# PREFACE TO THE EDITION

The **International Journal of Information Technology Research Studies (IJITRS)** is pleased to present its latest issue, bringing together a diverse collection of scholarly contributions that reflect the rapidly evolving landscape of information technology and its transformative influence across industries, education, governance, and society. As digital innovation continues to redefine organizational processes and human interaction, this issue showcases research that advances secure, intelligent, and responsible technological development.

The issue opens with a study on digital intelligence among youth, examining the multidimensional competencies required to navigate today's interconnected digital ecosystem. By exploring aspects such as digital literacy, communication, safety, security, emotional intelligence, and digital rights, the study highlights the growing importance of cultivating technologically proficient as well as ethically responsible digital citizens. Addressing one of the most critical concerns in intelligent transportation, the issue presents an advanced deep-learning-based intrusion detection framework for connected and autonomous vehicles. The proposed hybrid architecture demonstrates the potential of artificial intelligence to strengthen cybersecurity within in-vehicle communication networks while contributing to safer and more resilient automotive systems.

The expanding role of online education is explored through a comprehensive survey of personalized course recommendation systems for Massive Open Online Courses (MOOCs). The review synthesizes current recommendation methodologies, benchmark datasets, evaluation metrics, and emerging research directions, offering valuable insights into improving learner engagement, personalization, and educational effectiveness in digital learning environments. As quantum computing rapidly progresses, organizations face unprecedented cybersecurity challenges. This issue includes a timely contribution on post-quantum cryptography migration, presenting a practical roadmap for enterprises transitioning toward quantum-resistant security infrastructures. The study offers strategic guidance for adopting cryptographic agility while addressing the technical and operational challenges associated with future-ready security.

Artificial intelligence continues to reshape enterprise knowledge management, and this volume features a comprehensive examination of Retrieval-Augmented Generation (RAG) for trustworthy enterprise large language model assistants. By integrating retrieval mechanisms with generative AI, the work demonstrates approaches for improving factual accuracy, explainability, and secure enterprise deployment of intelligent assistants.

The issue concludes with an insightful discussion on self-service business intelligence and analytics in digital transformation, highlighting modern data architectures, governance frameworks, augmented analytics, and AI-assisted decision-making. The proposed maturity model provides organizations with practical guidance for building governed, scalable, and data-driven analytical ecosystems. Collectively, the articles featured in this issue illustrate the dynamic convergence of artificial intelligence, cybersecurity, enterprise computing, digital education, analytics, and emerging technologies. They address both foundational research questions and practical implementation challenges, offering valuable perspectives for researchers, academicians, industry professionals, policymakers, and technology practitioners. The breadth of topics demonstrates the increasing need for interdisciplinary collaboration in designing secure, intelligent, ethical, and sustainable digital systems.

We extend our sincere appreciation to all authors for their valuable scholarly contributions and to the reviewers whose expertise and dedication have ensured the quality and integrity of the published work. We also express our gratitude to the Editorial Board members and the publishing team for their continued commitment to maintaining the journal's academic excellence.

We hope this issue of the International Journal of Information Technology Research Studies (IJITRS) serves as an important resource for advancing research, encouraging innovation, and inspiring future developments that contribute to the continued evolution of information technology and its positive impact on society.

Dr. Mini T. V.

Chief Editor

CONTENTS


| SL. NO. | TITLE | AUTHOR | PAGE NO. |
|---|---|---|---|
| 1 | A Multi-Dimensional Analysis of Digital Intelligence: Understanding the Core Digital Competencies Among Youth in the Digital Era | Sreelatha P & C Karthik Deepa | 1-9 |
| 2 | Deep-Learning Intrusion Detection for Connected and Autonomous Vehicles | Ginne M James | 10-17 |
| 3 | Personalized Course Recommendation on Mooc Platforms: A Survey | Rejina P V | 18-25 |
| 4 | Post-Quantum Cryptography Migration for Enterprise Security | Mini T V | 26-32 |
| 5 | Retrieval-Augmented Generation for Trustworthy Enterprise Llm Assistants | Bini P B | 33-39 |
| 6 | Self-Service Business Intelligence and Analytics in Digital Transformation | Meena Jose Komban | 40-47 |

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209

# A Multi-Dimensional Analysis of Digital Intelligence: Understanding the Core Digital Competencies Among Youth in the Digital Era

Sreelatha P[1*], C Karthik Deepa[2]

[1] Research scholar, Department of Education, Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, India
[2] Assistant Professor, Department of Education, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

**Article information**

Received: 1st June 2026
Received in revised form: 24th June 2026
Accepted: 29th June 2026
Available online: 30th July 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0029

**Abstract**

Digital intelligence has become a vital ability for young people negotiating the complexities of the digital age in an increasingly connected society. This study investigates the multidimensional nature of digital intelligence by assessing eight fundamental components among college students: Digital Identity, Digital Use, Digital Safety, Digital Security, Digital Emotional Intelligence, Digital Communication, Digital Literacy, and Digital Rights. Using a sample of 100 students selected through stratified random sampling, the study aims to understand the level of digital competency and the relationships among these dimensions. Data were collected using a standardized Digital Intelligence Scale and analyzed through descriptive statistics and Pearson correlation techniques. Results revealed that Digital Literacy (Mean = 4.10) and Digital Use (Mean = 3.98) recorded the highest mean scores, indicating strong technological proficiency among students. In contrast, Digital Emotional Intelligence and Digital Rights scored comparatively lower, highlighting areas requiring further development. Significant positive correlations were found among several dimensions, demonstrating the interconnected nature of digital competencies.The findings suggest that while youth are functionally adept at using technology, competencies related to emotional awareness, ethical responsibility, and understanding of digital rights remain less developed. The study emphasizes the importance of holistic digital education strategies that foster balanced digital citizens who are not only technologically proficient but also responsible, empathetic, and aware of their rights and responsibilities in digital environments. This analysis serves as a foundational step for educators, policymakers, and digital platforms seeking to cultivate a digitally intelligent generation capable of thriving in a dynamic digital society.

**Keywords**:- Digital Intelligence, Youth Competency, Digital Literacy, Digital Rights, Digital Communication.

.

## I. INTRODUCTION

Digital intelligence is defined by eight key components: digital identity, digital use, digital safety, digital security, digital emotional intelligence, digital communication, digital literacy, and digital rights. These components provide a framework to evaluate an individual's readiness and resilience in the digital world. Recent studies indicate that while youth demonstrate proficiency in functional aspects such as device usage and internet

navigation, they often face challenges in critical areas including digital safety, misinformation recognition, online rights awareness, privacy management, and emotional regulation in digital interactions [1]. These findings highlight the need for a more comprehensive approach to digital competency development beyond technical skills.

Furthermore, the National Education Policy (NEP) 2020 emphasizes the integration of digital literacy and digital ethics into higher education curricula, underlining its importance for holistic student development. However, most institutions still focus on basic ICT training, leaving out nuanced aspects such as emotional, legal, and safety-related competencies. The lack of empirical data on students' digital intelligence, especially in the Indian higher education context, presents a critical gap.

By providing a multidimensional analysis of digital intelligence among college students, this study seeks to address that gap. The study examines students' levels of digital competency using a standardized framework encompassing all eight dimensions and identifies areas requiring targeted educational interventions. The findings are expected to support curriculum designers, educators, and policymakers in fostering digitally competent, ethically responsible, and emotionally intelligent citizens capable of navigating the complexities of contemporary digital environments [2].

## II. REVIEW OF LITERATURE

Digital intelligence, as a multidimensional construct, encompasses the competencies required to navigate, manage, and engage responsibly in digital environments. With the growing ubiquity of digital technology in education, communication, and daily life, scholars have increasingly emphasized the need to move beyond basic digital literacy toward a holistic understanding of digital intelligence.

Recent research has expanded the concept of digital competence beyond technical proficiency to include cognitive, social, ethical, and emotional dimensions necessary for effective participation in digital environments [3]. Building on this perspective, the DQ framework identifies eight core competencies of digital intelligence: digital identity, digital use, digital safety, digital security, digital emotional intelligence, digital communication, digital literacy, and digital rights [4].

Studies have shown that although young people are highly engaged with digital technologies, their competencies vary considerably across different dimensions of digital intelligence. Darling found that higher education students often demonstrate strong digital skills but require further development in digital citizenship, ethical behavior, and responsible online engagement [5].

Similarly, Satapaniganone, Tachaphahapong, and Methakunavudhi (2024) reported that university students tend to perform well in digital use and literacy while showing lower competencies in areas such as digital safety, emotional intelligence, and awareness of digital rights. These findings support the need for comprehensive digital intelligence education in higher education settings.

Kunkhong, Yongsorn, and Ponathong (2024) validated a multidimensional model of digital intelligence among higher education students and confirmed that digital intelligence consists of several interrelated competencies that collectively influence students' readiness for participation in digital societies.

Furthermore, Prinsloo, Khalil, and Slade (2024) emphasized the importance of digital wellbeing, emotional regulation, and ethical technology use in contemporary higher education environments, particularly as artificial intelligence and digital platforms become increasingly integrated into learning processes. Their findings suggest that digital intelligence should encompass not only technical capabilities but also responsible and emotionally aware engagement with digital technologies.

In the Indian context, rapid digitalization, online learning initiatives, and the implementation of the National Education Policy (NEP) 2020 have increased the importance of digital intelligence among college students. However, empirical research examining all eight dimensions of digital intelligence and their interrelationships among Indian higher education students remains limited.

Collectively, recent literature supports a multidimensional approach to evaluating digital intelligence among youth. However, there remains a gap in empirical research specifically assessing the interrelationships among all eight DQ components within diverse academic populations, especially in the Indian context, underscoring the relevance and necessity of the present study.

## III. OBJECTIVES

- To assess the level of digital intelligence across its eight core components among college students.
- To analyze the relationship between different components of digital intelligence to identify areas of strength and gaps.

## IV. OPERATIONAL DEFINITION

### A. Digital Intelligence:

In this study, digital intelligence refers to the measurable set of cognitive, technical, emotional, and ethical competencies that enable students to effectively and responsibly use digital technologies. Digital identity, use, safety, security, emotional intelligence, communication, literacy, and rights make up eight components here [6].

### B. Youth:

Youth in this study refers to students aged between 18 to 24 years, currently enrolled in undergraduate arts and science programs, who represent the emerging generation of digital users [7].

### C. Digital Comparison:

In this study, digital competencies refer to the essential set of knowledge, skills, attitudes, and values required to navigate, evaluate, create, and communicate effectively in digital environments. These competencies encompass information and media literacy, critical thinking, digital content creation, problem-solving, collaboration, and ethical use of technology [8].

### D. Digital Era:

In this study, the digital era refers to the contemporary period marked by the widespread integration of digital technologies into everyday life, education, and communication. It is characterized by rapid technological advancements, increased connectivity, the prevalence of smart devices, and the growing dependence on digital platforms for personal, academic, and professional activities [9].

## V. METHODOLOGY

### A. Research Design

The present study adopted a descriptive and correlational research design to explore and analyze the level and interrelationship of digital intelligence components among college students. The descriptive design helped in identifying the distribution and mean levels of each digital competency, while the correlational design allowed the researcher to assess how these competencies are interrelated within the digital behavior patterns of youth [10].

### B. Population and Sample

The study's target demographic consisted of arts and science college enrolled undergraduate students. A sample of 100 students was selected to represent the population adequately. The sample consisted of individuals aged between 18 and 24 years, with an equal representation of gender (50 males and 50 females) and course stream (50 from Arts and 50 from Science backgrounds). This balanced demographic distribution was ensured to capture diverse perspectives on digital intelligence among youth.

### C. Sampling Technique

The study used a stratified random sampling method. This approach was selected to keep proportional representation over the strata of gender and academic stream. Stratification ensured that subgroups were equally considered during participant selection, thereby enhancing the reliability and validity of the findings by minimizing sampling bias.

### D. Variables of the Study

The study concentrated on evaluating as variables eight elements of digital intelligence. Among these factors defining the main variables of interest were digital identity, digital use, digital safety, digital security, digital emotional intelligence, digital communication, digital literacy, digital rights. As this is a correlational study, no independent or manipulated variables were involved; rather, each digital intelligence domain was evaluated to examine its level and interrelation with other components [11].

### E. Tools Used for Data Collection

Data were collected using the Digital Intelligence (DQ) Assessment Scale, developed based on the Digital Intelligence (DQ) Framework proposed by the DQ Institute (2019). The instrument comprised structured items representing each of the eight components of digital intelligence, namely Digital Identity, Digital Use, Digital Safety, Digital Security, Digital Emotional Intelligence, Digital Communication, Digital Literacy, and Digital Rights. The scale was designed on a 5-point Likert format ranging from ‘Strongly Disagree’ to ‘Strongly Agree’, with each component represented by five to seven items.

The internal consistency reliability of the overall scale was established using Cronbach's Alpha, yielding a coefficient of α = 0.87, indicating good reliability. In addition to the overall reliability, Cronbach's Alpha was computed for each component to enhance transparency and internal consistency reporting: Digital Identity (α = 0.82), Digital Use (α = 0.79), Digital Safety (α = 0.84), Digital Security (α = 0.85), Digital Emotional Intelligence (α = 0.81), Digital Communication (α = 0.80), Digital Literacy (α = 0.83), and Digital Rights (α = 0.78). These values indicate acceptable to good internal consistency across all dimensions.Content validity of the instrument was established through expert review by specialists in digital education and social science research, along with pilot testing to ensure clarity, relevance, and appropriateness for the target population [12].

### F. Procedure for Data Collection

Prior to data collection, formal permission was obtained from the concerned academic institutions, and ethical clearance was ensured. Participants were informed about the purpose of the study, and their voluntary consent was sought. The survey was administered either in printed format or through a secure digital platform, depending on the convenience of the participants. Responses were collected within a stipulated period and compiled systematically for further analysis. All identifying information was kept confidential to maintain the anonymity of participants.

### G. Statistical Techniques Used

Descriptive and inferential statistical approaches were used to examine the data. The distribution of scores among every digital intelligence component was compiled using descriptive statistics like mean, standard deviation, frequency, and percentage. Using Pearson's correlation coefficient, the researcher was able to find noteworthy correlations between components and hence investigate their interactions. Should it be necessary, other methods such Principal Component Analysis (PCA) could be taken under consideration for the identification of dominating data clusters.

### H. Ethical Considerations

The study followed accepted moral research standards. Before they entered the study, every participant signed informed permission. Participants' identities were anonymised during data analysis and reporting; they were guaranteed of the confidentiality of their responses. The research process was conducted with integrity, ensuring transparency, voluntary participation, and respect for the rights of all individuals involved.

### I. Scope and Limitations

This study is limited in scope to analyzing the levels and relationships among digital intelligence components among college students aged 18 to 24 years. It does not extend to other variables such as mental health, socioeconomic status, or institutional digital policies. While the findings offer valuable insights into digital competencies among youth, the limited sample size (N = 100) and regional focus may affect the generalizability of the results. Nonetheless, the study provides a foundational understanding of digital intelligence as a multi-dimensional construct within the student population [13].

## VI. RESULTS AND INTERPRETATION

The results section presents the analysis of data collected from 100 students on eight components of digital intelligence. It includes descriptive statistics, correlation patterns among variables, and demographic insights. Visualizations such as bar graphs and heatmaps illustrate the levels and interrelationships of digital competencies across the sample group. (Naseer et al., 2025)

Table 1. Demographic Profile of Respondents

| Variable | Category | Frequency (n) | Percentage (%) |
|---|---|---|---|
| Age | 18–20 | 47 | 47.00% |
| | 21–22 | 33 | 33.00% |
| | 23–24 | 20 | 20.00% |
| Gender | Male | 50 | 50.00% |
| | Female | 50 | 50.00% |
| Course Stream | Arts | 50 | 50.00% |
| | Science | 50 | 50.00% |
| Year of Study | First Year | 30 | 30.00% |
| | Second Year | 37 | 37.00% |
| | Third Year | 33 | 33.00% |
| Internet Usage | <2 hours | 8 | 8.00% |
| | 2–4 hours | 33 | 33.00% |
| | 4–6 hours | 42 | 42.00% |
| | >6 hours | 17 | 17.00% |

The demographic profile was based on responses from 100 undergraduate students aged between 18 and 24 years. The age distribution revealed that 47% of participants were in the 18–20 age group, followed by 33% aged 21–22, and 20% aged 23–24, indicating that nearly half the respondents were in the early phase of their academic journey. Gender representation was perfectly balanced, with 50 males and 50 females, ensuring equitable insights into digital intelligence irrespective of gender.

Participants were equally split across academic streams, with 50 students each from Arts and Science, providing an opportunity to compare digital behaviors across different disciplines. With regard to the year of study, 37% of respondents were in their second year, 33% in third year, and 30% in first year, offering a well-distributed academic exposure and developmental diversity.

In terms of daily internet usage, 42% of students reported spending 4–6 hours online, while 33% used the internet for 2–4 hours. A smaller proportion, 17%, used it for more than 6 hours daily, and only 8% reported less than 2 hours of usage. These usage patterns confirm that the majority of the sample is actively engaged in the digital environment, making them well-suited for evaluating aspects of digital intelligence [22].

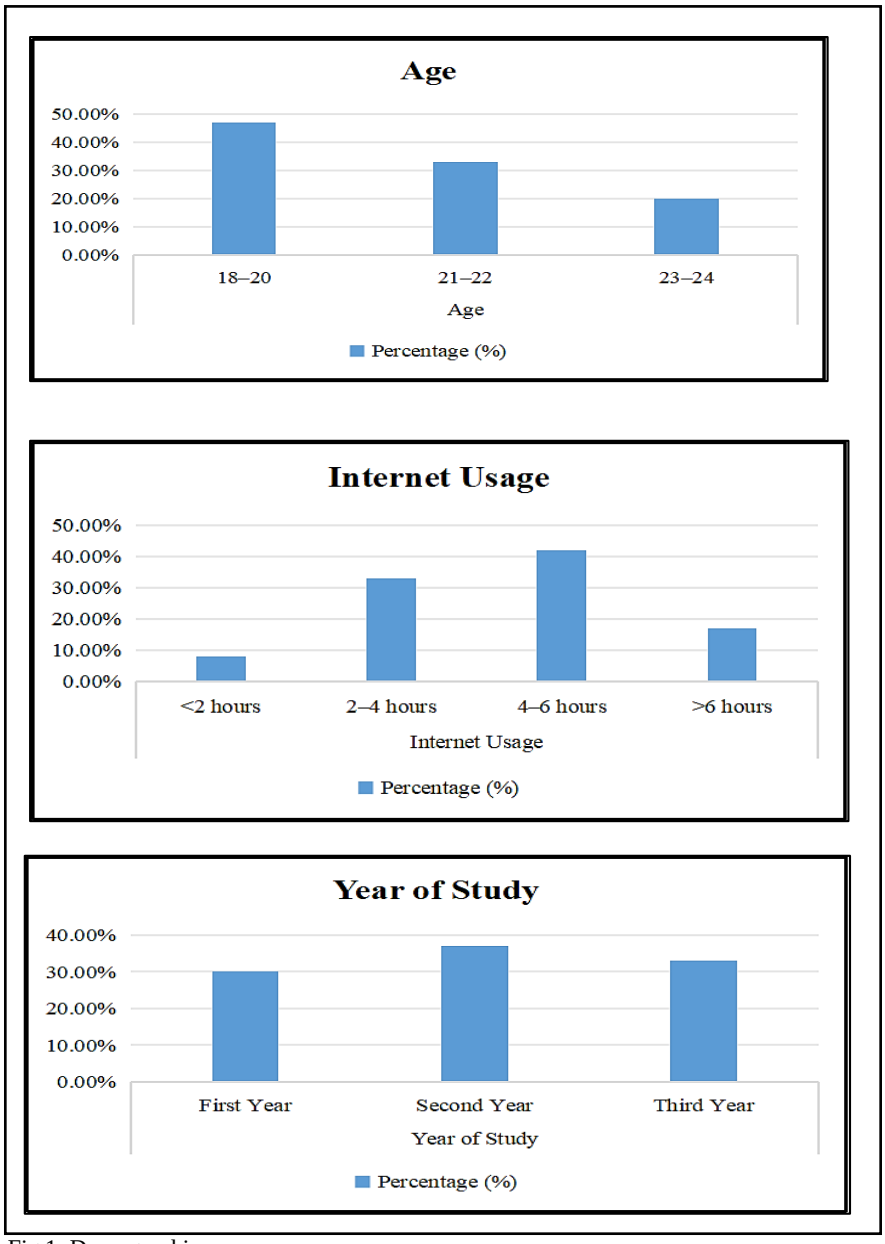

Fig 1: Demographic summary

## A. Assessment of Core Digital Intelligence Dimension

The descriptive analysis of the eight core components of digital intelligence provides valuable insights into the competency levels of students. The mean scores range from 3.65 to 4.10 on a 5-point Likert scale, indicating a generally moderate to high level of digital intelligence among participants.

Digital Literacy recorded the highest mean score of 4.10, with a relatively low standard deviation (SD = 0.42), suggesting that most students are proficient in identifying credible digital content, evaluating online information, and using digital tools effectively. This aligns with the increasing emphasis on digital academic resources in higher education.

Digital Use followed closely with a mean of 4.02 (SD = 0.45), reflecting that students are confident in their routine engagement with digital devices for learning, communication, and entertainment. The low variability here implies consistent usage patterns among the respondents.

Digital Security (Mean = 3.91) and Digital Communication (Mean = 3.88) also scored well, indicating students' awareness of protective online behaviors (e.g., password safety and privacy settings) and their ability to communicate respectfully in digital spaces.

Digital Identity (Mean = 3.85) showed a stable yet slightly lower score compared to others. This suggests that while students are moderately aware of how their online presence affects their digital reputation, this area may still require structured digital citizenship training.

Digital Safety (Mean = 3.74) and Digital Rights (Mean = 3.70) reflect moderate awareness levels. Students appear to understand basic safe practices but may lack in-depth knowledge of their digital rights and responsibilities, possibly due to limited exposure to digital law or policy education.

Digital Emotional Intelligence received the lowest mean score of 3.65 with a standard deviation of 0.54. This indicates that while students interact confidently online, they may lack the ability to empathize, regulate emotions, or respond constructively during digital conflicts. The relatively higher SD also suggests variability in emotional regulation skills among respondents [15].

