectrum of topics, ensuring comprehensive coverage of real-world communication scenarios.

However, only a limited number of datasets meet these demanding requirements. The datasets we investigated include:

- English-ASL Gloss Parallel Corpus 2012 (ASLG-PC12): A dataset mapping ASL gloss to formal English text[14]
- American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI): Contains video recordings with corresponding gloss annotations [15].
- MS-ASL Dataset: A large-scale dataset for isolated sign recognition[16].
- DAI - ASLLVD: A video dataset with ASL lexical items[17].
- ASL Finger Spelling Dataset: Focused on finger-spelling gestures[18].
- WLASL: A large-scale dataset for word-level American Sign Language recognition.[19]
- American Sign Language Lexicon Video Dataset: A comprehensive dataset with video recordings, gloss annotations, and English translations[20].

Among these datasets, the American Sign Language Lexicon Video Dataset proved to be the most suitable for our experiments, as it met all the aforementioned selection criteria. Its combination of video input, detailed gloss annotations, and high-quality English translations makes it an ideal resource for training and evaluating ASL-GLOSS-to-English translation models. Consequently, our experimental work primarily focuses on this dataset.

B. Large Language models

In our investigation, we selected a diverse range of language models to evaluate their performance on the ASL-GLOSS-to-English translation task. Given the rapid advancements in the field, with new models emerging regularly, compiling an exhaustive list is not feasible. However, our selection was guided by several key considerations to ensure a representative and comprehensive assessment.

The selected models fall into two broad categories:

- Large-Scale Proprietary Models: This category includes cutting-edge models such as Claude and ChatGPT, which are accessible exclusively through API-based interfaces. These models are considered among the most complex and sophisticated LLMs available, and we anticipated that their extensive training data and advanced architectures would yield the highest translation accuracy. Despite their closed-source nature, their performance serves as an upper-bound benchmark for comparison.
- Open-Source Models: We also included open-source models, such as LLaMA and DeepSeek. Although these models are typically less complex than their proprietary counterparts, their publicly available architectures and weights offer several advantages. Running these models on-premise enables greater control over execution environments, facilitates further optimization, and allows fine-tuning on domain-specific data. This flexibility is particularly valuable for tailoring models to the nuances of ASLGLOSS translation.

By evaluating models from both categories, we aim to balance performance, transparency, and practical deployability in our study. This comprehensive selection will provide insights into the trade-offs between accuracy and customizability, helping to identify the most suitable models for real-world sign language translation applications.

For the sake of reproducibility, all our experiments, including the code and detailed parameter setups, are available at the following GitHub link to ensure reproducibility: https://github.com/horan85/ASLGloss

IV. Results

Before presenting the experimental results, we summarize the comparative evaluation of various state-of-the-art language models in the ASL-GLOSS to English translation task. This analysis focuses on assessing how model architecture and prompting strategies influence translation quality. Our goal is to understand not only the absolute performance of these models but also how additional linguistic context affects their translation capabilities.

A. Model Comparisons

To systematically assess the performance of our translation models, we curated an evaluation dataset comprising 2,040 ASL-GLOSS-English sentence pairs sourced from the American Sign Language Lexicon Video Dataset. This dataset serves as a benchmark for measuring translation quality and the generalization capabilities of our models.

We conducted experiments with various models under two distinct prompting strategies. In the first setup, models received only a direct translation prompt, instructing them to generate an English sentence from a given ASL-GLOSS input. In the second setup, we supplemented the prompt with a brief explanation of the GLOSS structure (3,000 words in length) and a carefully selected set of twenty example translations to provide additional context and guidance.

As evaluation metrics, we selected the BLEU score and cosine similarity between the embedded representations of the translated and ground-truth sentences. For sentence embeddings, we utilized the CLIP-ViT-B/32 transformer model [21].

Our results for the models without additional descriptions are presented in Table II, which reports the mean performance along with the corresponding variances. To provide a more comprehensive view of the distribution, Figures 1 and 2 illustrate the detailed distributions of BLEU scores and cosine similarities, respectively. These visualizations offer deeper insights into the variability and consistency of model outputs across different evaluation metrics.

Our findings indicate that for more advanced and complex models, such as ChatGPT, Claude, and DeepSeek, the inclusion of structural information and example translations had minimal impact on overall translation quality. This suggests that these models inherently possess a strong ability to interpret ASL-GLOSS sequences and generate fluent English translations, even without explicit guidance on the source language structure.

In contrast, the LLaMA and ChatGPT-mini models exhibited moderate improvements, with increases of approximately 0.03 and 0.04 in cosine similarity and BLEU scores, respectively. However, further investigation is needed to determine whether this robustness extends to less frequent linguistic structures or more complex GLOSS annotations.

B. Model Consistency

Since LLMs generate probabilistic outputs, translation quality can vary due to multiple factors. Additionally, ASL-GLOSS sentences, when extracted without broader context, may have multiple valid interpretations. To assess whether generating multiple translation variants improves accuracy, we examined the effect of allowing GPT-based models to produce several alternative translations for each input.

While output variability can be adjusted by tuning the model's temperature parameter, we did not optimize this aspect. Instead, we instructed the models to generate five

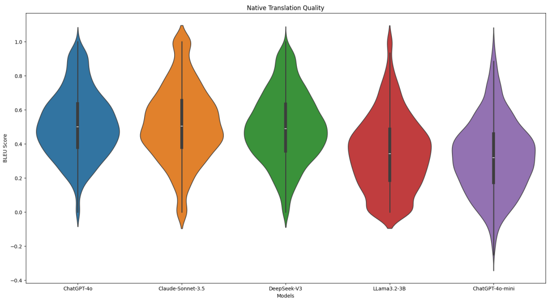

Figure 1. This figure depicts the Blue scores on the American Sign Language Lexicon Video Dataset using various LLM models as a GLOSS to Enlish translation task.

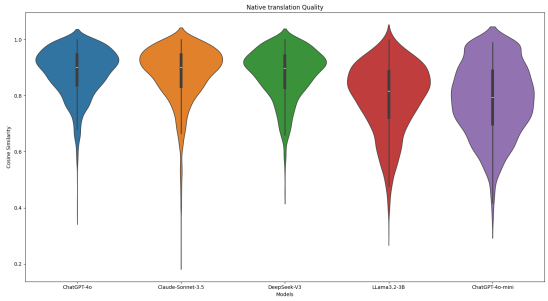

Figure 2. This figure depicts the Cosine similarity values on the American Sign Language Lexicon Video Dataset using various LLM models as a GLOSS to Enlish translation task.

