JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 19, No. 3
September 2026
ISSN: 1946-1836

In this issue:

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research and Analytics** (JISARA) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is four issues a year. The first date of publication was December 1, 2008. The original name of the journal was Journal of Information Systems Applied Research (JISAR).

JISARA is published online (https://jisara.org) in connection with the ISCAP (Information Systems and Computing Academic Professionals) Conference, where submissions are also double-blind peer reviewed. Our sister publication, the Proceedings of the ISCAP Conference, features all papers, teaching cases and abstracts from the conference. (https://iscap.us/proceedings)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%) and other submitted works. The non-award winning papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in JISAR. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisara.org or the publisher at publisher@jisara.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISARA.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

2026 JISARA Editorial Board

Household Digital Twin for Storm Preparedness and Response

Chaitali Bonke
chaitali.bonke01@utrgv.edu
University of Texas Rio Grande Valley
Edinburg, TX 78539

Jun Sun
jun.sun@utrgv.edu
University of Texas Rio Grande Valley
Edinburg, TX 78539

Abstract

Advancements in computational technologies are reshaping the landscape of sustainable smart city planning, enabling data-driven approaches to environmental resilience. A key innovation in this space is the integration of Artificial Intelligence (AI), the Artificial Intelligence of Things (AIoT), and Urban Digital Twin (UDT) technologies. Building on these foundations, this study presents a Household Digital Twin (HDT) system designed to enhance storm preparedness and disaster response at the individual home level. By creating dynamic, virtual replicas of households, the HDT platform enables tailored simulations of hurricane scenarios, offering personalized guidance to families before, during, and after extreme weather events. These simulations model the effects of varying storm intensities on homes and nearby trees, helping homeowners, local governments, and insurance providers assess vulnerabilities, optimize emergency planning, and implement targeted mitigation strategies. Through its ability to deliver proactive, context-specific insights, the HDT system contributes to greater community resilience and supports broader goals in climate adaptation and smart urban infrastructure.

Keywords: Household digital twin, storm preparedness, tree damage simulation, disaster risk reduction

Recommended Citation: Bonke, C., Sun, J., (2026). Household Digital Twin for Storm Preparedness and Response. *Journal of Information Systems Applied Research and Analytics*, v19(n3) pp 4-16.
DOI# https://doi.org/10.62273/HSCA1671

Household Digital Twin for Storm Preparedness and Response

Chaitali Bonke and Jun Sun

1. INTRODUCTION

Natural disasters strike abruptly and often leave devastating consequences in their wake, encompassing hydrological, geophysical, and climatological events. Over the past decade, nearly 290 such disasters in the United States have caused an estimated $1.3 trillion in damages, much of it attributed to residential and commercial property losses. In 2023 alone, severe storms accounted for nearly $50 billion in insured damages, and roughly $530 billion in total payouts, according to the Insurance Information Institute (III).

Among these disasters, hurricanes remain a persistent and growing threat, affecting both coastal and inland regions through high winds, storm surges, and secondary hazards such as tornadoes (Ayscue, 1996). One often underestimated but highly destructive consequence of hurricanes is falling trees. Tree impacts can puncture roofs, shatter windows, and compromise a home's structural integrity—frequently rendering homes uninhabitable and driving up repair costs (FLORIDA; Gilman et al., 2006). Beyond property damage, fallen trees disrupt power lines, impede emergency access, and pose significant safety risks, further complicating disaster response efforts (Salisbury & Koeser, 2023).

Traditional disaster management emphasizes response, recovery, and preparedness at broad scales. However, the increasing intensity and frequency of hurricanes underscore the need for localized, proactive mitigation strategies. Insurance providers, government agencies, and private stakeholders alike are increasingly supporting technologies that empower households to anticipate risks and take preemptive action.

Digital twin (DT) technology offers one such solution by creating dynamic, virtual replicas of physical environments to simulate and optimize responses to real-world conditions (Hughes et al., 2023). Embedded in disaster management systems, DTs support Intelligent Disaster Prevention and Mitigation Infrastructure (IDPMI) through improved risk assessment, scenario modeling, and decision support (Yu & He, 2022).

At the municipal level, urban digital twins (UDTs) have demonstrated value in enhancing hazard response and crisis management. These city-scale platforms often combine artificial intelligence (AI) and information and communication technologies (ICT) to model complex infrastructure systems and support data-driven planning (Fan, Zhang, Yahja, & Mostafavi, 2021). T However, these systems typically operate at a macro scale and do not offer the personalized insights needed by individual households.

This study introduces a Household Digital Twin (HDT) platform designed to address predictable natural disasters—particularly hurricanes—at the micro scale. Leveraging advanced simulation technologies, the HDT models the potential impacts of hurricanes on homes and surrounding trees, offering personalized, property-specific mitigation strategies. For example, homeowners can assess risks from nearby unstable trees and take preventive actions such as pruning, removing weak limbs, or reinforcing vulnerable structures. These models could be hosted by insurance providers or local governments, with potential cost savings from reduced claims and emergency relief.

Unlike UDTs, which focus on citywide planning and environmental sustainability (Bibri, Huang, & Krogstie, 2024; Mishra, 2023), the HDT centers on localized risk management and household resilience. It serves as a critical bridge between large-scale disaster simulations and individual preparedness strategies, contributing to a more holistic disaster management ecosystem.

In the broader context of increasing AI and IoT integration within smart cities (Gourisaria et al., 2022; Zaidi, Ajibade, Musa, & Bekun, 2023), the development of household-level digital twins represents a transformative shift in disaster preparedness. Between late 2022 and mid-2024, approximately 1.1% of U.S. households were displaced by disasters—primarily hurricanes (Paul et al., 2024). The high economic losses

and vulnerability often result from poor coordination and underinvestment in preparedness (U.S. Chamber of Commerce, 2025).

This research presents a scalable and intelligent Household Digital Twin platform that simulates hurricane impacts in advance, enabling homeowners, local authorities, and insurers to assess vulnerabilities and implement proactive risk mitigation strategies. By providing predictive insights and actionable recommendations, the HDT contributes to building resilient, disaster-ready communities and supports global efforts to harness digital technologies for climate adaptation and risk reduction.

2. RESEARCH BACKGROUND

Digital Twin (DT) technology, first conceptualized in *Mirror Worlds* by David Gelernter (1991), refers to the creation of dynamic, virtual replicas of physical systems or entities. These digital counterparts enable real-time simulation, monitoring, and data-driven decision-making. DTs are typically categorized into four types—Product Twins, Process Twins, System Twins, and Human Twins—each designed to optimize specific operations and interactions (Juarez, Botti, & Giret, 2021; Boschert, Heinrich, & Rosen, 2018). By integrating operational and engineering data, digital twins evolve in tandem with their physical counterparts, offering deep insights for system analysis, planning, and user engagement.

Within urban contexts, Urban Digital Twins (UDTs) serve as virtual replications of a city's social, environmental, and infrastructural systems. They are used to simulate, predict, and manage real-time urban processes—including transportation, energy distribution, and emergency response—thus enabling more informed governance and planning (Marçal Russo et al., 2025; Zhu & Jin, 2025). UDTs form a foundational component of smart city ecosystems, which leverage interconnected technologies such as the Internet of Things (IoT), Artificial Intelligence (AI), and digital platforms to improve urban sustainability, mobility, and citizen-centric services (Ersan, Irmak, & Colak, 2024; Mohammadi & Taylor, 2017; Vessali, Galal, & Dr Nowson, 2022).

Smart City Digital Twins (SCDTs) advance this vision by integrating data from IoT sensors, geospatial tools, and real-time communication systems to model city-wide dynamics. These platforms allow for proactive resource management, predictive analytics, and enhanced disaster response. Technologies such as LiDAR, drones, and augmented reality (AR) further extend their capabilities, supporting applications ranging from flood modeling to evacuation planning (Fan, Zhang et al., 2021; Faliagka et al., 2024).

An essential component of SCDTs is the ability to conduct scenario-based simulations. Tools like interactive GeoData Frames enable stakeholders to visualize the potential impacts of various emergencies at neighborhood or block levels (Gkontzis et al., 2024). Past disasters, such as Hurricane Katrina, underscore the importance of incorporating inclusive and community-oriented approaches into disaster planning frameworks (Patterson, Weil, & Patel, 2010). Models such as the Ontology-based Decision Model and Notation (oDMN) (Horita et al., 2016) and the Asian Disaster Preparedness Model (ADPC, 2015) highlight the need for integrated, system-level thinking in adaptive disaster response strategies.

SCDTs surpass traditional disaster management tools by enabling holistic views of urban risk and resilience. Projects like METACITIES exemplify the transformative potential of UDTs in optimizing urban transport, emergency response, and environmental management. However, despite their promise, widespread adoption of UDTs faces several challenges:

- Technical Barriers: Lack of standardized data models and limited interoperability across systems.
- Ethical and Governance Issues: Concerns around privacy, transparency, and data governance.
- Economic Constraints: High initial costs of digital infrastructure—such as intelligent sensors, 3D modeling tools, and data centers—pose barriers for disaster-prone but financially constrained regions (GovPilot, 2024; Hexagon, 2025; PwC, 2023; World Economic Forum, 2023).
- Stakeholder Coordination: Effective implementation requires cross-sector collaboration between technologists, policymakers, urban planners, and local communities.

To address these issues, hierarchical digital twin frameworks have been proposed. These structured architectures facilitate seamless data integration, interoperability, and real-time coordination across system layers—from

municipal to household levels (Ball, Lagadec, & Laval, 2025; Finke et al., 2023). At the same time, inclusive, people-centric approaches are critical for building trust, ensuring equity, and encouraging stakeholder buy-in (Bibri et al., 2024; Lu et al., 2020).

As urban populations continue to grow and climate-related disasters increase in frequency and intensity, integrating digital twin technologies into smart city infrastructure becomes increasingly essential. While most research and development has focused on city-scale applications, this study extends digital twin principles to the household level.

Trees are among the leading causes of property damage during hurricanes, often falling on homes, vehicles, and power lines due to high winds and waterlogged soil (Peterson, 2000; Gardiner et al., 2016). While hurricanes also cause flooding and structural failures, tree blowdowns are particularly suitable for simulation at the household level due to their visibility, measurability, and direct mitigation potential via pruning or removal. The focus on trees enables practical, user-driven interventions and supports early-stage validation of digital twin technology in disaster scenarios (Gilman et al., 2006). Future expansions of the HDT will incorporate additional hazards such as flooding and structural damage.

The following section presents a novel Household Digital Twin (HDT) platform—designed to enhance disaster preparedness through hyper-local risk modeling, community engagement, and proactive mitigation measures. By focusing on individual homes and surrounding environments, the HDT aims to fill a critical gap in existing disaster management systems and contribute to more resilient, technology-enabled communities.

3. HOUSEHOLD DIGITAL TWIN (HDT) DESIGN

The history of human intervention in complex systems highlights a critical lesson: sustainable and effective solutions require a deep understanding of the systems they intend to improve (Lawther, 2016). In disaster response, a promising path forward involves leveraging autonomous computational systems to manage complex, ethically charged, and time-sensitive tasks. One illustrative example is the Slándáil Social Media Monitor, a system that enhances disaster communication by integrating public input with official emergency management operations. It addresses the so-called "Good Will Hazard," where uncoordinated efforts by well-meaning individuals can lead to redundancy, resource waste, or critical service gaps (Hayes & Kelly, 2018). By aligning public contributions with official protocols, such systems increase overall response efficiency.

In this ecosystem, individual households are not passive entities—they are critical actors. Alongside governments, insurers, and utility providers, families play a direct role in preparing for and responding to disasters. To analyze and structure this interaction, Activity Theory (AT) serves as a robust framework. AT, a descriptive meta-theory of human behavior, examines how individuals and communities interact with tools, environments, and shared goals over time (Kaptelinin & Nardi, 2012; Kuutti, 1996). Its emphasis on tool-mediated, context-driven activity makes it particularly well suited for digital twin systems, which integrate technical interfaces, environmental data, and human decision-making.

3.1 Conceptual Design

Figure 1 illustrates the conceptual design of the household digital twin (HDT), developed for disaster preparedness. This design is grounded in Engeström's (1999) Activity System Model, highlighting the intricate relationships and interactions within the activity system.

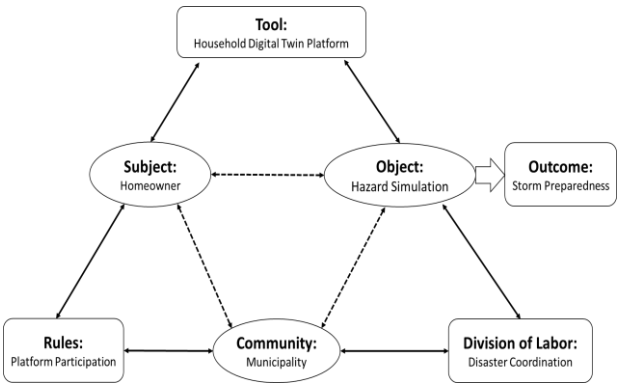


Figure 1. Conceptual design

Each household user can interact with the digital twin system, which simulates catastrophic events such as the impact of severe weather on trees and nearby areas. This interaction occurs through a web platform that is integral to the proposed HDT system. The community within a municipality, comprising households, local officials, utilities, and other stakeholders, collectively forms the operational ecosystem of this framework. The division of labor within the HDT system delineates the responsibilities of

each stakeholder in hazard simulation and disaster preparation. Cooperation between families and stakeholders is governed by disaster coordination regulations, ensuring a cohesive response. The system is driven by the overarching goal of providing effective response recommendations, facilitating proactive disaster mitigation.

3.2 Architectural Design

The origin of digital twin technology traces back to NASA's Apollo missions in the 1960s, where real-time digital simulations of spacecraft were used to troubleshoot failures from Earth. The formal concept was later introduced by Michael Grieves in 2002 in the context of Product Lifecycle Management (PLM) at the University of Michigan.

Today, smart homes and connected environments use similar principles—combining IoT devices, real-time sensors, and cloud-based platforms to model physical conditions. Figure 2 illustrates the architectural design of the HDT system, showing the key components and communication pathways between stakeholders.

A key component of the system is the core simulation module, which leverages a digital twin modeler to replicate the effects of catastrophic events on households and their immediate surroundings. This module analyzes a range of response strategies and provides users with actionable insights tailored to their specific conditions. For example, during hurricane scenarios, the web platform presents simulation outcomes to household users, visualizing potential impacts under varying storm intensities. These simulations are powered by data stored in the database module and are continuously updated with real-time inputs from users.

Household digital twin users typically consist of families or individuals living in smart homes equipped with IoT devices such as thermostats, environmental sensors, connected appliances, cameras, and health wearables. These devices enable real-time interaction with the system, supporting a seamless flow of data and feedback. When users report issues or complete recommended mitigation actions, service providers and municipal officials are automatically notified, allowing for prompt coordination and assistance. This collaborative framework enhances both individual preparedness and collective disaster resilience, fostering a proactive approach to risk reduction at the household level.

3.3 Logic Design

The logic design of the Household Digital Twin (HDT) system, as illustrated in Figure 3, outlines the operational flow from data collection to decision-making. The system leverages LiDAR technology to generate high-resolution digital maps of each region, capturing detailed structural and environmental features of residential properties. This geospatial data, combined with user-provided inputs about household conditions, is stored in a centralized database that serves as the foundation for hazard simulations. Using this integrated dataset, the simulation module forecasts the potential impacts of various hurricane scenarios—such as differing storm categories—well in advance of hurricane season. The simulation results are stored and passed to a recommender system, which generates personalized preparedness plans for each household.

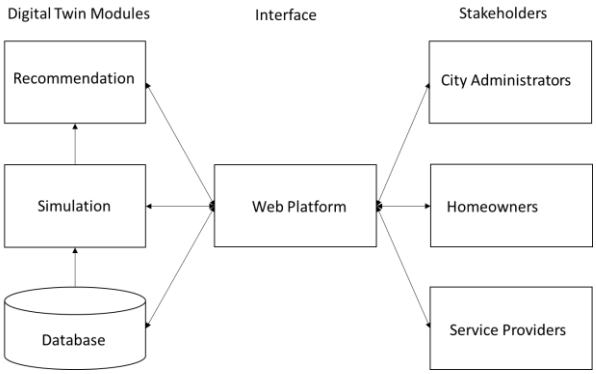


Figure 2. Architectural Design

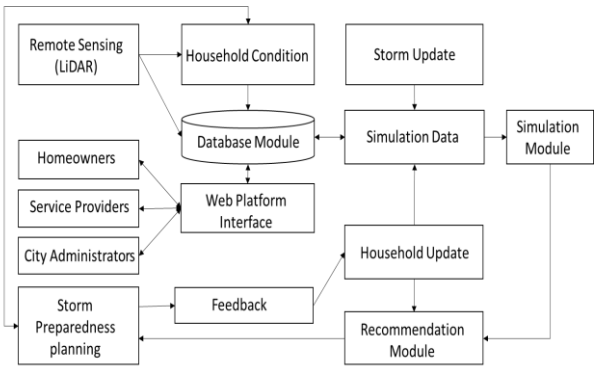


Figure 3. Logic Design

As users implement the suggested actions, such as reinforcing structures or trimming hazardous tree limbs, their feedback is recorded, updating the digital twin's state. This, in turn, retriggers simulations and enables the system to refine future recommendations based on real-time behavioral and environmental data. To ensure

that the system reflects current conditions, users can notify the platform upon completing tasks, prompting new assessments. In parallel, service providers, city officials, and homeowners can be alerted when follow-up actions—such as debris removal—are needed.

HDT functions as a mediating tool that links individual households with municipal systems, creating a coordinated ecosystem for storm preparedness. Each component reflects the principles of activity theory, in which the subject (household), object (disaster mitigation), and community (stakeholders) interact within a context-driven, tool-mediated structure to achieve collective resilience outcomes.