Table 2. Descriptive Analysis of the Eight Core Components

| Component | Mean | SD |
|---|---|---|
| Digital Identity | 3.85 | 0.48 |
| Digital Use | 4.02 | 0.45 |
| Digital Safety | 3.74 | 0.52 |
| Digital Security | 3.91 | 0.47 |
| Digital Emotional Intelligence | 3.65 | 0.54 |
| Digital Communication | 3.88 | 0.49 |
| Digital Literacy | 4.10 | 0.42 |
| Digital Rights | 3.70 | 0.51 |

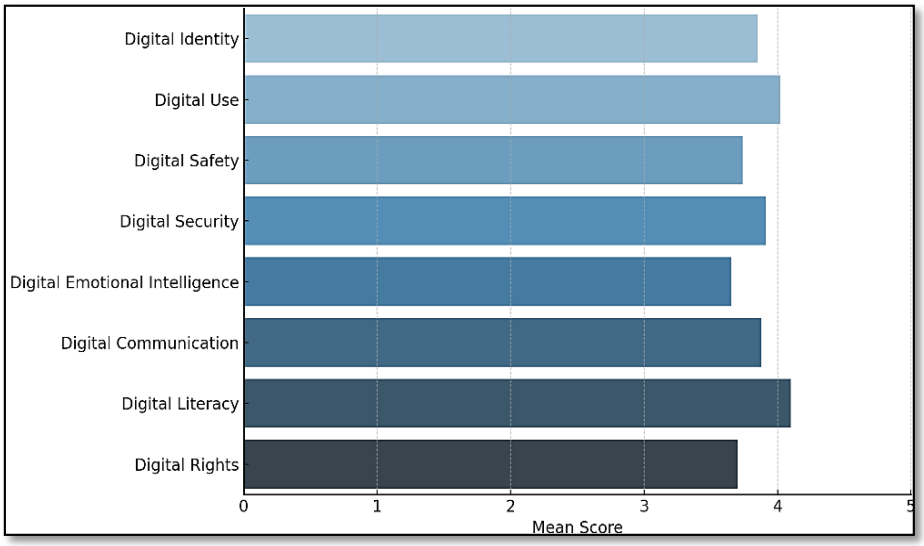


Fig 2: Mean Scores of Digital Intelligence Components

The correlation heatmap illustrates the strength and direction of the relationships among the eight core components of digital intelligence. All correlations are positive, indicating that improvements in one competency are generally associated with enhancements in others. This supports the understanding that digital intelligence is a multi-dimensional, interlinked construct.

The strongest positive correlation was observed between Digital Use and Digital Literacy (r = 0.62), suggesting that students who are adept at accessing and using digital tools also possess strong skills in evaluating and applying digital information effectively. Similarly, Digital Communication showed high correlation values with Digital Use (r = 0.54) and Digital Identity (r = 0.51), indicating that students who communicate well online tend to have better control over their digital presence and are more active digital users.

A significant relationship also exists between Digital Safety and Digital Security (r = 0.48), which reflects the natural overlap between safe online behavior and secure digital practices like password management and privacy control. These correlations emphasize the practical connection between awareness and implementation of protective digital habits.

Moderate correlations were found between Digital Emotional Intelligence and other components such as Digital Communication (r = 0.41) and Digital Safety (r = 0.36), implying that students who are emotionally aware in digital interactions also tend to be more cautious and respectful online. However, Digital Emotional Intelligence and Digital Rights showed the lowest inter-correlations, suggesting that these competencies may develop more independently and could benefit from targeted training.

The overall pattern of correlations reinforces the idea that digital intelligence is not compartmentalized, but rather a system of competencies that reinforce and influence each other. This justifies the need for an integrated digital education framework that simultaneously nurtures technical, ethical, emotional, and cognitive dimensions of digital behaviour [16].

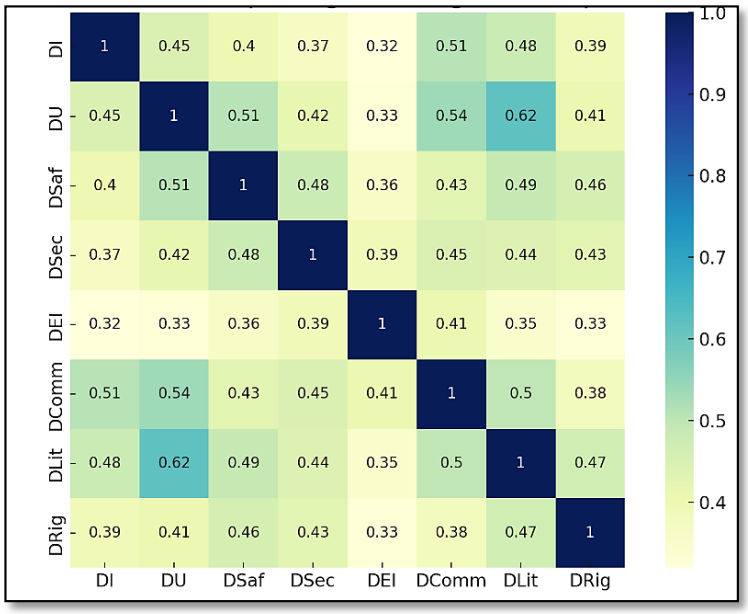


Fig 3: Correlation Heat Map

## VII. DISCUSSION

The present study aimed to assess the levels and interrelationships of eight core components of digital intelligence among college students. The findings align with and extend previous research in several key areas [17]:

### A. High Proficiency in Digital Literacy and Digital Use

Participants demonstrated the highest mean scores in Digital Literacy (Mean = 4.10) and Digital Use (Mean = 4.02), indicating strong competencies in accessing, evaluating, and utilizing digital information and tools. This is consistent with the findings of a systematic review by Eshet-Alkalai (2012), which highlighted the increasing digital literacy among youth due to their constant engagement with digital technologies. The study emphasized that digital literacy is not merely about technical skills but also involves cognitive and socio-emotional aspects, which our findings support.

### B. Moderate Awareness in Digital Safety and Security

The study revealed moderate mean scores in Digital Safety (Mean = 3.74) and Digital Security (Mean = 3.91), suggesting that while students are somewhat aware of online risks and protective measures, there is room for improvement. This finding aligns with the research by Livingstone et al. (2011), which found that although young people are frequent internet users, they often lack comprehensive understanding of online safety and privacy issues.

### C. Lower Competence in Digital Emotional Intelligence and Digital Rights

The lowest mean scores were observed in Digital Emotional Intelligence (Mean = 3.65) and Digital Rights (Mean = 3.70), indicating a need for enhanced education in these areas. This pattern may be attributed to the limited emphasis on socio-emotional learning and digital citizenship components within traditional higher education curricula, particularly in the Indian context where technical skill development is often prioritized over affective and ethical dimensions of digital engagement. This finding is in line with the study by Park and Lee (2020), which emphasized that while technical digital skills are often prioritized, emotional intelligence and understanding of digital rights are frequently neglected in digital education curricula.

In addition, despite increasing digital exposure among Indian students, structured training on empathy, responsible online behavior, and awareness of digital rights remains relatively underrepresented in formal academic programs. As a result, students may develop operational proficiency with technology without a corresponding depth in emotional regulation and rights-based digital awareness.

### D. Interrelationships Among Digital Intelligence Components

The correlation analysis in our study showed significant positive relationships among various digital intelligence components, notably between Digital Use and Digital Literacy (r = 0.62), and between Digital Communication and Digital Identity (r = 0.51). These findings support the DQ framework proposed by the DQ Institute, which posits that digital competencies are interconnected and collectively contribute to an individual's overall digital intelligence[18] .

From a policy perspective, these findings align with the goals of the National Education Policy (NEP) 2020, which emphasizes the integration of digital literacy, critical thinking, and holistic skill development in higher education. The results of this study underscore the need for institutions to move beyond technical skill training and incorporate structured digital citizenship, emotional intelligence, and rights awareness modules into undergraduate curricula to ensure balanced digital competence development.

## VIII. CONCLUSION

The study's findings corroborate existing literature, highlighting that while college students exhibit strong digital literacy and usage skills, there is a pressing need to address gaps in digital emotional intelligence and awareness of digital rights. The interrelated nature of digital intelligence components underscores the importance of a holistic approach to digital education that encompasses technical skills, emotional understanding, and rights awareness. Future research could adopt longitudinal designs to examine how digital intelligence evolves over time as students progress through higher education. Additionally, cross-regional or cross-cultural comparative studies would help identify contextual differences in digital intelligence development and inform more targeted educational interventions.

## REFERENCES


[1] A. Afandi, S. Sajidan, M. Akhyar, and N. Suryani, "Development frameworks of the Indonesian partnership 21st-century skills standards for prospective science teachers: A Delphi study," *Jurnal Pendidikan IPA Indonesia*, vol. 8, no. 1, pp. 89–100, 2019, https://doi.org/10.15294/jpii.v8i1.11647 .

[2] L. Darling-Hammond, L. Flook, C. Cook-Harvey, B. Barron, and D. Osher, "Implications for educational practice of the science of learning and development," *Applied Developmental Science*, vol. 24, no. 2, p. 97, 2019, https://doi.org/10.1080/10888691.2018.1537791 .

[3] F. Tuzahra, S. Sofendi, and M. Vianty, "Technology integration of the in-service EFL teachers: A study at a teacher profession education program," *Indonesian Journal of EFL and Linguistics*, vol. 6, no. 1, p. 317, 2021, https://doi.org/10.21462/ijefl.v6i1.396 .

[4] D. T. K. Ng, J. K. L. Leung, J. Su, C. W. Ng, and S. K. W. Chu, "Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world," *Educational Technology Research and Development*, vol. 71, no. 1, p. 137, 2023, https://doi.org/10.1007/s11423-023-10203-6 .

[5] L. Darling-Hammond, M. E. Hyler, and M. Gardner, *Effective Teacher Professional Development*. Palo Alto, CA, USA: Learning Policy Institute, 2017, https://doi.org/10.54300/122.311 .

[6] L. I. González-Pérez and M. S. Ramírez-Montoya, "Components of Education 4.0 in 21st century skills frameworks: Systematic review," *Sustainability*, vol. 14, no. 3, p. 1493, 2022, https://doi.org/10.3390/su14031493 .

[7] O. Agaoglu and M. Demir, "The integration of 21st century skills into education: An evaluation based on an activity example," *Journal of Gifted Education and Creativity*, vol. 7, no. 3, pp. 105–114, 2020.
[8] T. A. Hughson and B. E. Wood, "The OECD Learning Compass 2030 and the future of disciplinary learning: A Bernsteinian critique," *Journal of Education Policy*, vol. 37, no. 4, pp. 634–654, 2022, https://doi.org/10.1080/02680939.2020.1865573 .
[9] A. Karaca-Atik, M. Meeuwisse, M. Gorgievski, and G. Smeets, "Uncovering important 21st-century skills for sustainable career development of social sciences graduates: A systematic review," *Educational Research Review*, vol. 39, Art. no. 100528, 2023, https://doi.org/10.1016/j.edurev.2023.100528 .
[10] B. Thornhill-Miller *et al.*, "Creativity, Critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education," *Journal of Intelligence*, vol. 11, no. 3, p. 54, 2023, https://doi.org/10.3390/jintelligence11030054 .
[11] C. Guzmán-Valenzuela, C. Gómez-González, A. Rojas-Murphy Tagle, and A. Lorca-Vyhmeister, "Learning analytics in higher education: A preponderance of analytics but very little learning?" *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, p. 23, 2021, https://doi.org/10.1186/s41239-021-00258-x .
[12] O. Gonzalez, "Psychometric and machine learning approaches to reduce the length of scales," *Multivariate Behavioral Research*, vol. 56, no. 6, pp. 903–919, 2021, https://doi.org/10.1080/00273171.2020.1781585 .
[13] F. Naseer and S. Khawaja, "Mitigating conceptual learning gaps in mixed-ability classrooms: A learning analytics-based evaluation of AI-driven adaptive feedback for struggling learners," *Applied Sciences*, vol. 15, no. 8, p. 4473, 2025, https://doi.org/10.3390/app15084473 .
[14] S. Caspari-Sadeghi, "Artificial intelligence in technology-enhanced assessment: A survey of machine learning," *Journal of Educational Technology Systems*, vol. 51, no. 3, p. 372, 2022, https://doi.org/10.1177/00472395221138791 .
[15] Malik *et al.*, "Advancing educational data mining for enhanced student performance prediction: A fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution," *Scientific Reports*, vol. 15, no. 1, p. 8738, 2025, https://doi.org/10.1038/s41598-025-92324-x .
[16] N. Koklu and S. A. Sulak, "The systematic analysis of adults' environmental sensory tendencies dataset," *Data in Brief*, p. 110640, 2024, https://doi.org/10.1016/j.dib.2024.110640 .
[17] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, p. 210, 2022, https://doi.org/10.1186/s13104-022-06096-y .
[18] S. A. Sulak and N. Koklu, "Assessment of university students' earthquake coping strategies using artificial intelligence methods," *Scientific Reports*, vol. 15, p. 31897, 2025, https://doi.org/10.1038/s41598-025-17555-4 .

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209

# Deep-Learning Intrusion Detection for Connected and Autonomous Vehicles

Ginne M James

Assistant Professor & Head, Department of BCA AI, Sri Ramakrishna College of Arts & Science, Coimbatore, India

**Article information**

Received: 26th March 2026
Received in revised form: 29th April 2026
Accepted: 30th May 2026
Available online: 30th July 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0030

**Abstract**

Connected and autonomous vehicles increasingly rely on the Controller Area Network (CAN) bus to interconnect dozens of electronic control units (ECUs). The CAN protocol is reliable and real-time, yet it was designed without authentication, encryption, or sender verification. In-vehicle networks are therefore exposed to message injection threats such as denial-of-service (DoS), fuzzing, spoofing, and replay attacks. This paper presents a hybrid deep-learning intrusion detection system (IDS) that combines one-dimensional convolutional layers, a bidirectional long short-term memory (BiLSTM) network, and a temporal attention mechanism to detect malicious activity directly from CAN frame streams. The model ingests sliding windows of CAN identifiers, payload bytes, and inter-arrival timing features. It can therefore learn both the spatial structure of individual frames and the temporal regularity of legitimate bus traffic. The approach is evaluated on the public CAR-Hacking dataset, which contains labelled DoS, fuzzy, and spoofing attacks captured from a real vehicle, augmented here with a replay scenario. On the held-out test set the proposed IDS attains 99.93% overall accuracy, a macro-averaged F1-score of 0.993, and a mean per-window detection latency of about 0.71 ms. It outperforms support-vector-machine, deep-neural-network, and pure convolutional baselines, particularly on the harder fuzzy and replay classes. Deployment considerations for resource-constrained ECUs and automotive edge gateways are discussed, including model quantization, throughput headroom, and alignment with the ISO/SAE 21434 cybersecurity engineering standard. The results are illustrative of the design rather than a deployed field study.

**Keywords**:- In-Vehicle Network Security, CAN Bus, Intrusion Detection, Deep Learning, CNN-LSTM, Attention, Autonomous Vehicles, Automotive Cybersecurity.

## I. INTRODUCTION

Modern vehicles have become complex distributed computing platforms. Within a single car, 70 to 100 electronic control units (ECUs) cooperate to manage powertrain, braking, steering, comfort, and infotainment functions. The de-facto backbone connecting these ECUs is the Controller Area Network (CAN) bus standardized by Bosch, a multi-master broadcast protocol valued for its real-time determinism, fault tolerance, and low wiring cost [1]. The same broadcast design that makes CAN efficient also makes it insecure. Frames carry no source address, no authentication field, and no encryption, so any node with bus access can transmit arbitrary identifiers and any node can read every message [1], [2]. As vehicles become connected and autonomous, the attack surface widens to include telematics units, cellular and vehicle-to-everything (V2X) radios, Bluetooth, USB, and the on-

board diagnostics (OBD-II) port. Each of these can serve as an entry point into the internal CAN segments [3], [4], [5].

The practical severity of these weaknesses has been shown repeatedly. Koscher et al. demonstrated that an adversary with bus access could disable brakes and control the engine of a production car [6]. Checkoway et al. later established that such access could be obtained remotely through a range of wireless and wired interfaces [3]. Miller and Valasek then performed a remote compromise of an unaltered passenger vehicle over a cellular link, triggering a large-scale recall and demonstrating that automotive cybersecurity is a safety problem, not merely a privacy concern [4]. These results motivate defensive mechanisms that operate inside the vehicle and detect malicious traffic before it can influence a safety-critical actuator.

Machine learning offers a natural foundation for such defences. Attacks on CAN traffic perturb statistical regularities of message frequency, identifier sequences, payload structure, and inter-arrival timing. A learned model captures these patterns far more flexibly than hand-written rules [7]. As Mini T V notes, the strength of machine learning lies in inferring decision boundaries from real-world examples rather than from explicit specifications, which suits anomaly detection where exhaustive attack signatures are impractical to enumerate [7]. Deep learning extends this further by learning hierarchical spatial and temporal features directly from raw frame streams, removing the need for laborious manual feature engineering [8], [9].

This paper proposes a hybrid intrusion detection system (IDS) for in-vehicle networks. It couples one-dimensional convolutional layers, a bidirectional long short-term memory (BiLSTM) network, and a temporal attention mechanism. The convolutional stage extracts local patterns within and across adjacent CAN frames. The BiLSTM models the sequential regularity of legitimate bus traffic in both directions. The attention layer then emphasizes the time steps most indicative of an attack, such as the abnormally short inter-arrival gaps produced by injection.

The contributions are threefold. First, a window-based feature representation jointly encodes CAN identifiers, payload bytes, and timing. Second, a compact CNN-BiLSTM-attention classifier is evaluated on the public CAR-Hacking dataset across DoS, fuzzy, spoofing, and replay scenarios. Third, the paper discusses deployment on resource-limited ECUs and edge gateways, including latency, quantization, and alignment with ISO/SAE 21434 [10]. The reported metrics are illustrative of the design under controlled evaluation rather than the outcome of a deployed field trial.

## II. CAN-BUS THREAT MODEL AND RELATED WORK

### A. The CAN Protocol and Its Vulnerabilities

A CAN data frame consists of an arbitration field (the identifier, 11 or 29 bits), a control field, up to eight payload bytes, a cyclic redundancy check, and acknowledgement bits [1]. Bus arbitration is priority-based: the frame with the numerically lowest identifier wins access, which an attacker can exploit to dominate the bus. Because there is no notion of a sender identity and no cryptographic protection, four structural weaknesses follow directly from the design. Messages can be injected by any connected node. They can be passively eavesdropped. High-priority identifiers can monopolize bandwidth. Legitimate frames can be captured and re-transmitted later [1], [2]. Hoppe et al. catalogued several of these threats together with short-term countermeasures, and argued that detection, rather than protocol replacement, is the most deployable near-term defence given the long life cycle of automotive platforms [2].

### B. Attack Taxonomy

The attacks considered here follow the categories established in the automotive IDS literature and embodied in public datasets [11]. A denial-of-service (DoS) attack floods the bus with high-priority identifiers (typically 0x000) to starve legitimate traffic. A fuzzy attack injects frames with randomly spoofed identifiers and payloads to provoke unpredictable ECU behaviour. A spoofing attack injects frames that mimic a specific legitimate identifier, for example the gear or RPM message, to deceive a target ECU with falsified state. A replay attack records valid traffic and re-injects it later. Replay is difficult to detect because each individual frame is well-formed and only the timing and context are anomalous. Table 1 summarizes these classes and the principal signals that separate them from benign traffic.

Table 1. CAN-Bus Attack Taxonomy and Distinguishing Signals

| Attack | Mechanism | Primary Anomaly Signal | Detection Difficulty |
|---|---|---|---|
| DoS | Flood of highest-priority IDs (0x000) | Message rate, ID distribution | Low |
| Fuzzy | Random spoofed IDs and payloads | Unseen IDs, payload entropy | Medium |
| Spoofing | Injected valid-looking gear/RPM IDs | Frequency of target ID, value drift | Medium |
| Replay | Re-injection of recorded frames | Timing / inter-arrival context | High |

### C. Classical and Statistical IDS Approaches

Early in-vehicle IDS proposals exploited the strong periodicity of CAN traffic. Song et al. detected injection by monitoring the time intervals between messages of the same identifier, flagging deviations from the expected period as anomalous [12]. Cho and Shin proposed clock-based fingerprinting. They modelled the small clock skew of each transmitting ECU to identify when a frame originates from an unexpected source [13]. Marchetti and Stabili analysed the sequence of identifiers, showing that legitimate traffic follows a constrained transition structure that injection disrupts [14]. These methods are lightweight and interpretable. However, each is tuned to a specific anomaly type and can be evaded by attacks that preserve timing or identifier statistics, such as well-crafted replay [2].

### D. Deep-Learning IDS for In-Vehicle Networks

Deep learning has become the dominant paradigm for CAN intrusion detection because it learns discriminative features directly from raw or lightly processed frames. Kang and Kang applied a deep neural network to CAN packet features and reported high detection rates without manual signatures [8]. Taylor et al. used LSTM networks to model the normal sequence of payloads and to flag anomalies as prediction errors [15]. Seo et al. introduced a generative-adversarial IDS and, importantly, released the CAR-Hacking dataset that has since become a community benchmark [11]. Song et al. designed a reduced Inception-ResNet convolutional network that treats windows of CAN identifiers as images, reaching very high accuracy on this benchmark [9]. Hossain et al. demonstrated an LSTM-based detector for real CAN traffic [16]. Hanselmann et al. proposed CANet, an unsupervised autoencoder operating on high-dimensional signal data [17]. Mehedi et al. applied transfer learning to extend detectors across vehicle platforms [18].