TABLE II. COSINE SIMILARITIES AND BLEU SCORES WITH AND WITHOUT GLOSS DESCRIPTIONS

Model	**Cosine Similarity**	**BLEU Score**
ChatGPT-4o	0.881 ± 0.83	0.514 ± 0.192
ChatGPT-4o with GLOSS description	0.880 ± 0.87	0.512 ± 0.201
Claude Sonnet	0.879 ± 0.94	0.518 ± 0.219
Claude Sonnet with GLOSS description	0.880 ± 0.99	0.518 ± 0.244
DeepSeek V3	0.876 ± 0.83	0.496 ± 0.217
DeepSeek V3 with GLOSS description	0.876 ± 0.88	0.495 ± 0.227
Llama 3.2	0.793 ± 1.23	0.349 ± 0.203
Llama 3.2 with GLOSS description	0.824 ± 1.43	0.374 ± 0.486
ChatGPT-4o-mini	0.787 ± 1.26	0.324 ± 0.213
ChatGPT-4o-mini with GLOSS description	0.814 ± 1.67	0.365 ± 0.455

translation variants per input to evaluate whether this approach enhances translation quality.

Using the same dataset of 2,040 sentences, we selected the two best-performing models (ChatGPT and Sonnet) and tasked them with generating five distinct translations for each ASL-GLOSS input without providing detailed GLOSS descriptions. The BLEU scores for these translations are shown in Figure 3. Similar trends were observed in cosine similarity measurements, though these results are omitted due to space constraints.

Notably, our findings indicate that the first generated translation was consistently the most accurate. As the ranking progressed, translation quality gradually declined, though the differences were minor. This suggests that while generating multiple outputs introduces slight variations, the first translation is generally the most reliable.

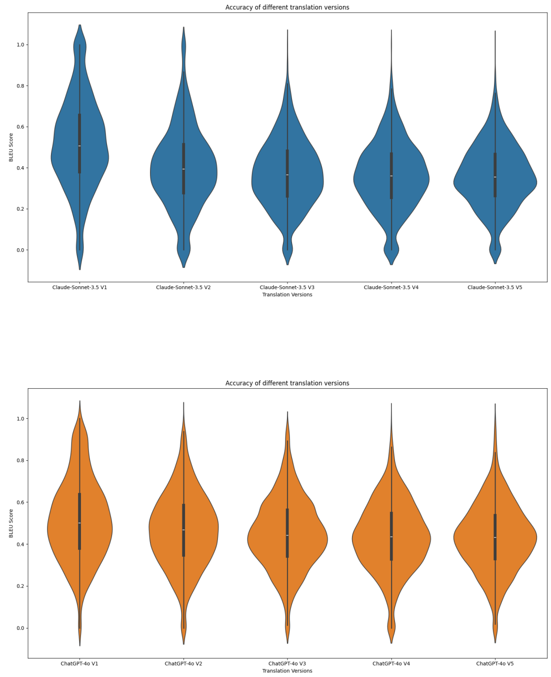

Figure 3. Translation quality (in terms of BLEU scores) when we asked the model to provide multiple variants for Claude-Sonnet (above) and Chat-GPT-4o (below)

-

TABLE III. COSINE SIMILARITIES AND BLEU SCORES ON THE REDUCED DATASET

Model	**Cosine Similarity**	**BLEU Score**
Translator	0.903 ± 0.361	0.582 ± 0.253
ChatGPT-4o	0.893 ± 0.49	0.548 ± 0.163
Claude Sonnet	0.901 ± 0.47	0.560 ± 0.182
DeepSeek V3	0.884 ± 0.78	0.523 ± 0.316
Llama 3.2	0.810 ± 1.23	0.377 ± 0.250

ISBN: 978-1-68558-287-6

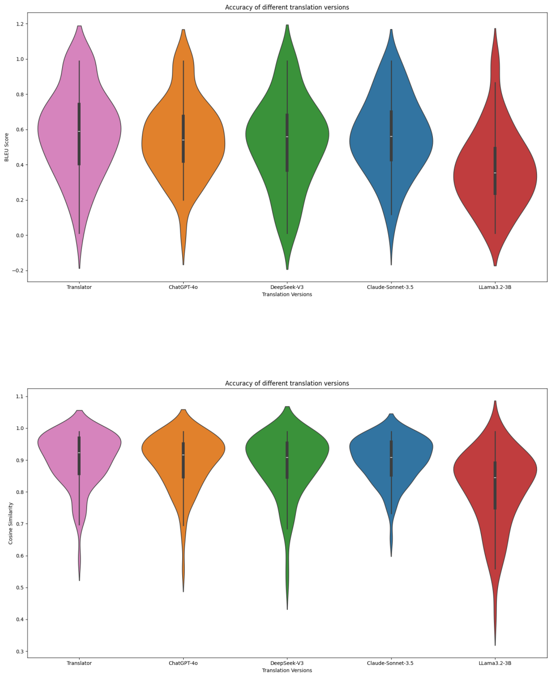

Figure 4. Cosine Similarities (above) and BLEU scores (below) on the 88-sentence dataset comparing a human translator's performance with various LLMs

V. Conclusion and Furute Work

In this paper, we demonstrated the capability of large language models (LLMs) to translate ASL-GLOSS to English without fine-tuning or architectural modifications. Our findings suggest that general-purpose pretrained transformers are viable for this task, achieving translation quality comparable to that of human experts.

A. Key Findings

- Zero-shot translation effectiveness: General-pretrained transformers can effectively translate ASL-GLOSS without additional fine-tuning, highlighting the strong zero-shot capabilities of modern LLMs in handling structured linguistic inputs like GLOSS.
- Limited impact of prompt engineering: While prompt engineering improves translation accuracy for simpler models, it has a negligible effect on more advanced LLMs. This suggests that state-of-the-art models already possess a robust understanding of GLOSS structures without explicit prompting strategies.
- Quality decline in multiple outputs: When LLMs were prompted to generate multiple translation variants, the first response was typically the most accurate, with subsequent translations exhibiting a gradual decline in quality. This suggests that probabilistic generation may introduce increasing errors when multiple outputs are requested.
- Near-human translation accuracy: LLMs achieve translation accuracy close to that of human experts. This underscores their potential to assist or even replace human translators in certain ASL-GLOSS translation tasks, improving scalability and accessibility.

While our results are promising, further research is needed to assess the robustness of LLM-based ASL-GLOSS translation across diverse linguistic structures and complex annotations. Future work could explore fine-tuning approaches, domain adaptation techniques, and real-world deployment scenarios to enhance translation reliability and applicability.

Acknowledgement

The support of the Alfréd Rényi Institute of Mathematics and the following grants: 2018-1.2.1-NKP00008: Exploring the Mathematical Foundations of Artificial Intelligence and TKP2021_02-NVA-27 – Thematic Excellence Program and also 2023-1.1.2-GYORSÍTÓSÁV-2024-00019 of the Hungarian National Research, Development and Innovation Office are gratefully acknowledged.

References

[1] X. Zhang, N. Rajabi, K. Duh, and P. Koehn, "Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora", in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 468–481.

[2] U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, and A. Abid, "Advances in machine translation for sign language: Approaches, limitations, and challenges", *Neural Computing and Applications*, vol. 33, no. 21, pp. 14 357–14 399, 2021.

[3] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7784–7793.

[4] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 023–10 033.

[6] P. Fayyazsanavi, A. Anastasopoulos, and J. Košecká, "Gloss2text: Sign language gloss translation using llms and semantically aware label smoothing", *arXiv preprint arXiv:2407.01394*, 2024.