4. PROTOTYPE DEVELOPMENT

The Household Digital Twin (HDT) prototype demonstrates how interactive storm simulations can assist homeowners in visualizing hurricane impacts on household trees and testing mitigation strategies to reduce damage. While traditional assessment methods offer broader and quicker evaluations, HDT provides granular, dynamic insights that complement these approaches by addressing complex, site-specific risks. To validate the system's design, a working HDT prototype was developed to assess interventions for reducing structural damage caused by tree blowdowns—an all-too-common consequence of hurricanes. These tropical storms frequently result in fallen trees and branches, causing significant property damage and power outages. As shown in Figure 4, the simulation interface enables homeowners to visualize storm impacts on trees adjacent to their property. Users can interactively adjust environmental variables such as wind speed, direction, and precipitation, and virtually prune trees to assess the effectiveness of different mitigation actions.

Figure 4. Simulator Layout

Simulations were conducted using the HDT system to model tree behavior under various hurricane conditions, using typical residential species such as oaks as representative cases. The simulation results provide homeowners with tailored recommendations—such as targeted pruning—to reduce tree-related risks during severe weather events. The current prototype focuses on individual tree modeling; scaling to simulate multiple trees or compound hazards will require more advanced modeling of tree-to-tree and tree-to-structure interactions. Emerging digital twin technologies that integrate high-resolution 3D models with physics-informed AI can enable these more complex simulations, allowing homeowners to manage risk from clustered trees or multiple environmental variables more effectively.

The HDT platform is hosted on a centralized application server that allows users to configure simulation parameters, review projected outcomes, apply recommended actions, and update tree status. Leveraging LiDAR data, homeowners can input or verify detailed attributes such as tree species, leaf shape and texture, number of branches, age, trunk diameter, height, and proximity to the house. Table 1 outlines the input parameters used in modeling a representative neighborhood oak tree for the simulation.

Parameter	Value	Model
Type of tree	Oak	
Shape of leaves	Oblong	
Type of leaves	Flat	
Texture of leaves	Smooth	
Trunk diameter	2 feet	
Number of branches	9	
Age of tree	12 years	
Distance from house	15 feet	
Height of tree	14 feet	

Table 1. Simulation parameters for a neighborhood Oak tree

LiDAR imaging was used to capture treetop structures and outline individual tree crowns. Techniques such as pouring algorithms and vector-based treetop identification were

employed to define crown geometry accurately (Koch et al., 2006). This data was used to generate community-scale digital models, as shown in Figure 5, representing both buildings and surrounding vegetation. These models enable realistic and localized storm impact simulations, adjusting for wind direction, speed, precipitation, and storm category.

Figure 5. LiDAR surface view of houses and surrounding trees

To quantify wind forces acting on trees, the simulation applies the formula:

$$F = C_d \cdot r \cdot a \cdot U^2,$$

where C_d is the momentum absorption coefficient (varies by tree type), r is air density, a is frontal area, and U is wind speed (Gardiner et al., 2016). Precipitation further affects tree vulnerability—saturated soil increases the likelihood of uprooting, while dry soil makes trees more prone to stem breakage (Peterson, 2000).

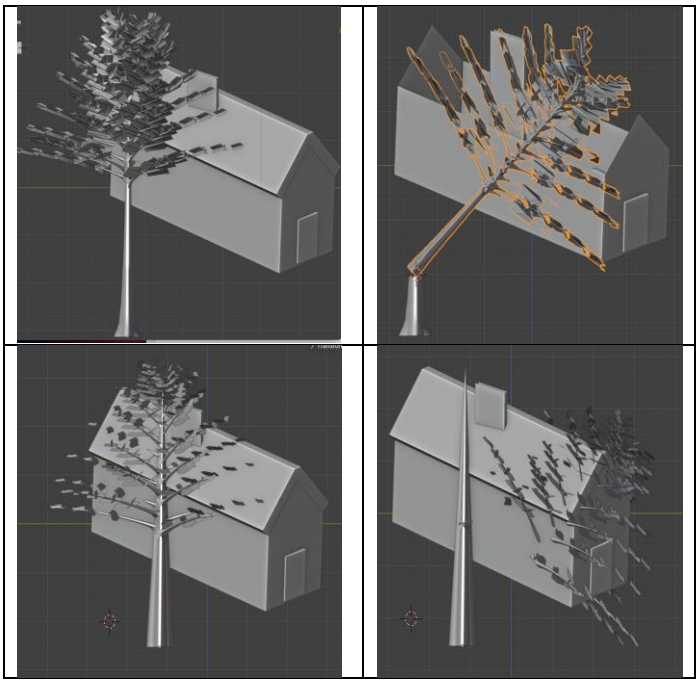
Table 2 presents calculated force thresholds and response outcomes for three common tree types—oak, pine, and palm—under Category 2 hurricane conditions (100 mph winds and 8-10 inches of precipitation). C_d values for pine and palm trees are lower than that of oak, indicating reduced resistance to wind momentum. Frontal area values also vary depending on trimming status.

Tree type	C_d	r	a	U	F= $Cd \cdot r \cdot a \cdot U^2$	Precip. (in)	Up-root	Stem Break
Oak	1	1	80%	100	8000	10	1	0
Pine	0.8	1	100%	100	8000	8	0	1
Palm	0.5	1	100%	100	5000	10	0	0

Table 2. Tree Damage Scenarios

Damage outcomes are determined by comparing calculated wind forces to predefined uprooting and stem breakage thresholds. For example, under 100 mph wind and 80% frontal density, the oak tree's wind force reaches 8000 units—sufficient for uprooting if the soil is saturated. These simulations help users assess which trees are most at risk and inform preventive actions such as trimming or removal. While current models focus on tree force thresholds, more comprehensive assessments would require integration of diverse environmental and structural parameters, along with scalability to broader ecosystems.

Tree simulations were developed using TreeIt and rendered using the Godot Game Engine, enabling lifelike reactions to wind forces across different species. As illustrated in Figure 6, each tree type reacts differently based on its structure, trimming, and environmental conditions.



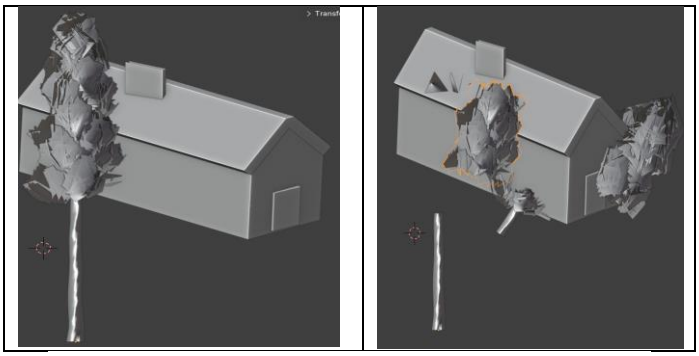


Figure 6. Simulation Demonstration

Figure 7 illustrates how tree trimming percentages (ranging from 10% to 50%) affect outcomes under varying hurricane categories. The trimming control in the simulation interface allows homeowners to observe risk levels before and after taking action.

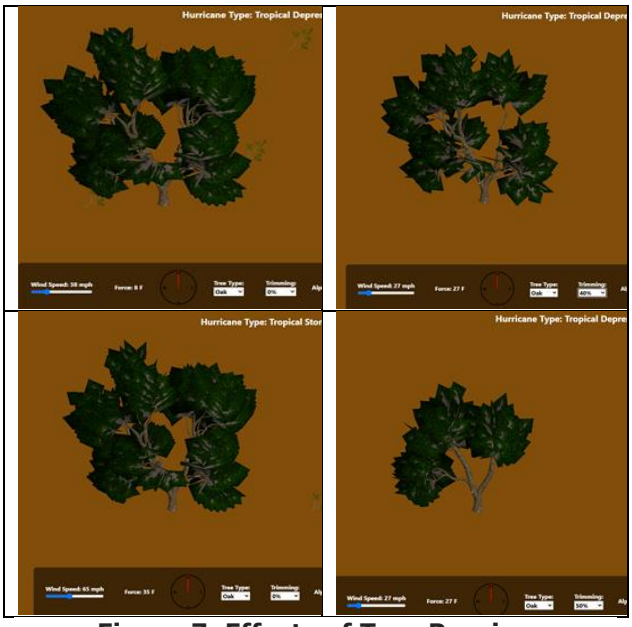


Figure 7. Effects of Tree Pruning

The web platform was developed using React, a JavaScript library well-suited for modular interfaces and real-time interactions. TailwindCSS was used for styling, and GreenSock Animation Platform (GSAP) powered the dynamic animations of tree movement. Lottie React was integrated to visualize house damage through real-time, animated sequences triggered by simulated wind speeds. Server-side rendering was handled by Next.js, ensuring optimized performance across devices.

The interface includes three primary controls: a wind speed slider (1–120 mph), a precipitation bar to model rainfall intensity, and a tree trimming percentage selector. Together, these elements enable users to explore complex scenarios interactively, reinforcing the importance of early mitigation strategies and informed decision-making at the household level.

To assess the accuracy of the simulation, the wind force outputs and resulting tree responses were cross-validated using damage thresholds reported in empirical studies of hurricane-related tree damage. For example, the calculated uprooting and stem breakage forces align with those observed in Gardiner et al. (2016) and Peterson (2000), who documented threshold wind speeds and precipitation levels associated with typical tree failures. Although full-scale real-world validation is ongoing, preliminary comparisons with historical damage from Hurricanes Katrina and Irma confirm that the force calculations and tree failure patterns simulated by the HDT prototype fall within reported empirical ranges (Byrne & Mitchell, 2013). Future work will include field deployment and post-hurricane observational studies to validate and refine simulation accuracy.

5. IMPLEMENTATION PATHWAY

Effective implementation of the Household Digital Twin (HDT) system requires coordinated collaboration among multiple stakeholders, each playing a pivotal role in reducing hurricane-induced tree damage to homes. These stakeholders include homeowners, insurance companies, local governments, tree care service providers, and community leaders. Within the HDT simulation framework, insurance companies assess property risk and incentivize mitigation; local governments oversee system deployment, resource allocation, and regulation enforcement; service providers implement recommended actions such as pruning or removal; and community leaders promote adoption, awareness, and stakeholder engagement to enhance disaster resilience.

A core strength of the HDT system is its ability to forecast tree-related vulnerabilities and guide timely interventions. For example, systems that monitor tree health in real time—using IoT-based sensors to track temperature, soil moisture, decay, and structural integrity—can be integrated with HDT simulations to enhance predictive accuracy (Patil et al., 2025). Similarly, advanced monitoring networks that model tree hydraulic responses to climate stressors provide insights into long-term growth dynamics and

vulnerability (Steppe & Schaepdryver, 2016). These data streams allow the HDT to move beyond static simulations, enabling dynamic, weather-informed decision support for both homeowners and municipal responders.

5.1 Stakeholders

Homeowners are the primary users and beneficiaries of the HDT platform. By engaging with the simulation interface, they can assess the vulnerability of trees based on wind exposure, species characteristics, and proximity to built structures. The system allows detailed input of tree-specific data—such as species, age, trunk diameter, canopy density, and distance from buildings—enabling personalized recommendations. Dense canopies, which behave like sails during high winds, are flagged by the system as high-risk. In such cases, HDT recommends targeted pruning to improve wind flow and tree stability. These actionable insights empower homeowners to proactively reduce the risk of tree-related property damage before a hurricane strikes.

Insurance companies benefit from HDT's data-driven risk models, which allow for more granular assessment of wind vulnerability on a per-property basis. Simulation data can be used to adjust premium pricing and encourage risk-reducing behavior among policyholders. Insurers may also offer discounts or incentives for proactive measures—such as tree trimming—aligned with HDT recommendations. By collaborating with local governments, insurance providers can help develop community-based mitigation frameworks and risk communication strategies.

Local governments play a critical role in institutionalizing the HDT system across communities. Simulation outputs at the neighborhood scale help identify high-risk areas and inform the prioritization of public mitigation efforts. Municipal authorities can also leverage the system to support urban forestry policies, establish safe tree-planting guidelines, and promote routine tree maintenance. Adoption can be further supported through educational campaigns, subsidies for tree services, and regulatory updates that mandate safe distances and trimming practices.

Tree care service providers are responsible for implementing HDT-generated recommendations. These professionals offer pruning, trimming, or removal services tailored to the simulation's output and help verify tree characteristics during input collection. Integration with the HDT platform supports efficient service scheduling, prioritization of high-risk cases, and documentation of completed work. Service providers also contribute valuable on-the-ground data—such as tree health indicators and regional climate patterns—that improve the simulation engine's local accuracy.

Community leaders are essential for building trust, raising awareness, and fostering grassroots engagement. They can organize public workshops, host neighborhood risk assessments, and facilitate collaboration among residents, service providers, insurers, and public officials. Their leadership helps ensure that HDT deployment is inclusive, resonates with local needs, and supports culturally sensitive disaster preparedness strategies.

Effective deployment of the HDT system relies on this multi-stakeholder alignment, creating an integrated ecosystem of data exchange, risk reduction, and responsive action. When stakeholders coordinate around shared resilience goals, HDT's impact expands—supporting policy reform, smarter urban planning, and resource-efficient disaster mitigation.

5.2 Proposed Deployment Strategy

A phased rollout of the HDT system is envisioned to ensure feasibility, adaptability, and stakeholder engagement across diverse contexts.

Phase 1 (0–6 months): Launch a small-scale pilot in a hurricane-prone municipality, targeting 50–100 households with varying tree profiles. Estimated costs of $100–200 per household will cover LiDAR mapping, app configuration, and coordination with tree service providers. Stakeholder feedback will be collected to refine system features and usability.

Phase 2 (6–18 months): Expand implementation within the same municipality based on Phase 1 outcomes. This phase includes formal partnerships with insurance providers and local governments. Municipalities may fund infrastructure and public outreach, while insurers offer premium discounts to incentivize homeowner participation. Broader public engagement and school-based tree awareness programs can be introduced.

Phase 3 (18–36 months): Regional scale-up through integration with smart city platforms and environmental monitoring systems. HDT will be linked with IoT-enabled weather stations and urban planning dashboards. Cost-benefit

projections suggest that a 10% reduction in tree-related claims during hurricanes could yield savings exceeding $1 million for insurers and municipal emergency services (PwC, 2023).

These phases are designed to balance upfront investment with measurable resilience and economic benefits. Pilot deployments will also serve as testbeds for comparing simulation outputs with actual storm damage records and tree failure data, creating a continuous feedback loop that strengthens model reliability.

Looking ahead, HDT will also inform long-term planning decisions, such as optimal tree spacing, species selection, and zoning strategies in residential areas. As environmental conditions evolve due to climate change, the platform will adapt through regular data collection, scenario testing, and stakeholder feedback.

5.3 Addressing Implementation Challenges

Despite its potential, successful HDT deployment must overcome several barriers. Scalability remains a key concern, as the system must be tailored to diverse urban, suburban, and rural contexts. Integration with existing smart home systems, insurance frameworks, and municipal planning tools requires standardization and interoperability. Predictive accuracy may vary across different tree species, topographies, and environmental conditions, necessitating continuous calibration and validation. Additionally, widespread adoption depends on addressing concerns around cost, digital accessibility, data privacy, and stakeholder trust. Ensuring equitable participation—particularly in historically underserved communities—will be essential to realizing the full societal benefits of the HDT system.

In summary, the HDT implementation pathway emphasizes an ecosystem approach—blending advanced simulation technologies with real-world stakeholder collaboration. Through phased deployment, adaptive feedback, and integration with broader resilience initiatives, HDT offers a transformative tool for reducing hurricane-related tree damage and strengthening community preparedness.

6. CONCLUSION AND FUTURE RESEARCH

This study addresses the growing need to reduce house damage caused by tree blowdowns during hurricanes—an increasingly critical issue as climate change amplifies the severity of tropical storms. Hurricanes often result in trees or large branches being uprooted, causing extensive property damage, blocked access routes, and widespread power outages. In response to this threat, a Household Digital Twin (HDT) prototype was developed to simulate how trees respond to different hurricane conditions when located near residential structures. Using the Godot game engine, the prototype enabled interactive visualization of hurricane scenarios, allowing users to test how factors such as wind speed, precipitation, and pruning influence the likelihood of tree failure and damage to homes.

The prototype represents a foundational step in demonstrating how digital twins can be used for personalized, site-specific disaster mitigation. By integrating user-defined parameters and enabling real-time interaction, the HDT platform empowers homeowners to understand vulnerabilities and take preventive measures. As the system evolves, incorporating additional tree species, regional vegetation types, and structural layouts will expand its applicability. Over time, data collected from user interactions can be used to refine model behavior and enhance simulation accuracy. Furthermore, the validation of simulation results against historical hurricane damage—following methodologies used in prior studies (Byrne & Mitchell, 2013; Gardiner et al., 2016; Guan et al., 2022)—will be essential to ensure predictive credibility and stakeholder confidence.

Future research will focus on several key areas to extend the capabilities of the HDT platform. First, the development of more detailed and biologically accurate tree models is necessary to reflect species-specific responses to wind and precipitation. These models should incorporate variables such as growth stage, canopy density, soil saturation, and trunk flexibility, which influence tree stability under storm situations. Second, expanding customization options for different home types and lot configurations will enable more tailored risk assessments and more relevant recommendations for diverse users.

Another important direction involves simulating interactions between physical infrastructure and social systems. Understanding how tree-related disruptions cascade into impacts on transportation, emergency response, and energy networks will provide a more holistic comprehensive view of disaster resilience. Incorporating these dynamics into the HDT framework can help inform coordinated community-level planning and preparedness strategies.

Additionally, integrating real-time environmental and smart home data will significantly enhance the responsiveness and situational awareness of the system. Drawing on inputs from IoT sensors—such as wind gauges, rain monitors, and soil moisture detectors—can transform the HDT into an adaptive tool capable of updating forecasts and recommendations as weather conditions change. This real-time functionality will be particularly valuable for households and emergency managers preparing for imminent storms.