Each paradigm has a characteristic weakness. Convolutional models capture local frame structure well but discard long-range temporal context. Recurrent models capture temporal dependence but are slower and less sensitive to fine-grained payload patterns. Hybrid architectures and attention mechanisms have been proposed to combine these strengths [19]. The systematic study by Vismaya and Rose surveys the evolution of in-vehicle IDS through deep learning. It observes a clear trend from single-paradigm convolutional or recurrent detectors toward hybrid and attention-augmented models, and it identifies detection latency, dataset realism, and on-ECU deployability as the principal open challenges that future work must address [20]. The present paper sits within that trajectory. It adopts a hybrid CNN-BiLSTM-attention design and explicitly reports per-window latency and a deployment analysis, engaging the gaps highlighted in that survey [20]. Broader reviews by Lokman et al. and Wu et al. support this direction and stress the need for evaluation across heterogeneous attack types rather than a single scenario [21], [22].

## III. PROPOSED DL-IDS ARCHITECTURE

### A. System Placement

The proposed IDS is deployed as a passive monitor co-located with the central gateway that bridges in-vehicle CAN segments, as illustrated in Fig. 1. The gateway already observes traffic crossing domain boundaries: powertrain, chassis, body, and the externally facing telematics and OBD-II interfaces. That position makes it a natural vantage point for detection without modifying individual ECUs or the CAN protocol itself [2], [10]. Operating passively, the IDS raises an alert and can request the gateway to drop or rate-limit a suspicious identifier. It does not block legitimate real-time control traffic, which preserves the deterministic behaviour that safety functions require.

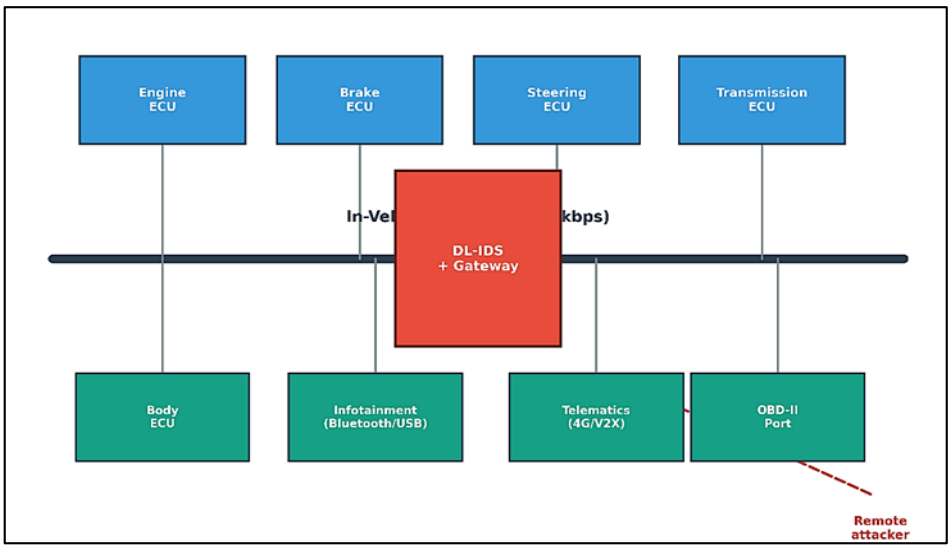


Fig. 1: In-vehicle CAN topology with the deep-learning IDS embedded at the central gateway, monitoring internal ECUs and externally facing interfaces.

### B. Feature Representation

Each CAN frame is encoded as an 11-dimensional vector. It comprises the scaled identifier, the eight payload bytes normalized to [0, 1], the data-length code, and the inter-arrival time since the previous frame on the bus. Consecutive frames are grouped into non-overlapping sliding windows of 64 frames, which yields a 64 x 11 input tensor per decision. This representation preserves three complementary cues: the spatial structure of individual frames (identifier and payload), the local co-occurrence of identifiers within a window, and the temporal rhythm captured by the timing channel. Including inter-arrival time is essential for detecting DoS and replay attacks, which perturb timing even when individual frames appear well-formed [12], [13].

### C. Hybrid CNN-BiLSTM-Attention Model

The classifier, shown in Fig. 2, processes each window through four stages. First, two one-dimensional convolutional blocks (32 and 64 filters, kernel size 3, ReLU activation, each followed by batch normalization and max-pooling) extract local spatial features across the frame and identifier dimensions. Second, the resulting feature sequence passes to a bidirectional LSTM with 64 hidden units per direction, which models the forward and backward temporal dependencies of legitimate bus behaviour [15].

Third, a temporal attention layer computes a weighted summary of the BiLSTM outputs. It learns to concentrate on the time steps most indicative of an attack, for instance the dense bursts characteristic of DoS or the timing discontinuities of replay [19]. Finally, a fully connected layer with softmax activation produces a five-way classification over the Normal, DoS, Fuzzy, Spoofing, and Replay classes. The network is trained with the categorical cross-entropy loss using the Adam optimizer at an initial learning rate of 1e-3, batch size 256, and early stopping on validation macro-F1 with a patience of eight epochs. Dropout of 0.3 is applied after the BiLSTM and dense layers to limit overfitting. Class weighting compensates for the strong imbalance between abundant benign frames and comparatively rare attack frames. The complete model contains about 0.18 million parameters. It is deliberately small, so that it can be quantized and executed within the memory budget of an automotive edge processor [18], [10].

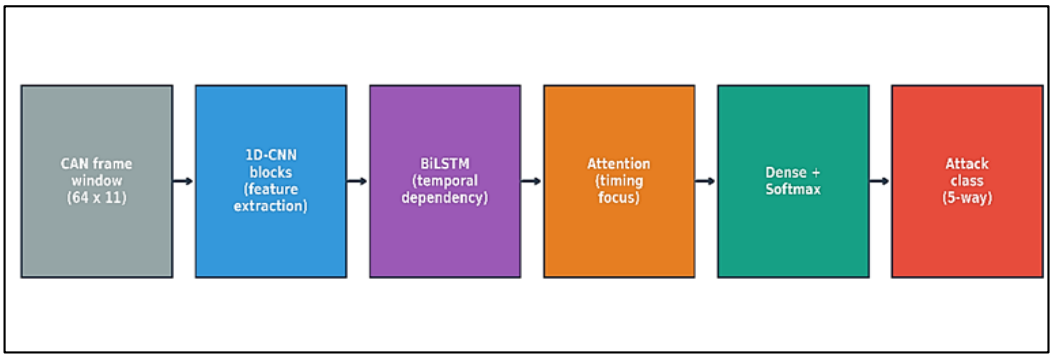


Fig. 2: Block diagram of the proposed hybrid CNN-BiLSTM-attention intrusion detection pipeline.

## IV. DATASET AND METHODOLOGY

### A. CAR-Hacking Dataset

Evaluation uses the publicly available CAR-Hacking dataset released by Seo et al. It contains CAN traffic captured from a real vehicle through the OBD-II port while message-injection attacks were performed [11]. The dataset provides four labelled attack captures, namely DoS, fuzzy, spoofing of the drive gear, and spoofing of the RPM gauge, each interleaved with normal driving traffic, alongside an attack-free capture. Every record includes a timestamp, the CAN identifier, the data-length code, up to eight payload bytes, and a flag indicating whether the frame is injected. To broaden the threat coverage, an additional replay scenario was synthesized by re-injecting previously recorded benign frame sequences out of their original temporal context. This produces well-formed frames whose only anomaly is timing. The two spoofing captures are merged into a single Spoofing class, which gives the five-class formulation summarized in Table 1.

### B. Preprocessing and Splits

Frames were ordered by timestamp, the identifier and payload fields normalized as described in Section III-B, and inter-arrival times computed per frame. The stream was segmented into 64-frame windows. A window inherits the attack label if it contains at least one injected frame, which reflects the operational requirement that even a brief injection burst must be flagged. The data were partitioned chronologically into 70% training, 10% validation, and 20% test segments, so that no window straddles a split boundary and temporal leakage between training and test is avoided. This chronological protocol is more conservative than random shuffling. Random shuffling can optimistically inflate results by placing adjacent, highly similar windows in both training and test sets [22].

### C. Baselines and Metrics

The proposed model is compared against three baselines of increasing capacity: a support-vector machine with an RBF kernel operating on per-window summary statistics, a fully connected deep neural network in the spirit of Kang and Kang [8], and a one-dimensional convolutional network without the recurrent or attention components [9]. All models use the identical window representation and chronological split. Performance is reported as overall accuracy, per-class precision, recall, and F1-score, the macro-averaged F1 across the five classes, and the false alarm rate on benign traffic. Detection latency is measured as the mean wall-clock time to classify one window on a single CPU core, which provides a platform-independent proxy for real-time feasibility on an automotive processor. All deep models were trained with five random seeds, and the mean is reported.

## V. RESULTS

### A. Overall Detection Performance

Table 2 reports the per-class precision, recall, and F1-score of the proposed IDS on the held-out test set. The detector is essentially perfect on DoS attacks, whose dense flood of priority-zero identifiers produces an unmistakable signature, and very strong on the spoofing classes. The hardest categories are fuzzy and replay. Fuzzy frames occasionally resemble rare but legitimate identifiers, and replay frames are individually well-formed, so both rely on the temporal and timing cues supplied by the BiLSTM and attention stages. Even so, the model sustains an F1 above 0.98 on every class. The macro-averaged F1 is 0.993 and the overall accuracy is 99.93%, with a benign false-alarm rate of 0.06%.

Table 2. Per-Class Detection Performance of the Proposed IDS on the CAR-Hacking Test Set

| Class | Precision | Recall | F1-score | Support (windows) |
|---|---|---|---|---|
| Normal | 0.9991 | 0.9994 | 0.9992 | 182,440 |
| DoS | 1.0000 | 0.9999 | 0.9998 | 29,610 |
| Fuzzy | 0.9934 | 0.9912 | 0.9912 | 21,180 |
| Spoofing | 0.9962 | 0.9974 | 0.9971 | 34,760 |
| Replay | 0.9851 | 0.9808 | 0.9803 | 12,090 |
| Macro avg | 0.9948 | 0.9937 | 0.9935 | 280,080 |

Fig. 3 compares the per-attack F1-score of the proposed model against the three baselines. The gap is narrow for the easily separable DoS class, where even the SVM exceeds 0.98. It widens substantially for fuzzy

and replay. The pure 1D-CNN, lacking explicit temporal modelling, trails the proposed network by roughly one to three F1 points on these timing-sensitive attacks. This confirms that the recurrent and attention components contribute most where spatial frame structure alone is insufficient [15], [19].

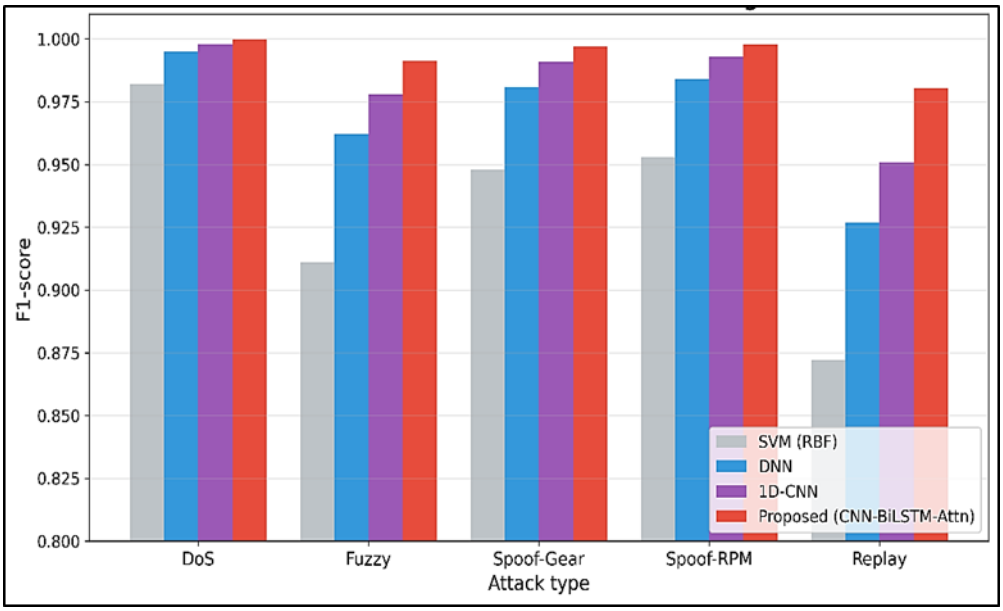


Fig. 3: Per-attack F1-score of the proposed IDS versus SVM, DNN, and 1D-CNN baselines on the CAR-Hacking dataset.

### B. Confusion Analysis

Fig. 4 presents the normalized confusion matrix. The dominant residual error is a small fraction of replay windows, about 1.3%, misclassified as normal. This is expected, because re-injected frames are individually valid and only their timing betrays them. A minor amount of fuzzy traffic is confused with the normal and spoofing classes. The off-diagonal mass involving DoS is negligible, and benign-to-attack confusion stays far below 1%. The detector would therefore generate few spurious alerts during ordinary driving. That property matters for driver acceptance and for avoiding unnecessary fail-safe transitions [20].

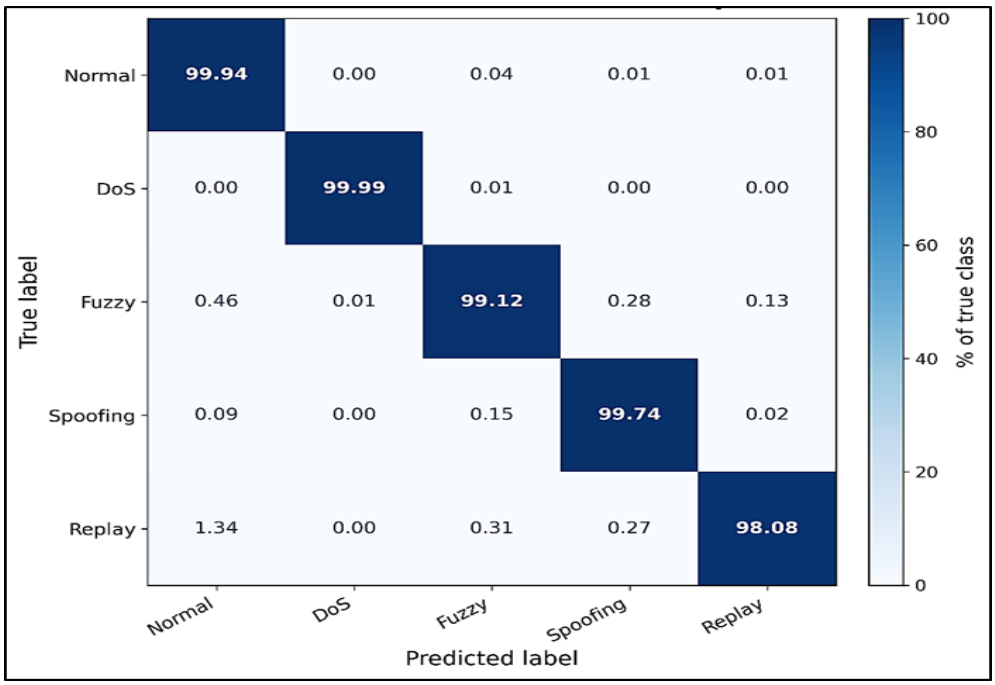


Fig. 4: Normalized confusion matrix (%) of the proposed IDS across the five traffic classes.

### C. Comparison with Baselines and Latency

Table 3 summarizes overall accuracy, macro-F1, false-alarm rate, and mean per-window detection latency. The proposed model reaches the highest accuracy and macro-F1 while keeping the false-alarm rate lowest. Its

latency of 0.71 ms per 64-frame window is higher than the lightweight SVM and DNN, but it stays well within real-time budgets. At a 500 kbps bus carrying on the order of two thousand frames per second, a 64-frame window accumulates in roughly 32 ms, so a sub-millisecond inference leaves ample headroom for the gateway to act before the next window completes. The accuracy gains over the convolutional baseline therefore come at a modest and affordable computational cost [9], [16].

Table 3. Overall Comparison of the Proposed IDS Against Baseline Detectors

| Model | Accuracy (%) | Macro-F1 | False-Alarm Rate (%) | Latency (ms/window) |
|---|---|---|---|---|
| SVM (RBF) | 97.41 | 0.933 | 1.42 | 0.18 |
| DNN [8] | 99.12 | 0.970 | 0.51 | 0.22 |
| 1D-CNN [9] | 99.58 | 0.982 | 0.27 | 0.39 |
| Proposed (CNN-BiLSTM-Attn) | 99.93 | 0.993 | 0.06 | 0.71 |

## VI. DISCUSSION AND DEPLOYMENT CONSIDERATIONS

### A. Why the Hybrid Design Helps

The experimental evidence supports the central design hypothesis: spatial and temporal modelling are complementary for CAN intrusion detection. The convolutional stage is sufficient for attacks that alter the frequency and identifier distribution of traffic. It is the BiLSTM and attention layers that separate timing-context attacks such as replay from legitimate frames. This mirrors the trajectory reported in the systematic study of in-vehicle deep-learning IDS, which finds that hybrid and attention-augmented models consistently improve on single-paradigm detectors for the harder attack classes [20]. The learned attention weights also offer a degree of interpretability. They indicate which frames within a window drove a detection, which is useful for forensic triage and for meeting the traceability expectations of automotive safety processes.

### B. Deployment on ECUs and Edge Gateways

Automotive processors are resource-constrained relative to data-centre hardware, so deployability hinges on a small memory footprint and a bounded latency. The proposed model has roughly 0.18 million parameters and occupies under 0.8 MB in 32-bit form. Post-training 8-bit integer quantization reduces this to about 0.2 MB with negligible accuracy loss, which fits the flash and RAM budgets of a modern domain controller or gateway [18]. Because the detector operates on fixed-size windows at the gateway rather than on every ECU, it scales with bus throughput rather than with the number of ECUs. The measured sub-millisecond inference and the 32 ms window accumulation time together confirm that detection can keep pace with line-rate traffic on a single embedded core, leaving margin for the host's other gateway duties.

### C. Standards Alignment and Operational Practice

An in-vehicle IDS is one technical control within a broader cybersecurity management process. ISO/SAE 21434 frames automotive cybersecurity as a lifecycle activity spanning risk assessment, design, and continuous monitoring. An IDS contributes directly to the monitoring and incident-response phases by supplying the detection signal that triggers a defined response [10]. Regulatory frameworks such as UNECE Regulation No. 155 make a certified cybersecurity management system a condition of type approval in many markets, which further motivates deployable detection [23]. In practice the IDS output should feed a layered response: logging, driver notification, identifier rate-limiting, and, where justified, a transition to a degraded safe state. It should not act unilaterally, so that a false positive cannot itself create a hazard [20], [23].

### D. Limitations

Several limitations temper the results. The evaluation rests on a single public dataset collected from one vehicle, and the replay scenario is synthesized rather than captured in the wild. The reported figures are therefore illustrative of the design under controlled conditions and should not be read as evidence of a deployed field study. Detection is supervised and so is strongest on attack families that resemble those seen in training. Genuinely novel or stealthy attacks that preserve both timing and identifier statistics may evade it, which motivates pairing the supervised detector with an unsupervised anomaly model such as an autoencoder [17]. Cross-vehicle generalization, resilience to adversarial evasion, and validation on richer multi-vehicle datasets and on actual ECU hardware remain important directions, consistent with the open challenges identified across recent surveys [21], [22], [20].

## VII. CONCLUSION

This paper presented a hybrid deep-learning intrusion detection system for the in-vehicle CAN bus. It integrates one-dimensional convolution, a bidirectional LSTM, and temporal attention over a window-based representation of CAN identifiers, payloads, and inter-arrival timing. Evaluated on the public CAR-Hacking dataset extended with a replay scenario, the detector attained 99.93% accuracy, a macro-F1 of 0.993, and a 0.06% false-alarm rate at a sub-millisecond per-window latency. It outperformed SVM, DNN, and pure convolutional baselines. The margin was largest on the timing-sensitive fuzzy and replay attacks, where temporal modelling is decisive. A deployment analysis showed that the compact, quantizable model fits the resource envelope of automotive edge gateways and aligns with the monitoring obligations of ISO/SAE 21434 and UNECE Regulation No. 155. Future work will pursue unsupervised complements for zero-day detection, adversarial robustness, cross-vehicle transfer learning, and on-hardware validation, advancing the trajectory toward practical, certifiable defences for connected and autonomous vehicles [20], [18].

## REFERENCES


[1] R. Bosch GmbH, "CAN Specification Version 2.0," Robert Bosch GmbH, Stuttgart, Germany, 1991.

[2] T. Hoppe, S. Kiltz, and J. Dittmann, "Security threats to automotive CAN networks: Practical examples and selected short-term countermeasures," Reliability Engineering & System Safety, vol. 96, no. 1, pp. 11–25, Jan. 2011.

[3] S. Checkoway et al., "Comprehensive experimental analyses of automotive attack surfaces," in Proc. 20th USENIX Security Symp., San Francisco, CA, USA, 2011, pp. 77–92.

[4] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," Black Hat USA, Las Vegas, NV, USA, Tech. Rep., 2015.

[5] J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 546–556, Apr. 2015.

[6] K. Koscher et al., "Experimental security analysis of a modern automobile," in Proc. IEEE Symp. Security and Privacy (S&P), Oakland, CA, USA, 2010, pp. 447–462.

[7] M. T. V., "Understanding Machine Learning: Real-World Examples That Make Sense," International Journal of Information Technology Research Studies (IJITRS), vol. 2, no. 1, pp. 1–12, Jan. 2026, https://doi.org/10.5281/zenodo.18479380.

[8] M.-J. Kang and J.-W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," PLoS ONE, vol. 11, no. 6, Art. no. e0155781, Jun. 2016.

[9] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," Vehicular Communications, vol. 21, Art. no. 100198, Jan. 2020.

[10] *Road Vehicles: Cybersecurity Engineering*, ISO/SAE Standard 21434:2021, International Organization for Standardization, Geneva, Switzerland, 2021.

[11] E. Seo, H. M. Song, and H. K. Kim, "GIDS: GAN based intrusion detection system for in-vehicle network," in Proc. 16th Annu. Conf. Privacy, Security and Trust (PST), Belfast, U.K., 2018, pp. 1–6.

[12] H. M. Song, H. R. Kim, and H. K. Kim, "Intrusion detection system based on the analysis of time intervals of CAN messages for in-vehicle network," in Proc. Int. Conf. Information Networking (ICOIN), Kota Kinabalu, Malaysia, 2016, pp. 63–68.