[7] J. Porta, F. López-Colino, J. Tejedor, and J. Colás, "A rule-based translation from written spanish to spanish sign language glosses", *Computer Speech & Language*, vol. 28, no. 3, pp. 788–811, 2014.

[8] B. David and P. Bouillon, "Prototype of automatic translation to the sign language of french-speaking belgium evaluation by the deaf community", *Modelling, Measurement and Control C*, vol. 79, no. 4, pp. 162–167, 2018.

[9] A. Othman and M. Jemni, "Statistical sign language machine translation: From english written text to american sign language gloss", *arXiv preprint arXiv:1112.0168*, 2011.

[10] Y. Madhuri, G. Anitha, and M. Anburajan, "Vision-based sign language translation device", in *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, IEEE, 2013, pp. 565–568.

[11] M. Ahmed, M. Idrees, Z. ul Abideen, R. Mumtaz, and S. Khalique, "Deaf talk using 3d animated sign language: A sign language interpreter using microsoft's kinect v2", in *2016 SAI Computing Conference (SAI)*, IEEE, 2016, pp. 330–335.

[12] S. B. Rizwan, M. S. Z. Khan, and M. Imran, “American sign language translation via smart wearable glove technology”, in *2019 International Symposium on Recent Advances in Electrical Engineering (RAEE)*, IEEE, vol. 4, 2019, pp. 1–6.

[13] L. Leeson and H. Haaris, “Signall: A european partnership approach to deaf studies via new technologies”, in *INTED2009 Proceedings*, IATED, 2009, pp. 1270–1279.

[14] A. Othman and M. Jemni, “English-asl gloss parallel corpus 2012: Aslg-pc12”, in *Sign-lang@ LREC 2012*, European Language Resources Association (ELRA), 2012, pp. 151–154.

[15] C. Neidle and S. Sclaroff, “American sign language linguistic research project”, Report, Tech. Rep., 1998.

[16] H. R. V. Joze and O. Koller, “Ms-asl: A large-scale data set and benchmark for understanding american sign language”, *arXiv preprint arXiv:1812.01053*, 2018.

[17] C. Neidle, A. Thangali, and S. Sclaroff, “Challenges in development of the american sign language lexicon video dataset (asllvd) corpus”, in *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, Language Resources and Evaluation Conference (LREC) 2012*, 2012.

[18] N. Pugeault and R. Bowden, “Spelling it out: Real-time asl fingerspelling recognition”, in *2011 IEEE International conference on computer vision workshops (ICCV workshops)*, Ieee, 2011, pp. 1114–1119.

[19] D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison”, in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459–1469.

[20] V. Athitsos *et al.*, “The american sign language lexicon video dataset”, in *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, IEEE, 2008, pp. 1–8.

[21] A. Radford *et al.*, “Learning transferable visual models from natural language supervision”, in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

Estimating the Risk of Failing Physics Courses through the Monte Carlo Simulation

Isaac Caicedo-Castro[*][†][‡] Rubby Castro-Púche[†][§] Samir Castaño-Rivera[*][‡]
[*]Socrates Research Team
[†]Research Team: Development, Education, and Healthcare
[‡]Department of Systems and Telecommunications Engineering
[§]Department of Social Science
University of Córdoba
Carrera 6 No. 76-103, 230002, Montería, Colombia
e-mail: {isacaic|rubycastro|sacastano}@correo.unicordoba.edu.co

***Abstract*—This research is conducted in the context of the Systems Engineering undergraduate program at the University of Córdoba in Colombia, aiming to calculate the risk of failing physics courses, which are considered particularly challenging for students. At this university, the academic semester is divided into three sessions, each equally weighted in the final grade. Our goal is to estimate the failure risk based on student performance in the earlier sessions. To this end, we collected a dataset comprising the session grades and final results of students enrolled in Physics I, II, and III during 2024. We then implemented a Monte Carlo simulation to calculate the absolute and relative risk of course failure. The results show that failing early sessions is strongly associated with a higher probability of failing the course, especially in Physics I and III. These insights can support lecturers in adjusting the syllabus and designing interventions to reduce dropout rates and improve student outcomes.**

***Keywords*-Monte Carlo simulation; educational innovation; computational social science.**

I. Introduction

Nowadays, academic success is a primary concern for universities worldwide. As a consequence, identifying strategies and conducting research to predict the risk of academic failure has become an active area of study within social computing, particularly through educational data mining approaches. These methods are commonly used to predict student dropout, delayed graduation [1]–[3], and the likelihood of course failure or withdrawal [4]–[13].

Our research focuses on the academic context of the University of Córdoba in Colombia, where each academic semester is split into three sessions, each lasting six weeks and contributing equally to the final grade. The final course grade is calculated as the mean of the student's grades across the three sessions. Within each session, no single assessment can exceed 40% of the session grade, meaning that each student undergoes at least nine evaluations during a semester.

This structure is designed to achieve several pedagogical goals: reducing the pressure of final exams, diversifying assessment strategies, encouraging consistent study habits, enabling continuous monitoring of learning progress, facilitating early interventions, and providing timely support to students. This approach is supported by several educational theories and instructional strategies, including constructivism, formative assessment, multiple intelligences theory, cognitive load theory, active learning, and outcome-based education.

According to constructivist theory, students build their understanding through interaction with their environment. Frequent evaluations help instructors monitor this evolving understanding and adjust teaching strategies accordingly.

Formative assessment emphasizes the use of ongoing evaluations throughout the instructional period to monitor learning, identify challenges, and guide teaching. This method offers continuous feedback to both students and instructors, aligning well with the university's evaluation strategy.

The theory of multiple intelligences posits that students possess diverse talents and learning preferences. A variety of assessments throughout the semester provides a more inclusive way to evaluate these varied strengths.

Cognitive load theory suggests that students learn more effectively when information is presented in manageable segments. Multiple evaluations distributed over time align with this principle by reducing cognitive overload.

Active learning promotes student engagement through problem-solving, discussions, and hands-on activities. Multiple evaluations throughout the semester can reinforce this approach by encouraging students to actively engage with the material.

At the University of Córdoba, Outcome-Based Education (OBE) is the foundational approach. It emphasizes clearly defined learning outcomes and assessments aligned with those outcomes. Dividing the semester into multiple sessions allows for a more granular alignment of evaluations with specific goals.

The university's OBE model is integrated with the Structure of the Observed Learning Outcomes (SOLO) taxonomy, which categorizes learning into five levels:

1) Prestructural (0.0–2.0): The student has not yet grasped the key concepts.
2) Unistructural (2.1–2.9): The student understands a single aspect of the task.
3) Multistructural (3.0–3.7): The student understands several aspects, but without integration.
4) Relational (3.8–4.5): The student can integrate multiple aspects meaningfully.
5) Extended Abstract (4.6–5.0): The student demonstrates deep understanding and applies concepts to new contexts.