Finally, continued focus on model reliability, interpretability, and data ethics will be critical for long-term success. Implementing validation protocols, improving user understanding of simulation outputs, and adopting robust data governance practices will help ensure that HDTs are not only accurate but also trusted and accessible.

In conclusion, the HDT prototype presented in this study offers a compelling proof-of-concept for how digital twin technology can be applied to reduce property damage from tree blowdowns during hurricanes. With further refinement, including expanded modeling capabilities, real-time data integration, and stakeholder engagement, HDTs hold significant promise as tools for advancing disaster preparedness and urban resilience. As climate risks grow, systems like HDT can play a transformative role in building safer, smarter, and more adaptive communities.

7. REFERENCES

ADPC (Asian Disaster Preparedness Center). 2015. "Handbook for disaster recovery practitioners." Accessed May 20, 2019. http://www.adpc.net /igo/category/ID809/doc/2015-yDTg8K-ADPC-tgllhandbook.pdf.

Ayscue, J. K. (1996). Hurricane damage to residential structures: Risk and mitigation. National Hazards Research and Applications Information Center, Institute of Behavioral Science, University of Colorado, Boulder, Colorado.

Ball, M. F., Lagadec, L., & Laval, J. (2025). Hierarchical System of Digital Twins: A Holistic Architecture for Swarm System Analysis.

Bibri, S. E., Alexandre, A., Sharifi, A., & Krogstie, J. (2023). Environmentally sustainable smart cities and their converging AI, IoT, and big data technologies and solutions: an integrated approach to an extensive literature review. Energy Informatics, 6(1), 9.

Bibri, S. E., Huang, J., & Krogstie, J. (2024). Artificial intelligence of things for synergizing smarter eco-city brain, metabolism, and platform: Pioneering data-driven environmental governance. Sustainable Cities and Society, 108, 105516.

Bibri, S. E., Huang, J., Jagatheesaperumal, S. K., & Krogstie, J. (2024). The synergistic interplay of artificial intelligence and digital twin in environmentally planning sustainable smart cities: a comprehensive systematic review. Environmental Science and Ecotechnology, 100433.

Boschert, S., Heinrich, C., & Rosen, R. (2018, May). Next generation digital twin. In Proc. tmce (Vol. 2018, pp. 7-11). Las Palmas de Gran Canaria, Spain.

Byrne, K. M., & Mitchell, S. J. (2013). A model for predicting wind damage in forest ecosystems: Validation and comparisons. Canadian Journal of Forest Research, 43(7), 677–686. https://doi.org/10.1139/cjfr-2013-0072

Engeström, Y. (1999). "Activity theory and individual and social transformation." Perspectives on activity theory 19(38): 19-30.

Ersan, M., Irmak, E., & Colak, A. M. (2024, May). Applications, Insights and Implications of Digital Twins in Smart City Management. In 2024 12th International Conference on Smart Grid (icSmartGrid) (pp. 378-383). IEEE.

Faliagka, E., Christopoulou, E., Ringas, D., Politi, T., Kostis, N., Leonardos, D., Tranoris, C., Antonopoulos, C. P., Denazis, S., & Voros, N. (2024). Trends in digital twin framework architectures for smart cities: A case study in smart mobility. *Sensors*, *24*(5), 1665. https://doi.org/10.3390/s24051665

Fan, C., Zhang, C., Yahja, A., & Mostafavi, A. (2021). Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. International journal of information management, 56, 102049.

Finke, C., Groth, M., Schumann, M., Dewitz, P., Gehrke, J., & Marahrens, T. (2023). Design and Implementation of Hierarchical Digital Twins in Industrial Production Environments.

FLORIDA, H. I. EFFECT OF VEGETATION ON RESIDENTIAL BUILDING DAMAGE FROM HURRICANE ANDREW, AUGUST 1992.

Gardiner, B., Berry, P., & Moulia, B. (2016). Wind impacts on plant growth, mechanics and damage. Plant science, 245, 94-118.

Gilman, E. F., Duryea, M. L., Kampf, E., Partin, T. J., Delgado, A., & Lehtola, C. J. (2006). Assessing damage and restoring trees after a hurricane. University of Florida, Institute of Food and Agricultural Sciences: Gainesville, FL, USA.

Gkontzis, A. F., Kotsiantis, S., Feretzakis, G., & Verykios, V. S. (2024). Enhancing urban resilience: smart city data analyses, forecasts, and digital twin techniques at the neighborhood level. Future Internet, 16(2), 47.

Gourisaria, M. K., Jee, G., Harshvardhan, G. M., Konar, D., & Singh, P. K. (2022). Artificially intelligent and sustainable Smart Cities. In Sustainable Smart Cities: Theoretical Foundations and Practical Considerations (pp. 237-268). Cham: Springer International Publishing.

GovPilot. (2024, November 10). The rise of digital twins: How cities are creating virtual models to better serve residents.

https://www.govpilot.com/blog/the-rise-of-digital-twins-how-cities-are-creating-virtual-models-govpilot

Guan, D., He, H. S., & Wang, X. (2022). Empirical validation of airflow velocity attenuation in forest canopies under varying conditions. Forest Ecology and Management, 525, 120184. https://doi.org/10.1016/j.foreco.2022.120184

Hayes, P., & Kelly, S. (2018). Distributed morality, privacy, and social media in natural disaster response. Technology in Society, 54, 155-167.

Hexagon. (2025, August 8). Hexagon's urban digital twin platform. https://hexagon.com/go/sig/urban-digital-twin

Horita, F., D. Link, J. P. Albuquerque, and B. Hellingrat. 2016. "oDMN: An integrated model to connect decision-making needs to emerging data sources in disaster management." In Proc., 49th Hawaii Int. Conf. on System Science. New York: IEEE. https://doi.org/10.1109/HICSS .2016.361.

Juarez, M. G., Botti, V. J., & Giret, A. S. (2021). Digital twins: Review and challenges. Journal of Computing and Information Science in Engineering, 21(3), 030802.

Hughes, W., Lu, Q., Ding, Z., & Zhang, W. (2023). Modeling tree damages and infrastructure disruptions under strong winds for community resilience assessment. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 9(1), 04022057.

Kaptelinin, V., & Nardi, B. A. (2012). Activity theory in HCI: Fundamentals and reflections (Vol. 13). Morgan & Claypool Publishers.

King, E. (2025). Digital twins as space media. New Media & Society, 27(8), 4533-4548.

Koch, B., Heyder, U., & Weinacker, H. (2006). Detection of individual tree crowns in airborne lidar data. Photogrammetric Engineering & Remote Sensing, 72(4), 357-363.

Kuutti, K. (1996). "Activity theory as a potential framework for human-computer interaction research." Context and consciousness: Activity theory and human-computer interaction 1744: 9-22.

Lawther, P. M. (2016). Towards a natural disaster intervention and recovery framework. Disasters, 40(3), 494-517.

Lu, Q., Parlikad, A. K., Woodall, P., Don Ranasinghe, G., Xie, X., Liang, Z., Konstantinou, E., Heaton, J., & Schooling, J. (2020). Developing a digital twin at building and city levels: Case study of West Cambridge campus. Journal of Management in Engineering, 36(3), 05020004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000763

Patil, S. Kothari, M., Patrawala, M. & Madrewar, O. (2025). Tree health monitoring and management system using IoT. International Journal of Novel Research and Development, 10(5), d912–d917. https://ijnrd.org/papers/IJNRD2505393.pdf

Mishra, B. K. (Ed.). (2023). Handbook of Research on Applications of AI, Digital Twin, and Internet of Things for Sustainable Development. IGI Global.

Mishra, P., & Singh, G. (2023). Artificial intelligence for sustainable smart cities. In Sustainable Smart Cities: Enabling Technologies, Energy Trends and Potential

Applications (pp. 119-142). Cham: Springer International Publishing.

Patterson, O., Weil, F., & Patel, K. (2010). The role of community in disaster response: conceptual models. Population Research and Policy Review, 29, 127-141.

Paul, N., Galasso, C., Baker, J., & Silva, V. (2024). A predictive model for household displacement duration after disasters. Risk Analysis.

Peterson, C. J. (2000). Catastrophic wind damage to North American forests and the potential impact of climate change. Science of the total Environment, 262(3), 287-311.

PwC. (2023). How digital twins can make smart cities
better (Report). https://www.pwc.com/m1/en/publications/documents/how-digital-twins-can-make-smart-cities-better.pdf

Salisbury, B., & Koeser, A. (2023). Improving community hurricane resilience: A focus on trees and urban forests. [PDF]. U.S. Department of Homeland Security.

Smith, C. M., Langor, D. W., Myrholm, C., Weber, J., Gillies, C., & Stuart-Smith, J. (2013). Changes in white pine blister rust infection and mortality in limber pine over time. Canadian Journal of Forest Research, 43(10), 919-928.

Steppe, K., Von der Crone, J. S., & De Pauw, D. J. (2016). TreeWatch. net: a water and carbon monitoring and modeling network to assess instant tree hydraulics and carbon status. Frontiers in plant science, 7, 993.

U.S. Chamber of Commerce. (2025). Beyond the payoff: How investments in resilience and disaster preparedness protect communities. https://www.uschamber.com/security/beyond-the-payoff-how-investments-in-resilience-and-disaster-preparedness-protect-communities

Weil, C., Bibri, S. E., Longchamp, R., Golay, F., & Alahi, A. (2023). Urban digital twin challenges: A systematic review and perspectives for sustainable smart cities. Sustainable Cities and Society, 99, 104862.

World Economic Forum. (2023). Digital twin cities: Key insights and recommendations (Report). https://www3.weforum.org/docs/WEF_Digital_Twin_Cities_2023.pdf

Vessali, K., Galal, H., & Dr Nowson, S. (2022). How digital twins can make smart cities better. Real-time simulations can create a bridge between physical and virtual worlds', PwC.

Zaidi, A., Ajibade, S. S. M., Musa, M., & Bekun, F. V. (2023). New insights into the research landscape on the application of artificial intelligence in sustainable smart cities: a bibliometric mapping and network analysis approach. International Journal of Energy Economics and Policy, 13(4), 287-299.

Zhu, M., & Jin, J. (2025). Data-Driven Urban Digital Twins and Critical Infrastructure Under Climate Change: A Review of Frameworks and Applications. Urban Planning, 10.

AI & Machine Learning Deployment: Best Practices, Costs and Priorities

Nicholas Williams
williams.nicholas2100@gmail.com

Jeff Cummings
cummingsj@uncw.edu

Yu Wang
wangyu@uncw.edu

Yasin Emre Gokce
gokcey@uncw.edu

University of North Carolina Wilmington
Wilmington, NC 28405

Abstract

As artificial intelligence (AI) models become more prevalent across all fields, streamlined development of these models is becoming increasingly necessary. Deploying an AI model consists of integration with existing systems, monitoring various metrics related to the model, and maintenance of the model to keep it functional and up to date. Thus, successful deployment ensures value and sustainability. The objective of this research is to (1) identify the best practices within the phases of deployment, (2) explore cost requirements as well as strategies for savings, and (3) identify priorities during deployment. To explore these objectives, an exploratory study using semi-structured interview paradigm was developed and conducted with AI professionals with a qualitative analysis performed on the resulting transcripts. The analysis showed that participants emphasized explainable models that were accessible to users. Deployment costs were highly dependent on where the model was hosted and whether the model was developed in house or acquired from the commercial market. Finally, priorities were dependent on the type of model being developed, the users it would interact with, and the data it was handling. Regardless of these factors, all participants highlighted the importance of explainability, accessibility, and cost. These factors were prioritized by participants during model deployment.

Keywords: Artificial intelligence, machine learning, deployment priorities, best practices

Recommended Citation: Williams, N., Cummings, J., Wang, Y., Gokce, Y., (2026). AI & Machine Learning Deployment: Best Practices, Costs and Priorities. *Journal of Information Systems Applied Research and Analytics*, v19(n3) pp 17-27. DOI# https://doi.org/10.62273/ZVNR9663

AI & Machine Learning Deployment: Best Practices, Costs, and Priorities

Nicholas Williams, Jeff Cummings, Yu Wang and Yasin Emre Gokce

1. INTRODUCTION

Artificial Intelligence (AI) has quickly become technology's most enticing frontier. Driven by their ability to extract value from large volumes of data (Challoumis, 2024), AI tools have made this power more accessible across industries with increasing availability. However, despite its promises, AI implementation comes with its own set of challenges. One frequently cited challenge is rooted in the deployment of a trained model (Benbya, Davenport, & Pachidi, 2020; Paleyes, Urma, & Lawrence, 2023; Shankar, Garcia, Hellerstein, & Parameswaran, 2022). Even after successful deployments, it can be difficult to understand what a model needs to remain useful for an organization. This often requires monitoring and maintenance strategies that understand what to look for and how to respond effectively (Schober, 2022; Schröder & Schulz, 2022).

AI covers any time a machine tries to mimic human intelligence (Benbya, Davenport, & Pachidi, 2020), from rule-based expert systems to large language models. Technology in this category has been shown to make use of the copious amounts of data that are available in the modern world. AI demonstrates exceptional abilities in data analysis and augmenting human performance. It is able to assist with many repetitive workplace tasks and therefore, frees up people to take on more creative and innovative roles (Challoumis, 2024).

This paper focuses on AI, as well as a particular subset of AI classified as machine learning (ML). ML has numerous types of models that can be trained on data to make predictions. Ashmore et al. (2019) highlight the main steps in the ML process as follows: Data management, model learning, model verification, and model deployment. While data management, model training, and model verification are the foundation of a productive, well-performing model, this is not the end of the ML process. No matter what type of model is used, the deployment step is crucial to its initial and continued success.

While deployment is considered a crucial step, there have been a variety of approaches and suggestions to how this can be accomplished as well as the costs associated with it. This leads to the following research questions:

- How should companies effectively approach integration, monitoring, and maintenance of ML models?
- What costs are associated with implementing these practices and are they worth it?
- How should companies prioritize tasks within deployment?

2. BACKGROUND

ML, a powerful subset of AI that can be leveraged by many, is the practice of using algorithms to draw predictions from data. There are different types of ML (e.g., supervised, unsupervised, deep learning), yet they all follow the same basic steps as outlined by Ashmore et al. (2019). The first step is data management, which collects the data the ML model will be learning from. Next, the model learns to make predictions based on available data. Once satisfied with model predictions, the next step is model verification, which involves evaluating the model's performance on previously unseen data.

Once these steps are complete, the final step is deployment. Deployment involves 3 phases: (1) model integration, (2) performance monitoring over time, and (3) regular model maintenance and updates with new data. This study centers on deployment, with each phase discussed further in the subsequent sections.

Integration

Integration is the process of incorporating the ML model (i.e., the stored predictions) into the system it is to be providing information for. An organization must consider how an ML model will be integrated before attempting to deploy it. Even a well-trained and highly accurate model is useless if it cannot be incorporated into existing infrastructure. ML practitioners will encounter more infrastructure issues than expected during this initial deployment phase (Google Developers, n.d.).

A good first step is deciding whether the model

will be making live predictions or producing outputs that can be stored and periodically updated. This can guide an organization to realize what type of infrastructure will best fit its needs. Zinkevich (n.d) also notes that once an infrastructure has been decided on, a model should be deployed to test how well it integrates. At the time of this initial deployment, value gain from predictions should not be prioritized (Google Developers, n.d.); instead, the focus should be on ensuring compatibility and seamless integration. This initial setup may require upskilling of existing employees or adoption of new technology that can make model predictions readily available to the application or data scientists (Benbya, Davenport, & Pachidi, 2020; Challoumis, 2024). Because of the cost both in time and money, organizations need to have an effective approach to this phase of deployment.

Monitoring

Once a model has been integrated, the next phase is monitoring. A common approach to monitoring the model is in-house. Schröder & Schulz (2022) present a variety of metrics that may be relevant to any given ML model: performance, robustness, confidence, economic, interpretability, and ethical metrics. Among these, performance metrics are most commonly used, providing insight into the accuracy of the model's predictions. The type of model and what it predicts will determine the performance metric (e.g., accuracy, precision, etc.).

Monitoring model performance not only allows a user to see how well a model is doing, but can also be a strong counter to both problems related to drift (i.e., data and concept drift). Data drift is the concept that the data given to an ML model for it to make predictions will change over time (Ackerman et al., 2021). This is a very common occurrence within ML, as the world is ever changing. Concept drift is a phenomenon in which the performance of an ML model decreases over time, despite having the necessary up-to-date data (Schober, 2022).

While data drift is more specific to what data the model sees during training, concept drift describes a scenario in which the model was trained with a certain target in mind, but the target changes. This change in target could be due to a sudden change in the environment, a change in standards, or many other reasons. Concept drift is countered via retraining, just as data drift is. To deal with concept drift, the target must be defined in a way that demonstrates what is expected of the model. By seeing data that captures the changes to the desired outcome, the model can learn what predictions will be accurate for this new target (Schober, 2022). Thus, one area that needs to be explored is how practitioners approach drift across different domains.

Beyond the detection and prevention of drift, monitoring also has other utilities. Keeping track of how quickly a model's performance degrades can enlighten ML practitioners to how often a model should be retrained, allowing for the creation of scheduled retraining. Keeping track of metrics that are important to an organization can also catch any mistakes a model may be producing before it goes live and has these errors reported by users. Even if a model appears to be performing correctly in its accuracy checks, an ethics check may reveal a bias that would damage the organization's reputation if customers were to see this.

Maintaining and Updating

Over time models can degrade (Patel, 2025), thus strategies to maintain or update a model must be considered. To update a model, two approaches are widely used: scheduled regular retraining and continual learning (Paleyes, Urma, & Lawrence, 2023). Scheduled retraining decides on a fixed amount of time between retraining the model with the latest data. This schedule should balance training frequency (enough to mitigate performance loss due to drift) and resource usage so that the retraining does not consume more resources than necessary.

Continual learning, by contrast, updates the model as it gets new data in. Many recommender systems use continual learning, as things change frequently with new data availability (Lee & Lee, 2020). The task an ML model is designed for will guide which type of updating it will work best with.