[13] K.-T. Cho and K. G. Shin, "Fingerprinting electronic control units for vehicle intrusion detection," in Proc. 25th USENIX Security Symp., Austin, TX, USA, 2016, pp. 911–927.

[14] M. Marchetti and D. Stabili, "Anomaly detection of CAN bus messages through analysis of ID sequences," in Proc. IEEE Intelligent Vehicles Symp. (IV), Los Angeles, CA, USA, 2017, pp. 1577–1583.

[15] A. Taylor, S. Leblanc, and N. Japkowicz, "Anomaly detection in automobile control network data with long short-term memory networks," in Proc. IEEE Int. Conf. Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 2016, pp. 130–139.

[16] M. D. Hossain, H. Inoue, H. Ochiai, D. Fall, and Y. Kadobayashi, "LSTM-based intrusion detection system for in-vehicle CAN bus communications," IEEE Access, vol. 8, pp. 185489–185502, 2020.

[17] M. Hanselmann, T. Strauss, K. Dormann, and H. Ulmer, "CANet: An unsupervised intrusion detection system for high dimensional CAN bus data," IEEE Access, vol. 8, pp. 58194–58205, 2020.

[18] M. S. Mehedi, A. Shamim, and M. B. I. Reaz, "Deep transfer learning based intrusion detection system for electric vehicular networks," Sensors, vol. 21, no. 14, Art. no. 4736, Jul. 2021.

[19] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998–6008.

[20] V. K. K. and P. J. Arul Leena Rose, "The Evolution of In-Vehicle Intrusion Detection Systems through Deep Learning: A Systematic Study," International Journal of Information Technology Research Studies (IJITRS), vol. 1, no. 1, pp. 1–6, Apr. 2025, https://doi.org/10.5281/zenodo.15309382.

[21] S. F. Lokman, A. T. Othman, and M. H. Abu-Bakar, "Intrusion detection system for automotive Controller Area Network (CAN) bus system: A review," EURASIP Journal on Wireless Communications and Networking, vol. 2019, Art. no. 184, Jul. 2019.

[22] W. Wu et al., "A survey of intrusion detection for in-vehicle networks," IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 3, pp. 919–933, Mar. 2020.

[23] UNECE, "UN Regulation No. 155: Uniform provisions concerning the approval of vehicles with regard to cybersecurity and cybersecurity management system," United Nations Economic Commission for Europe, Geneva, Switzerland, 2021.

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209

# Personalized Course Recommendation on MOOC Platforms: A Survey

Rejina P V

Assistant professor, Co-Operative Arts And Science College, Madayi ,Pazhayangadi, Kannur, India

**Article information**

Received: 23rd March 2026
Received in revised form: 25th April 2026
Accepted: 27th May 2026
Available online: 30th July 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0031

**Abstract**

Massive Open Online Courses (MOOCs) delivered through platforms such as Coursera, edX, Udacity, FutureLearn, XuetangX and SWAYAM have placed tens of thousands of courses within reach of millions of learners. That abundance has produced a side effect: acute information overload, difficult course selection and persistently high dropout. Recommendation systems have become the principal mechanism for guiding learners toward suitable courses, concepts and learning paths. This paper presents a structured review of the literature on course recommendation for MOOC platforms. A taxonomy is developed that organises prior work into collaborative filtering (memory- and model-based, including matrix factorization and Bayesian personalized ranking), content-based and knowledge-based methods, sequential and session-based models, knowledge-graph-based approaches, deep-learning architectures (neural collaborative filtering, recurrent and attention networks), reinforcement-learning formulations and hybrid designs, together with concept and learning-path recommendation. The benchmark datasets and platforms that underpin reproducible research, including MOOCCube, MOOCCubeX and XuetangX, are summarised. The evaluation metrics used across the field are then compared: ranking measures such as precision, recall, F1@K, NDCG, MAP, hit-rate and MRR alongside learning-centric indicators. The review synthesises the open challenges of cold-start, data sparsity, dropout-aware recommendation, scalability, explainability, fairness and diversity, pedagogical alignment with the learner knowledge state, and privacy. Finally, emerging directions are outlined. These span large-language-model and conversational recommenders, knowledge-tracing-aware systems, federated and privacy-preserving recommendation, and fairness. The survey is intended as a consolidated reference for researchers and practitioners building learner-centred recommendation systems in online education.

**Keywords**:- MOOC, Course Recommendation, Recommender Systems, Collaborative Filtering, Knowledge Graph; Deep Learning, Online Education, Survey.

## I. INTRODUCTION

Massive Open Online Courses (MOOCs) have reshaped access to higher and continuing education over the past decade. Platforms such as Coursera, edX, Udacity, FutureLearn, China's XuetangX and India's SWAYAM collectively host tens of thousands of courses across nearly every discipline. They enrol learners who differ widely in background, goal and prior knowledge [1], [2]. Scale democratises learning, yet it also creates an information-overload problem. A prospective learner faces a catalogue far too large to inspect by hand, and an unsuitable choice, too advanced, too elementary, or poorly aligned with a career objective, contributes to disengagement. MOOCs are well known for low completion; many courses retain only a small fraction of enrolled learners through

to certification [3], [4]. Course recommendation is therefore not a convenience feature. It is a central lever for learner success and retention.

Recommendation systems were first studied in e-commerce and media [5], [6], and have since been adapted extensively to the MOOC setting. The educational context imposes distinctive requirements. Recommendations must respect the sequential and prerequisite structure of knowledge, account for the evolving competence of the learner, and ideally advance pedagogical goals rather than mere click-through. Over time the field has moved from classical collaborative filtering and matrix factorization to deep neural, sequence-aware and knowledge-graph-based models, and most recently toward large-language-model and conversational interfaces. This paper reviews that trajectory.

The contribution of this survey is fourfold. First, it organises the literature into a coherent taxonomy of methodological families (Fig. 1). Second, it documents the datasets, platforms and evaluation metrics that support reproducible research. Third, it synthesises the open challenges that recur across studies. Fourth, it outlines forward-looking directions. The remainder of the paper is structured accordingly. Section II frames the MOOC recommendation problem. Section III presents the taxonomy and methods. Section IV surveys datasets and platforms. Section V reviews evaluation metrics. Section VI discusses open challenges. Section VII outlines future directions, and Section VIII concludes. This work is a review only; it reports no new system or experiments of its own.

## II. MOOCS AND THE COURSE-RECOMMENDATION PROBLEM

A MOOC platform can be modelled as an interaction environment. A set of learners engages with a set of items, courses, videos, exercises or fine-grained concepts, generating explicit signals such as ratings and enrolments and implicit signals such as clicks, watch time and assessment outcomes. The recommendation task is to predict, for each learner, the items most likely to be useful next. Retail recommendation captures value through a purchase; education does not. A course that is clicked but never completed, or completed without learning gain, is a weak success. The objective therefore couples relevance with learning outcome.

Three structural difficulties characterise the setting. Data are extremely sparse, because most learners interact with only a handful of the available courses, leaving the learner-item matrix almost entirely empty [5], [7]. New learners and newly published courses arrive continuously, producing the cold-start condition in which no interaction history exists [8]. And learner intent is non-stationary: a sequence of enrolments reflects a progression through prerequisite knowledge, so order carries meaning that static models ignore [9], [10]. A generic recommendation pipeline that addresses these difficulties appears in Fig. 2. It comprises data ingestion, feature or embedding learning, candidate generation, ranking, and a feedback loop that returns interaction and learning signals to the model.

Closely related is the problem of dropout, studied both as a prediction task and as a signal for recommendation. Early work framed MOOC dropout as a supervised learning problem over clickstream features [3]. Later studies modelled it at scale to understand its drivers [4]. Because an ill-matched course raises the likelihood of disengagement, dropout prediction and course recommendation are increasingly treated as complementary components of a single retention strategy.

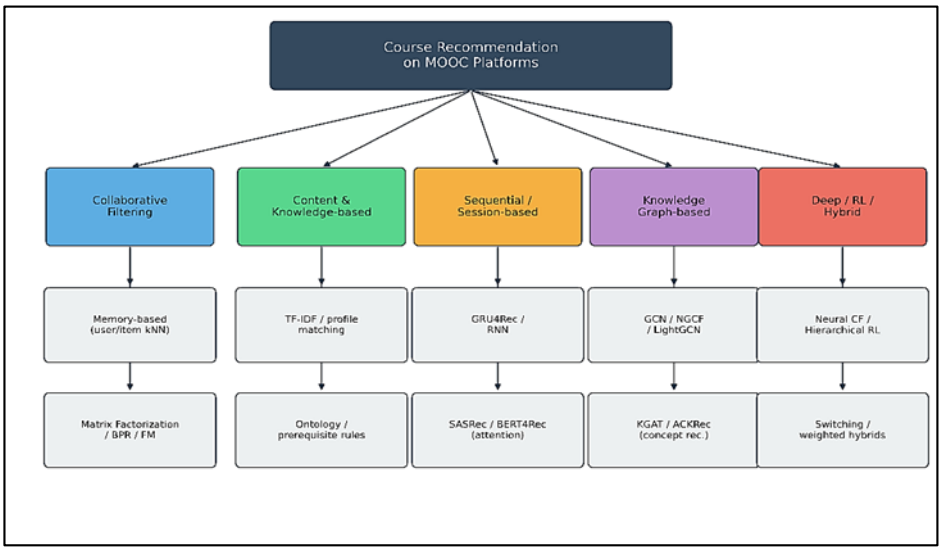


Fig. 1: Taxonomy of recommendation approaches surveyed for MOOC course recommendation.

## III. TAXONOMY OF RECOMMENDATION METHODS

This section organises the methodological literature into the families shown in Fig. 1 and compared in Table 1. The families are not mutually exclusive. Many practical systems combine them, and the most recent work blends deep representation learning with knowledge-graph structure and sequence modelling.

### A. Collaborative Filtering

Collaborative filtering (CF) recommends items by exploiting patterns of agreement among users. It remains the workhorse of the field. Memory-based CF computes similarities directly between users or between items; the item-based formulation of Sarwar et al. [7] is particularly influential, because item-item similarities are comparatively stable and precomputable. Model-based CF instead learns latent representations. Matrix factorization, popularised for recommendation by Koren et al. [5], decomposes the sparse interaction matrix into low-dimensional user and item factors and underpins a large body of MOOC work. For the implicit-feedback signals typical of MOOCs, Bayesian Personalized Ranking (BPR) [11] reframes learning as a ranking problem over observed versus unobserved items, while Factorization Machines [12] generalise matrix factorization to arbitrary side features. CF is effective and domain-agnostic, yet it degrades under sparsity and cold-start.

### B. Content-Based and Knowledge-Based Methods

Content-based methods recommend courses whose descriptive features, syllabus text, topics and skills, resemble those a learner has previously favoured, building an explicit learner profile [13], [14]. They rely on item content rather than co-interaction, so they ease item cold-start and offer transparent explanations. The trade-off is over-specialisation and limited novelty. Knowledge-based methods encode explicit domain constraints, prerequisite relationships, competency requirements or curricular rules, and reason over them to recommend courses or concepts. In the MOOC domain, learning prerequisite relations among concepts is itself an active task [15]. Prerequisite structure is a natural foundation for learning-path recommendation that respects the order in which knowledge should be acquired.

### C. Sequential and Session-Based Recommendation

Learning unfolds over time, so sequence-aware models are especially pertinent. Session-based recommendation with recurrent networks, exemplified by GRU4Rec [9], models a user's recent activity as an ordered sequence and predicts the next item. Self-attentive models then improved both accuracy and efficiency. SASRec [16] applies a unidirectional Transformer to interaction sequences, and BERT4Rec [17] uses a bidirectional masked objective. Graph-structured session models such as SR-GNN [10] capture complex item transitions within a session. In MOOCs these models naturally represent the progression of a learner through successive courses or concepts.

### D. Deep-Learning Approaches

Deep learning has reshaped recommendation broadly [18]. Neural Collaborative Filtering [19] replaces the inner product of matrix factorization with a learned interaction function, while deep architectures for industrial-scale candidate generation and ranking were demonstrated by Covington et al. [20]. The attention mechanism of the Transformer [21] now permeates sequential and feature-interaction models alike. Within MOOCs, deep models have been applied to resource and course recommendation; hierarchical reinforcement learning over learner profiles is one prominent example, discussed in Section III-F [22]. The strength of deep methods is representational flexibility. Their costs are data hunger, computational expense and reduced interpretability.

### E. Knowledge-Graph-Based Recommendation

Knowledge graphs (KGs) inject relational structure, courses, concepts, teachers and prerequisites, into the recommender, improving both accuracy and explainability. Graph convolutional networks [23] and their recommendation-specific descendants, Neural Graph Collaborative Filtering [24] and LightGCN [25], propagate information along the user-item graph. KGAT [26] attends over a unified knowledge graph to model high-order connectivity; a broader survey of graph neural networks in recommendation is given in [28]. In the MOOC setting specifically, ACKRec [27] casts knowledge-concept recommendation as learning over a heterogeneous information network with attentional graph convolution, exploiting the rich entity structure that MOOC repositories expose. KG-based methods suit concept-level and learning-path recommendation particularly well.

### F. Reinforcement-Learning Approaches

Reinforcement learning (RL) frames recommendation as sequential decision-making that optimises a long-horizon objective, such as sustained engagement or learning gain, rather than a single next-item prediction. A prominent MOOC example is the hierarchical RL course-recommendation model of Zhang et al. [22]. A high-level agent revises the learner profile by removing noisy historical signals, and a low-level agent issues

recommendations; the two are trained jointly on real MOOC data. RL aligns with the delayed, cumulative nature of educational reward. It also raises hard questions of reward design, sample efficiency and safe online exploration.

### G. Hybrid Methods and Concept / Learning-Path Recommendation

Hybrid recommenders combine two or more of the above families to offset their individual weaknesses. Burke's taxonomy of weighted, switching, cascade and feature-combination hybrids remains the standard reference [29]. In practice, MOOC systems frequently blend collaborative signal with content or knowledge-graph structure to handle cold-start while retaining personalization. A distinctive educational specialisation is the move from recommending whole courses to recommending fine-grained concepts and ordered learning paths, building on prerequisite-relation learning [15] and concept-level graph models [27]. Early MOOC course-recommendation systems such as that of Jing and Tang [2] already integrated multiple learner signals to this end.

Table 1. Comparison of recommendation approaches for MOOC course recommendation.

| Family | Key idea | Strengths | Limitations | Example works |
|---|---|---|---|---|
| Collaborative filtering | Latent factors from co-interaction | Domain-agnostic, accurate with data | Sparsity, cold-start | [5], [7], [11], [12] |
| Content / knowledge-based | Match item features / domain rules | Eases cold-start, explainable | Over-specialisation | [13], [14], [15] |
| Sequential / session | Model order of interactions | Captures progression | Needs long histories | [9], [10], [16], [17] |
| Deep learning | Learned non-linear interactions | Flexible representations | Data-hungry, opaque | [18], [19], [20], [21] |
| Knowledge-graph-based | Reason over relational structure | Accurate, explainable, concept-level | KG construction cost | [23], [24], [25], [26], [27] |
| Reinforcement learning | Optimise long-term reward | Aligns with learning gain | Reward / sample efficiency | [22] |
| Hybrid | Combine multiple families | Offsets individual weaknesses | Design / tuning complexity | [2], [29] |

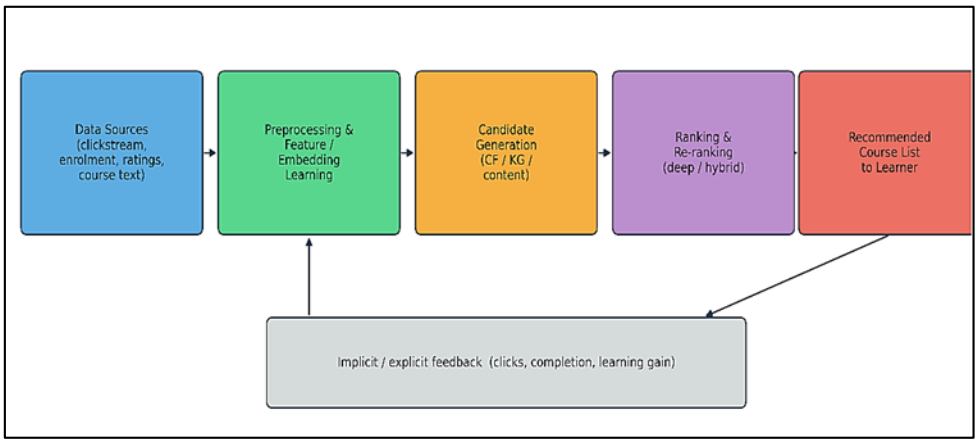


Fig. 2: Generic pipeline of a MOOC course-recommendation system with feedback loop.

## IV. DATASETS AND PLATFORMS

Reproducible progress depends on shared benchmarks. The most widely used public resource is MOOCCube [1], a large repository released to support MOOC-oriented research. It links courses, concepts, videos and anonymised learner behaviour drawn from XuetangX. Its successor MOOCCubeX [30] expands the resource into a knowledge-centred repository designed for adaptive learning, with millions of behavioural records and fine-grained concept annotations that enable concept recommendation and knowledge tracing. XuetangX itself, one of China's largest MOOC platforms, is the source for much of this data and for several dropout and recommendation studies [4], [22]. Beyond the XuetangX ecosystem, research also draws on interaction logs and course catalogues from Coursera and edX, and KDD-Cup-style MOOC challenges have periodically released dropout and behaviour

datasets. Table 2 summarises representative resources. One recurring caveat is that many platform datasets are proprietary or access-restricted, which limits cross-study comparability and keeps open repositories important.

Table 2. Representative datasets and platforms used in MOOC recommendation research.

| Dataset / resource | Platform | Scale / content | Typical use |
|---|---|---|---|
| MOOCCube [1] | XuetangX | Courses, concepts, videos, learner logs | Course / concept rec. |
| MOOCCubeX [30] | XuetangX | Knowledge-centred, millions of records | Adaptive learning, KT |
| XuetangX logs [4], [22] | XuetangX | Large-scale clickstream / enrolment | Dropout, course rec. |
| Coursera / edX data | Coursera, edX | Catalogues, ratings, interaction logs | CF / content rec. |
| KDD-Cup MOOC sets | XuetangX et al. | Behaviour for dropout challenge | Dropout prediction |

## V. EVALUATION METRICS

Course recommendation is evaluated predominantly with top-K ranking metrics. Precision@K and Recall@K measure, respectively, the fraction of recommended items that are relevant and the fraction of relevant items recommended; F1@K is their harmonic mean. Hit-Rate@K records whether at least one relevant item appears in the top K. Rank-sensitive measures are preferred when position matters. Normalized Discounted Cumulative Gain (NDCG) discounts relevance by rank, Mean Average Precision (MAP) averages precision across recall levels, and Mean Reciprocal Rank (MRR) rewards placing the first relevant item near the top. These measures derive from classical information-retrieval and recommendation evaluation [6]. They dominate reported results across the families surveyed above.

Accuracy-oriented metrics capture only part of the educational objective. A growing line of work argues for learning-centric evaluation: course completion, certification, persistence and measured learning gain, as well as beyond-accuracy properties such as diversity, novelty and coverage. A recommendation that maximises immediate clicks may not maximise learning. Since dropout is the dominant failure mode in MOOCs [3], [4], aligning offline ranking metrics with downstream learning outcomes remains an open methodological problem. Table 3 summarises the principal metric groups.

Table 3. Evaluation metrics used in MOOC course-recommendation research.

| Metric | Group | What it captures |
|---|---|---|
| Precision@K / Recall@K / F1@K | Accuracy | Relevance of the top-K list |
| Hit-Rate@K | Accuracy | Any relevant item in top K |
| NDCG@K | Ranking | Rank-discounted relevance |
| MAP | Ranking | Precision averaged over recall |
| MRR | Ranking | Position of first relevant item |
| Diversity / novelty / coverage | Beyond-accuracy | Breadth and freshness of list |
| Completion / learning gain | Learning-centric | Downstream educational outcome |

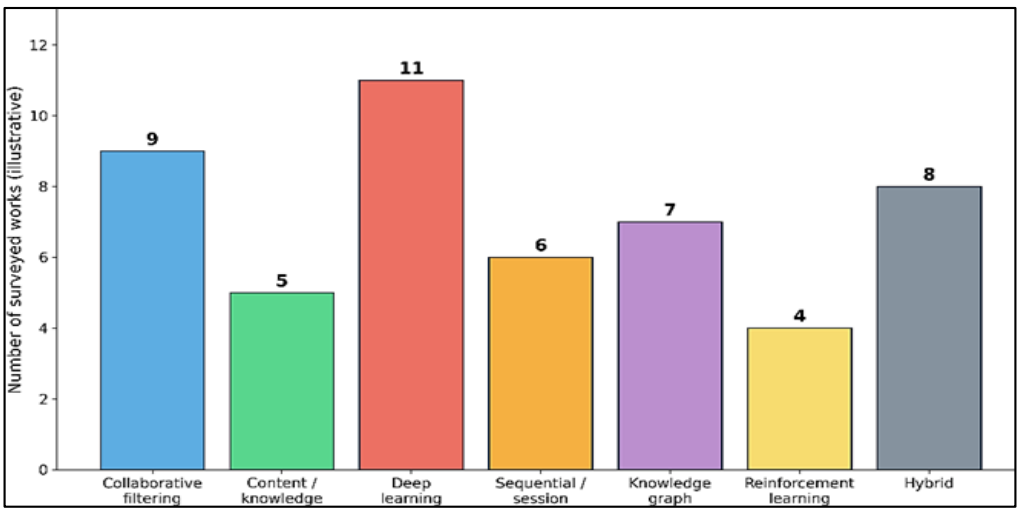


Fig. 3: Illustrative distribution of methodological families across the surveyed literature.

## VI. OPEN CHALLENGES

Cold-start and data sparsity persist as the most cited obstacles. Most learners touch only a few of many thousands of courses, so the interaction matrix is almost entirely empty, and new learners or courses arrive without history [8]. Content, knowledge-graph and cross-domain signals are the common remedies, though none fully resolves the problem. Scalability is a related concern. Platforms serve millions of learners, so candidate generation and ranking must stay efficient, which has favoured precomputable item similarities [7] and lightweight graph models [25].