To pass an evaluation, a student must achieve at least the multistructural level, corresponding to a grade above 3.0.

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-287-6

By structuring learning outcomes around SOLO taxonomy principles, the curriculum offers a coherent and progressively challenging learning experience. This structure emphasizes the development of deeper understanding as students progress. However, despite this pedagogical framework, physics courses remain particularly challenging for systems engineering students, who often struggle to reach the relational or extended abstract levels. For example, in 2024, the average final grades for Physics I, II, and III were 3.17, 3.30, and 3.35, respectively, suggesting limited integration of concepts or application to real-world contexts.

In an endeavor to mitigate failure and dropout, the university assumes students at risk of failing a course if they fail either of the first two sessions. This leads us to pose the following research questions:

- What is the risk of failing a physics course if a student fails the first session?
- What is the risk if a student fails the second session but passed the first?
- What is the risk if a student fails both the first and second sessions?

To the best of our knowledge, no prior research has directly addressed these questions. To fill this gap, we simulate all possible grade scenarios using the Monte Carlo numerical method, informed by historical academic performance data. This method has been used in similar educational contexts, for instance, to evaluate curriculum effectiveness [14] or to estimate students' motivation in learning scientific computing [15].

Our simulations reveal the following findings:

- For Physics I, approximately 42 out of 100 students are at risk of failing if they failed the first session; 44 out of 100 if they failed the second session; and 63 out of 100 if they failed both.
- For Physics II, about 28 out of 100 students are at risk if they failed the first session; 14 out of 100 if they failed the second; and 49 out of 100 if both were failed.
- For Physics III, around 49 out of 100 students are at risk if they failed the first session; 11 out of 100 if they failed the second; and 14 out of 100 if they failed both.

These insights contribute to implement early intervention strategies and improve academic support in physics courses.

Finally, the rest of this article is outlined as follows: in Section II, we present the research and simulation methodology adopted in this research, while we present and discuss the results in Section II. The article concludes in Section IV.

II. Research Methodology

We adopted a quantitative approach, collecting the session and final grades of 100 students enrolled in physics courses at the University of Córdoba in 2024. Specifically, 36 students were enrolled in Physics I, 32 in Physics II, and 32 in Physics III. The relatively small dataset size reflects the recent implementation of the previously described Outcome-Based Education (OBE) framework at the institution.

Given the limited number of students and the sparsity of failure cases in certain session combinations (as shown in Table I), direct estimation of absolute and relative risks from empirical data would be statistically unreliable. For example, the dataset contains no instances of students failing the Physics I course after failing the second or third session, provided they passed the first. This type of data sparsity presents a challenge for risk estimation.

TABLE I. Number of students who failed a physics when they have failed at least one session (S1, S2, and S3).

Course	Failed S1	Failed S2	Failed S3	Failed Students
Physics 1	Yes	Yes	Yes	1
	Yes	Yes	No	7
	Yes	No	No	2
	No	Yes	Yes	0
	No	Yes	No	0
Physics 2	Yes	Yes	Yes	2
	Yes	Yes	No	3
	Yes	No	No	0
	No	Yes	Yes	0
	No	Yes	No	0
Physics 3	Yes	Yes	Yes	0
	Yes	Yes	No	1
	Yes	No	No	0
	No	Yes	Yes	0
	No	Yes	No	0

To address this, we employed the Monte Carlo simulation method [16] to explore the full probability space of possible student performance outcomes. Instead of relying solely on the small number of observed cases, we reconstructed the grade distribution using a parametric model, specifically, a normal distribution, with parameters (mean and standard deviation) derived from the original dataset. Grades were clipped to fall within the [0, 5] scale, as the normal distribution might otherwise generate implausible values in the tails. This allowed us to simulate large numbers of plausible student grade combinations and estimate the associated risks under uncertainty.

In essence, the Monte Carlo simulation serves as a data-informed method for approximating risk in underrepresented or unobserved configurations, enabling generalization beyond the empirical observations while remaining grounded in the observed statistical characteristics of the data.

Thus, the probability that a student fails a physics course given that they failed the jth session is denoted as $P(y < 3 \mid x_j < 3)$, where y is the final course grade. A final grade below 3.0 indicates course failure, as previously explained. The variable x_j represents the grade the student obtained in the jth session, with $j = 1, 2, 3$. Thus, $x \in \mathcal{X} \subseteq [0, 5]^3$ is a real-valued three-dimensional vector containing the grades from each session, all within the range [0, 5]. A session grade below 3.0 ($x_j < 3$) indicates failure in that session.

Since the final grade y is the arithmetic mean of the three session grades, it is computed as:

$$y = \frac{1}{3} \sum_{j=1}^{3} x_j \qquad (1)$$

The *Absolute Risk (AR)* of failing the course given failure

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-287-6

in session jth is defined as:

$$AR(y < 3 \mid x_j < 3) = \int_{\mathcal{X}} \frac{P(y < 3, x_j < 3)}{P(x_j < 3)}\, dx \tag{2}$$

Similarly, the absolute risk of failing the course given that the student did *not* fail session jth is:

$$AR(y < 3 \mid x_j \geq 3) = \int_{\mathcal{X}} \frac{P(y < 3, x_j \geq 3)}{P(x_j \geq 3)}\, dx \tag{3}$$

The *Relative Risk (RR)* is defined as the ratio of these two quantities:

$$RR(y < 3 \mid x_j < 3) = \frac{AR(y < 3 \mid x_j < 3)}{AR(y < 3 \mid x_j \geq 3)} \tag{4}$$

To estimate these quantities via the Monte Carlo method, we generate an $N \times 3$-dimensional matrix $X \in [0, 5]^{N\times3}$, where its component $X_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$ is normally distributed with mean μ_j and standard deviation σ_j computed from the historical grades of students in session j of each physics course.

The absolute risk $AR(y < 3 \mid x_j < 3)$ is approximated as:

$$AR(y < 3 \mid x_j < 3) \approx \frac{\sum_{i=1}^{N} \mathbf{1}(y_i < 3 \wedge X_{ij} < 3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij} < 3)} \tag{5}$$

where $\mathbf{1}(u) = 1$ if the condition u is true, and 0 otherwise. Similarly, the absolute risk for students who did *not* fail session j is:

$$AR(y < 3 \mid x_j \geq 3) \approx \frac{\sum_{i=1}^{N} \mathbf{1}(y_i < 3 \wedge X_{ij} \geq 3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij} \geq 3)} \tag{6}$$

Finally, the relative risk is calculated as:

$$RR(y < 3 \mid x_j < 3) \approx \frac{\frac{\sum_{i=1}^{N} \mathbf{1}(y_i<3 \wedge X_{ij}<3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij}<3)}}{\frac{\sum_{i=1}^{N} \mathbf{1}(y_i<3 \wedge X_{ij}\geq3)}{\sum_{i=1}^{N} \mathbf{1}(X_{ij}\geq3)}} \tag{7}$$

This simulation-based approach enables us to estimate the conditional risks associated with failing individual sessions and provides a probabilistic understanding of academic outcomes based on partial performance.