Cost

The value AI/ML provides can be very alluring, but it comes at a cost. There are many expenses that are outside the scope of this paper (e.g., data acquisition and storage, initial model training, personnel hours, etc.). The estimated cost of developing and implementing an AI/ML solution varies greatly from $10,000 to $1,000,000 (Shashkina, 2025). This large variation in price is largely chalked up to the complexity of the model and how long a company is willing to spend in the exploration phase.

Organizations have the choice to buy commercially available products or create their own models. There is also the decision of where to host, train, and update their ML model, either locally or via a cloud service. Each approach comes with benefits and issues, especially across the three phases of deployment.

Monitoring how much a model is costing is a wise and responsible step to take. If the ML model is being housed in a cloud vendor such as AWS or GCP, budgets can be set. These types of services usually provide a way to keep track of spending and offer automatic alerts. If the model is being housed on-site, metrics to keep track of compute time and other associated costs can be set up (Schröder & Schulz, 2022).

Deployment Best Practices and Priorities

Because of the vast differences discussed in the phases and costs associated with deployment, the goal of our study is to help develop the best practices and priorities that impact the cost and effectiveness of organizational AI/ML. Without proper implementation, anything the model provides becomes inaccessible and useless. Without proper monitoring, problems will go unnoticed, and the value provided by the model will begin to decline. By knowing what leads to smooth deployment operations, effective and enduring AI models can be incorporated anywhere.

3. METHODOLOGY

To address the research questions, a qualitative interview approach was used. Because of the challenges to capture the knowledge of experts in complex domains (i.e., AI) (Vasileiou, et al., 2018), the decision was made to focus on depth and conceptual understanding of the process by targeting specific roles (e.g., machine learning engineers) that may lead to more detailed and context-specific insights as opposed to interviewing everyone involved in AI (Turner & Hagstrom-Schmidt, 2022). Thus, the study focused on extensive interviews with 3 expert participants in the field with varying backgrounds and industries to provide diverse insights. Prior research has found that when participants are selected based on a high degree of expertise and role similarity, thematic saturation (i.e., identification of most major themes) may occur with as few as 3 interviews, especially when the topic is narrowly focused (Guest et al., 2006).

Participants

Participants were AI professionals with the responsibility of deploying AI/ML models. Participant 1 (P1) holds a managerial role with 3 years of AI experience, overseeing AI projects and a team of software engineers. Participant 2 (P2) is a data scientist with 2 years of professional AI experience. Participant 3 (P3) is an AI architect with 1 year of professional AI experience. All participants have directly contributed to the deployment of at least 2 AI projects.

Protocol

Interviews were conducted in a semi-structured fashion. Interview questions were constructed based on the research questions and best practices (i.e., open-ended, neutral, and clear (McNamara, n.d.; Turner & Hagstrom-Schmidt, 2022). The list of interview questions can be found in Appendix A. Interviews were conducted via Zoom, with an interaction time of 30-60 minutes. Conversations were not recorded for the privacy of the individual, but full transcriptions were collected and stored for analysis. Transcriptions were redacted of personally identifiable information and stored securely. This is based on similar interview paradigms implemented in past research (Shankar et al., 2022).

Analysis

The standard for qualitative research analysis is transcript coding (i.e., qualitative content analysis) (Shankar et al., 2022). This practice extracts common themes across interviews. MaxQDA, a commonly used qualitative data analysis software, was employed in this study. Deidentified transcripts recorded from Zoom were imported into the software. Coding passes were then performed on each transcript using a top-down approach. 8 codes were derived from the research questions and literature and applied to relevant segments within each interview. A list of codes can be found in Table 1. In total, 139 segments across the 3 interviews were given codes. Common themes, unique approaches, and surprising contrasts are presented in the following section.

Codes	Segments	%
Cost	39	28.06
Priorities	24	17.27
Maintenance	18	12.95
Monitoring	18	12.95
Integration	16	11.51
Demographic	10	7.19
AI process	8	5.76
Strategy & Governance	6	4.32

Table 1. Qualitative Content Analysis Codes

4. RESULTS & DISCUSSION

In this section, the results of the qualitative content analysis are presented and discussed. First, best practices related to the phases of deployment are covered. Next, cost considerations and management strategies are discussed. Finally, we elaborate on what should be prioritized when looking to deploy AI/ML models.

Integration
The goal of integration is to make the AI/ML model or its outputs accessible to users. In the case of an AI chatbot, this would be making sure users are able to speak with it, whereas for a sales forecasting ML model, this would be making sure the predictions can be seen by the stakeholders to make relevant business decisions.

Participants identified integration as a critical first step in AI/ML deployment. Participants specifically focused on the explainability of the models, which refers to the ability to understand how a model generates its outputs. It did not matter if an interviewee was referring to a pre-trained AI chat model or an in-house trained ML prediction model; understanding how an output was reached was always cited as an important factor.

"You don't really want to focus in on one variable when you're explaining the model, you kind of want to tell a story about all the significant variables at once." P2

P2 continued to describe how understanding a model's thought process can easily translate into providing logical, data-driven justifications for model outputs that domain experts can understand and agree with. In this case, ML models are strong tools for pattern recognition to assist human judgment. P2 also valued explainability above model performance, indicating that performance can be increased if the model can be explained, but a model that cannot be explained will be much more difficult to improve. This sentiment concerning explainability was echoed by the other participants. P1 cited tools like LangGraph and LangSmith for their ability to demystify a model's chain of reasoning. Being able to see where things went wrong allows the developers to adjust that parameter in a way that steers the model toward the desired outcome.

P2 was the only participant to talk about security. This is an important factor to consider, especially during integration. This is the phase where a model is about to be accessible to more than just the data scientists. Many AI/ML projects deal with potentially sensitive data. The handling of the data being fed to the model must be secure, and the outputs of the model must also be secure. Secure practices should be implemented every step of the way, including during deployment.

Other best practices mentioned by the participants included combining models to make an ensemble model to increase decision confidence and performing stress testing to evaluate system resilience under expected amount of traffic during integration.

Monitoring
As previously mentioned, performance is a metric that predictive models use to know how well they are doing their task. While this has been cited in previous research as the most important due to its direct link to the model's value, participants in our study claim that explainability metrics are even more important. Knowing how a model arrives at its final output allows data scientists to give raw data to support correct decisions, as well as debug incorrect decisions.

Other features that are directly tracked include: how long users interact with chatbot AI models, latency of model response, token usage, and cost. For indirect monitoring, P1 explains an interesting process in which an AI model is integrated; users provide feedback on their experience with it, then the development team recruits a separate AI model to perform sentiment analysis on user feedback. This allows a way to quickly get a feeling for what is and is not meeting users' expectations, allowing for rapid fixes that satisfy people who use the AI model. P3 mentions that they track what users are engaging with their AI chatbot model for. This is done to identify any common tasks that users may frequently ask the model to perform. Once those are identified, the model can be tuned to handle those common tasks more efficiently, decreasing compute costs.

Maintenance
Maintenance covers the steps needed to keep a model working properly, up-to-date, and expanding functionality. Just like all technology, new pre-trained, commercially available models are regularly being released. Participants stated that they did not always immediately update to the latest model. Instead, they evaluate if the increase in performance is worth the increase in

price. If not, they continue to use the slightly older model that is still providing satisfactory performance. Conversely, it was also mentioned that money can potentially be saved by updating to the latest model if the performance and cost differences of a new model warrant the switch.

P1 also mentioned the usage of an "automated control test suite". This is an automated test for a group of use cases for which the model should provide accurate results every time. It is used any time an update is made to ensure that basic functionality has not been broken.

For in-house models, participants mentioned that models should be retrained regularly to avoid any type of drift, with retraining dependent on the task being performed. P2 gave the example of a model that is used to assist in stock trading, which should be updated daily, at a minimum. This could be juxtaposed to a model that is used for annual sales forecasting, which may only need to be retrained once per quarter. P2 also discussed backtesting for model updates. This process trains a model on historical data. The model is then tested on more historical data so its performance can be immediately evaluated. This is useful in maintenance because it would be bad if an updated model were integrated but then performed worse than its previous iteration.

Backtesting is a way to vet an updated model before presenting it to end users. P3 indicated that the models they built did not yet have a need for maintenance. The models were all under 3 months old and used for tasks where new data was not greatly different than old data. This should factor into developing a maintenance strategy.

Deployment Best Practices Summary

Based upon the interviews conducted, the following summarizes the results across the phases of deployment. Organizations should approach integration, monitoring, and maintenance of models by focusing on accessibility, explainability, and security, while tailoring strategies to specific use cases.

Explainability should be prioritized during integration to help stakeholders understand and trust the decisions generated by the models. Security during integration safeguards sensitive data, while advanced practices like ensemble modeling and stress testing improve reliability. Monitoring tracks many different aspects, such as performance metrics, user feedback, and engagement patterns, to refine models and optimize their functionality.

Explainability, again, plays a key role in debugging and decision-making, surpassing the importance of performance alone. Maintenance strategies should involve regular retraining to prevent drift, automated testing to verify functionality, and cost-effective evaluation of updates to pre-trained models. By adopting these approaches, companies can ensure that their AI models remain valuable decision-making tools.

Cost

Cost is one of the main factors that a business will consider when using AI models. The interview participants had very diverse approaches to powering their companies with AI, leading to very different allocations of resources. The two factors that determined where money was focused the most were the origin of the model and how the model was hosted.

Models can originate from within the company or be acquired from an outside vendor (i.e., build vs. buy). P1 uses pre-trained, commercially available LLM models that can be tuned and adjusted to the specific task they require. The main justification for choosing to buy instead of build was that the AI landscape is changing at a pace that is very difficult to keep up with:

"We're subject to the leapfrog effect, right? So, by the time you have invested the time and resources to train a model, the commercially available models have already bypassed you … again and again, we've seen that happen and we've seen competitors try [to keep up with] that and then fall short." P1

Using commercially available models was described as "pay as you go". Cost scales with token usage, latency/speed of response, and amount of data transferred. One benefit of this approach is that most of the cost is upfront. Once the model has been paid for, monitoring and maintenance are inexpensive, as they are provided by the vendor.

Another benefit of buying a commercially available model is that it is ready for production much sooner. Many commercial models have out-of-the-box capabilities, providing value as soon as they are purchased. This trade-off of high up-front cost with immediate usage vs the lower costs but slower time to market of models built in-house is one to consider. Opting to buy commercially available models appears to be worth considering if the business plans for

scaling up in size over time would benefit from having the most recent models available, or if immediate responses are necessary.

In contrast, P2 chose to build models in-house instead of buying commercially available models. Importantly, these in-house models were predictive ML models, not LLMs like those used by P1. P2 claims that integration is cheaper, while maintenance is more expensive. This is in direct contrast to P1, who said their greatest expense was in integration, and maintenance was not very costly. Since P2 does model retraining using cloud computing, the computing cost must be paid every time a model is updated.

P3 has an entirely different experience with cost, as they buy hardware, completely avoiding cloud computing costs. There is a cost in acquiring the hardware which must be covered every time the organization scales up by adding a new AI project. However, by training and maintaining their models locally, they do not have to pay every single time they want to update their models. This approach is best suited for those who can acquire hardware cheaply or who do not plan to have a vast amount of AI products in their organization.

Participants suggested an artful balance must be struck between processing power and time. This is especially true for those who decide to use cloud computing to power their AI models. Better processing can complete tasks in a shorter amount of time, at the cost of a higher rate. If time is not a critical factor, less powerful processing can be used, resulting in a cheaper rate, but increased time to complete the task.

Participants also mentioned multiple tradeoffs where there was potential to save money. P1 leverages batch jobs to save money when time is not a critical factor. Thus, if you can pay over time, you do not need to spend as much money. Another option to consider is looking for open-source software. As the AI landscape develops daily, more and more solutions are becoming available. P1 cited this as a strategy that is considered when possible and reported that money was saved when these solutions were employed.

	P1	P2	P3
Model origin	Commercial	In house	In house
Hosted	Cloud	Cloud	Local
Most expensive phase	Integration	Maintenance	Integration

Table 2. Participant Experience & Cost

Priorities

There are many variables to consider when looking to implement an AI project. The first to consider is whether the model will be built in-house or outsourced via a commercially available model. Important factors to consider when making this decision include current availability of resources, the pacing of the development team, and how important accessibility is.

As previously mentioned, using commercially available models incurs greater cost up front. If this can be afforded, paying for a commercially available model is a viable route. It has also been mentioned that the AI world is constantly evolving, making it difficult to keep up with. If the business must be using the most advanced, cutting-edge AI models, they will either need to have a team to support this rapid development cycle or turn to commercially available models. The availability of models is a unique problem that was only mentioned by P3. This interviewee works in a remote, rural area where natural disasters frequently cut off communication to the outside world. Since some of the models built here provide important information, cloud hosting was not an option. This is an important reminder that any models that have outputs or interactions with critical systems must be available and not reliant on the cloud.

There are some common goals for models, regardless of the domain or particular use case they are applied to. These include successful integration, explainability, alignment with business goals, and security. Integration has been covered extensively, but the importance of users being able to interact with the model or its outputs cannot be understated. Explainability has also been discussed in depth, as participants highlighted it as the most important factor of a deployed model. To have an explainable model means that undesired outputs can be traced to the point of failure and subsequently adjusted.

Furthermore, if a model's steps can be traced, that can be translated into valuable information that non-technical members can benefit from

hearing. From here, priorities become domain-specific. If a model is going to be interacting with customers, latency must be considered, as they do not tolerate slow response times, according to participants. Alternatively, if a model is only going to be used internally, response speed may not be as great a concern.

The domain in which the model will be operating also provides context for maintenance. This was a difference observed between two of the interviewees. P1 had a customer-facing model that needed to be kept up to date. In contrast, P3 had internal models that did not work with data that changed frequently. This led P1 to prioritize maintenance more than P3, who said, "Model drift is not a … concern for us currently." They continued to explain that due to the invariability over time of their data, there is not much pressure to regularly retrain or update models.

Security is a concern that must be addressed at every phase of a model's life. participants briefly mentioned security, mostly related to ensuring the data is only accessible to the necessary parties. Security must also be checked to ensure that the model is not able to communicate any sensitive information to people who should not have access to it.

Automation is an important part of AI models, as it saves a lot of time and keeps things up to date. When first releasing a model, automation does not have to be a priority, as the model will still serve its purpose without automatic updates. Participants described how this practice is not vital for release, but will quickly become important, as the model's lifecycle continues.

With these priorities considered, a strategy can be devised that will promote initial success and provide steps for a model to have a long, sustainable life. Some priorities are dependent on what domain the model is operating in, such as latency, which is important for customer-facing models. However, successful integration and being able to explain how a model got to its output are steps that are crucial for success, regardless of the operating domain.

5. CONCLUSION

In this study, results were presented from a semi-structured interview of AI deployment practitioners. Findings suggest is paramount, while monitoring and maintenance are later, yet still important concerns. Costs within deployment were dependent on whether a model was built in-house or by a commercial vendor. In-house models result in a slow increase in price when hosted via a cloud provider, as maintenance and updating incur compute costs. Commercially available models have the bulk of the cost upfront, as maintenance and updating are handled by the provider. Priorities for any type of model revolved around accessibility, explainability, and cost.

Overall, the study provides valuable insights into the practical aspects of AI/ML deployment and identifies approaches for organizations looking to implement these technologies effectively. The findings contribute to the existing literature by providing a detailed analysis of real-world experiences and challenges faced by AI/ML professionals.

6. FUTURE WORK AND LIMITATIONS

Future work could build on this work by examining the power of explainability and researching strategies to ensure this is achieved. This study was limited to a small sample size. Future studies could expand on these ideas across multiple industries. A broader scope could be used to reinforce the findings presented.

A framework for necessary and highly important steps and decisions within deployment could be developed. Part of this framework could include AI auditing, which involves reviewing algorithms for fairness, compliance, accountability, etc. While this topic was beyond the scope of this study, it should be researched and considered in a responsible deployment framework.

Security could be researched as it relates to the steps of deployment, with secure practices laid out for practitioners. Security may also fit into the previously mentioned potential concerns about using commercially available models. If sensitive data must be communicated to third-party vendors, there could be potential for vulnerabilities to arise.

This study was limited to interviews of the experiences of a few people. Future research could take a quantitative approach, especially when examining cost. This could incorporate the data of a great number of individuals and report concrete numbers. By addressing these areas, future research can further enhance our understanding of AI/ML deployment and contribute to the development of more effective and sustainable AI solutions.

7. REFERENCES

Ackerman, S., Raz, O., Zalmanovici, M., & Zlotnick, A. (2021). Automatically detecting data drift in machine learning classifiers (arXiv:2111.05672). arXiv. https://doi.org/10.48550/arXiv.2111.05672

Ashmore, R., Calinescu, R., & Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges (arXiv:1905.04223). arXiv. https://doi.org/10.48550/arXiv.1905.04223

Benbya, H., Davenport, T., & Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. MIS Quarterly Executive, 19, 9–21. https://doi.org/10.2139/ssrn.3741983

Challoumis, C. (2024, October). The economics of AI - How machine learning is driving value creation. https://doi.org/10.5281/zenodo.13929032

Deepchecks Community. (2025, February 6). Model versioning for ML models: A comprehensive guide. Deepchecks. https://www.deepchecks.com/model-versioning-for-ml-models/

Google Developers. (n.d.). *Rules of machine learning: Best practices for ML engineering*. Google. Retrieved June 15, 2025, from https://developers.google.com/machine-learning/guides/rules-of-ml

Guest, G., Bunce, A., & Johnson, L. (2006). *How many interviews are enough? An experiment with data saturation and variability*. Field Methods, 18(1), 59–82. https://doi.org/10.1177/1525822X05279903 Lee, C. S., & Lee, A. Y. (2020). Clinical applications of continual learning machine learning. The Lancet Digital Health, 2(6), e279–e281. https://doi.org/10.1016/S2589-7500(20)30102-3

McNamara, C. (n.d.). General guidelines for conducting research interviews. Retrieved June 1, 2025, from https://management.org/businessresearch/interviews.htm

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2023). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1–29. https://doi.org/10.1145/3533378

Patel, H. (2025, February 11). ML model deployment challenges. Censius. https://censius.ai/blogs/challenges-in-deploying-machine-learning-models

Schober, A. (2022, September). *What is concept drift and how to detect it*. Motius. Retrieved [June 1, 2025], from https://www.motius.com/post/what-is-concept-drift-and-how-to-detect-it.