Integrating dropout prediction with recommendation is an increasingly prominent challenge. A poorly matched course raises the risk of disengagement, so jointly modelling who is likely to leave and what to recommend offers a path to retention-aware systems [3], [4]. Explainability and trust matter acutely in education, where learners and instructors benefit from understanding why a course was suggested; knowledge-graph models give a structural basis for such explanations [26], [27]. Fairness and diversity raise a further risk. Popularity-driven recommenders can entrench advantage, under-serve niche subjects or disadvantage particular learner groups, which motivates explicit diversity and fairness objectives.

The deepest challenge is pedagogical alignment. General recommenders optimise relevance, but education requires respecting the learner's evolving knowledge state and recommending material within the zone of proximal development, neither trivially easy nor unreachably hard [31]. This connects recommendation to knowledge tracing, the modelling of learner mastery over time, from the classical Bayesian formulation of Corbett and Anderson [32] to Deep Knowledge Tracing [33]. A recommender that ignores mastery may suggest courses for which a learner is unprepared. Privacy is a final, growing constraint: behavioural learning data are sensitive, and centralised collection conflicts with data-protection expectations, which motivates privacy-preserving designs.

## VII. FUTURE DIRECTIONS

Several directions are gaining momentum, summarised on the evolutionary timeline in Fig. 4. The first is the use of large language models. Pretrained language models [34] and few-shot-capable large models [35] enable conversational, instruction-following course advisors. Such a system can interpret a learner's stated goals in natural language, reason over course descriptions, and justify its suggestions, moving recommendation from a ranked list toward an interactive dialogue. The second is knowledge-tracing-aware recommendation. An estimate of the learner's mastery from knowledge-tracing models [32], [33] conditions the recommender so that suggestions target the zone of proximal development [31] and advance a coherent learning path rather than maximising click-through alone.

A third direction is federated and privacy-preserving recommendation. It keeps sensitive behavioural data on learner devices or institutional servers and shares only model updates, aligning personalization with data-protection requirements. A fourth is the explicit treatment of fairness and diversity as first-class objectives, ensuring equitable exposure across subjects, providers and learner populations. Across all four, a unifying theme is the shift from accuracy-only optimisation toward learner-centred objectives that couple relevance with measured learning outcomes, and toward systems whose recommendations are transparent and accountable. Realising this agenda will require benchmarks that report learning-centric metrics alongside ranking accuracy, extending the repositories surveyed in Section IV.

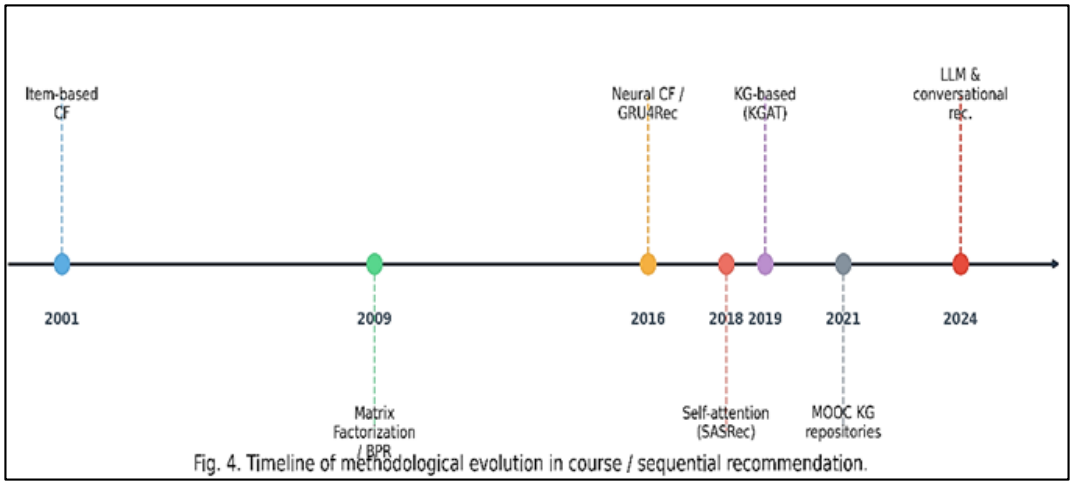


Fig. 4: Timeline of methodological evolution in course / sequential recommendation.

## VIII. CONCLUSION

MOOC platforms have made education abundantly available, but abundance without guidance leaves learners overloaded and prone to dropout. This review traced the development of course-recommendation systems for MOOCs. The path runs from memory-based collaborative filtering and matrix factorization, through content, knowledge-based, sequential, deep-learning, knowledge-graph and reinforcement-learning methods, to hybrid and concept-level designs. The survey covered the datasets and platforms, principally the MOOCCube family and XuetangX, that enable reproducible study, and the ranking and learning-centric metrics by which systems are judged. The synthesis of open challenges, cold-start, sparsity, dropout integration, scalability, explainability, fairness, pedagogical alignment and privacy, indicates a field maturing from a purely accuracy-driven discipline toward a learner-centred one. The most promising directions, large-language-model and conversational recommenders, knowledge-tracing-aware systems, federated and privacy-preserving recommendation, and fairness, point toward recommenders that not only predict what a learner will click but help a learner genuinely progress. This consolidated reference is offered in support of that goal.

## REFERENCES


[1] J. Yu, G. Luo, T. Xiao, Q. Zhong, Y. Wang, W. Feng, J. Luo, C. Wang, L. Hou, J. Li, Z. Liu, and J. Tang, "MOOCCube: A large-scale data repository for NLP applications in MOOCs," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 3135–3142.

[2] X. Jing and J. Tang, "Guess you like: Course recommendation in MOOCs," in *Proc. Int. Conf. Web Intelligence (WI)*, 2017, pp. 783–789.

[3] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods," in *Proc. EMNLP Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*, 2014, pp. 60–65.

[4] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in MOOCs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 517–524, 2019.

[5] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web (WWW)*, 2001, pp. 285–295.

[8] I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proc. 25th ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2002, pp. 253–260.

[9] Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[10] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 346–353, 2019.

[11] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell. (UAI)*, 2009, pp. 452–461.

[12] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2010, pp. 995–1000.

[13] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2011, pp. 73–105.

[14] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web*. Berlin, Germany: Springer, 2007, pp. 325–341.

[15] L. Pan, C. Li, J. Li, and J. Tang, "Prerequisite relation learning for concepts in MOOCs," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2017, pp. 1447–1456.

[16] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2018, pp. 197–206.

[17] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2019, pp. 1441–1450.

[18] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–38, 2019.

[19] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 173–182.

[20] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst. (RecSys)*, 2016, pp. 191–198.

[21] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.

[22] J. Zhang, B. Hao, B. Chen, C. Li, H. Chen, and J. Sun, "Hierarchical reinforcement learning for course recommendation in MOOCs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 435–442, 2019.

[23] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[24] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2019, pp. 165–174.
[25] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2020, pp. 639–648.
[26] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "KGAT: Knowledge graph attention network for recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2019, pp. 950–958.
[27] J. Gong, S. Wang, J. Wang, W. Feng, H. Peng, J. Tang, and P. S. Yu, "Attentional graph convolutional networks for knowledge concept recommendation in MOOCs in a heterogeneous view," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2020, pp. 79–88.
[28] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: A survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, 2023.
[29] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Model. User-Adapt. Interact.*, vol. 12, no. 4, pp. 331–370, 2002.
[30] J. Yu, Y. Wang, Q. Zhong, G. Luo, Y. Mao, K. Sun, W. Feng, W. Xu, S. Cao, K. Zeng, Z. Yao, L. Hou, Y. Lin, P. Li, J. Zhou, B. Xu, J. Li, J. Tang, and M. Sun, "MOOCCubeX: A large knowledge-centered repository for adaptive learning in MOOCs," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2021, pp. 4643–4652.
[31] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA, USA: Harvard Univ. Press, 1978.
[32] T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994.
[33] Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 505–513.
[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.
[35] T. B. Brown, B. Mann, N. Ryder, *et al*., "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209

# Post-Quantum Cryptography Migration for Enterprise Security

Mini T V

Associate Professor, Department of Computer Science, Sacred Heart College (Autonomous), Chalakudy, India

**Article information**

Received: 3rd April 2026
Received in revised form: 8th May 2026
Accepted: 10th June 2026
Available online: 30th July 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0032

**Abstract**

Large-scale quantum computers threaten the public-key cryptography that protects enterprise data, communications, and digital identity. Shor's algorithm solves integer factorization and discrete logarithms in polynomial time, which would break RSA, Diffie-Hellman, and elliptic-curve schemes once a cryptographically relevant quantum computer exists. The danger is not only future. Adversaries can record encrypted traffic today and decrypt it later, a tactic known as harvest-now-decrypt-later. In August 2024 the U.S. National Institute of Standards and Technology published its first post-quantum standards: FIPS 203 (ML-KEM, derived from CRYSTALS-Kyber), FIPS 204 (ML-DSA, from CRYSTALS-Dilithium), and FIPS 205 (SLH-DSA, from SPHINCS+). This paper presents a practical migration framework for enterprises. It reviews the quantum threat and Mosca's inequality for timing the transition, summarizes the new standards with their key and signature sizes, and proposes a five-phase roadmap built on crypto-agility: discovery and inventory, risk prioritization, agility engineering, hybrid deployment, and validation. Performance trade-offs are analyzed using documented parameter sizes and illustrative TLS handshake figures. ML-KEM-768 carries a 1,184-byte public key against 64 bytes for an elliptic-curve key, while SPHINCS+ signatures can exceed 17 kilobytes. The work does not report a real deployment. It consolidates published parameters and field guidance into an actionable plan, and it highlights open challenges in certificate sizing, hardware support, and long-lived embedded systems.

**Keywords**:- Post-Quantum Cryptography, ML-KEM, ML-DSA, SPHINCS+, Crypto-Agility, Hybrid TLS, Shor's Algorithm, Enterprise Security Migration.

.

## I. INTRODUCTION

Public-key cryptography underpins almost every secure digital interaction. RSA, Diffie-Hellman, and elliptic-curve cryptography (ECC) protect web sessions, virtual private networks, software updates, email, and the certificates that bind identities to keys [1], [2]. Their security rests on mathematical problems that are hard for classical computers: factoring large integers and computing discrete logarithms. A sufficiently powerful quantum computer changes that assumption.

In 1994 Peter Shor described a quantum algorithm that factors integers and solves discrete logarithms in polynomial time [3], [4]. Run on a cryptographically relevant quantum computer (CRQC), it would defeat RSA-2048 and ECC P-256 outright. No such machine exists today. Resource estimates suggest that factoring a 2,048-bit RSA modulus could require on the order of twenty million noisy physical qubits running for hours [5], far beyond current hardware. The timeline is uncertain. The trajectory of quantum engineering is steady, though, and the cost of being unprepared is high.

Two factors make the threat urgent rather than distant. The first is harvest-now-decrypt-later: an adversary captures and stores encrypted traffic now, then decrypts it once a CRQC becomes available [6], [7]. Any data with a confidentiality lifetime measured in years, such as health records, financial archives, state secrets, and intellectual property, is already at risk. The second is the long replacement cycle of cryptography itself. Embedded devices, industrial controllers, and certificate hierarchies can stay in service for a decade or more. Migrating them takes time.

Mosca framed the question precisely [6]. Let X be the number of years data must stay confidential, Y the years needed to migrate systems to quantum-resistant cryptography, and Z the years until a CRQC arrives. If X plus Y exceeds Z, the organization is already exposed. For many enterprises X and Y together span fifteen years or more, which leaves little margin even under optimistic estimates of Z. Migration should begin before the machine exists, not after.

This paper offers enterprise security teams a structured path. Section II reviews the quantum threat and Shor's algorithm. Section III summarizes the 2024 NIST post-quantum standards and their parameters. Section IV presents a five-phase migration roadmap centered on crypto-agility. Section V analyzes performance trade-offs using published sizes and illustrative handshake figures. Section VI discusses open challenges, and Section VII concludes. The work does not describe a live deployment. It assembles verified parameters and standards-body guidance into a plan that organizations can adapt.

## II. QUANTUM THREAT AND SHOR'S ALGORITHM

### A. Why Public-Key Schemes Break

RSA security depends on the difficulty of factoring a product of two large primes. ECC and Diffie-Hellman depend on the hardness of the discrete logarithm problem in finite groups. Classically, the best known algorithms for both run in sub-exponential or exponential time, which keeps 2,048-bit RSA and 256-bit elliptic curves safe in practice [1], [2]. Shor's algorithm collapses these problems to polynomial time on a quantum computer by using the quantum Fourier transform to find the period of a modular function [3], [4]. Once that period is known, the factors or the discrete logarithm follow directly. The consequence is stark. A working CRQC retroactively breaks every RSA and ECC key it can obtain, including keys already in use.

### B. Symmetric Cryptography and Grover's Algorithm

Symmetric ciphers and hash functions fare much better. Grover's search algorithm gives a quadratic speedup against an n-bit key, reducing brute-force effort from 2^n to roughly 2^(n/2) [8]. The fix is simple: double the key length. AES-256 retains about 128 bits of effective security against a quantum attacker, and SHA-384 or SHA-512 remain sound. The hard problem is therefore confined to public-key primitives, which is exactly where the new standards focus [9], [10].

### C. Estimating the Timeline

Predicting when a CRQC will exist is inherently uncertain. Public roadmaps from hardware vendors and expert surveys place a meaningful probability on the 2030s [6], [11]. Logical qubits must be assembled from many error-corrected physical qubits, and a factoring attack on RSA-2048 is estimated to need millions of physical qubits sustained over hours of coherent operation [5]. Security planning should not hinge on a single forecast. Because migration is slow and data lifetimes are long, the prudent posture is to act now and treat the arrival date as a risk variable rather than a fixed deadline.

## III. NIST POST-QUANTUM CRYPTOGRAPHY STANDARDS

NIST ran an open, multi-round evaluation from 2016 onward to select quantum-resistant algorithms [9], [12]. After analysis by the global cryptographic community, three standards were published on 13 August 2024. Each is a Federal Information Processing Standard derived from a finalist of the competition.

### A. FIPS 203: ML-KEM (Key Encapsulation)

FIPS 203 specifies the Module-Lattice-Based Key-Encapsulation Mechanism, ML-KEM, derived from CRYSTALS-Kyber [13], [14]. Its security rests on the Module Learning With Errors problem over structured lattices. ML-KEM replaces ECDH for establishing shared secrets. It defines three parameter sets: ML-KEM-512 at NIST security category 1, ML-KEM-768 at category 3, and ML-KEM-1024 at category 5. ML-KEM-768 is the common default for general use. Its public key is 1,184 bytes and its ciphertext is 1,088 bytes, against 32 to 64 bytes for an elliptic-curve key. Encapsulation and decapsulation are fast, often faster than ECDH per operation, so the cost is bandwidth rather than CPU time.

### B. FIPS 204: ML-DSA (Primary Signature)

FIPS 204 specifies the Module-Lattice-Based Digital Signature Algorithm, ML-DSA, derived from CRYSTALS-Dilithium [15], [17]. It is the recommended general-purpose signature scheme and also rests on module lattice problems. ML-DSA-65, the category 3 set, has a 1,952-byte public key and a 3,309-byte signature. Verification is efficient. The cost relative to ECDSA is again size: an ECDSA P-256 signature is 64 bytes, so the lattice signature is roughly fifty times larger.

### C. FIPS 205: SLH-DSA (Conservative Signature)

FIPS 205 specifies the Stateless Hash-Based Digital Signature Algorithm, SLH-DSA, derived from SPHINCS+ [16], [23]. Its security depends only on the strength of its underlying hash function, which makes it a conservative hedge. It does not rely on the relatively young hardness assumptions of structured lattices. The trade-off is size and speed. Public keys are tiny, just 32 bytes at the 128-bit level, but signatures are large and signing is slow. The small variant SLH-DSA-128s produces a 7,856-byte signature, and the fast variant SLH-DSA-128f produces a 17,088-byte signature. SLH-DSA suits low-frequency, high-assurance uses such as firmware and root certificate signing.

Table 1 lists the parameters of the three standards alongside classical baselines. Fig. 1 visualizes the same data, making the size gap between lattice schemes and classical keys immediately clear.

Table 1. Post-Quantum and Classical Parameter Sizes (*classical schemes are quantum-vulnerable, shown for comparison)

| Algorithm (Standard) | Type | NIST Cat. | Public Key (B) | Ciphertext/Sig (B) | Secret Key (B) |
|---|---|---|---|---|---|
| RSA-2048 (classical) | KEM/Sig | broken* | 256 | 256 | 256 |
| ECDH/ECDSA P-256 (classical) | KEM/Sig | broken* | 64 | 64 | 32 |
| ML-KEM-512 (FIPS 203) | KEM | 1 | 800 | 768 | 1,632 |
| ML-KEM-768 (FIPS 203) | KEM | 3 | 1,184 | 1,088 | 2,400 |
| ML-KEM-1024 (FIPS 203) | KEM | 5 | 1,568 | 1,568 | 3,168 |
| ML-DSA-44 (FIPS 204) | Sig | 2 | 1,312 | 2,420 | 2,560 |
| ML-DSA-65 (FIPS 204) | Sig | 3 | 1,952 | 3,309 | 4,032 |
| ML-DSA-87 (FIPS 204) | Sig | 5 | 2,592 | 4,627 | 4,896 |
| SLH-DSA-128s (FIPS 205) | Sig | 1 | 32 | 7,856 | 64 |
| SLH-DSA-128f (FIPS 205) | Sig | 1 | 32 | 17,088 | 64 |

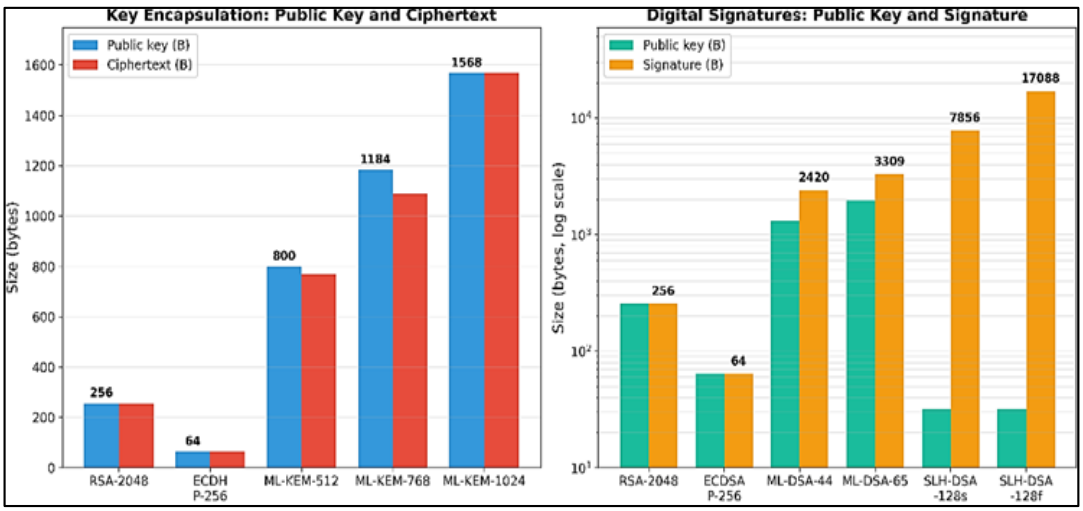


Fig. 1: Public-key, ciphertext, and signature sizes for classical and NIST post-quantum algorithms (signatures on a log scale).

## IV. MIGRATION ROADMAP AND CRYPTO-AGILITY

A successful transition is a program, not a single project. The U.S. National Cybersecurity Center of Excellence and national agencies recommend a structured, phased approach grounded in crypto-agility, the ability to change cryptographic algorithms quickly and with low disruption [7], [18]. Fig. 2 shows the five-phase roadmap proposed here. Fig. 4 shows the supporting architecture.

### A. Phase 1: Discovery and Inventory

An organization cannot protect what it cannot see. The first phase builds a Cryptographic Bill of Materials (CBOM): every place RSA, ECC, or Diffie-Hellman is used across applications, libraries, protocols, hardware security modules, and third-party services [7]. Automated scanning of code, network traffic, and certificate stores feeds this inventory. Each entry records the algorithm, key size, data sensitivity, and owner. The inventory becomes the foundation for every later decision.

### B. Phase 2: Risk Prioritization

Not everything migrates at once. Assets are ranked using the Mosca inequality from Section I: systems protecting long-lived secrets, or exposed to traffic capture, move first [6]. A short-lived session token matters far less than a twenty-year archive of patient records. Prioritization weighs data lifetime, exposure to harvest-now-decrypt-later, regulatory obligations, and the difficulty of changing each component. The output is a sequenced backlog.

### C. Phase 3: Crypto-Agility Engineering

Many systems hard-code algorithm choices deep in their logic, which makes any change expensive. This phase removes that rigidity. Cryptographic operations are placed behind a clean abstraction layer so the algorithm in use is a configuration choice, not a code rewrite [18], [25]. Certificate and key formats are reviewed for the larger objects that lattice schemes require. Done well, agility lets an enterprise swap a primitive again if a future weakness is found, which is valuable on its own merit.

### D. Phase 4: Hybrid Deployment

Few teams want to bet solely on young algorithms during a transition. The pragmatic answer is hybrid key establishment: combine a classical exchange such as X25519 with a post-quantum KEM such as ML-KEM-768, feeding both shared secrets through a key-derivation function [19], [20], [25]. The session stays secure as long as either component holds. This protects against both a CRQC and any undiscovered flaw in the new scheme. Industry trials of hybrid TLS by major browser and infrastructure providers confirmed it works at scale with modest overhead [22]. Deployment starts with internal pilots, then expands to external-facing services. Fig. 3 depicts the hybrid combiner inside a crypto-agile stack.

### E. Phase 5: Validation and Decommission

The final phase tests interoperability, monitors performance, and confirms that fallbacks behave correctly. Larger handshakes can interact badly with old middleboxes and fragmentation limits, so validation in realistic network conditions matters [20], [21]. Once post-quantum protection is confirmed, legacy quantum-vulnerable primitives are retired where policy allows. The CBOM is updated continuously, so the organization keeps an accurate, living picture of its cryptographic posture.