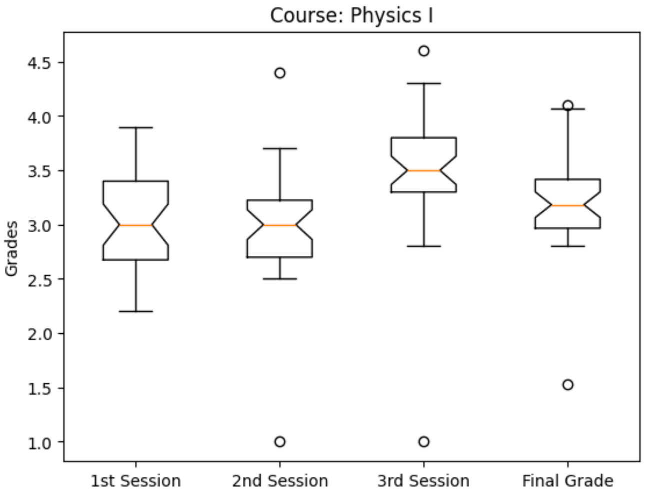

Figure 1. Grades of the students enrolled in the physics I course in 2024

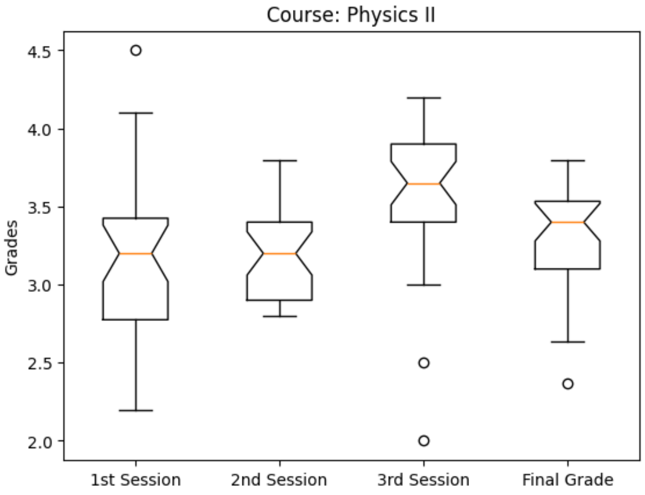

Figure 2. Grades of the students enrolled in the physics II course in 2024

The simulation was implemented in Python using the NumPy and Matplotlib libraries. The anonymized dataset and corresponding source code are available upon request from the first author.

III. The Research Results and Discussion

Based on the collected dataset, the mean grades for the first, second, and third sessions in the Physics I course were 3.03, 2.98, and 3.50, respectively, with corresponding standard deviations of 0.43, 0.53, and 0.58. As shown in Figure 1, the box plot corresponding to the final grade illustrates that students rarely failed the course outright or achieved exceptionally high grades. Furthermore, there appears to be a general trend of improved performance in the final session.

Similarly, the mean grades for Physics II were 3.17, 3.20, and 3.55 for the first, second, and third sessions, respectively, with standard deviations of 0.57, 0.27, and 0.47. Figure 2 demonstrates a performance pattern comparable to Physics I, where both failures and outstanding performances were infrequent. However, a notable difference emerges in the second session: while performance in Physics I declined slightly from 3.03 to 2.98, Physics II showed a slight improvement, with the average grade increasing from 3.17 to 3.20. This suggests a possible difference in instructional design or assessment difficulty between the two courses during that session.

Additionally, the dataset reveals that the mean grades for the first, second, and third sessions in the Physics III course are 3.12, 3.28, and 3.66, respectively, with standard deviations of 0.41, 0.36, and 0.31. Figure 3 illustrates that student grades in this course follow a pattern similar to the previous two physics courses.

TABLE II. EXPECTED FINAL GRADES BY COURSE OBTAINED FROM THE MONTE CARLO SIMULATION RESULTS.

Course	Expected Grade	Standard Error	95% CI
Physics 1	3.178	1.2×10^{-4}	[3.178, 3.179]
Physics 2	3.305	10^{-4}	[3.30498, 3.305]
Physics 3	3.354	1.1×10^{-4}	[3.353, 3.354]

The results of the numerical simulation show that the expected final grades for Physics I, II, and III are 3.17, 3.31, and 3.35, respectively (see Table II). Figures 4–6 demonstrate how the Monte Carlo simulations converge to these values, which are consistent with the histograms presented in Figures 7–9, displaying the distribution of the final grades for each course.

TABLE III. ABSOLUTE AND RELATIVE RISK BY SESSION AND COURSE.

Course	Session(s) Failed	Absolute Risk (%)	Relative Risk (%)	95% CI (RR)
Physics I	S1	41.66	2.77	[2.762, 2.777]
	S2	43.52	4.05	[4.032, 4.059]
	S1 and S2	62.93	4.18	[4.172, 4.195]
Physics II	S1	28.11	12.62	[12.532, 12.703]
	S2	22.76	2.49	[2.484, 2.505]
	S1 and S2	49.03	22.00	[21.851, 22.159]
Physics III	S1	10.61	17.93	[17.601, 18.267]
	S2	14.30	8.11	[8.021, 8.196]
	S1 and S2	33.05	55.86	[54.827, 56.913]

A summary of the relative and absolute risk estimates derived from the Monte Carlo simulation is presented in Table III. The corresponding relative risks for each course are depicted using forest plots in Figures 10–12. As expected, failing the first two sessions corresponds to the highest relative risk of failing a physics course. It is noteworthy that for Physics I, failing the second session alone is associated with a higher relative risk than failing the first session. This pattern differs from Physics II and III, where failing the first session presents a greater relative risk. Notably, the risk of failing Physics I after failing only the second session is nearly equivalent to the risk of failing after both the first and second sessions.

The absolute risk of course failure among students exposed to session failures versus those unexposed is compared in Table IV. The results of the simulation reveal that there is an absolute risk of 41.66% that students fail the Physics I course if they fail the first session. This corresponds to a risk difference of 26.62 percentage points, with a 95% confidence interval of [26.554%, 26.688%], compared to an absolute risk of 15.04% for students who do not fail the first session. This difference is statistically significant, indicating a meaningful association between failing the first session and ultimately failing Physics I. Furthermore, the relative risk is 2.77, suggesting that students who fail the first session are 2.77 times more likely to fail the course than those who do not (see Figure 10).