Schröder, T., & Schulz, M. (2022). Monitoring machine learning models: A categorization of challenges and methods. Data Science and Management, 5(3), 105–116. https://doi.org/10.1016/j.dsm.2022.07.004

Serban, A., van der Blom, K., Hoos, H., & Visser, J. (2020, October). Adoption and effects of software engineering best practices in machine learning. In Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (pp. 1–12). https://doi.org/10.1145/3382494.3410681

Shankar, S., Garcia, R., Hellerstein, J. M., & Parameswaran, A. G. (2022, September 16). Operationalizing machine learning: An interview study (arXiv:2209.09125). arXiv. https://doi.org/10.48550/arXiv.2209.09125

Shashkina, V. (2025, February 11). Machine learning (ML) costs: Price factors and real-world estimates. ITRex. https://itrexgroup.com/blog/machine-learning-costs-price-factors-and-estimates/

Turner, D. W., III, & Hagstrom-Schmidt, N. (2022, January). Appendix: Qualitative interview design. https://odp.library.tamu.edu/howdyorhello/back-matter/appendix-qualitative-interview-design/

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. BMC Medical Research Methodology, 18(1), 148. https://doi.org/10.1186/s12874-018-0594-7

Yasenchak, E. (2025, February 3). The challenges of deploying machine learning models: Best practices. Medium. https://medium.com/@emyasenc/the-challenges-of-deploying-machine-learning-

models-best-practices-7c616a5a07d2

Zinkevich, M. (n.d.). *Rules of machine learning: Best practices for ML engineering*. Google Developers. https://developers.google.com/machine-learning/guides/rules-of-ml

APPENDIX A.

Semi-Structured Interview Questions

1. What type of AI do you use?
2. Can you give a general overview of your AI process from conception to deployment? (A, B)
3. How do you implement model predictions into your system? (A)
4. Do you track any aspects of your AI after it has been deployed? (A)
    a. What metrics do you use to monitor each aspect? (A)
5. Do you update your AI? If so, how? (A)
    a. What prompts the need for an update? (A, B)
6. Have you used any techniques that ended up being a waste of resources in hindsight? (A, B, C)
7. Were there any practices/tools/strategies that were costly to implement (C), but worth the spending in the long run? (A, B, C)
8. Could the model be live in production without one or more of these steps? (A, B, C)
    a. Would you see the same amount of value if those steps were skipped?
9. What are the priorities when deploying an AI model? (B)

Computer-Supported Collaborative Learning: An Analysis of the Relationship Between Human Critical Thinking and the Use of Artificial Intelligence (AI)

Temitope Elijah
te06466@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

Hayden Wimmer
hayden.wimmer@gmail.com
Georgia Southern University
Atlanta, GA 30302

Carl Rebman Jr
carlr@sandiego.edu
University of San Diego
San Diego, CA 92110

Abstract

This study explores the relationship between human critical thinking and the use of AI in CSCL environments. The integration of Artificial Intelligence (AI) into Computer-Supported Collaborative Learning (CSCL) environments represents a transformative development in education, reshaping traditional learning paradigms. CSCL leverages technology to facilitate collaboration among learners, enhancing critical thinking, problem-solving, and a deeper understanding of educational material. The introduction of AI tools such as intelligent tutors, adaptive learning systems, and automated feedback mechanisms further enhances these processes by offering personalization and improving group dynamics. The findings from this study revealed that AI tools can significantly enhance collaborative processes, including teamwork, communication, and conflict resolution. These results emphasize the need for balanced integration of AI tools to ensure they complement rather than replace human cognitive engagement. This research seeks to offer valuable insights for educators, policymakers, and developers aiming to optimize the role of AI in education while safeguarding critical thinking and learner autonomy.

Keywords: Artificial intelligence, Computer-Supported collaborative learning, Critical thinking, Education, Collaboration, Communication.

Recommended Citation: Elijah, T., Wimmer, H., Rebman Jr., C.M., (2026). Computer-Supported Collaborative Learning: An Analysis of the Relationship Between Human Critical Thinking and the Use of Artificial Intelligence (AI). *Journal of Information Systems Applied Research and Analytics,* v19(n3) pp 28-40. DOI# https://doi.org/10.62273/VDKT7359

Computer-Supported Collaborative Learning: An Analysis of the Relationship Between Human Critical Thinking and the Use of Artificial Intelligence (AI)

Temitope Elijah, Hayden Wimmer and Carl Redman

1. INTRODUCTION

Education is continuously evolving with advancements in technology, and one of the most transformative developments in recent years is the integration of Artificial Intelligence (AI) into collaborative learning environments. Computer-Supported Collaborative Learning (CSCL) is a teaching approach that leverages technology to facilitate collaboration among learners, fostering critical thinking, problem-solving, and deeper understanding. By enabling interaction and collaboration beyond physical and temporal barriers, CSCL has revolutionized traditional learning paradigms, making education more accessible and inclusive. Hernández-Sellés et al. (2019). The intersection of CSCL and AI presents significant opportunities and challenges. AI tools such as intelligent tutors, adaptive learning systems, and automated feedback mechanisms have enhanced the ability of learners to engage in meaningful discussions and collaborative activities. However, their role in enhancing or decreasing human critical thinking remains an area of active investigation. Critical thinking characterized by the ability to analyze, evaluate, and synthesize information to form reasoned judgment is a cornerstone of effective learning. Understanding how AI influences these processes is crucial to optimizing educational practices and technology design. Tedla and Chen (2024)

This research is an exploratory study which aims to analyze the relationship between human critical thinking and the use of AI in CSCL environments. It investigates how AI tools shape learners' cognitive and collaborative behaviors, whether they enhance or diminish critical thinking capabilities, and how educators can strike a balance between AI support and human cognitive autonomy. A quantitative methodology was adopted, employing a survey of 105 participants to assess perceptions of AI's impact on cognitive and collaborative processes. Data analysis was conducted using SPSS, focusing on key variables such as the influence of AI on critical thinking, collaborative learning, and teacher involvement.

The findings offer valuable insights into how AI tools enhance collaborative processes while posing potential risks to independent cognitive engagement. The integration of AI tools into CSCL environments has a profound impact on both collaborative learning and critical thinking. Participants highlighted the benefits of AI in promoting effective communication, enabling equal participation, and providing real-time feedback. AI tools facilitated smoother group interactions and improved problem-solving capabilities, indicating their potential to enhance the overall learning experience. However, the findings also underscore the dual nature of AI's influence. While AI simplifies complex tasks and offers structured guidance, over-reliance on these tools can reduce opportunities for deep cognitive engagement. Participants noted a tendency for AI to replace some aspects of critical analysis, which risks diminishing the development of independent thinking skills essential for academic and professional growth.

2. BACKGROUND

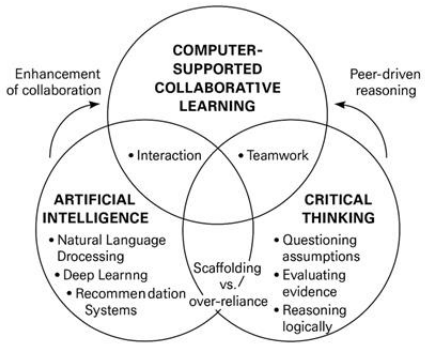


Figure 1: Conceptual framework showing the interplay between CSCL.

Figure 1 highlights how CSCL, AI, and Critical Thinking intersect. CSCL fosters collaboration through interaction, teamwork, and shared problem-solving. AI contributes by enhancing collaboration with tools such as NLP, deep learning, and recommendation systems**,** offering

scalability and personalization, but also raising concerns of over-reliance. Critical Thinking develops as learners question assumptions, evaluate evidence, and reason logically, supported by CSCL's peer-driven dialogue. The overlap emphasizes that while AI can scaffold learning, its integration must safeguard autonomy and deeper cognitive engagement

The integration of AI into CSCL holds immense potential to revolutionize collaborative learning by enhancing personalization, efficiency, and scalability. However, understanding how AI impacts learners' critical thinking, group dynamics, and overall educational outcomes remains underexplored. Critical thinking (CT) is a vital competency for success in academic, professional, and social contexts. It involves the ability to question assumptions, evaluate evidence, and reason logically. Within CSCL, CT is cultivated through interactions that challenge learners to defend their ideas, critique others' reasoning, and refine their arguments (Tedla & Chen, 2024).

While AI tools offer scaffolding that supports these processes, there is a risk of learners becoming passive recipients of AI-generated content rather than active participants in the learning process. Studies suggest that excessive dependence on AI may lead to surface-level engagement and hinder the deeper cognitive processes required for critical thinking. Thus, understanding the relationship between AI and critical thinking in CSCL is imperative for designing systems that empower, rather than decreasing cognitive growth and critical thinking (Warsah et al., 2021).

3. LITERATURE REVIEW

Computer-Supported Collaborative Learning (CSCL) has undergone significant advancements with the integration of Artificial Intelligence (AI), The integration of technology in collaborative learning environments has transformed educational practices, reshaping the way critical thinking and collaborative learning improve learning outcomes Rosé et al. (2008). A range of studies has examined various aspects of CSCL and the role of AI in enhancing collaborative processes. This literature review provides an in-depth analysis of the relationship between human critical thinking and the use of AI in CSCL environments, synthesizing findings from various studies.

Liu et al. (2023) focused on the use of technology in educational environments, particularly in facilitating collaborative learning through tools like concept mapping. The researcher focused on whether students' attitudes play a role in learning processes. The findings indicated a positive correlation between collaborative perceptions and knowledge acquisition, though factual knowledge understanding remained unaffected (Cress et al., 2015). This suggests that fostering positive collaborative attitudes is vital for deeper engagement and learning outcomes.

Ada (2009) explored the role of CSCL in enhancing higher-order thinking skills within the context of textile studies. Higher-order thinking skills are cognitive processes such as analysis, evaluation, synthesis and creativity, which go beyond basic memorization or understanding and indicated that technological integration in educational settings improves learning outcomes. Collaborative environments were found to nurture these skills through high levels of social interaction and co-creation of knowledge, suggesting that integrating technology enhances cognitive processes essential for critical thinking. The results showed a positive link between the quality of group collaboration and the development of cognitive skills. High levels of social interaction and collaboration contributed to the establishment of a community of learning, nurturing a space for fostering higher order thinking through co-creation of knowledge processes (Radkowitsch et al., 2020).

Tedla and Chen (2024) conducted a meta-analysis on CSCL's impact on students' critical thinking, finding a moderate to large effect size. Factors such as group size and task complexity influenced outcomes, demonstrating that structured and interactive digital tools significantly promote critical thinking. The meta-analysis findings show that CSCL has a moderate effect on students' critical thinking skills. The results suggest that interactive, collaborative elements are effective components to promote critical thinking in computer-mediated environments across be many educational levels and subjects, under four specific conditions including group size and task complexity which might modulate the effectiveness of those effects. The overall ES estimate of the impacts of CSCL on CT was assessed using a random-effects model, and it was recorded as large (ES=0.854).

Warsah et al. (2021) examined and the impact of collaborative learning (CL) on learners critical thinking skills in addressing Islamic radicalism

and their critical thinking retention and, investigated learners' perspectives on collaborative learning by using a mixed approach of 40 learners. The findings show that Learners taught by using collaborative learning experienced better critical thinking improvement and have good retention of their critical thinking skills compared to those taught means of lecturing.
Hu et al. (2022) compared different collaborative learning patterns and investigated effective patterns of group collaborative learning that were used in a digital AI course to promote fourth graders creative thinking and explored the difference between the four patterns in their promotion of students creative thinking in a seven-week teaching practice. Their results indicated that students that engaged in more interactive and collaborative groups demonstrate higher levels of critical thinking skills compared to those students that engaged in less interactive groups.

McLaren et al. (2010) examined the use of artificial intelligence (AI) techniques to support collaborative learning and e-discussions in educational settings. The study highlighted the importance of collaborative learning, where students work together to solve problems and discuss concepts. However, facilitating effective collaboration and ensuring productive discussions among students can be challenging, as instructors may struggle to manage and assess student interactions. To address this, the authors propose that AI can play a role in monitoring, guiding, and assessing collaborative learning environments. The result shows that AI plays a vital role in supporting collaborative learning by enhancing interaction quality, assisting instructors in managing discussions, and ultimately contributing to improved educational outcomes (Järvelä et al., 2015).

Ramirez (2021) investigated the effects of a CSCL environment with and without question-asking scripting on the development of conceptual understanding and critical thinking in science among middle school students. By comparing these CSCL approaches, the research sought to identify whether scripting activity specifically contributes to improved learning outcomes in critical thinking and science comprehension. The result shows that the CSCL environment significantly impacted students' conceptual understanding and critical thinking skills, with the greatest benefits observed in the group exposed to CSCL with question-asking scripting (Chu et al., 2024).

Chou et al. (2022) explored how human-computer interaction experiences and ICT self-efficacy influence the effectiveness of AI-based learning technologies and discusses the importance of artificial intelligence (AI) in education, particularly in information and communications technology (ICT) teaching, as AI technology becomes increasingly integral to modern learning environments. Their results indicated that the human-computer interaction experience significantly and positively relates to the effectiveness of AI-based technology applications, and that students' ICT self-efficacy has an indirect correlation with the learning effectiveness of AI-based technology application through the human-computer interaction experience (Andersen et al., 2022).

CSCL has been recognized as an effective approach to fostering higher-order cognitive skills such as analysis, evaluation, and synthesis (Ada, 2009). The dynamic and interactive nature of CSCL enables learners to co-construct knowledge through social interaction, which has been found to promote deeper understanding and critical thinking Altinay and Paraskevas (2007). By creating collaborative environments, CSCL allows students to engage in problem-solving and reflective activities that are essential for cognitive development (Hernández-Sellés et al., 2019).

McLaren et al. (2010) developed an AI-driven tool for e-discussions, which facilitated meaningful interactions and improved critical thinking outcomes by identifying unproductive patterns and offering real-time feedback. Similarly, Lee (2015) demonstrated that structured collaboration scripts, guided by AI, significantly enhanced students' reading literacy and critical engagement, suggesting that AI can adaptively refine collaboration frameworks to suit learners' needs.

4. METHODOLOGY

The research design was using quantitative methodology where we utilize a survey as the primary data collection tool to explore the relationship between human critical thinking and the use of artificial intelligence (AI) in computer-supported collaborative learning (CSCL). Subjects were recruited from the university student community. Standard testing was conducted for factor analysis and homogeneity of variance prior to proceeding with the analysis presented here within.

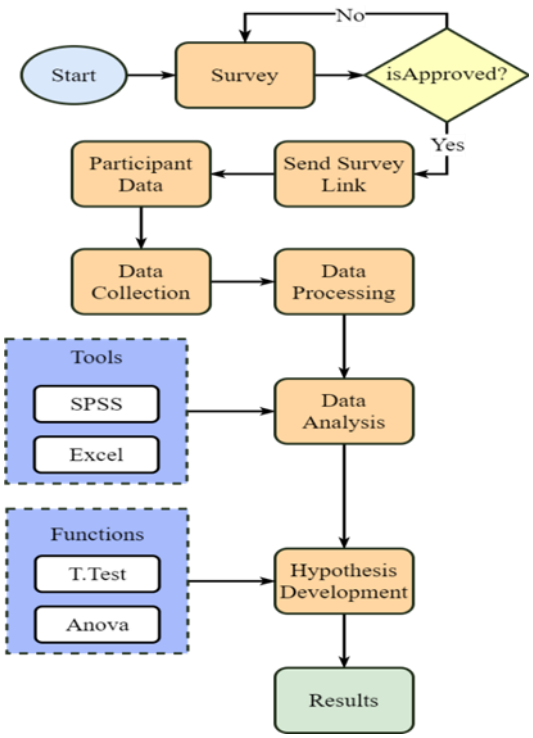


Figure 2: CSCL Flowchart

4.1 Hypothesis Development

This research aims to explore the relationship between human critical thinking and the use of artificial intelligence in a collaborative learning environment. Based on the research questions, the following hypotheses are formulated:

Research question 1: *Does the use of AI reduce critical thinking or enhance critical thinking?*
Research question 2: *Does AI tools enhance collaborative learning in CSCL?*
Research question 3: *Does Teachers involvement in collaborative learning have an impact on learner's learning skills in a collaborative environment?*
Research question 4: *Does the use of AI impact collaborative learning?*
Research question 5: *Does collaborative learning influence academic achievement among learners in a collaborative learning environment?*

These hypotheses form the basis for statistical tests to follow, including T-tests to examine binary group differences and ANOVA to explore differences across multiple group means.

4.2 Survey Development

The survey was approved by the Institutional Review Board (IRB) ensuring that all ethical guidelines for conducting research with human participants were strictly followed. Participants were informed of the study's purpose, their rights, and the confidentiality of their responses before providing informed consent. The survey questions were developed by combining key questions from seven papers on CSCL, human critical thinking, and AI. These papers were selected based on their relevance and contribution to the field and contribution to the IS literature.

The questionnaire was designed based on a 7-point Likert style scale on Qualtrics platform. The participants were recruited from university students and followed university recruitment protocols. Questions were adapted to align with the study objectives, ensuring clarity and relevance. The final survey consisted of six main sections:

- Demographics and Background Information
- Impact of Collaborative learning on Academic achievement
- Role of Technology and AI in Collaborative Learning.
- Impact of AI tools in collaborative learning and critical thinking
- AI's influence on critical thinking and group dynamics in collaborative learning
- Teachers' involvement in collaborative learning environments

5. RESULTS

The dataset for this research was designed and collected through a survey on Qualtrics platform aimed at examining and analyzing the relationship between human critical thinking and the use of artificial intelligence in a collaborative learning environment. A total of 105 responses were used for analysis to ensure the reliability and consistency of the dataset. The dataset consists of both demographic and behavioral variables; The participants spanned diverse age groups, Gender, Education level and Present Employment.