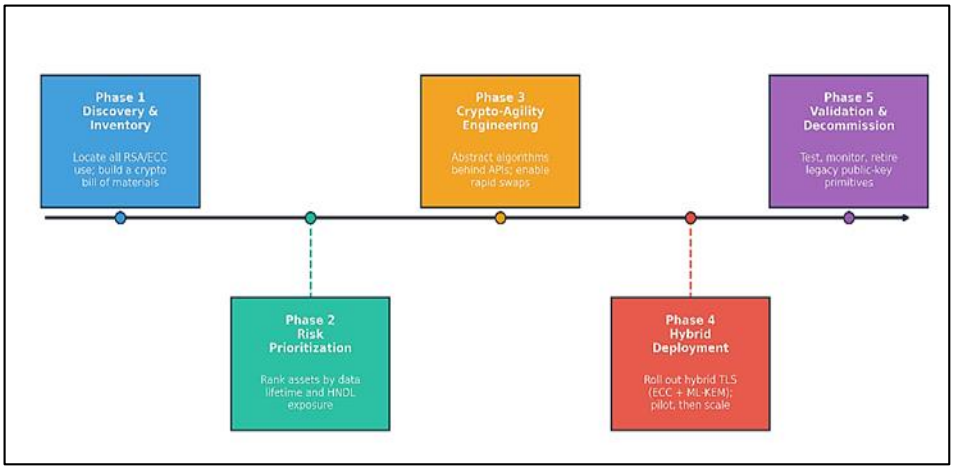


Fig. 2: Five-phase enterprise roadmap for migrating to post-quantum cryptography.

Table 2. Migration Phase Checklist

| Phase | Primary Activity | Key Artifact | Typical Owner |
|---|---|---|---|
| 1. Discovery | Inventory all public-key usage | Cryptographic Bill of Materials | Security architecture |
| 2.Prioritization | Rank assets by Mosca risk | Sequenced migration backlog | Risk and compliance |
| 3. Agility | Abstract algorithms behind APIs | Crypto-agility library/policy | Engineering |
| 4. Hybrid Deploy | Roll out hybrid TLS (ECC + ML-KEM) | Pilot and production configs | Platform/Ops |
| 5. Validation | Test, monitor, retire legacy | Interop report; updated CBOM | QA and Operations |

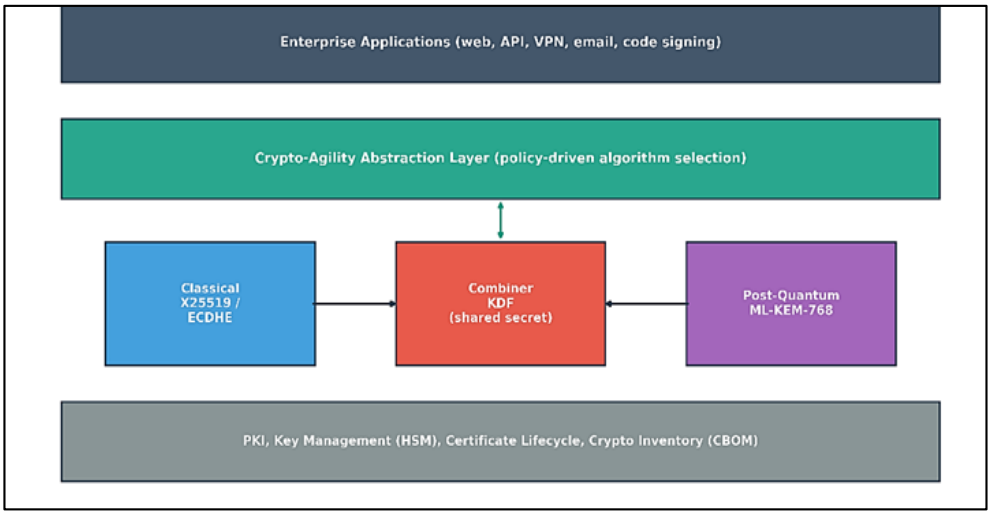


Fig. 3: Crypto-agile hybrid TLS architecture combining a classical and a post-quantum key exchange.

## V. PERFORMANCE ANALYSIS

### A. Bandwidth, the Dominant Cost

The headline trade-off of post-quantum migration is data volume, not computation. As Table 1 and Fig.1 show, lattice public keys and signatures are one to two orders of magnitude larger than their classical counterparts. An ECDH P-256 public key is 64 bytes. An ML-KEM-768 public key is 1,184 bytes and its ciphertext is 1,088 bytes. A 64-byte ECDSA signature becomes a 3,309-byte ML-DSA-65 signature, and a hash-based SLH-DSA signature can exceed 17 kilobytes. Where many signatures or certificates travel together, as in a TLS certificate chain, these increases add up quickly.

### B. Handshake Latency

Per-operation CPU cost for ML-KEM is competitive with, and sometimes better than, ECDH. The extra bytes are the main driver of any latency change, because a larger handshake may need additional packets [19], [20]. Fig. 4 presents illustrative TLS 1.3 handshake figures over a low-latency link, derived from published benchmark trends rather than a new study. A classical ECDHE handshake completes in roughly 3.1 ms. A hybrid X25519 plus ML-KEM-768 handshake takes about 3.9 ms while moving from roughly 1.4 KB to about 3.0 KB on the wire. On fast networks the overhead is small. On constrained or lossy links, and on low-power devices, the larger transcript can matter more, and engineering attention shifts to packet fragmentation and buffer sizes.

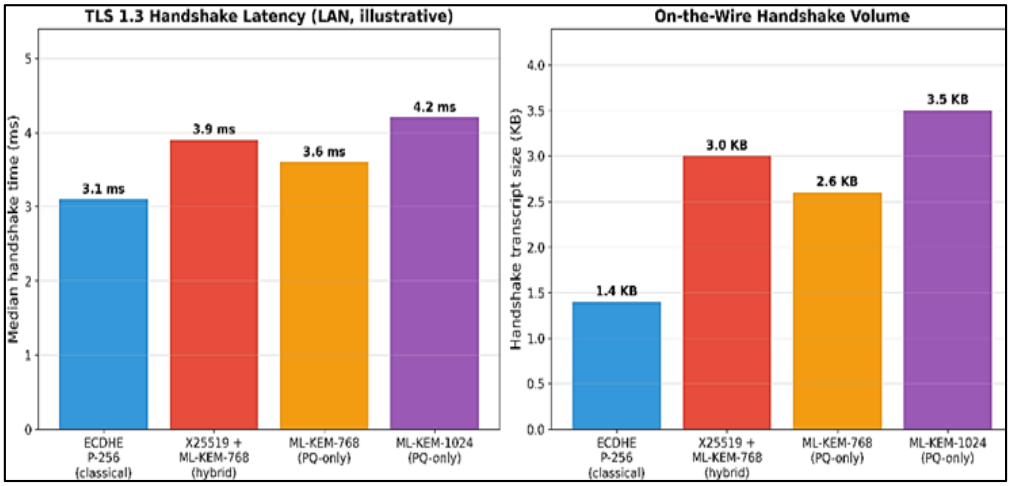

Fig. 4: Illustrative TLS 1.3 handshake latency and transcript size for classical, hybrid, and post-quantum configurations.

Two practical lessons follow. First, hybrid mode adds bytes but buys strong safety during the transition, which is usually worth the cost. Second, ML-KEM is a good fit for interactive protocols such as TLS, where its compact ciphertext and fast operations help, while large hash-based signatures are best reserved for infrequent, high-value signing. Matching the algorithm to the workload keeps the performance penalty manageable.

## VI. CHALLENGES AND OPEN ISSUES

### A. Certificate and Protocol Sizing

Larger keys and signatures inflate X.509 certificates and complete chains. Some protocols carry implicit size assumptions, and parts of the deployed network still mishandle handshakes that exceed familiar limits [21]. Certificate chain compression and careful protocol tuning help, but adapting the wider ecosystem of clients, servers, and intermediaries will take time.

### B. Hardware, Embedded, and Long-Lived Systems

Hardware security modules, smart cards, and Internet-of-Things devices often have fixed memory and slow processors. Storing kilobyte-scale signatures or running lattice arithmetic can strain them. Devices with a deployed life of ten to twenty years, common in industrial control and critical infrastructure, are especially hard to update in the field [2], [7]. Some will need full hardware replacement rather than a software patch.

### C. Algorithm Maturity and Defense in Depth

Lattice schemes are newer than RSA and have had fewer decades of public scrutiny. Standardization reflects strong confidence, yet history counsels humility about young assumptions [10], [23]. Two responses help. Hybrid deployment keeps a classical algorithm in the loop during the transition, and the conservative hash-based SLH-DSA offers a fallback whose security depends only on hash strength. Crypto-agility itself is the deeper insurance, because it lets an organization replace any primitive quickly if analysis later reveals a weakness.

### D. Governance and Compliance

National guidance is converging on firm timelines. The U.S. National Security Agency's Commercial National Security Algorithm Suite 2.0 sets dates for adopting ML-KEM and ML-DSA in national security systems, and it mandates SLH-DSA for firmware signing [24]. Enterprises in regulated sectors should expect comparable expectations and align their migration backlog with emerging requirements rather than waiting for a final mandate.

## VII. CONCLUSION

The quantum threat to public-key cryptography is credible, and harvest-now-decrypt-later makes it a present concern for any data with a long confidentiality horizon. The 2024 NIST standards, ML-KEM, ML-DSA, and SLH-DSA, give enterprises vetted, quantum-resistant tools. Their larger keys and signatures shift the main cost to bandwidth and storage rather than computation, and that cost is manageable when algorithms are matched to workloads.

Migration should start now. The five-phase roadmap presented here (discovery, prioritization, agility engineering, hybrid deployment, and validation) turns a daunting transition into a sequence of tractable steps. Crypto-agility is the central idea. It eases the present move to post-quantum schemes and prepares the organization

for whatever comes next. This paper consolidates verified parameters and standards guidance rather than reporting a deployment. Future work should measure hybrid TLS at scale across diverse networks, profile post-quantum primitives on constrained hardware, and refine automated discovery tooling for accurate cryptographic inventories.

## REFERENCES


[1] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, vol. 21, no. 2, pp. 120–126, Feb. 1978.

[2] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 6, pp. 644–654, Nov. 1976.

[3] P. W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in *Proc. 35th Annu. Symp. Found. Comput. Sci. (FOCS)*, Santa Fe, NM, USA, 1994, pp. 124–134.

[4] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM J. Comput.*, vol. 26, no. 5, pp. 1484–1509, Oct. 1997.

[5] C. Gidney and M. Ekerå, "How to factor 2048-bit RSA integers in 8 hours using 20 million noisy qubits," *Quantum*, vol. 5, Art. no. 433, Apr. 2021.

[6] M. Mosca, "Cybersecurity in an era with quantum computers: Will we be ready?," *IEEE Security Privacy*, vol. 16, no. 5, pp. 38–41, Sep./Oct. 2018.

[7] W. Barker, W. Polk, and M. Souppaya, *Getting Ready for Post-Quantum Cryptography: Exploring Challenges Associated with Adopting and Using Post-Quantum Cryptographic Algorithms*, NIST Cybersecurity White Paper, Apr. 2021.

[8] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proc. 28th Annu. ACM Symp. Theory Comput. (STOC)*, Philadelphia, PA, USA, 1996, pp. 212–219.

[9] L. Chen *et al*., *Report on Post-Quantum Cryptography*, Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, USA, NISTIR 8105, Apr. 2016.

[10] D. J. Bernstein and T. Lange, "Post-quantum cryptography," *Nature*, vol. 549, no. 7671, pp. 188–194, Sep. 2017.

[11] G. Alagic *et al*., *Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process*, Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, USA, NISTIR 8413, Sep. 2022.

[12] D. Moody *et al*., *Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process*, Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, USA, NISTIR 8309, Jul. 2020.

[13] National Institute of Standards and Technology, *Module-Lattice-Based Key-Encapsulation Mechanism Standard*, FIPS PUB 203, Gaithersburg, MD, USA, Aug. 2024.

[14] J. Bos *et al*., "CRYSTALS-Kyber: A CCA-secure module-lattice-based KEM," in *Proc. IEEE Eur. Symp. Security Privacy (EuroS&P)*, London, U.K., 2018, pp. 353–367.

[15] National Institute of Standards and Technology, *Module-Lattice-Based Digital Signature Standard*, FIPS PUB 204, Gaithersburg, MD, USA, Aug. 2024.

[16] National Institute of Standards and Technology, *Stateless Hash-Based Digital Signature Standard*, FIPS PUB 205, Gaithersburg, MD, USA, Aug. 2024.

[17] L. Ducas *et al*., "CRYSTALS-Dilithium: A lattice-based digital signature scheme," *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, vol. 2018, no. 1, pp. 238–268, Feb. 2018.

[18] D. Ott, C. Peikert, and other workshop participants, "Identifying research challenges in post-quantum cryptography migration and cryptographic agility," *arXiv* preprint arXiv:1909.07353, Sep. 2019.

[19] D. Sikeridis, P. Kampanakis, and M. Devetsikiotis, "Post-quantum authentication in TLS 1.3: A performance study," in *Proc. Netw. Distrib. Syst. Security Symp. (NDSS)*, San Diego, CA, USA, 2020, pp. 1–16.

[20] C. Paquin, D. Stebila, and G. Tamvada, "Benchmarking post-quantum cryptography in TLS," in *Proc. Int. Conf. Post-Quantum Cryptography (PQCrypto)*, Paris, France, 2020, pp. 72–91.

[21] P. Kampanakis, P. Panburana, E. Daw, and D. Van Geest, "The viability of post-quantum X.509 certificates," *IACR Cryptol. ePrint Arch.*, Paper 2018/063, 2018.

[22] K. Kwiatkowski and L. Valenta, "The TLS post-quantum experiment," *The Cloudflare Blog*, Oct. 2019.

[23] D. J. Bernstein, A. Hülsing, S. Kölbl, R. Niederhagen, J. Rijneveld, and P. Schwabe, "The SPHINCS+ signature framework," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, Toronto, ON, Canada, 2019, pp. 2129–2146.

[24] National Security Agency, *Announcing the Commercial National Security Algorithm Suite 2.0 (CNSA 2.0)*, Cybersecurity Advisory, Sep. 2022.

[25] D. Stebila and M. Mosca, "Post-quantum key exchange for the Internet and the Open Quantum Safe project," in *Proc. Int. Conf. Sel. Areas Cryptogr. (SAC)*, St. John's, NL, Canada, 2016, pp. 14–37.

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209

# Retrieval-Augmented Generation for Trustworthy Enterprise LLM Assistants

Bini P B

Assistant Professor, Department of Computer Science, CCSIT Dr John Matthai Center, Thrissur, India

**Article information**

Received: 13th April 2026
Received in revised form: 16th May 2026
Accepted: 18th June 2026
Available online: 30th July 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0033

**Abstract**

Large language models (LLMs) have changed enterprise knowledge work. Their value, however, is capped by three failures: they hallucinate, their parametric memory is frozen and grows stale, and they cannot read the proprietary data that holds most business answers. Retrieval-Augmented Generation (RAG) targets all three. It grounds generation in passages fetched at inference time from an external, continuously updatable corpus, so answers become verifiable and citation-backed without any model retraining. This paper presents a technical synthesis of RAG for trustworthy enterprise assistants. The end-to-end pipeline is described in full: document chunking, embedding, vector indexing, retrieval, cross-encoder re-ranking, and grounded generation with inline citations. Advanced variants are then surveyed, namely hybrid sparse-dense retrieval, Hypothetical Document Embeddings (HyDE), graph-based RAG, and agentic iterative retrieval. A RAGAS-style evaluation method quantifies faithfulness, answer relevance, and context precision and recall. On an illustrative enterprise question-answering scenario, an advanced configuration that combines hybrid retrieval with cross-encoder re-ranking lifts faithfulness from 0.71 to 0.91 and context precision from 0.62 to 0.84 over naive dense-only RAG. Agentic retrieval reaches 0.95 faithfulness, but pays for it in latency. Enterprise concerns, including document-level access control, data security, cost, and latency budgets, are treated as first-class design constraints. The reported metrics are illustrative. They characterise representative trade-offs rather than a specific deployed study.

**Keywords**:- Retrieval-Augmented Generation, Large Language Models, Hallucination, Dense Retrieval, Vector Database, Re-ranking, Faithfulness, Enterprise AI.

## I. INTRODUCTION

Large language models (LLMs) such as the GPT and LLaMA families read and write text with remarkable fluency. They summarize, answer questions, and reason across long documents, and they now sit behind many enterprise assistants [1], [2]. Fluency is not accuracy. Three structural limits hold these models back in serious business use. First, they hallucinate, asserting confident statements that no source supports [3]. Second, their knowledge is baked into the weights at training time and goes stale as products, policies, and regulations change. Third, they cannot see the proprietary material, the internal wikis, support tickets, contracts, and engineering specifications, where most enterprise answers actually live.Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [4], attacks all three limits at once. It pairs a parametric generator with a non-parametric, searchable knowledge store. At inference time the user query retrieves relevant passages from an external corpus, and those passages are handed to the LLM as grounding context. Knowledge lives outside the weights. It can therefore be refreshed continuously without retraining, kept private inside an organisational boundary, and attributed through

citations so users can check each claim [5]. Separating reasoning from knowledge is exactly what an enterprise assistant needs to be both current and accountable.

This paper gives a technical synthesis of RAG for trustworthy enterprise LLM assistants. The contributions are fourfold:

- A structured account of the end-to-end RAG pipeline and its design choices;
- A survey of advanced retrieval and generation variants, namely hybrid sparse-dense retrieval, HyDE, graph-based RAG, and agentic iterative RAG;
- A RAGAS-style evaluation method with illustrative results that characterise the faithfulness and relevance of each configuration; and
- An analysis of enterprise deployment concerns such as access control, security, cost, and latency.

All quantitative figures are illustrative. They convey representative trade-offs, not results from a specific production deployment.

## II. BACKGROUND AND RELATED WORK

### A. LLMs and the Hallucination Problem

Modern LLMs are built on the Transformer architecture [6], which uses self-attention to model long-range dependencies, and are trained on web-scale text by next-token prediction followed by instruction tuning and alignment from human feedback [7]. The weights hold a vast but lossy compression of that training data. Ask about facts that are rare, recent, or simply absent, and the model interpolates a plausible answer instead of abstaining. That is a hallucination. Ji et al. [3] survey the phenomenon across generation tasks and separate intrinsic hallucinations, which contradict the supplied source, from extrinsic ones, which cannot be checked against any source. For an assistant handling financial, legal, or medical content, neither is acceptable.

### B. Retrieval and Knowledge-Grounded Generation

Augmenting models with retrieval predates instruction-tuned LLMs. REALM [8] trained a retriever and a masked language model jointly, so retrieval was optimised end to end for downstream accuracy. Dense Passage Retrieval (DPR) [9] then showed that a dual-encoder trained with contrastive learning beats sparse lexical baselines on open-domain question answering. Fusion-in-Decoder [10] demonstrated that a generative reader conditioned on many passages scales gracefully with the passage count. Lewis et al. [4] unified retrieval and generation under the RAG name, marginalising the generator over retrieved documents, and set the template that current systems extend.

### C. The Modern RAG Landscape

Gao et al. [5] survey the field and organise it into naive, advanced, and modular paradigms, cataloguing the optimisation points across indexing, retrieval, and generation. The practical appeal is direct: RAG converts the open-ended goal of factual accuracy into a more tractable retrieval-and-grounding problem while leaving the underlying LLM frozen. That economics suits enterprises well. A general-purpose model can be wired to a private corpus instead of fine-tuning or pre-training a bespoke model [4], [5].

## III. RAG PIPELINE ARCHITECTURE

A production RAG system has two paths. An offline indexing path ingests enterprise documents and builds a searchable index. An online query path answers user queries by retrieving and grounding. Fig. 1 shows both. The subsections below walk through each component.

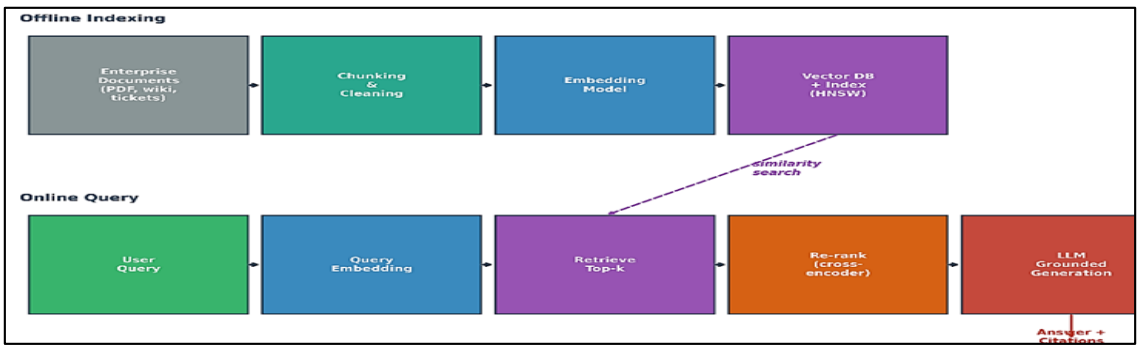


Fig. 1: End-to-end RAG pipeline showing the offline indexing path (chunking, embedding, vector indexing) and the online query path (retrieval, re-ranking, grounded generation with citations).

### A. Chunking and Embedding

Source documents are split into chunks. A chunk must be small enough to fit the retriever and generator budget yet large enough to stay semantically coherent. Common strategies include fixed-size token windows with overlap, recursive splitting on document structure such as headings and paragraphs, and semantic chunking that groups sentences by embedding similarity. Each chunk is mapped to a dense vector by an embedding model, usually a sentence-level bi-encoder such as Sentence-BERT [11], which places similar text near itself in vector space. Chunk size, overlap, and metadata (source, section, access tags) materially affect retrieval quality and are key tuning parameters.