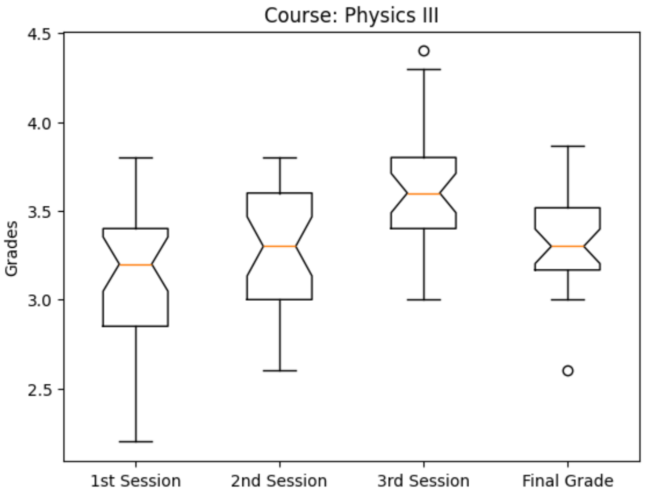

Figure 3. Grades of the students enrolled in the physics III course in 2024

TABLE IV. COMPARISON OF ABSOLUTE RISK (AR) OF COURSE FAILURE BETWEEN STUDENTS EXPOSED AND UNEXPOSED TO FAILING PREVIOUS SESSIONS, WITH CORRESPONDING RISK DIFFERENCES (RD)

Course	Session(s) Failed	AR (%) exposed	AR (%) unexposed	RD (%)	95% CI (RD)
Physics I	S1	41.66	15.66	26.62†	[26.554, 26.688]
	S2	43.52	10.76	32.76†	[32.696, 32.823]
	S1 and S2	62.93	15.04	47.89†	[47.805, 47.975]
Physics II	S1	28.11	2.23	25.89†	[25.829, 25.944]
	S2	22.76	9.12	13.63†	[13.563, 13.707]
	S1 and S2	49.03	2.23	46.80†	[46.673, 46.934]
Physics III	S1	10.61	0.59	10.02†	[9.961, 10.071]
	S2	14.30	1.76	12.54†	[12.455, 12.623]
	S1 and S2	33.05	0.59	32.45†	[32.276, 32.634]

† (p-value < 0.05)

The absolute risk of failing the Physics I course increases to 43.52% if students fail the second session. In this case, the risk difference is 32.76 percentage points, with a 95% confidence interval of [32.696%, 32.823%], compared to an absolute risk of 10.76% among students who pass the second session. The relative risk in this scenario is 4.05, indicating that students who fail the second session are over four times more likely to fail the course than those who succeed (see Figure 10).

When students fail both the first and second sessions of Physics I, the absolute risk of failing the course increases to 62.93%. The associated risk difference is 47.89 percentage points, with a 95% confidence interval of [47.805%, 47.975%], compared to the 15.04% absolute risk observed among those who do not fail the first two sessions. The relative risk of 4.18 further highlights the increased likelihood of course failure under these conditions (see Figure 10).

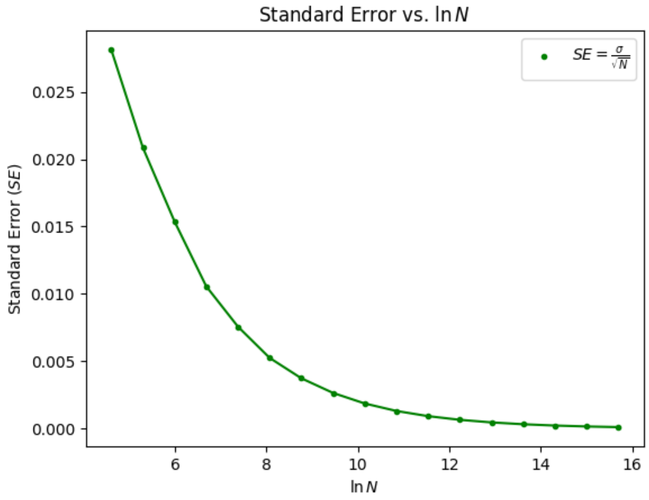

Figure 4. The simulation converges to the expected final grade of 3.178 in the Physics I course as $N = 6,553,600$, with a standard error of 1.2×10^{-4}. The result lies within the 95% confidence interval of [3.178, 3.179].

Regarding the Physics II course, the simulation shows that students who fail the first session have an absolute risk of 28.11% of failing the course. This results in a risk difference of 25.89 percentage points, with a 95% confidence interval of [25.829%, 25.944%], compared to an absolute risk of just 2.23% for those who do not fail the first session. The relative risk of 12.62 indicates that students who fail the first session are over 12 times more likely to fail Physics II (see Figure 11).

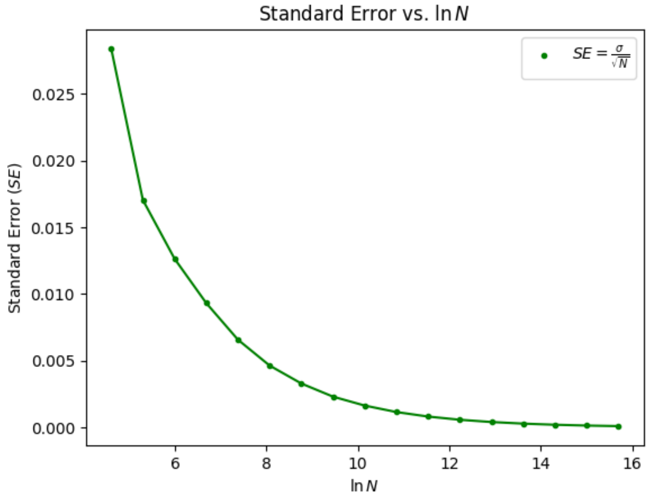

Figure 5. The simulation convergences to the expected final grade of 3.305 in the Physics II course as $N = 6,553,600$, with a standard error of 10^{-4}. The result lies within the 95% confidence interval of [3.30498, 3.305].

Failing the second session in Physics II results in an absolute risk of 22.76%, with a risk difference of 13.63 percentage points and a 95% confidence interval of [13.563%, 13.707%], compared to an absolute risk of 9.12% for those who do not fail the second session. The relative risk of 2.49 indicates a significantly increased likelihood of failing the course for these students (see Figure 11).

When students fail both the first and second sessions in Physics II, the absolute risk of failing the course rises sharply to 49.03%. This is associated with a risk difference of 46.80 percentage points and a 95% confidence interval of [46.673%, 46.934%], compared to the same 2.23% absolute risk for students who succeed in both sessions. The relative risk of 22 underscores the very strong association between poor performance in the initial sessions and course failure (see Figure 11).

In the case of Physics III, students who fail the first session have an absolute risk of 10.61% of failing the course. The risk difference in this case is 10.02 percentage points, with a 95% confidence interval of [9.961%, 10.071%], compared to an absolute risk of 0.59% among students who pass the first session. The relative risk is 17.93, indicating a very strong link between failing the first session and failing the course (see Figure 12).

For students who fail the second session in Physics III, the absolute risk of course failure is 14.30%, compared to 1.76% among those who pass that session. This results in a risk difference of 12.54 percentage points, with a 95% confidence interval of [12.455%, 12.623%].The corresponding relative risk of 8.11 suggests that failing the second session in Physics III is associated with a higher likelihood of course failure than the same condition in Physics I and II (see Figure 12).

Finally, for students who fail both the first and second sessions in Physics III, the absolute risk of failing the course increases to 33.05%. The risk difference is 32.45 percentage points, with a 95% confidence interval of [32.276%, 32.634%], in contrast to the absolute risk of 0.59% for students who succeed in both sessions. The relative risk of 55.86 implies an exceptionally high likelihood of failure under these circumstances (see Figure 12).