The behavioral variables were divided into five main sections; each section comprises seven questions focusing on a specific aspect of collaborative learning and AI integration

- **Impact of Collaborative Learning on Academic Achievement**: This evaluates how collaborative learning influences academic excellence.
- **Role of Technology and AI in Collaborative Learning**: This examines the

integration of technology and AI in enhancing collaborative learning environments.

- **Impact of AI Tools in Collaborative Learning and Critical Thinking**: This analyzes the impact of AI in collaborative learning and assesses the effectiveness of AI tools in enhancing or degrading critical thinking skills in a collaborative learning environment.
- **AI's Influence on Critical Thinking and Group Dynamics in Collaborative Learning**: This Analyzes how AI affects group interactions and critical thinking within collaborative learning environment.
- **Teachers' Involvement in Collaborative Learning Environments**: Examines the role of educators in facilitating collaborative learning using AI.

5.1 Data Cleaning and Analysis

To ensure the quality and integrity of the data, incomplete responses were excluded to maintain the reliability of the findings. The data analysis for this research was conducted using SPSS and excel to address six research questions that focus on Collaborative learning, Critical thinking and the role of artificial intelligence (AI) in a collaborative learning environment. Table 1 lists the specific questions and is presented in Appendix A. The results of the research questions are as follows:

5.2 The impact of AI on human critical thinking in a CSCL environment.

Research question 1: Does the use of AI reduce critical thinking or enhance critical thinking?

		Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
INF1_1	Between Groups	69.902	6	11.65	3.424	**0.004**
	Within Groups	333.488	98	3.403		
	Total	403.39	104			
INF1_2	Between Groups	22.654	6	3.776	2.349	**0.037**
	Within Groups	157.536	98	1.608		
	Total	180.19	104			
INF1_4	Between Groups	18.182	6	3.03	2.41	**0.032**
	Within Groups	123.247	98	1.258		
	Total	141.429	104			
INF1_6	Between Groups	28.285	6	4.714	2.752	**0.016**
	Within Groups	167.849	98	1.713		
	Total	196.133	104			

Table 2: The impact of AI on human critical thinking in a CSCL environment.

The impact of the use of AI on human critical thinking in a CSCL environment was measured using SPSS one way ANOVA. The results of the data analysis are presented in Table 2 below. Following Table 2 is a discussion of each of the factors.

INF1-1: "I believe AI would replace human critical thinking in collaborative learning environments". There is a significant F-value (3.424, **P = 0.004**), this result shows that people who engaged in collaborative learning using AI tools believe that AI reduces critical thinking.

INF1-2: "The use of AI tools in collaborative learning has helped me improve my critical thinking skills". There is a significant F-value (3.349, **P=0.037**), this result shows that reliance on AI tool in CSCL discourages independent thought.

INF1-4: "AI tools in collaborative learning allow me to better analyze the argument presented in group discussions". There is a significant F-value (2.41, **P=0.032**). This result shows that using AI to simplify analytical processes makes them less engaged in deep critical reasoning.

INF1-6: "AI facilitates equal participation among teammates during collaborative learning activities". There is a significant F-value (2.752, **P=0.016**). This result shows that AI enhances collaborative learning.

Based on the factors and the results in Table 2 it appears that AI tools have both positive and negative perceptions as regards their impact on critical thinking. For example, while many participants recognize the benefits of AI in collaborative learning, they also believe it has an impact on reducing critical thinking in the long run.

5.3 Relationship between AI and Collaborative Learning.

Research question 2: Does AI tools enhance collaborative learning in CSCL?

Table 3 presents evidence that indicates the use of AI tools in a collaborative learning environment can significantly enhance collaborative learning by improving problem-solving, communication, idea generation, conflict resolution and corporation Following Table 3 is a discussion of each of the factors.

		Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
RT1_1	Between Groups	8.981	3	2.994	5.75	**0.001**
	Within Groups	52.581	101	0.521		
	Total	61.562	104			
RT1_2	Between Groups	10.558	3	3.519	4.416	**0.006**
	Within Groups	80.49	101	0.797		
	Total	91.048	104			
RT1_3	Between Groups	9.491	3	3.164	6.893	**<.001**
	Within Groups	46.356	101	0.459		
	Total	55.848	104			
RT1_4	Between Groups	12.314	3	4.105	6.49	**<.001**
	Within Groups	63.877	101	0.632		
	Total	76.19	104			
RT1_5	Between Groups	5.781	3	1.927	4.122	**0.008**
	Within Groups	47.21	101	0.467		
	Total	52.99	104			
RT1_6	Between Groups	7.725	3	2.575	5.841	**0.001**
	Within Groups	44.523	101	0.441		
	Total	52.248	104			

Table 3: Relationship between AI and collaborative learning.

RT1-1: "Role of Technology and AI in Collaborative Learning - Having access to computer-supported collaborative Learning has helped me to continue my studies to completion". This shows there is a significant difference in how often people perceive collaboration to enhance their learning and those who frequently engage in collaborative learning find it more beneficial

RT1-2: "Role of Technology and AI in Collaborative Learning - The use of AI tools in collaborative learning have helped me to collaborate more effectively". This shows there is a significant difference in how people view AI's ability to enhance effective teamwork and confirms that AI tools improve teamwork.

RT1-3: "Role of Technology and AI in Collaborative Learning - The collaborative learning forum allowed a fluid exchange of information". The result shows that there is a significant difference in perception of how AI influences problem-solving in collaboration environments and they believe that AI enhances problem solving by providing analytical tools.

RT1-4: "Role of Technology and AI in Collaborative Learning - The collaborative learning forum allowed a fluid exchange of information". This shows there is a significant difference between how AI facilitates discussions and generating ideas in group activity.

RT1-5: "Role of Technology and AI in Collaborative Learning - The collaborative learning environment has allowed me to establish personal connections with my teammates". The result shows that there is a significant difference between how AI tools is used to improve communication and task collaboration in a collaborative learning environment.

RT1-6: "Role of Technology and AI in Collaborative Learning - Collaborative learning has contributed to making me feel more involved in studying". The result shows that there is a significant difference in the AI ability to reduce conflict and brings corporation during group work.

5.4 Teacher's involvement in collaborative learning environment.

Research question 3: Does Teachers involvement in collaborative learning have an impact on learner's learning skills in a collaborative environment?

Based on the analysis as shown in Table 4 since TI1-2 and TI1-4 is ($P < 0.05$), there is a significant difference indicating that teacher involvement has an impact on learners' skills in collaborative learning environment particularly in areas that require active participation and guidance. Following Table 4 is a discussion of each of the factors.

		Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
TI1_2	Between Groups	18.52	3	6.173	7.555	**<.001**
	Within Groups	82.528	101	0.817		
	Total	101.048	104			
TI1_4	Between Groups	21.616	3	7.205	8.952	**<.001**
	Within Groups	81.298	101	0.805		
	Total	102.914	104			

Table 4: Teacher's involvement in collaborative learning environment.

TI1-2: "The teachers accompanied the students in an appropriate way to favor learning within collaborative environments". The result shows there is a significant difference in learners' perception of teachers' involvement in collaborative learning, it shows that teacher's involvement in collaborative learning enhances learners' ability to work effectively in a group.

TI1-4: "The teachers contributed to developing links with the learning community formed by each team and with other students". The result shows that there is a significant difference in learner's perception of how teachers' involvement impacts their skills. It confirms that teachers' involvement in collaborative learning enhances critical thinking.

5.5 Impact of AI on Collaborative Learning.

Research question 4: Does the use of AI impact collaborative learning?

Based on the analysis shown in Table 5, the

overall result shows that there is a statistically significant difference in the impact of collaborative learning on academic achievement and the role of AI in collaborative learning environments. Following Table 5 is a discussion of each of the factors.

	t	*df*	*Significance*		*Mean Difference*	*95% Confidence Interval of the Difference*	
			One-Sided p	**Two-Sided p**		**Lower**	**Upper**
CL1_1	24.837	104	**<.001**	**<.001**	1.571	1.45	1.7
RT1_2	19.296	104	**<.001**	**<.001**	1.762	1.58	1.94

Table 5: Impact of AI on collaborative learning.

CL1_1: "Impact of collaborative learning on academic achievement - I often engage in collaborative learning". The result shows there is a statistically significant difference between collaborative learning on academic achievement. The mean difference of 1.571 and P value < 0.05 shows that AI tools significantly enhance collaborative learning.
RT1_2: "Role of technology and AI in collaborative learning - The use of AI tools in collaborative learning has helped me collaborate more effectively". The result shows that there is a significant difference between the impact of AI tools on the effectiveness of collaboration in learning environment. The mean difference of 1.762 and P value <0.05 shows that AI tools make collaboration more effective.

5.6 Impact of AI on Collaborative Learning.
Research question 5: Does collaborative learning influence academic achievement among learners in a collaborative learning environment?

		Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
CL1_2	Between Groups	14.474	3	4.825	11.355	**<.001**
	Within Groups	42.916	101	0.425		
	Total	57.39	104			
CL1_3	Between Groups	19.884	3	6.628	13.574	**<.001**
	Within Groups	49.316	101	0.488		
	Total	69.2	104			
CL1_4	Between Groups	6.925	3	2.308	6.604	**<.001**
	Within Groups	35.303	101	0.35		
	Total	42.229	104			
CL1_5	Between Groups	22.152	3	7.384	5.586	**0.001**
	Within Groups	133.505	101	1.322		
	Total	155.657	104			
CL1_6	Between Groups	9.367	3	3.122	4.801	**0.004**
	Within Groups	65.681	101	0.65		
	Total	75.048	104			

Table 6: Learner's perspectives on collaborative learning

Based on the analysis shown in Table 5, it appears that collaborative learning significantly impacts academic achievement and enhances knowledge sharing and interaction in a collaborative learning environment. Following Table 6 is a discussion of each of the factors.

CL1_2: "My team members have given me support, help and support through collaborative learning". Since **P < 0.05**, this shows support provided by team members in a collaborative environment has a great impact in positively influencing academic achievement.
CL1_3: "Collaborative learning has helped me achieve good academic achievement and development". Since **P < 0.05**, this shows that there is a significant difference in learners' perceptions of the role of collaborative learning in achieving academic success and development.
CL1_4: "Teamwork has allowed me to complement my knowledge with that of my teammates". Since **P < 0.05,** this shows there is a significant difference in learners' perspectives on how teamwork complements their knowledge in a collaborative learning environment
CL1_5: "The collaborative learning environment has allowed me to establish personal connections with my teammates". Since **P < 0.05,** this shows that there is a significant difference between when learning through interaction with teammates compared to when studying alone.
CL1_6: "The teachers guided their students in the process of forming the collaborative learning environment". Since **P < 0.05,** this shows that there is a significant difference between the time

spent on collaborative learning and the benefits in achieving academic excellence.

6. DISCUSSION

This study advances current knowledge on the intersection of Artificial Intelligence (AI), Computer-Supported Collaborative Learning (CSCL), and critical thinking. Consistent with prior research Ada (2009), Tedla and Chen (2025), the results confirm that collaborative learning environments enriched with technology foster critical thinking by promoting communication, interaction, and co-construction of knowledge. Participants in this study reported that AI tools facilitated equal participation, smoother group interaction, and improved problem-solving, which aligns with McLaren et al. (2010), who demonstrated that AI-driven systems can guide and enhance collaborative discussions. Similarly, Hu et al. (2022) found that highly interactive group patterns supported by digital tools significantly enhanced creative and critical thinking, reinforcing the positive role of AI-enabled CSCL identified here.

Also, this research highlights a tension less frequently emphasized in prior studies: the potential of AI to diminish deep cognitive engagement. While (Liu et al., 2023) and Ramirez (2021) stress the benefits of structured AI-supported environments for enhancing conceptual understanding, our findings suggest that over-reliance on AI may discourage independent reasoning and reduce opportunities for learners to practice critical analysis. Participants noted that simplification of analytical processes by AI sometimes led to passivity, which contrasts with the uniformly positive conclusions drawn by Warsah et al. (2021) regarding collaborative learning's impact on critical thinking. This divergence underscores the importance of considering not just whether AI improves collaboration, but how it may simultaneously reshape learners' engagement with higher-order thinking.

Teacher involvement emerged as a critical moderating factor in our study, with results indicating that active facilitation by educators balanced the risks of AI dependency. This finding complements Kasepalu et al. (2022), who noted that AI-supported pedagogical interventions require human oversight to ensure authentic cognitive engagement. Yet, our results go further by quantifying this role, showing that teacher participation significantly influenced learners' perceptions of skill development and critical engagement. This emphasis on the interplay between AI and human facilitation distinguishes the present study from earlier work, which often evaluated AI or CSCL in isolation.

Overall, the dual findings that AI simultaneously enhances collaboration and risks diminishing independent thinking position this research at the forefront of debates on the "double-edged" role of AI in education. Unlike previous studies that primarily documented benefits, this study makes a unique contribution and offers a more nuanced perspective, stressing the need for balance. it frames AI not as an unequivocal enhancer of critical thinking, but as a tool whose value depends heavily on context, user agency, and teacher mediation.

6. CONCLUSIONS

This research highlights the relationship between human critical thinking and the use of artificial intelligence in a collaborative learning environment. While AI tools improve collaboration and real-time feedback, they also risk diminishing critical thinking if over-relied upon. Effective integration requires a balanced approach, where educators play a central role in guiding cognitive and collaborative processes. Future research should focus on developing adaptive AI systems that support rather than supplant critical thinking and address ethical concerns such as data privacy and algorithmic bias. The insights from this study offer a roadmap for designing AI-enhanced CSCL systems that empower learners and optimize educational outcomes. By balancing technological advancements with human oversight, educators and developers can create learning environments that not only enhance collaboration but also empower learners to think critically, innovate, and excel in an increasingly digital world.

9. REFERENCES

Ada, M. (2009). Computer supported Ada, M. (2009). Computer supported collaborative learning and higher order thinking skills: A case study of textile studies. Interdisciplinary Journal of E-Learning and Learning Objects, 5(1), 145-167. https://doi.org/10.28945/69

Altinay, L., & Paraskevas, A. (2007). A computer-supported collaborative learning (CSCL) approach in teaching research methods. International Journal of Hospitality Management, 26(3), 623-644.

https://doi.org/10.1016/j.ijhm.2006.05.005

Andersen, R., Mørch, A. I., & Litherland, K. T. (2022). Collaborative learning with block-based programming: investigating human-centered artificial intelligence in education. Behaviour & Information Technology, 41(9), 1830-1847.
https://doi.org/10.1080/0144929x.2022.2083981

Chou, C.-M., Shen, T.-C., Shen, T.-C., & Shen, C.-H. (2022). Influencing factors on students' learning effectiveness of AI-based technology application: Mediation variable of the human-computer interaction experience. Education and Information Technologies, 27(6), 8723-8750.
https://doi.org/10.1007/s10639-021-10866-9

Chu, H.-C., Hwang, G.-J., & Chang, C.-Y. (2024). An Integrative Review with Word Cloud Analysis of Computer-Supported Collaborative Learning. Journal of Science Education and Technology, 1-14.
https://doi.org/10.1007/s10956-024-10156-2

Cress, U., Stahl, G., Ludvigsen, S., & Law, N. (2015). The core features of CSCL: Social situation, collaborative knowledge processes and their design. International Journal of Computer-Supported Collaborative Learning, 10(2), 109-116.
https://doi.org/10.1007/s11412-015-9214-2

Hernández-Sellés, N., Muñoz-Carril, P.-C., & González-Sanmamed, M. (2019). Computer-supported collaborative learning: An analysis of the relationship between interaction, emotional support and online collaborative tools. Computers & Education, 138, 1-12.
https://doi.org/10.1016/j.compedu.2019.04.012

Hu, X., Liu, Y., Huang, J., & Mu, S. (2022). The effects of different patterns of group collaborative learning on Fourth-Grade students' creative thinking in a digital artificial intelligence course. Sustainability, 14(19), 12674.
https://doi.org/10.3390/su141912674

Järvelä, S., Kirschner, P. A., Panadero, E., Malmberg, J., Phielix, C., Jaspers, J., Koivuniemi, M., & Järvenoja, H. (2015). Enhancing socially shared regulation in collaborative learning groups: Designing for CSCL regulation tools. Educational Technology Research and Development, 63(1), 125-142.
https://doi.org/10.1007/s11423-014-9358-1

Kasepalu, R., Prieto, L. P., Ley, T., & Chejara, P. (2022). Teacher Artificial Intelligence-Supported Pedagogical Actions in Collaborative Learning Coregulation: A Wizard-of-Oz Study. Frontiers in Education 7 (2022). URL: https://www. frontiersin. org/articles/10.3389/feduc.
https://doi.org/10.3389/feduc.2022.736194

Lee, Y.-H. (2015). Facilitating critical thinking using the C-QRAC collaboration script: Enhancing science reading literacy in a computer-supported collaborative learning environment. Computers & Education, 88, 182-191.
https://doi.org/10.1016/j.compedu.2015.05.004

Liu, S., Kang, L., Liu, Z., Fang, J., Yang, Z., Sun, J., Wang, M., & Hu, M. (2023). Computer-supported collaborative concept mapping: The impact of students' perceptions of collaboration on their knowledge understanding and behavioral patterns. Interactive Learning Environments, 31(6), 3340-3359.
https://doi.org/10.1080/10494820.2021.1927115

McLaren, B. M., Scheuer, O., & Mikšátko, J. (2010). Supporting collaborative learning and e-discussions using artificial intelligence techniques. International Journal of Artificial Intelligence in Education, 20(1), 1-46.
https://doi.org/10.3233/jai-2010-0001

Ouyang, F., & Zhang, L. (2024). AI-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review. Educational Research Review, 44, 100616.
https://doi.org/10.1016/j.edurev.2024.100616

Radkowitsch, A., Vogel, F., & Fischer, F. (2020). Good for learning, bad for motivation? A meta-analysis on the effects of computer-supported collaboration scripts. International Journal of Computer-Supported Collaborative Learning, 15(1), 5-47.
https://doi.org/10.1007/s11412-020-09316-4

Ramirez, H. J. M. (2021). Facilitating Computer-Supported Collaborative Learning with Question-Asking Scripting Activity and its Effects on Students' Conceptual

Understanding and Critical Thinking in Science. International Journal of Innovation in Science and Mathematics Education, 29(1).
https://doi.org/10.30722/ijisme.29.01.003

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning, 3(3), 237-271.
https://doi.org/10.1007/s11412-007-9034-0

Tedla, Y. G., & Chen, H.-L. (2024). The impacts of computer-supported collaborative learning on students' critical thinking: a meta-analysis. Education and Information Technologies, 1-30.
https://doi.org/10.1007/s10639-024-12857-y

Tedla, Y. G., & Chen, H.-L. (2025). The impacts of computer-supported collaborative learning on students' critical thinking: a meta-analysis. Education and Information Technologies, 30(2), 1487-1516.
https://doi.org/10.1007/s10639-024-12857-y

Warsah, I., Morganna, R., Uyun, M., Afandi, M., & Hamengkubuwono, H. (2021). The impact of collaborative learning on learners' critical thinking skills. International Journal of Instruction, 14(2), 443-460.
https://doi.org/10.29333/iji.2021.14225a
.