### B. Vector Indexing and Retrieval

Chunk embeddings are stored in a vector database that supports approximate nearest-neighbour (ANN) search. FAISS [12] provides GPU-accelerated similarity search over billions of vectors, while graph-based indices such as HNSW [13] give logarithmic-time search at high recall. At query time the question is embedded with the same model, and the top-k most similar chunks are returned. Dense retrieval captures semantic similarity but can miss exact-match signals such as product codes or error identifiers. Sparse lexical retrieval such as BM25 [14] excels at those exact matches yet ignores synonymy. This complementarity is what motivates the hybrid retrieval of Section IV.

### C. Re-ranking

First-stage retrieval optimises recall over a large corpus, so it returns some loosely relevant passages. A second-stage re-ranker raises precision by scoring each query-passage pair jointly. The bi-encoder embeds query and passage independently. A cross-encoder re-ranker based on BERT [15], [16] instead attends over the concatenated pair, yielding far more accurate relevance scores at higher compute cost. ColBERT [17] sits between the two through late interaction over token-level embeddings. Re-ranking the top 50 to 100 candidates down to the top three to five passages is one of the highest-impact moves for faithfulness.

### D. Grounded Generation and Citation

The re-ranked passages are placed into a prompt template that tells the LLM to answer using only the supplied context and to attribute every claim to its source passage. Grounding the generator in retrieved evidence sharply cuts hallucination, and inline citations let users verify answers, which builds trust in the assistant [4], [5]. A good prompt also instructs the model to abstain, replying that the answer is not in the knowledge base, when retrieved context is insufficient. In an enterprise setting, an honest abstention beats a confident fabrication.

## IV. ADVANCED RAG TECHNIQUES

Naive RAG, dense retrieval followed by single-shot generation, leaves accuracy on the table. Table 1 summarises advanced techniques that strengthen retrieval and reasoning, and Fig. 2 contrasts naive and advanced configurations across the core metrics.

### A. Hybrid Sparse-Dense Retrieval

Hybrid retrieval fuses lexical BM25 [14] scores with dense bi-encoder [9] similarities, usually through reciprocal rank fusion or a weighted score combination. The fusion captures exact-match terms and semantic paraphrase together. In enterprise corpora thick with identifiers, acronyms, and code, it is among the most dependable improvements over dense-only retrieval [5].

### B. Query Transformation and HyDE

User queries are often short and underspecified next to the documents that answer them. Hypothetical Document Embeddings (HyDE) [18] close that gap. The LLM drafts a hypothetical answer to the query, and that synthetic document is embedded to retrieve real passages with similar content. Related query-transformation methods rewrite, expand, or decompose the query before retrieval, shrinking the lexical and semantic distance between question and evidence.

### C. Graph-based RAG

Flat passage retrieval struggles when an answer is scattered across many documents. GraphRAG [19] builds a knowledge graph of entities and relationships extracted from the corpus, then summarises communities within that graph to answer global, query-focused questions. Traversing explicit relationships supports multi-hop reasoning and corpus-level sensemaking that similarity search alone cannot deliver.

### D. Agentic and Iterative RAG

Agentic RAG treats retrieval as an action the model can invoke repeatedly inside a reasoning loop. Self-RAG [20] trains the model to decide when to retrieve and to critique the relevance and support of retrieved passages using reflection tokens. IRCoT [21] interleaves retrieval with chain-of-thought reasoning, so each reasoning step issues a fresher, sharper query. FLARE [22] triggers retrieval whenever the model is about to generate low-confidence content. These iterative methods reach the highest faithfulness on complex multi-hop questions. The price is extra retrieval rounds and added latency.

Table 1. Comparison of RAG Variants

| RAG Variant | Key Mechanism | Strength | Primary Cost |
|---|---|---|---|
| Naive (dense) | Bi-encoder top-k + single-shot generation | Simple, low latency | Misses exact match; weak precision |
| Hybrid | BM25 + dense fusion | Exact match + semantic recall | Fusion tuning |
| + Re-ranking | Cross-encoder re-scoring | High context precision | Extra inference per passage |
| HyDE | LLM drafts hypothetical doc to embed | Bridges short-query gap | One extra LLM call |
| GraphRAG | Entity graph + community summaries | Multi-hop, global queries | Graph construction cost |
| Agentic / Iterative | Retrieve-reason-critique loop | Highest faithfulness | Multiple rounds, latency |

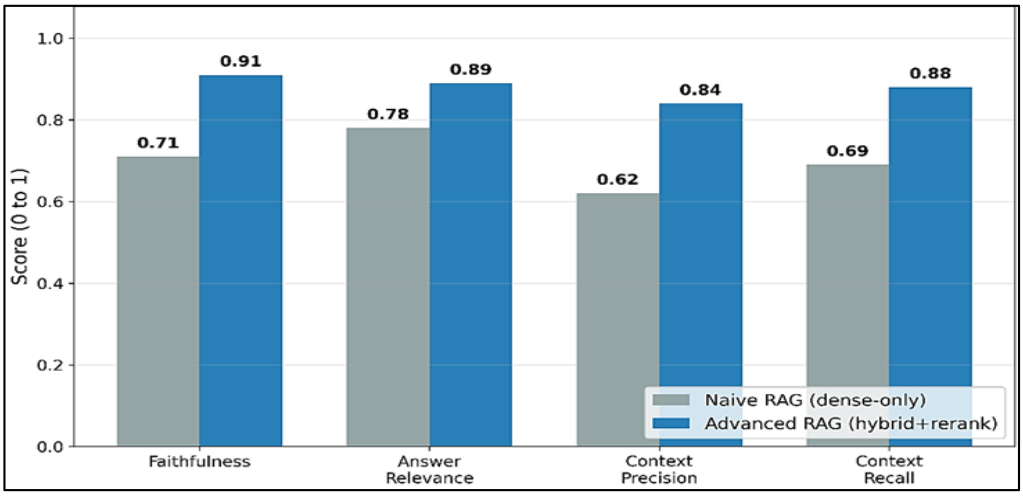


Fig. 2: Naive (dense-only) versus advanced (hybrid retrieval with cross-encoder re-ranking) RAG on RAGAS-style metrics (illustrative).

## V. EVALUATION METHODOLOGY AND RESULTS

### A. Metrics

A RAG system has to be judged on two axes, retrieval and generation. Following the RAGAS framework [23], four reference-light metrics are used. Faithfulness, or groundedness, is the fraction of claims in the answer that the retrieved context entails; it measures hallucination directly. Answer relevance captures how well the answer addresses the question. Context precision is the share of retrieved passages that are actually relevant, and context recall is the share of the information needed to answer that retrieval found. Each can be estimated with an LLM-as-judge, which avoids large hand-labelled reference sets [23].

### B. Experimental Configurations

Six configurations are compared on an illustrative enterprise question-answering scenario over a mixed corpus of policy documents, support tickets, and technical manuals. They are an ungrounded LLM-only baseline, dense RAG, hybrid RAG, hybrid RAG with cross-encoder re-ranking, that configuration plus HyDE, and an agentic iterative variant. Retrieval uses an HNSW index [13]; re-ranking uses a cross-encoder [16]; faithfulness

and relevance are scored by an LLM judge [23]. The reported values are illustrative, chosen to reflect trends that recur in the literature rather than a single measured deployment.

Fig. 3 plots faithfulness and answer relevance across the six configurations. The ungrounded baseline is least faithful at 0.52, confirming that an unaided LLM often asserts unsupported claims. Dense retrieval lifts faithfulness to 0.71, hybrid retrieval to 0.79, and cross-encoder re-ranking to 0.91. The lesson is blunt: the precision of the supplied context, not merely its presence, governs groundedness. HyDE and agentic retrieval add further gains, to 0.93 and 0.95. Table 2 reports the full metric set.

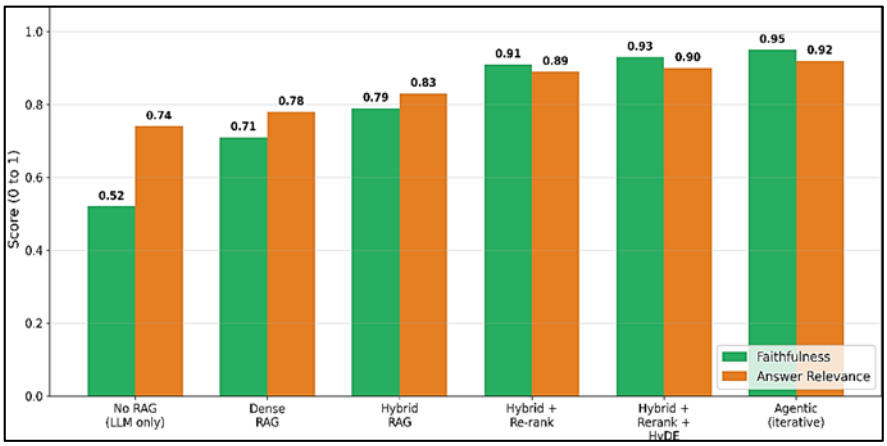


Fig. 3: Faithfulness and answer relevance across six RAG configurations (illustrative).

Table 2. RAGAS-style Evaluation Across Configurations (Illustrative)

| Configuration | Faithfulness | Answer Relevance | Context Precision | Context Recall |
|---|---|---|---|---|
| LLM only (no RAG) | 0.52 | 0.74 | n/a | n/a |
| Dense RAG | 0.71 | 0.78 | 0.62 | 0.69 |
| Hybrid RAG | 0.79 | 0.83 | 0.71 | 0.81 |
| Hybrid + Re-rank | 0.91 | 0.89 | 0.84 | 0.88 |
| Hybrid + Rerank + HyDE | 0.93 | 0.90 | 0.86 | 0.90 |
| Agentic (iterative) | 0.95 | 0.92 | 0.88 | 0.93 |

### C. Retrieval Depth and Latency Trade-offs

Fig. 4(a) shows how precision and recall move with the number of retrieved passages k. Precision falls and recall climbs as k grows, and hybrid retrieval dominates dense-only retrieval at every depth. The generator context window and the re-ranker cost both scale with k. Practical systems therefore retrieve a wide candidate set, re-rank it, and present only the top few passages. Fig. 4(b) plots faithfulness against end-to-end latency. Re-ranking buys a large faithfulness gain for modest added latency. The agentic loop's extra faithfulness, by contrast, carries a steep latency premium. That tension is the central engineering trade-off an enterprise has to budget for.

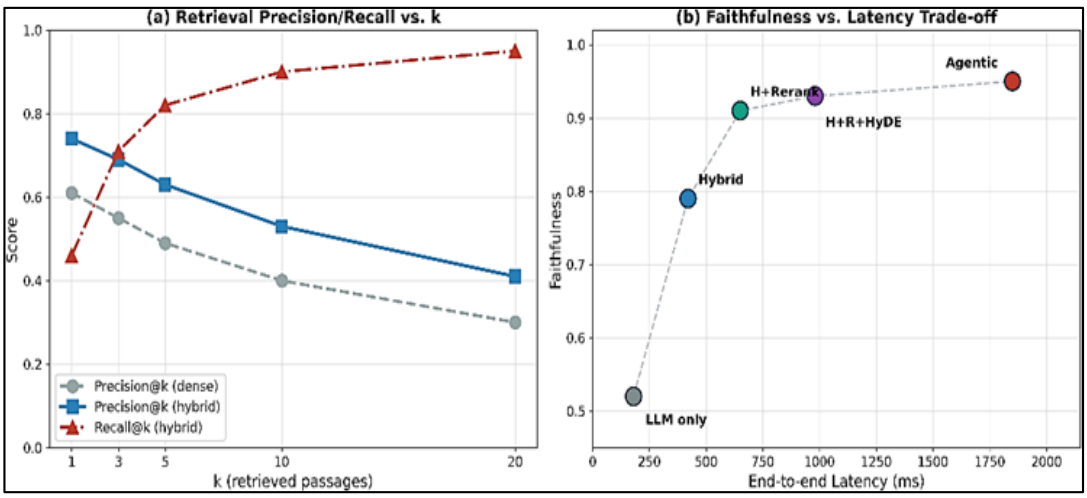


Fig. 4: (a) Retrieval precision and recall versus k for dense and hybrid retrieval; (b) faithfulness versus end-to-end latency across configurations (illustrative).

## VI. ENTERPRISE DEPLOYMENT CHALLENGES

### A. Access Control and Data Security

Enterprise corpora hold documents with mixed sensitivity and per-user access rights. A correct RAG system enforces those rights at retrieval time, filtering the candidate set to documents the requesting user is authorised to see before any passage reaches the LLM. Access tags stored as chunk metadata make such pre-retrieval filtering possible. Skip this step and a serious leakage channel opens, because the generator will faithfully surface whatever it is handed. Sensitive data must also be protected in transit and at rest, and personally identifiable information may need redaction during ingestion [5].

### B. Cost and Latency

Every advanced component adds compute. Embedding and re-ranking cost per-query inference, HyDE and agentic loops add extra LLM calls, and a larger k inflates the generation context and the token bill. Fig. 4(b) makes the implication plain. Teams should pick the cheapest configuration that hits their faithfulness target, not chase accuracy without limit. Caching frequent queries and their retrieved contexts, distilling re-rankers, and right-sizing k are the usual levers for staying inside latency and cost budgets.

### C. Maintenance and Governance

RAG knowledge lives outside the model, so the index has to stay synchronised with source systems as documents are created, updated, and deleted. Stale or deleted content that lingers in the index reintroduces the very staleness RAG was meant to cure. Continuous evaluation with the metrics of Section V, human review of low-confidence answers, and audit logs that record which passages grounded each response are all needed for governance and regulatory compliance [23].

## VII. CONCLUSION

Retrieval-Augmented Generation has become the default architecture for trustworthy enterprise LLM assistants because it grounds generation in an external, updatable, access-controlled corpus and supports verifiable citations. This paper synthesised the end-to-end RAG pipeline, surveyed advanced variants (hybrid retrieval, HyDE, GraphRAG, and agentic iterative RAG), and described a RAGAS-style evaluation method. One finding stands out across the illustrative results. Context precision is the dominant lever for faithfulness: hybrid retrieval with cross-encoder re-ranking raises faithfulness from 0.71 to 0.91 over naive dense RAG, while agentic retrieval reaches 0.95 at a real latency cost [4], [5], [23].

The practical message for enterprises is simple. Retrieval quality, access control, and latency budgets are first-class design constraints, not afterthoughts, and the right configuration is the cheapest one that meets a target faithfulness threshold. Future directions include multimodal RAG over tables and figures, end-to-end optimisation of retrievers against generation quality, stronger automatic faithfulness verifiers, and tighter integration of structured knowledge graphs with dense retrieval for complex multi-hop enterprise reasoning [5], [19], [20].

## REFERENCES


[1] T. B. Brown *et al*., "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[2] H. Touvron *et al*., "LLaMA: Open and efficient foundation language models," *arXiv* preprint arXiv:2302.13971, 2023.

[3] Z. Ji *et al*., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, pp. 1–38, Mar. 2023.

[4] P. Lewis *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[5] Y. Gao *et al*., "Retrieval-augmented generation for large language models: A survey," *arXiv* preprint arXiv:2312.10997, 2023.

A. Vaswani *et al*., "Attention Is All You Need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[6] L. Ouyang *et al*., "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 27730–27744, 2022.

[7] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 3929–3938.

[8] V. Karpukhin *et al*., "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6769–6781.

[9] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2021, pp. 874–880.

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[11] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[12] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020.
[13] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.
[15] R. Nogueira and K. Cho, "Passage re-ranking with BERT," *arXiv* preprint arXiv:1901.04085, 2019.
[16] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2020, pp. 39–48.
[17] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2023, pp. 1762–1777.
[18] D. Edge *et al*., "From local to global: A graph RAG approach to query-focused summarization," *arXiv* preprint arXiv:2404.16130, 2024.
A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
[19] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2023, pp. 10014–10037.
[20] Z. Jiang *et al*., "Active retrieval augmented generation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023, pp. 7969–7992.
[21] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval-augmented generation," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL): Syst. Demonstrations*, 2024, pp. 150–158.

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209

# Self-Service Business Intelligence and Analytics in Digital Transformation

Meena Jose Komban

Assistant Professor, Department of Computer Science, YuvakshatraInstitute of Management Studies (YIMS), Mundur, Kerala, India

**Article information**

Received: 20th April 2026
Received in revised form: 22nd May 2026
Accepted: 24th June 2026
Available online: 30th July 2026

Volume: 2
Issue: 3
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0034

**Abstract**

The digital-transformation era has changed how organizations generate, govern, and consume analytical insight. This paper analyzes the shift from centralized, IT-centric reporting toward self-service business intelligence (BI), in which business users author governed analyses without continuous engineering intervention. The discussion traces the architectural move from on-premises data warehouses to cloud data warehouses and lakehouse platforms that separate compute from storage and unify structured and semi-structured data under open table formats. A layered reference architecture is proposed. It comprises integration, storage and processing, augmented analytics, semantic and governance, and consumption layers. The paper then examines augmented analytics and the use of large language models for natural-language querying and natural-language-to-SQL translation. It pairs these capabilities with the governance controls that make self-service trustworthy at scale: semantic metric layers, row-level security, data catalogs, and lineage. A structured comparison of four leading platforms, namely Microsoft Power BI, Tableau, Google Looker, and Qlik, is presented across authoring, semantic modeling, cloud scalability, augmented analytics, and governance dimensions using illustrative capability scores. A five-level maturity model is introduced to help organizations benchmark their analytical capability and sequence investment. The analysis concludes that durable value comes not from tool selection alone but from coupling self-service freedom with a governed semantic foundation that treats data as a managed product. The metrics reported here are illustrative and synthesized from public industry sources, not from a single deployed study.

**Keywords**:- Self-Service Business Intelligence, Digital Transformation, Cloud Data Warehouse, Lakehouse, Augmented Analytics, Natural Language to SQL, Data Governance, Semantic Layer, Analytics Maturity.

.

## I. INTRODUCTION

Digital transformation has recast data. What was once a passive by product of operations is now a strategic asset that shapes competitive position across nearly every sector. As organizations digitize customer journeys, supply chains, and internal processes, the volume, velocity, and variety of available data have grown well beyond what centrally produced reports can serve [1]. A coherent transformation strategy must specify not only which technologies to adopt but also how decision-making itself becomes data-driven, and it must avoid common traps such as technology-first initiatives disconnected from measurable outcomes [2]. Business intelligence, the set of technologies and practices for collecting, integrating, analyzing, and presenting business information, sits at the center of this shift [3].

For much of its history, BI was an IT-centric discipline. Specialists modeled data into enterprise warehouses, and a central team produced reports in response to formal requests [4], [5]. The model delivered consistency and control, but it created a structural bottleneck. Business users waited days or weeks for new reports. The backlog of analytical requests grew faster than central teams could clear it. A persistent gap opened between the questions decision-makers needed answered and the speed at which the organization could answer them [6].

Self-service BI emerged as a response to this bottleneck. By equipping business analysts with governed datasets, intuitive visual authoring tools, and curated semantic models, organizations let domain experts explore data and build their own analyses without continuous engineering support [7]. Three forces accelerated the transition: the migration of analytical workloads to elastic cloud platforms, the maturation of augmented analytics that automates parts of insight generation, and, most recently, large language models (LLMs) that let users query data in plain language [8]. Self-service freedom without governance, however, reintroduces the inconsistency that centralized BI was built to prevent. The central technical problem of the modern era is to reconcile autonomy with trust.

This paper offers a technical analysis of self-service BI in the digital-transformation era. Its contributions are fourfold. First, it synthesizes the architectural evolution from data warehouses to cloud lakehouses. Second, it proposes a layered reference architecture that positions augmented analytics and governance as first-class layers. Third, it presents a structured, multi-dimensional comparison of four leading platforms, Power BI, Tableau, Looker, and Qlik, drawing on a recent technical vendor analysis [9]. Fourth, it introduces a five-level maturity model with associated governance and adoption guidance. All quantitative figures are illustrative and synthesized from public industry sources; the paper does not report a single deployed empirical study.

## II. EVOLUTION OF BI AND RELATED WORK

### A. From Decision Support to Enterprise Data Warehousing

The intellectual roots of BI lie in decision-support systems and, later, in the data-warehousing movement of the 1990s. Inmon described the corporate information factory, in which a normalized, subject-oriented, integrated, time-variant, and non-volatile warehouse acts as the single source of truth feeding downstream data marts [4]. Kimball offered a complementary, bottom-up philosophy centered on dimensional modeling, where conformed dimensions and fact tables organized as star schemas optimize query performance and readability for business users [5]. Both methodologies remain foundational. The dimensional model in particular still shapes how semantic layers and self-service datasets are designed today.

Davenport and colleagues later reframed analytics as a source of strategic differentiation. They argued that organizations could compete on analytics by embedding fact-based decisions into their core processes and culture [10]. That perspective shifted the conversation from infrastructure to value. It also foreshadowed the later emphasis on democratizing access to data, so that analytical capability is not confined to a specialist elite.

### B. The Rise of Self-Service and Visual Analytics

The 2010s saw the ascent of visual, self-service analytics tools that lowered the technical barrier to data exploration. Interactive visualization, grounded in principles set out by Few and others, made patterns and outliers visible to non-technical users, while drag-and-drop authoring removed the dependence on hand-written queries [7], [11]. Independent industry evaluations document this transition. The Gartner Magic Quadrant for Analytics and Business Intelligence Platforms, in particular, repeatedly identifies ease of use, governed self-service, and augmented capabilities as the primary axes of differentiation among vendors [12].

### C. Cloud Data Warehouses and the Lakehouse

A decisive architectural change came with the move to the cloud. Cloud-native data warehouses introduced an architecture that separates compute from storage, letting each scale independently and enabling elastic, pay-per-use processing over very large datasets [13]. The data-lake paradigm, in parallel, offered inexpensive object storage for raw, semi-structured, and unstructured data [14]. Early data lakes, however, lacked the transactional guarantees and schema management of warehouses. The lakehouse architecture was proposed to unify the two worlds. It implements warehouse-grade management features, including ACID transactions, schema enforcement, and time travel, directly on low-cost object storage through open table formats [15]. Open formats such as Delta Lake and Apache Iceberg provide the transactional metadata layer that makes this convergence practical, separating the storage of data from the engines that query it [16], [17]. For self-service BI, the lakehouse matters because it supports governed, performant analytics over a single copy of data, with no cost or latency penalty for replicating that data into a separate warehouse.