IV. Conclusion and Perspective

We adopted Monte Carlo simulation because the collected dataset is small and statistically unstable or undefined (i.e., division by zero or nearly zero) to estimate absolute and relative risk causing even high variance. Thereby the Monte Carlo method provides a data-informed but smoothed approximation of what outcomes would look like using a larger dataset with similar distributional properties of the collected dataset.

We draw the following conclusions from the results:

- Teaching staff and lecturers may consider reorganizing the syllabus to reduce the risk of course failure by incorporating the observed probabilities of failure at each session.
- In Physics II and III, failing the first session is associated with a higher risk of overall course failure than failing the second session. This pattern might be driven by psychological or motivational factors; students who begin the course with poor performance often experience discouragement, reduced engagement, and diminished resilience in response

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-287-6

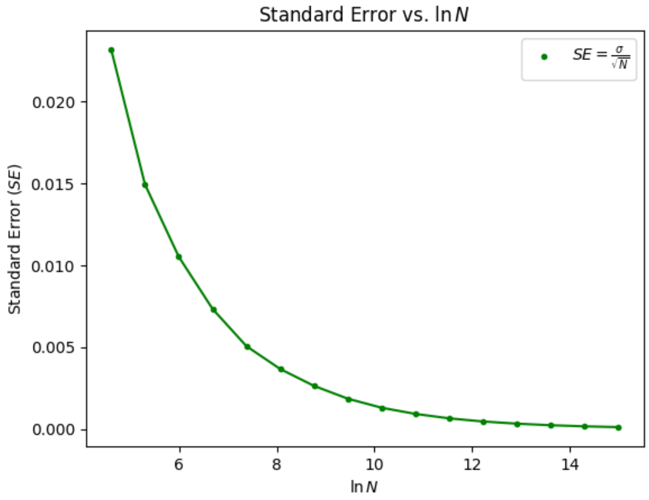

Figure 6. The simulation convergences to the expected final grade of 3.354 in the Physics III course as $N = 3,276,800$, with a standard error of 1.1×10^{-4}. The result lies within the 95% confidence interval of [3.353, 3.354].

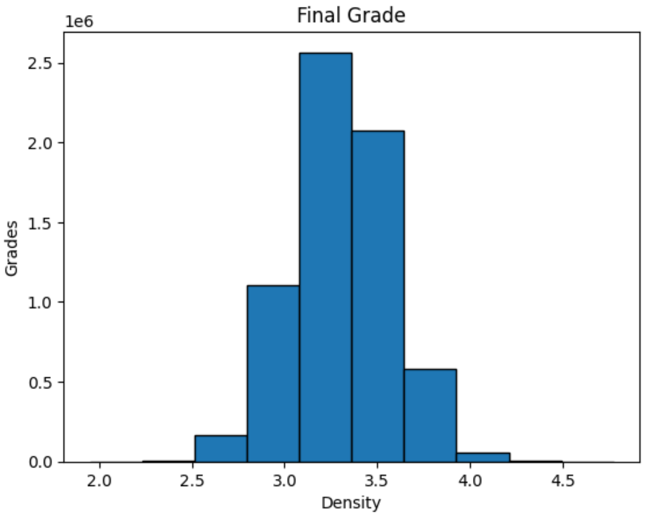

Figure 8. Distribution of final grades obtained from the simulation for the Physics II course.

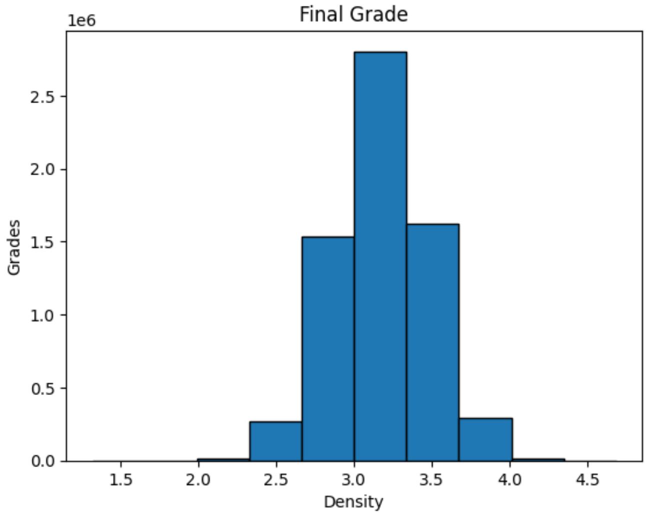

Figure 7. Distribution of final grades obtained from the simulation for the Physics I course.

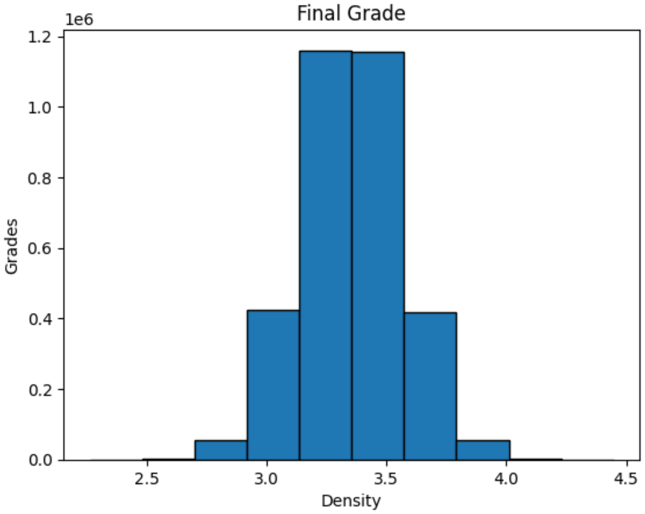

Figure 9. Distribution of final grades obtained from the simulation for the Physics III course.

to subsequent academic challenges [17], [18]. More broadly, performance in the first or second session is strongly associated with final course outcomes. Beyond mere statistical correlation, early academic struggles may serve as indicators of underlying motivational or behavioral challenges, making them valuable triggers for early intervention and academic support strategies. Further research is needed to design and implement targeted measures that might help students recover from early setbacks and improve their overall performance trajectory.