Appendices and Annexures

Table 1: Questionnaire

Sources	Questions	Question Codes
Hernández-Sellés et al. (2019)	I often engage in collaborative learning.	CL_1
	My team members have given me support, help and support through collaborative learning.	CL_2
	Collaborative learning has helped me achieve good academic achievement and development.	CL_3
	Teamwork has allowed me to complement my knowledge with that of my teammates.	CL_4
	I have learnt more interacting with my teammates than when I study alone.	CL_5
	The teachers guided their students in the process of forming the collaborative learning environment.	CL_6
	The Teachers guided their students in the process of forming a collaborative learning environment.	TI1_1
	The teachers accompanied the students in an appropriate way to favor learning within collaborative environments.	TI1_2
	The teachers guided their students to develop teamwork skills that allow them to work more effectively.	TI1_3
	The teachers contributed to developing links with the learning community formed by each team and with other students.	TI1_4
Kasepalu et al. (2022) and Ouyang and Zhang (2024)	Interacting with my teammates has improved my grades than when I was studying alone	CL_7
	Having access to computer-supported collaborative Learning has helped me to continue my studies to completion.	RT1_1
	The use of AI tools in collaborative learning has helped me to collaborate more effectively.	RT1_2
	The collaborative learning forum allowed a fluid exchange of information.	RT1_3
	The team's discussion in a collaborative learning forum allowed establishing personal links.	RT1_4
	The collaborative learning environment has allowed me to establish personal connections with my teammates.	RT1_5
	Collaborative learning has contributed to making me feel more involved in studying.	RT1_6

Sources	Questions	Question Codes
Ada (2009)	I believe AI could suppress human critical thinking.	RT1_7
	I frequently use AI-supported tools in a collaborative learning environment.	IM1_1
	I believe AI tools enhance critical thinking by providing better insights and suggestions in collaborative tasks.	IM1_2
	AI improves my decision-making process in group work.	IM1_3
Tedla and Chen (2024), Hu et al. (2022), Warsah et al. (2021)	I believe AI supported tools reduce critical thinking	IM1_4
	I believe AI tools make collaborative learning more efficient.	IM1_5
	I believe AI tools affect collaborative learning processes by automating too much.	IM1_6
	The use of AI tools in collaborative learning would create bias opportunities.	IM1_7
	I believe AI would replace human critical thinking in collaborative learning environments.	INF1_1
	The use of AI tools in collaborative learning has helped me improve my critical thinking skills.	INF1_2
	The use of AI tools in collaborative learning environments encourages deeper discussions during group work.	INF1_3
	AI tools in collaborative learning allow me to better analyze the argument presented in group discussions.	INF1_4
	I feel more empowered to make critical decisions within the teammates when using AI tools in a learning environment.	INF1_5
	AI facilitates equal participation among teammates during collaborative learning activities.	INF1_6
	AI provides personalized feedback that helps me improve my critical thinking in group tasks.	INF1_7

Enhancing Programming Productivity for Individuals with ADHD Through Generative Artificial Intelligence: An Inductive Analysis

Lionel Mew
lmew@richmond.edu
University of Richmond
Richmond, VA 23173

Abstract

This inductive analysis examines how generative artificial intelligence (AI) can enhance programming productivity for individuals with attention-deficit/hyperactivity disorder (ADHD) by addressing executive function deficits that impair coding performance. Through systematic analysis of 45 peer-reviewed studies spanning ADHD interventions, programming productivity research, and AI-assisted development, we identify four primary mechanisms by which generative AI tools mitigate ADHD-related programming challenges: (1) cognitive scaffolding through automated pattern recognition and workflow optimization that compensates for executive function deficits, (2) task decomposition that breaks complex algorithms into manageable, discrete components, (3) real-time contextual support that reduces attention-switching costs by eliminating the need for external documentation searches, and (4) personalized learning systems that adapt to individual ADHD presentations and work patterns. Our analysis reveals that AI-assisted programming demonstrates particular efficacy for entry-level and junior programmers with ADHD, with documented productivity increases of up to 55% in code generation tasks. However, implementation requires careful consideration of code quality validation demands, potential skill development dependency, and privacy concerns. We propose a three-phase implementation framework (Assessment-Integration-Optimization) and discuss implications for computer science education and workplace accommodations. This research contributes evidence-based guidance for leveraging generative AI as cognitive support technology hin neurodiversity-inclusive programming environments.

Keywords: ADHD, programming productivity, generative artificial intelligence, executive function, neurodiversity, cognitive support

Recommended Citation: Mew, L., (2026). Enhancing Programming Productivity for Individuals with ADHD Through Generative Artificial Intelligence: An Inductive Analysis. *Journal of Information Systems Applied Research and Analytics*, v19(n3) pp 41-49. DOI# https://doi.org/10.62273/HMAD7698

Enhancing Programming Productivity for Individuals with ADHD Through Generative Artificial Intelligence: An Inductive Analysis

Lionel Mew

1. INTRODUCTION

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder that affects approximately 5-7% of children and 2.5-3.4% of adults worldwide (Faraone et al., 2021; Thomas et al., 2015). Characterized by persistent patterns of inattention, hyperactivity, and impulsivity that interfere with functioning and development (American Psychiatric Association [APA], 2013), ADHD presents unique challenges in professional contexts requiring sustained cognitive effort, such as computer programming.

The intersection of ADHD and computer programming has garnered increasing attention as the technology sector continues to grow and diversify. Programming requires sustained attention, meticulous organization, and executive functioning—cognitive skills that are often adversely impacted by ADHD (Barkley, 2015). Individuals with ADHD often experience difficulties in maintaining focus on a single task for extended periods, leading to challenges in completing complex programming tasks which require sustained cognitive effort (Fuermaier et al., 2015). These challenges are compounded by the precision and accuracy demands inherent in software development.

Throughout this paper, we use "programming" to refer primarily to the act of writing code, while "software development" encompasses the broader process including design, implementation, testing, and maintenance. While ADHD impacts both activities, our analysis focuses primarily on the coding and implementation aspects of software development where generative AI tools provide the most direct support.

Simultaneously, generative artificial intelligence (AI) has emerged as a transformative force in software development, with productivity measured by the number of lines of code produced increased by 55% for the group using the Large Language Model (LLM) (Gambacorta et al., 2024). The rapid adoption of generative AI tools presents an unprecedented opportunity to address ADHD-related challenges in programming by providing cognitive support, task automation, and personalized assistance.

This paper presents an inductive analysis examining how generative AI can enhance programming productivity for individuals with ADHD. We synthesize current research on ADHD's impact on programming performance, review traditional intervention strategies, and analyze the potential of AI-powered tools to mitigate ADHD-related challenges in software development contexts.

2. METHODOLOGY

This inductive analysis synthesizes research across three domains: ADHD and executive functioning (n=18 studies), generative AI programming tools (n=15 studies), and cognitive intervention strategies (n=12 studies). Literature was identified through systematic searches of ACM Digital Library, PubMed, and IEEE Xplore using keywords: "ADHD," "executive function," "programming," "generative AI," "coding assistants," and "productivity" (2010-2024).

Following established inductive analysis frameworks, we employed a three-phase coding process:

1. Open coding: Identified 87 initial codes describing ADHD challenges and AI affordances in programming contexts

2. Axial coding: Grouped codes into 12 categories linking AI features to executive function support

3. Selective coding: Identified four core mechanisms explaining how AI addresses ADHD-specific programming challenges

Analysis was guided by Cognitive Load Theory (Sweller et al., 2011) as a theoretical lens for understanding mechanism effectiveness. This approach allowed systematic identification of patterns across disparate literature domains without requiring primary data collection.

3. LITERATURE REVIEW

ADHD and Executive Functioning in Programming Contexts

ADHD significantly impacts executive functioning, which encompasses the cognitive processes necessary for planning, organization, working memory, and self-regulation (Diamond, 2013). Programming requires sustained attention, meticulous organization, and executive functioning—skills that can be adversely impacted by ADHD. Research indicates that poor skills in prioritizing and organizing workloads significantly hinder adults with ADHD in their workplace, resulting in occupational and educational underachievement (Wang et al., 2020).

The cognitive demands of programming exacerbate ADHD symptoms in several ways. In many instances, programmers with ADHD experience increased impulsivity and disorganization, leading to mistakes in coding and debugging processes. These challenges manifest as inefficient coding practices, difficulty tracking code changes, and problems managing multiple tasks simultaneously—all essential components of software development.

Furthermore, ADHD also affects interpersonal skills, critical in team-based programming environments in today's collaborative technology industry. Communication difficulties and reduced adaptability to unexpected changes or stressors can significantly impact collaborative programming efforts (Pollak et al., 2021; Tarver et al., 2021).

Traditional Intervention Strategies for ADHD in Programming

Multiple intervention strategies have demonstrated efficacy in supporting individuals with ADHD in professional contexts. Cognitive-behavioral therapy (CBT) has shown promising results, with research indicating that CBT can lead to significant symptom reduction among adults by addressing difficulties in executive functioning and managing daily tasks (Knouse & Safren, 2010). Tailored CBT programs help individuals develop strategies for time management, organization, and task prioritization—vital skills for coding and software development.

Neurofeedback represents another innovative approach. Lim et al. describe a brain-computer interface (BCI) based attention training program that can improve attention in individuals with ADHD, with research showing that a 20-session BCI attention training program improved ADHD symptoms (Lim et al., 2012). These techniques focus on training users to achieve specific brain activity patterns, potentially leading to better focus and control over impulsivity.

Mindfulness-based interventions (MBIs) have also gained recognition as effective practices. Tan and Jones conducted a scoping review discussing the advantages of mindfulness in improving emotional regulation and executive functioning in adolescents with ADHD (Tan & Jones, 2024). Evidence suggests that mindfulness practices can reduce stress and anxiety, which may otherwise exacerbate programming challenges.

Traditional productivity enhancement strategies encompass multiple domains:
Environmental Modifications: Research demonstrates that structured workspaces with minimal visual distractions help individuals with ADHD maintain focus for longer periods (Hallberg et al., 2020). Sensory input management represents another important environmental consideration, with individuals with ADHD often benefiting from controlling auditory stimulation (Kooij et al., 2019).

Time Management Strategies: The Pomodoro Technique, a time management method involving focused work intervals separated by short breaks, has shown particular promise for individuals with ADHD. Research by Lindsley and Brass (2018) found that implementing the Pomodoro Technique led to significant improvements in task completion and reduced self-reported stress among adults with ADHD.

Organizational Systems: Color-coding has emerged as a particularly effective organizational strategy, with research showing that color-coded filing systems improved homework management and completion for adolescents with ADHD (Langberg et al., 2011).

Generative AI in Programming and ADHD Support

The intersection of generative AI and ADHD support represents an emerging area of significant interest. Studies suggest that generative AI can significantly alleviate the cognitive load on developers, which is particularly beneficial for those with ADHD who often experience difficulties with focus and task management (Sauvola et al., 2024; Damyanov et al., 2024).

Research highlights that programming environments augmented by generative AI can foster an inclusive educational framework, thereby improving accessibility for individuals with ADHD. The utilization of tools like ChatGPT facilitates not just coding proficiency but also overall engagement, allowing programmers to ask questions and receive immediate assistance in a conversational format.

These AI-powered interactions can break tasks into smaller, manageable parts, making it easier for users with ADHD to stay on track (Ekellem, 2024; Zhao et al., 2024). This capability is particularly valuable given that breaking tasks into smaller components improved task initiation and completion rates among college students with ADHD (Reaser et al., 2019).

However, challenges exist in AI implementation. The variable quality of AI-generated code raises concerns regarding reliability and correctness, which necessitates a level of programming competency to evaluate the output critically (Wills et al., 2024; Idrisov & Schlippe, 2024). For programmers with ADHD, who may already struggle with attention to detail, distinguishing between accurate and erroneous AI outputs can compound existing challenges.

Current Applications of AI for ADHD Management

Recent developments in AI applications for ADHD management demonstrate promising trends. AI-driven chatbots provide customer support and assistance, while virtual assistants can help with organization, prioritization, and time management tasks. ChatGPT can benefit ADHD adults who use it as an AI executive function support tool, helping simplify day-to-day tasks involving organization, prioritization, and time management.

Generative AI tools can help people with ADHD break down big tasks into smaller, more manageable steps, with applications like Goblin. tools offering features such as "magic to-do" lists that automatically break down complex tasks into manageable components (Associated Press, 2024).

Theoretical Framework: Cognitive Load Theory and AI Support

Cognitive Load Theory provides a useful framework for understanding how generative AI can support programmers with ADHD. The theory distinguishes between intrinsic cognitive load (inherent to the task), extraneous cognitive load (imposed by instructional design), and germane cognitive load (contributing to learning and automation) (Sweller et al., 2011).

For programmers with ADHD, traditional programming tasks often create excessive cognitive load due to the need to simultaneously manage multiple aspects: problem decomposition, syntax recall, debugging logic, and code organization. Generative AI can reduce this load by:

1. Automating routine coding tasks (reducing intrinsic load)
2. Providing structured guidance and templates (reducing extraneous load)
3. Enabling focus on higher-order problem-solving (optimizing germane load)

4. DISCUSSION

4.1 Four Mechanisms of AI Support for ADHD Programmers

Building on the literature reviewed above, research on AI productivity gains shows significant promise. Studies indicate that LLMs can significantly boost productivity among programmers, with productivity measured by the number of lines of code produced increased by 55% for the group using the LLM (Gambacorta et al., 2024). Notably, the productivity gains were statistically significant primarily among junior staff, with a less pronounced effect on senior employees, suggesting benefits for entry-level programmers who may be developing foundational skills while managing ADHD symptoms. Our analysis identifies four mechanisms that explain these productivity benefits:

Cognitive Scaffolding: AI assistants can detect patterns in work habits to suggest optimal times for focused work, breaks, and transitions between activities—all common pain points for those with ADHD. This external regulation compensates for executive function deficits characteristic of ADHD.

Task Decomposition: Generative AI tools can help people with ADHD break down big tasks into smaller, more manageable steps. In programming contexts, this translates to breaking complex algorithms into discrete functions, providing step-by-step implementation guidance, and suggesting logical code organization structures.

Real-time Support: Unlike traditional programming resources, AI assistants provide immediate, contextual support without requiring

programmers to shift attention to external documentation or references. This reduces attention switching costs, which are particularly challenging for individuals with ADHD.

Personalized Learning: AI apps learn user habits and priorities to generate personalized to-do lists, schedules, and reminders, studying how individuals work best and adapting to keep them on track. This personalization addresses the heterogeneous nature of ADHD presentation.

4.2 Implementation Framework

Based on our analysis, we propose a framework for implementing AI-enhanced programming support for individuals with ADHD:

Phase 1: Assessment and Customization
- Evaluate individual ADHD presentation and programming skill level
- Configure AI tools to match specific attention patterns and preferences
- Establish baseline productivity metrics

Phase 2: Guided Integration
- Introduce AI tools gradually to prevent cognitive overload
- Provide training on effective prompt engineering and AI collaboration
- Establish protocols for validating AI-generated code

Phase 3: Adaptive Optimization
- Monitor productivity outcomes and adjustment patterns
- Refine AI configurations based on usage data and performance metrics
- Develop personalized workflows that maximize AI benefits

4.3 Benefits and Limitations

Primary Benefits:
1. **Reduced Cognitive Load**: AI handles routine tasks, allowing focus on creative problem-solving
2. **Enhanced Organization**: Automated project structure and code organization support
3. **Improved Error Detection**: AI-assisted debugging reduces time spent on detail-oriented tasks
4. **Flexible Pacing**: On-demand assistance accommodates varying attention spans

Key Limitations:
1. Code Quality Concerns: AI-generated code requires critical evaluation, which may challenge individuals with attention deficits
2. Dependency Risk: Over-reliance on AI may impede skill development
3. Privacy Considerations: Using AI chatbots creates privacy issues when providing personal information, emails, calendar, and personal writings to big companies
4. Cost and Access: Premium AI tools may not be accessible to all programmers

4.4 Implications for Education and Industry

Educational Implications:
- Computer science curricula should incorporate AI-assisted programming techniques
- Instructors should receive training on supporting neurodivergent learners with AI tools
- Assessment methods should account for AI-enhanced productivity

Industry Implications:
- Organizations should consider AI tool provision as workplace accommodation
- Training programs should address AI collaboration skills
- Performance evaluation criteria may need adjustment for AI-enhanced workflows.