## III. SELF-SERVICE BI REFERENCE ARCHITECTURE

A dependable self-service capability rests on a layered architecture. Each layer has one clear responsibility and a governed contract with the layers above and below it. Figure 1 presents the proposed reference architecture. It comprises six layers: source systems, integration, storage and processing, augmented analytics, semantic and governance, and consumption. The design principle is consistent throughout: push freedom upward to business users while concentrating control in the semantic and governance layer.

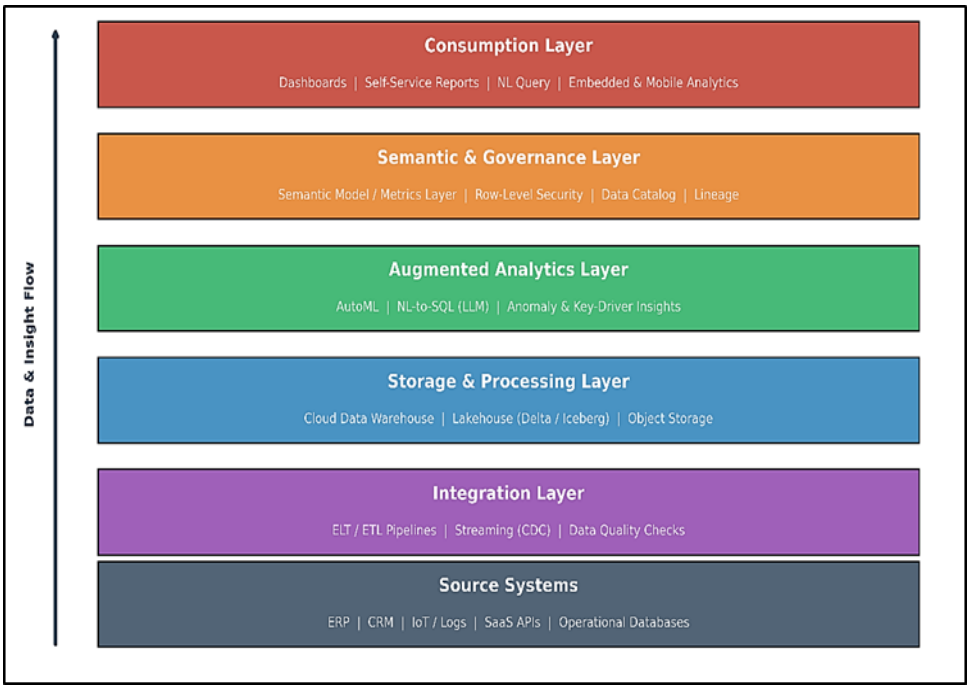


Fig. 1: Proposed layered reference architecture for governed self-service BI, in which a semantic and governance layer mediates between cloud storage and business consumption.

### A. Integration, Storage, and Processing

The integration layer ingests data from operational sources: enterprise resource planning, customer relationship management, software-as-a-service applications, logs, and event streams. It uses batch extract-load-transform (ELT) pipelines and change-data-capture streaming [18]. The contemporary ELT pattern loads raw data into the cloud platform first and transforms it in place, exploiting elastic compute and keeping an auditable raw history [13]. The storage and processing layer is realized as a cloud data warehouse or a lakehouse on open table formats, supplying the transactional guarantees and elastic scale that downstream analytics depend on [15], [16]. Data-quality and validation checks belong here. Placed at this layer, they stop defects from propagating into self-service datasets, where the defects would be far more costly to detect and correct.

### B. The Semantic and Governance Layer

The semantic layer is the architectural keystone of trustworthy self-service. It encodes business definitions, including metrics, dimensions, hierarchies, and relationships, as governed, reusable objects. A term such as "active customer" or "net revenue" then resolves to one agreed calculation, no matter who builds the report. By centralizing metric logic, the semantic layer prevents the proliferation of subtly inconsistent definitions that erodes trust in self-service environments [6], [7]. The governance controls sit alongside it: role-based and row-level security, so users see only the data they are entitled to; a data catalog that makes certified datasets discoverable; and lineage that traces every figure back to its sources for audit and impact analysis [19]. This layer is what lets the consumption layer above it stay genuinely open without becoming chaotic.

### C. The Consumption Layer

The consumption layer is where business value is realized. It hosts interactive dashboards, self-service report authoring, natural-language query interfaces, and analytics embedded into operational applications and mobile experiences. Every artifact in this layer draws on certified semantic objects. As a result, analyses produced independently by different teams stay mutually consistent. Embedding analytics directly into the workflows where decisions are made, rather than confining them to a standalone BI portal, is a defining trait of mature, transformation-era deployments [3], [11].

## IV. COMPARISON OF SELF-SERVICE BI PLATFORMS

Four platforms dominate enterprise self-service BI: Microsoft Power BI, Tableau (Salesforce), Google Looker, and Qlik. Each is consistently positioned among the leaders in independent market evaluations, yet each embodies a distinct architectural philosophy [12]. The comparison that follows synthesizes a recent comprehensive technical analysis of enterprise BI platforms [9] with vendor documentation and public evaluations. Capability scores in Figure 2 and Table 1 are illustrative. They convey relative emphasis rather than exact rankings, which shift with every product release.

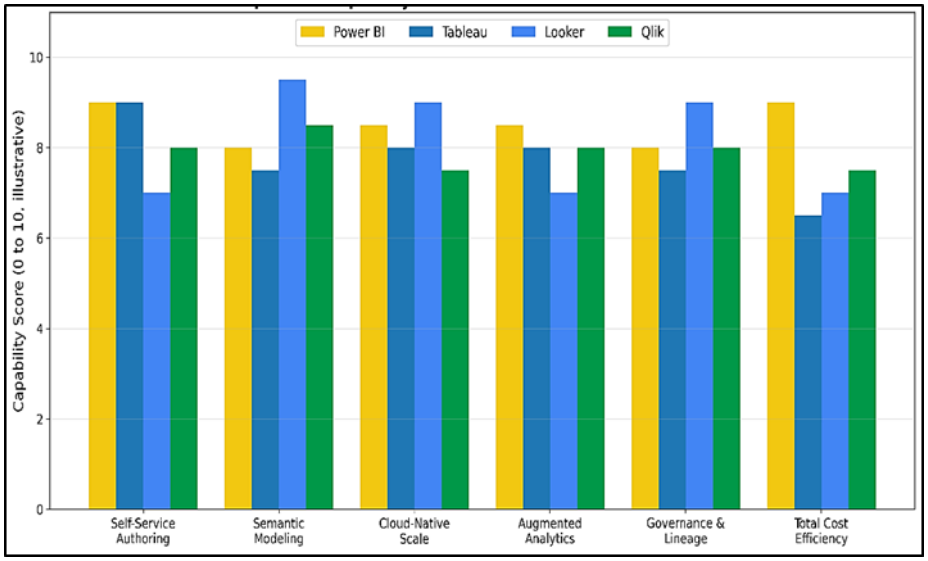


Fig. 2: Illustrative comparative capability assessment of Power BI, Tableau, Looker, and Qlik across six self-service dimensions (scores 0 to 10).

Power BI pairs strong self-service authoring with deep integration into the Microsoft ecosystem and an aggressive cost position. For enterprises already invested in that ecosystem, it is often the default choice [9]. Tableau is widely admired for the depth and fluidity of its visual-analytics experience, and it excels where exploratory, visually driven analysis is paramount [11]. Looker is distinguished by its code-defined semantic model, which expresses metrics as version-controlled, reusable definitions and aligns naturally with lakehouse-centric, in-database analytics. Qlik differentiates itself through an associative in-memory engine that lets users explore relationships across all data without predefined drill paths. Table 1 summarizes the comparison across the dimensions most relevant to governed self-service.

Table 1. Illustrative Feature Comparison of Leading Self-Service BI Platforms

| Dimension | Power BI | Tableau | Looker | Qlik |
|---|---|---|---|---|
| Self-service authoring | Excellent | Excellent | Good | Very good |
| Semantic / metrics layer | Very good | Good | Excellent | Very good |
| Cloud-native scalability | Very good | Very good | Excellent | Good |
| Augmented analytics | Very good | Very good | Good | Very good |
| Governance & lineage | Very good | Good | Excellent | Very good |
| Core engine | VertiPaq in-memory | Hyper in-memory | In-database LookML | Associative in-memory |
| Typical strength | Ecosystem + cost | Visual exploration | Governed semantics | Associative discovery |

No single platform is uniformly superior. The appropriate choice depends on existing technology investments, the relative priority placed on visual exploration versus governed semantics, and the target

deployment model. One trend cuts across all four: each is increasingly architected to query cloud warehouses and lakehouses directly. That reflects the broader shift toward a shared, governed data foundation beneath whichever consumption tool an organization selects [9], [15].

## V. AUGMENTED ANALYTICS AND LARGE LANGUAGE MODELS

Augmented analytics applies machine learning and natural-language technologies to automate parts of data preparation, insight discovery, and explanation that once required a data scientist. Its capabilities include automatic detection of anomalies and trends, key-driver analysis that surfaces the factors most associated with an outcome, and automated machine learning that fits and ranks candidate models [20]. By lowering the analytical skill threshold, augmented analytics extends self-service beyond power users to a wider population of decision-makers [8].

### A. Natural-Language Querying and NL-to-SQL

The most consequential recent development is the application of large language models to the natural-language interface. Natural-language-to-SQL translation lets a user pose a question in plain language, for example, "what was quarter-over-quarter revenue growth for the western region?", and have the system generate and execute the corresponding query [8]. Research on semantic parsing and text-to-SQL, exemplified by large cross-domain benchmarks, established the task and its evaluation. Modern LLMs have since sharply improved generation accuracy on realistic schemas [21], [22]. When an LLM is grounded in the governed semantic layer rather than left to interpret raw tables, its generated queries inherit certified metric definitions and security rules. That grounding substantially reduces the risk of plausible but wrong answers.

### B. Reliability, Grounding, and Limitations

Progress has been rapid, yet LLM-driven analytics carries well-documented risks. Models can hallucinate columns or join conditions that do not exist. They can misread an ambiguous business term. They can silently apply an incorrect aggregation. Three practical mitigations keep a human in the loop: grounding generation in the semantic layer, constraining outputs to the catalog of certified metrics, and returning the generated SQL for inspection [19], [22]. The aim is not to replace the analyst. It is to compress the distance between a business question and a trustworthy, governed answer. Seen this way, augmented analytics and self-service governance are complementary, not competing, concerns.

## VI. GOVERNANCE, CHALLENGES, AND MATURITY MODEL

### A. Governance and Data-as-a-Product

Governance is the discipline that makes self-service sustainable. It reaches beyond access control to the certification of trusted datasets, stewardship roles with clear accountability, and policies for metric definition and change management [19]. A growing organizational pattern treats curated datasets as managed products, each with documented owners, service expectations, and quality guarantees. Consumers can then rely on such a dataset as they would on any other product. This data-as-a-product orientation pushes responsibility for quality toward the domains that understand the data best, while a federated governance framework keeps definitions and controls globally consistent.

### B. Adoption Challenges

Many obstacles to self-service adoption are organizational rather than technical. Poor data quality undermines trust in any analysis. Inconsistent metric definitions generate conflicting numbers that drain confidence. Uneven data literacy limits how far self-service can spread. Without a semantic layer, an unchecked proliferation of redundant reports can recreate the very fragmentation that self-service was meant to resolve [6], [7]. Figure 3 summarizes two things: the trajectory of the wider BI and analytics market, and the benefits that organizations most often associate with self-service adoption. These figures are illustrative syntheses of public industry sources, not measurements from a specific deployment.

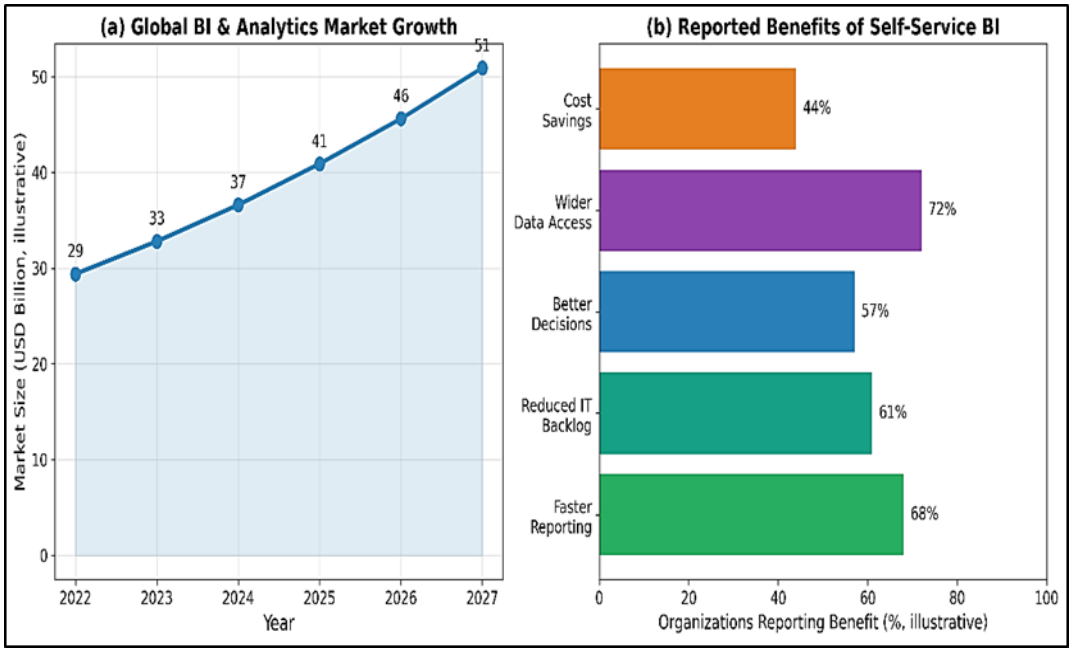


Fig. 3: (a) Illustrative growth trajectory of the global BI and analytics market and (b) frequently reported benefits of self-service BI adoption.

### C. A Five-Level Maturity Model

To help organizations benchmark capability and sequence investment, Figure 4 presents a five-level maturity model. At Level 1 (Initial), analysis is ad-hoc and spreadsheet-driven, with no single source of truth. Level 2 (Managed) adds a central warehouse and standardized, IT-produced reports. Level 3 (Self-Service) establishes governed datasets and a semantic layer that empower business authoring. Level 4 (Augmented) brings automated insight, natural-language querying, and embedded analytics. Level 5 (Pervasive) reflects a genuine data culture, in which governed, real-time analytics is woven into everyday decisions across the organization. Progression is cumulative. Each level builds on the governance foundation of the one below it.

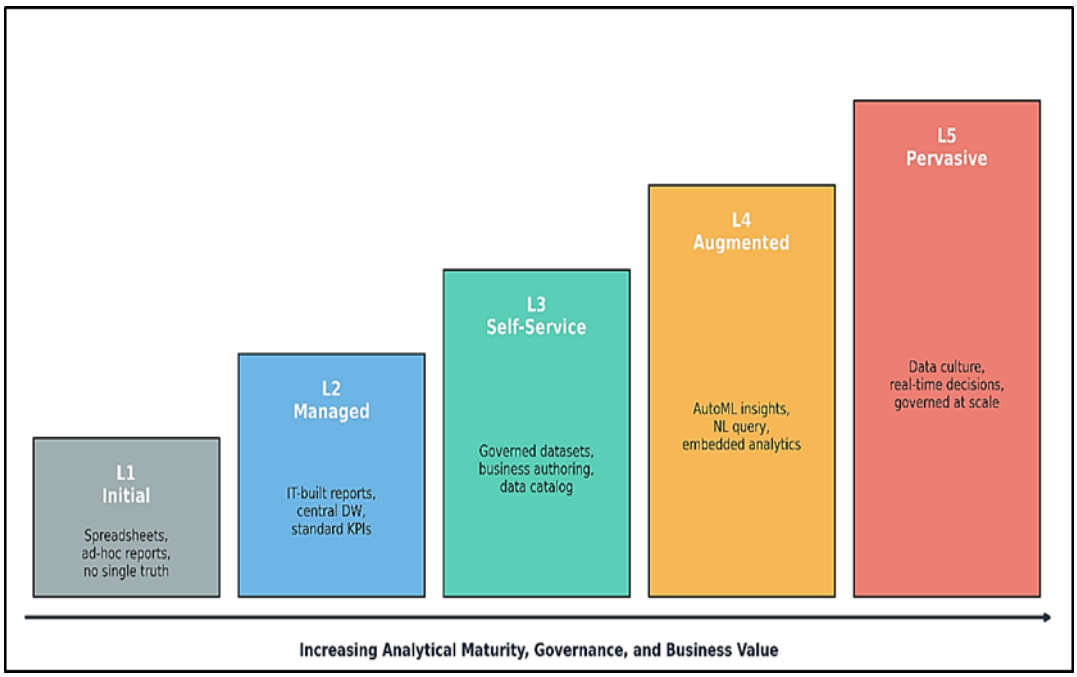


Fig. 4: Five-level self-service BI maturity model, from ad-hoc reporting to pervasive, governed analytics.

Table 2. Characteristics of the Self-Service BI Maturity Levels

| Level | Stage | Data Foundation | Primary Authors | Key Risk |
|---|---|---|---|---|
| L1 | Initial | Spreadsheets / silos | Individuals | No single truth |
| L2 | Managed | Central warehouse | IT / BI team | Report backlog |
| L3 | Self-Service | Governed semantic layer | Business analysts | Definition sprawl |
| L4 | Augmented | Lakehouse + ML/LLM | Analysts + automation | Ungrounded AI output |
| L5 | Pervasive | Federated data products | Whole organization | Governance at scale |

The maturity model also clarifies sequencing. Deploying augmented analytics (Level 4) before a governed semantic layer (Level 3) tends to amplify inconsistency rather than insight, because automated and LLM-generated outputs inherit whatever definitional ambiguity sits beneath them. A disciplined path secures a governed foundation first, then layers automation on top [2], [19].

## VII. CONCLUSION

Self-service business intelligence is a defining capability of the digital-transformation era. It shifts analytical authorship from a central IT function to the business users closest to each decision. This paper traced the architectural arc from enterprise data warehouses to cloud lakehouses, proposed a layered reference architecture that gives the semantic and governance layer a first-class role, compared the four leading platforms (Power BI, Tableau, Looker, and Qlik), and examined how augmented analytics and large language models are reshaping the analytical interface through natural-language querying and NL-to-SQL translation.

The recurring lesson is simple. Self-service freedom and governance are complementary, not opposing, forces. Durable value comes not from selecting a single tool but from establishing a governed semantic foundation, treating curated data as a managed product, and grounding automation in certified definitions. The maturity model offered here gives organizations a practical way to benchmark capability and sequence investment, so they advance toward pervasive, trustworthy analytics. Future work should validate the proposed architecture and maturity model empirically across deployments and quantify the reliability gains obtained by grounding LLM-based querying in the semantic layer. The metrics presented in this paper are illustrative syntheses of public sources and are not drawn from a single deployed study [2], [9], [12].

## REFERENCES


[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.

[2] S. Charly, "Digital Transformation Strategy: Where to Start and What to Avoid," Int. J. Inf. Technol. Res. Stud. (IJITRS), vol. 1, no. 3, pp. 180–189, Oct. 2025, https://doi.org/10.5281/zenodo.17510445 .

[3] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," MIS Quart., vol. 36, no. 4, pp. 1165–1188, Dec. 2012.

[4] W. H. Inmon, *Building the Data Warehouse*, 4th ed. Indianapolis, IN, USA: Wiley, 2005.

[5] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.

[6] B. Evelson, "Topic Overview: Business Intelligence," Forrester Research, Cambridge, MA, USA, 2008.

[7] C. Imhoff and C. White, "Self-Service Business Intelligence: Empowering Users to Generate Insights," TDWI Best Practices Report, The Data Warehousing Institute, Renton, WA, USA, 2011.

[8] R. Sallam, C. Howson, *et al*., "Augmented Analytics Is the Future of Data and Analytics," Gartner Research, Stamford, CT, USA, 2017.

[9] K. Abraham, "Business Intelligence Tools Comparison: Tableau, Power BI, and Alternatives: A Comprehensive Technical Analysis of Enterprise BI Platforms," Int. J. Inf. Technol. Res. Stud. (IJITRS), vol. 2, no. 1, pp. 34–41, Jan. 2026, https://doi.org/10.5281/zenodo.18490659 .

[10] T. H. Davenport and J. G. Harris, *Competing on Analytics: The New Science of Winning*. Boston, MA, USA: Harvard Business Review Press, 2007.

[11] S. Few, *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Burlingame, CA, USA: Analytics Press, 2009.

[12] Gartner, "Magic Quadrant for Analytics and Business Intelligence Platforms," Gartner Research, Stamford, CT, USA, 2024.

[13] B. Dageville *et al*., "The Snowflake elastic data warehouse," in *Proc. 2016 ACM SIGMOD Int. Conf. Management of Data*, San Francisco, CA, USA, 2016, pp. 215–226.

[14] P. Russom, “Data Lakes: Purposes, Practices, Patterns, and Platforms,” TDWI Best Practices Report, The Data Warehousing Institute, Renton, WA, USA, 2017.
[15] M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, “Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics,” in *Proc. 11th Conf. Innovative Data Systems Research (CIDR)*, 2021.
[16] M. Armbrust *et al*., “Delta Lake: High-performance ACID table storage over cloud object stores,” *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 3411–3424, Aug. 2020.
[17] Apache Software Foundation, “Apache Iceberg: A high-performance open table format for huge analytic datasets,” *Apache Iceberg Documentation*, 2023. [Online]. Available: https://iceberg.apache.org
[18] P. Vassiliadis, “A survey of extract-transform-load technology,” *Int. J. Data Warehousing Mining*, vol. 5, no. 3, pp. 1–27, 2009.
[19] J. Ladley, *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*, 2nd ed. Cambridge, MA, USA: Academic Press, 2019.
[20] G. Shmueli and O. R. Koppius, “Predictive analytics in information systems research,” *MIS Quart.*, vol. 35, no. 3, pp. 553–572, 2011.
[21] T. Yu *et al*., “Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task,” in *Proc. 2018 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, 2018, pp. 3911–3921.
[22] J. Li *et al*., “Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs (BIRD),” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.