- In Physics III, students who pass the first two sessions have an almost negligible risk of failing the course. Consequently, they may become complacent and neglect the final session. In contrast, students who fail the first two sessions face a significantly high risk of failing the course. This discrepancy suggests an imbalance in the difficulty and weight of the course sessions. Simulating alternative scenarios may help to redesign the course structure and improve student success rates.
- A consistent trend of improved student performance is observed from Physics I to Physics III, as indicated by higher average grades and lower absolute risk in the later courses. This pattern might reflect students' adaptation to course demands or the development of stronger academic skills over time. Nevertheless, in the Systems Engineering program, students are not strictly required to follow prerequisite sequencing. For instance, a student may enroll in Physics III without having previously taken Physics I or II. Although most students typically follow the intended

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-287-6

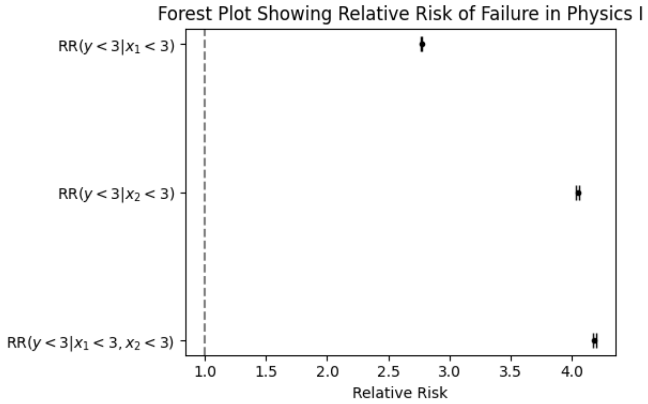

Figure 10. Forest plot showing the relative risk (RR) of failing the Physics I course. $RR(y < 3 \mid x_1 < 3) = 2.77$, with a 95% confidence interval of [2.762, 2.777]; $RR(y < 3 \mid x_2 < 3) = 4.05$, with a 95% confidence interval of [4.032, 4.059]; and $RR(y < 3 \mid x_3 < 3) = 4.18$, with a 95% confidence interval of [4.172, 4.195]. In all cases, the Wald test p-value is less than 0.05.

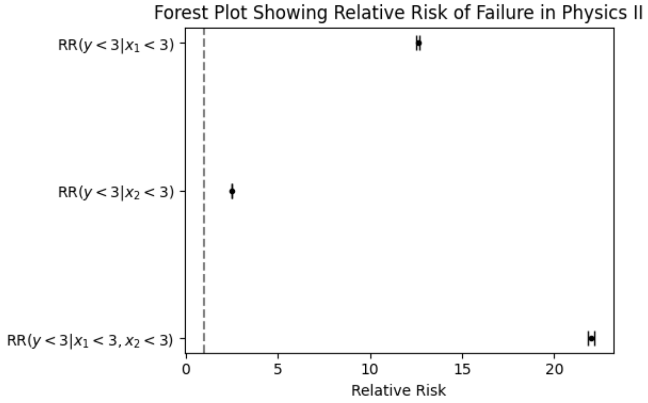

Figure 11. Forest plot showing the relative risk (RR) of failing the Physics II course. $RR(y < 3 \mid x_1 < 3) = 12.62$, with a 95% confidence interval of [12.532, 12.703]; $RR(y < 3 \mid x_2 < 3) = 2.49$, with a 95% confidence interval of [2.484, 2.505]; and $RR(y < 3 \mid x_3 < 3) = 22$, with a 95% confidence interval of [21.851, 22.159]. In all cases, the Wald test p-value is less than 0.05.

curricular progression, exceptions do occur. In this study, information about such cases was not available.

- The simulation based on the Monte Carlo numerical method has proven to be a valuable tool for estimating the absolute and relative risks of course failure. It might support evidence-based decision-making in academic planning and policy design. Grades were simulated using a normal distribution, with parameters estimated from observed student data. While our dataset includes relatively few course failures, we modeled grades probabilistically to reflect the empirical distribution, ensuring that rare but plausible outcomes (e.g., failing scenarios) were represented.

As directions for further research, we propose the following:

- We shall collect additional data to apply this methodology to other courses and broaden the scope of academic risk analysis.

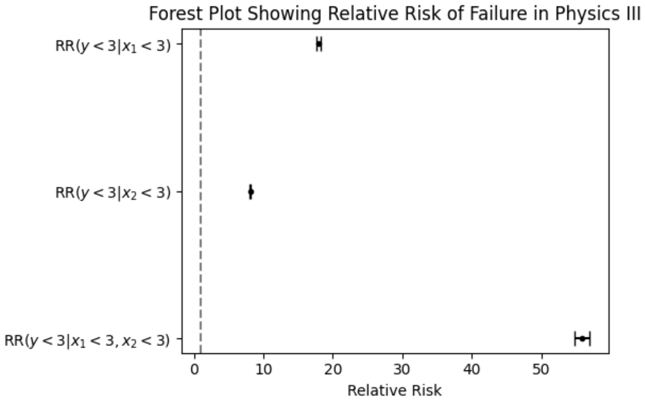

Figure 12. Forest plot showing the relative risk (RR) of failing the Physics III course. $RR(y < 3 \mid x_1 < 3) = 17.93$, with a 95% confidence interval of [17.601, 18.267]; $RR(y < 3 \mid x_2 < 3) = 8.11$, with a 95% confidence interval of [8.021, 8.196]; and $RR(y < 3 \mid x_3 < 3) = 55.86$, with a 95% confidence interval of [54.827, 56.913]. In all cases, the Wald test p-value is less than 0.05.

- We shall extend the simulation to incorporate the specific coursework or evaluation structure assigned in each session, aiming to estimate risk with greater accuracy.
- We shall adapt the simulation to assume an non-uniform weighting of sessions when calculating final grades, in order to reduce the risk of failure.
- We shall incorporate bootstrap resampling to estimate the variability of simulation parameters (i.e., mean and standard deviation) in order to strengthening the robustness of the risk estimates under data scarcity.

ACKNOWLEDGMENT

Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. The authors thank Universidad de Córdoba in Colombia for supporting this study. Finally, the authors thank the anonymous referees for their comments that contributed to improve the quality of this article.

REFERENCES

[1] 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, February 2022.

[2] 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.

[3] 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, 01 2023.

[4] 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.

[5] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Applied Sciences*, vol. 11, p. 3130, 04 2021.

[6] 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. 100066, 2022.

[7] 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. IATED, 2022, pp. 5831–5837.

[8] 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. IATED, 2022, pp. 5872–5878.

[9] 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. IARIA: International Academy, Research, and Industry Association, 2023, pp. 40–51.

[10] I. Caicedo-Castro, M. Macea-Anaya, and S. Castaño-Rivera, "Early Risk Detection of Bachelor's Student Withdrawal or Long-Term Retention," in *The 2023 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*. International Academy, Research, and Industry Association, 2023, pp. 177–187.

[11] 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.

[12] ——, "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.

[13] ——, "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.

[14] D. Torres, J. Crichigno, and C. Sanchez, "Assessing Curriculum Efficiency Through Monte Carlo Simulation," *Journal of College Student Retention*, vol. 22, no. 4, pp. 597–610, 2021.

[15] 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*. International Academy, Research, and Industry Association, 2025, pp. 15–22.

[16] N. Metropolis and S. Ulam, "The Monte Carlo Method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949.

[17] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student Attrition*, 2nd ed. Chicago, IL: University of Chicago Press, 1994.

[18] M. Yorke, "Retention, Persistence and Success in On-Campus Higher Education, and their Enhancement in Open and Distance Learning," *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 19, no. 1, pp. 19–32, 2004.

ISBN: 978-1-68558-287-6