5. FUTURE RESEARCH

Our analysis identifies several critical areas for future investigation:

Longitudinal Effectiveness Studies
Long-term studies are needed to assess the sustained benefits of AI-assisted programming for individuals with ADHD. Research should examine whether productivity gains persist over time and how AI usage patterns evolve with experience.

Personalization Algorithms
Development of AI systems specifically tailored to ADHD presentations requires research into:
- Attention pattern recognition and adaptive response mechanisms
- Customizable intervention strategies based on individual ADHD profiles
- Integration with existing ADHD management approaches

Comparative Effectiveness Research
Studies comparing AI-enhanced programming with traditional interventions (CBT, medication, environmental modifications) will inform evidence-based practice recommendations. Hybrid approaches combining AI tools with established interventions merit particular attention.

Neurocognitive Mechanisms
Neuroimaging studies could elucidate how AI-assisted programming affects neural networks associated with executive function and attention in ADHD. Understanding these mechanisms could inform more targeted AI tool development.

Equity and Access Research
Investigation of how factors such as socioeconomic status, educational background, and cultural considerations affect access to and benefits from AI programming tools is essential for ensuring equitable implementation.

Ethical Frameworks
Development of ethical guidelines for AI use in ADHD support, addressing issues of autonomy, skill development, and data privacy, requires interdisciplinary collaboration between technologists, clinicians, and ethicists.

6. SUMMARY

This inductive analysis demonstrates that generative artificial intelligence presents significant opportunities to enhance programming productivity for individuals with ADHD. By providing cognitive scaffolding, task decomposition support, and personalized assistance, AI tools can address many of the core challenges that ADHD presents in programming contexts.

The evidence suggests that AI-assisted programming is particularly beneficial for entry-level and junior programmers, where productivity gains of up to 55% have been documented. These tools help mitigate executive function deficits by reducing cognitive load, automating routine tasks, and providing real-time guidance.

However, successful implementation requires careful consideration of individual needs, appropriate training, and awareness of limitations such as code quality validation requirements and potential dependency concerns. The heterogeneous nature of ADHD necessitates personalized approaches to AI tool configuration and usage.

7. CONCLUSIONS

This inductive analysis identified four distinct mechanisms through which generative artificial intelligence addresses executive function deficits in programming contexts: cognitive scaffolding, task decomposition, real-time contextual support, and personalized adaptive learning. These mechanisms operate synergistically to reduce cognitive load, maintain attention, and support skill development, with evidence suggesting particular benefit for entry-level and junior programmers with ADHD.

The convergence of generative AI and neurodiversity awareness represents a paradigm shift toward more inclusive programming environments. As AI tools become increasingly sophisticated and accessible, their application extends beyond individual productivity gains to encompass broader organizational advantages through neurodivergent talent inclusion and diverse problem-solving approaches.

Future success in this domain requires continued collaboration between technology developers, ADHD researchers, educational institutions, and industry stakeholders. Empirical validation of the mechanisms identified here through controlled studies and longitudinal research will be essential. By embracing AI as cognitive support technology rather than mere automation, we can create programming environments that empower individuals with ADHD to reach their full potential while contributing unique perspectives to software development.

The principles of cognitive scaffolding, task decomposition, real-time support, and personalized learning identified in this analysis provide a foundation for broader applications of AI in neurodiversity accommodation across diverse professional domains. As we continue to develop and refine these tools, evidence-based implementation that centers the experiences and needs of individuals with ADHD remains paramount.

8. ACKNOWLEDGEMENTS

Generative AI (Claude) was used to draft initial sections and summarize literature, with all content reviewed and revised by the author. Grok was used to provide comprehensive feedback.

9. REFERENCES

Adamou, M., Arif, M., Asherson, P., Aw, T. C., Bolea, B., Coghill, D., ... & Young, S. (2013). Occupational issues of adults with ADHD. BMC Psychiatry, 13(1), 59. https://doi.org/10.1186/1471-244X-13-59

Adamou, M., Graham, K., MacKeith, J., Burns, S., & Emerson, L. M. (2018). Advancing

services for adult ADHD: The development of the ADHD Star as a framework for multidisciplinary interventions. BMC Health Services Research, 18(1), 184. https://doi.org/10.1186/s12913-018-2988-z

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Associated Press. (2024, May 24). People with ADHD are turning to AI apps to help with tasks. Experts say try it cautiously. *Associated Press*. https://apnews.com/article/adhd-apps-attention-deficit-hyperactivity-disorder-e455a921062dea5e0d5900f993f5d11f

Ayyash, H. (2017). The outcome of an ADHD parenting group training programme (APEG) in the Peterborough neurodevelopmental service (NDS). Clinical Journal of Nursing Care Practice, 1(1), 013-019.

Barkley, R. A. (2015). Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment (4th ed.). Guilford Press.

Bikic, A., Reichow, B., McCauley, S. A., Ibrahim, K., & Sukhodolsky, D. G. (2017). Meta-analysis of organizational skills interventions for children and adolescents with attention-deficit/hyperactivity disorder. Clinical Psychology Review, 52, 108-123. https://doi.org/10.1016/j.cpr.2016.12.004

Cooper, N., Clark, A., Lecomte, N., Qiao, H., & Ellison, A. (2024). Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. Methods in Ecology and Evolution, 15(10), 1757-1763.

Daley, D., Oord, S., Ferrín, M., Cortese, S., Danckaerts, M., Döepfner, M., ... & Sonuga-Barke, E. (2017). Practitioner review: current best practice in the use of parent training and other behavioural interventions in the treatment of children and adolescents with attention deficit hyperactivity disorder. Journal of Child Psychology and Psychiatry, 59(9), 932-947.

Damyanov, I., Tsankov, N., & Nedyalkov, I. (2024). Applications of generative artificial intelligence in the software industry. TEM Journal, 2724-2733.

Diamond, A. (2013). Executive functions. Annual Review of Psychology, 64, 135-168.

Ekellem, E. (2024). Leveraging conversational AI for adult ADHD management: enhancing social skills and daily functioning.

Faraone, S. V., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M. A., ... & Wang, Y. (2021). The World Federation of ADHD International Consensus Statement: 208 evidence-based conclusions about the disorder. Neuroscience & Biobehavioral Reviews, 128, 789-818.

Fuermaier, A. B. M., Tucha, L., Koerts, J., Aschenbrenner, S., Kaunzinger, I., Hauser, J., Weisbrod, M., Lange, K. W., & Tucha, O. (2015). Sustained attention in adult ADHD: Time-on-task effects of various measures of attention. Journal of Neural Transmission, 122(12), 1681-1690. https://doi.org/10.1007/s00702-015-1426-0

Gambacorta, L., Qiu, H., Shan, S., & Rees, D. (2024). Generative AI and labour productivity: a field experiment on coding. BIS Working Papers, 1208. https://www.bis.org/publ/work1208.htm

News reports. (2024, August-September). Multiple news sources covering AI applications for ADHD management. Various outlets including AP News, VOA Learning English, US News & World Report.

Hallberg, E., Klingberg, T., & Seckler, P. (2020). Working memory training in adults with ADHD: A randomized controlled trial. Journal of Attention Disorders, 24(6), 918-927. https://doi.org/10.1177/1087054718756728

Idrisov, B., & Schlippe, T. (2024). Program code generation with generative AIs. Algorithms, 17(2), 62.

Jongh, M., Wium, A., & Basson, W. (2019). The piloting of a specific support programme for grade R teachers on attention deficit hyperactivity disorder: the process of development. South African Journal of Communication Disorders, 66(1).

Knouse, L. E., & Safren, S. A. (2010). Current status of cognitive behavioral therapy for

adult attention-deficit hyperactivity disorder. Psychiatric Clinics of North America, 33(3), 497-509. https://doi.org/10.1016/j.psc.2010.04.001

Kooij, J. J., Bijlenga, D., Salerno, L., Jaeschke, R., Bitter, I., Balázs, J., ... & Asherson, P. (2019). Updated European Consensus Statement on diagnosis and treatment of adult ADHD. European Psychiatry, 56(1), 14-34. https://doi.org/10.1016/j.eurpsy.2018.11.001

Langberg, J. M., Epstein, J. N., Becker, S. P., Girio-Herrera, E., & Vaughn, A. J. (2011). Evaluation of the Homework, Organization, and Planning Skills (HOPS) intervention for middle school students with attention deficit hyperactivity disorder as implemented by school mental health providers. School Psychology Review, 40(3), 477-492. https://doi.org/10.1080/02796015.2011.12087714

Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., ... & Vinyals, O. (2022). Competition-level code generation with AlphaCode. Science, 378(6624), 1092-1097.

Lim, C. G., Lee, T. S., Guan, C., Fung, D. S. S., Zhao, Y., Teng, S. S. W., ... & Krishnan, K. R. R. (2012). A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder. PloS One, 7(10), e46692.

Lindsley, M., & Brass, K. (2018). Time management interventions for adults with ADHD: The Pomodoro Technique vs. traditional scheduling. Journal of Applied Psychology, 103(4), 358-372.

Liu, T., Hsiao, R., Chou, W., & Yen, C. (2024). Prospective and cross-sectional factors predicting caregiver motivation to vaccinate children with attention-deficit/hyperactivity disorder against covid-19: a follow-up study. Vaccines, 12(5), 450.

Magon, R., Latheesh, B., & Müller, U. (2015). Specialist adult ADHD clinics in east anglia: service evaluation and audit of nice guideline compliance. BJPsych Bulletin, 39(3), 136-140.

Pollak, Y., Shoham, R., Dayan, H., Gabrieli-Seri, O., & Berger, I. (2021). Symptoms of ADHD predict lower adaptation to the COVID-19 outbreak: financial decline, low adherence to preventive measures, psychological distress, and illness-related negative perceptions. Journal of Attention Disorders, 26(5), 735-746.

Reaser, A., Prevatt, F., Petscher, Y., & Proctor, B. (2019). The learning and study strategies of college students with ADHD. Psychology in the Schools, 56(1), 82-94. https://doi.org/10.1002/pits.22149

Sauvola, J., Tarkoma, S., Klemettinen, M., Riekki, J., & Doermann, D. (2024). Future of software development with generative AI. Automated Software Engineering, 31(1).

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. Springer.

Tan, L., & Jones, M. (2024). Hyped-up or meditate: a scoping review of mindfulness-based group interventions for adolescents with attention deficit hyperactivity disorder. Clinical Child Psychology and Psychiatry, 29(4), 1383-1399.

Tan, L. (2015). A critical review of adolescent mindfulness-based programmes. Clinical Child Psychology and Psychiatry, 21(2), 193-207.

Tarver, J., Daley, D., & Sayal, K. (2021). A self-help version of the new forest parenting programme for parents of children with attention deficit hyperactivity disorder: a qualitative study of parent views and acceptability. Child and Adolescent Mental Health, 27(3), 215-222.

Thomas, R., Sanders, S., Doust, J., Beller, E., & Glasziou, P. (2015). Prevalence of attention-deficit/hyperactivity disorder: a systematic review and meta-analysis. Pediatrics, 135(4), e994-e1001.

Vujoviæ, D. (2024). Generative AI: riding the new general purpose technology storm. Ekonomika Preduzea, 72(1-2), 125-136.

Wang, Q., Xu, R., & Volkow, N. (2020). Increased risk of COVID-19 infection and mortality in people with mental disorders: analysis from electronic health records in the

united states. World Psychiatry, 20(1), 124-130.

Wang, Z. (2017). Neurofeedback training intervention for enhancing working memory function in attention deficit and hyperactivity disorder (ADHD) Chinese students. NeuroQuantology, 15(2).

Wills, S., Poon, S., Salili-James, A., & Scott, B. (2024). The use of generative AI for coding in academia. Methods in Ecology and Evolution, 15(12), 2189-2191.

Zhao, Q., Li, H., Yu, X., Huang, F., Wang, Y., Liu, L., ... & Wang, Y. (2017). Abnormal resting-state functional connectivity of insular subregions and disrupted correlation with working memory in adults with attention deficit/hyperactivity disorder. Frontiers in Psychiatry, 8.

Zhao, X., Cox, A., & Chen, X. (2024). Disabled students' use of generative AI in higher education.

AI-Enhanced Interview Preparation: A Comprehensive Review of Technical, Behavioral, and Immersive Training Systems

Silvia Sanjana
ssanjana@students.kennesaw.edu
Kennesaw State University
Marietta, GA 30060

Yi Li
joy.li@kennesaw.edu
Kennesaw State University
Marietta, GA 30060

Selena He
she4@kennesaw.edu
Kennesaw State University
Marietta, GA 30060

Abstract

This systematic review synthesizes advances in AI-driven job interview systems, examining their technological foundations, pedagogical alignment, and effectiveness in supporting professional skill development. Existing platforms range from rule-based and NLP-enhanced systems to immersive VR/AR and avatar-based simulations. Recent progress in large language models (LLMs), multimodal feedback, and photorealistic virtual agents has significantly improved adaptivity, realism, and user engagement. However, current systems remain fragmented, with most focusing on either technical or behavioral skills, offering limited emotional responsiveness, insufficient curriculum alignment, and minimal support for integrated coding environments or plagiarism-aware assessments. By analyzing 44 high-quality studies, this review identifies critical gaps in personalization, real-time multimodal evaluation, and pedagogically structured feedback. It proposes a unified framework incorporating emotionally adaptive avatars, interactive coding simulations, domain-aware questioning, and learner-centered dashboards. This work provides the first systematic integration of technical, behavioral, and immersive dimensions of AI-based interview systems, establishing a foundation for designing intelligent, inclusive, and context-aware platforms that better align academic preparation with real-world hiring expectations.

Keywords: Job interview systems, soft skills, coding interviews, AI in education, avatar-based training, virtual reality

Recommended Citation: Sanjana, S., Li, Y., He, S., (2026). AI-Enhanced Interview Preparation: A Comprehensive Review of Technical, Behavioral, and Immersive Training Systems. *Journal of Information Systems Applied Research and Analytics*, v19(n3) pp 50-67. DOI# https://doi.org/10.62273/GJOI3460

AI-Enhanced Interview Preparation: A Comprehensive Review of Technical, Behavioral, and Immersive Training Systems

Silvia Sanjana, Yi Li and Selena He

1. INTRODUCTION

This paper provides a systematic review of the existing job interview systems. Interviewing is a critical competency across diverse domains including technical (Qin et al., 2019), engineering (Senthilkumar et al., 2025), business (Takeuchi & Koda, 2021), healthcare (Lee et al., 2020), education (Geng et al., 2024), and law (Hassan et al., 2023). Employers now expect proficiency not only in domain expertise but also in soft skills such as communication, collaboration, and critical thinking. Yet, formal training remains minimal, with existing programs often costly and accessible primarily to affluent individuals (Nofal et al., 2025). This inequity has driven interest in intelligent and automated systems capable of replicating interviews and providing personalized feedback.

Recent advancements in artificial intelligence and virtual reality (VR) have facilitated the creation of robust interview training systems designed to simulate real interview scenarios. These systems take diverse forms, like AI-powered chatbots (Røed et al., 2023) that engage people in simulated interviews, embodied virtual interviewer avatars sometimes within immersive VR settings (Hassan et al., 2023), and mixed-reality simulations. Chou et al. (2022) typically utilize adaptive feedback such as real-time analysis of a candidate's responses, nonverbal cues, or coding solutions to facilitate user improvement. Significantly, these systems encompass various fields. Virtual interview tutors have been developed for commercial and technical job interviews, medical and healthcare training, such as practicing patient interviews (Rädel-Ablass et al., 2025), legal and law enforcement situations, and soft skill management (Luo et al., 2024). Moreover, as the global employment landscape grows more competitive, there is an urgent necessity to create inclusive, accessible, and efficient strategies for interview preparation and evaluation, allowing candidates from varied backgrounds, including individuals with disabilities or anxiety disorders, to effectively demonstrate their competencies.

Traditional academic assessments such as written exams, multiple-choice tests, and project-based coursework often emphasize theoretical knowledge or technical problem-solving in controlled settings. While effective for measuring content mastery, these forms of evaluation rarely capture the interactive, adaptive, and high-pressure dynamics of job interviews. Similarly, offerings from career development offices, though valuable, are typically optional and resource-limited, prompting the rise of digital platforms for interview preparation. However, these interviewing systems vary greatly in design and focus, with many limited to either technical or behavioral training. A comprehensive understanding of their effectiveness and limitations remains lacking, and few offer personalized, adaptive, or realistic feedback, leaving key aspects of professional readiness under addressed.

This review is motivated by the need to identify what has been achieved, what limitations persist, and how future systems can better support the transition from academic to professional environments. It is guided by five research questions (RQ1–RQ5) concerning technologies, pedagogical strategies, real-world alignment, professional skill support, existing gaps, and technical, pedagogical, and ethical challenges. By consolidating prior work, this study highlights critical gaps including absent integration of technical and behavioral assessments, limited emotional adaptivity, inadequate coding simulations with plagiarism detection, and insufficient personalization and outlines directions for future research.

The rest of the paper is organized as follows: Section 2 outlines the methodology for literature selection and analysis. Section 3 reviews and categorizes existing job interview systems outlining key future research directions. Section 4 analyzes findings based on the research questions. Finally, Section 5 concludes by summarizing the key insights.

2. METHODOLOGY

To ensure rigor and transparency, this review adopts a structured methodology comprising three components: the search strategy (scope and reproducibility), the inclusion and exclusion criteria (quality boundaries), and the selection process (multi-stage screening).

A. Database and Search Strategy

A comprehensive keyword search strategy was designed to identify the most relevant and recent studies for this review. It focused on system types, embodied interaction (e.g., avatars), assessed skills (e.g., soft and technical), interaction modalities (e.g., AR/VR), and feedback mechanisms. A structured query was applied across academic databases to capture high-quality studies, with priority given to peer-reviewed publications and selective inclusion of preprints when they addressed recent advances aligned with the review objectives with potential contributions for their relevancy.

Sample Keywords

Category	Keywords
System Type	"Intelligent job interview system" OR "interviewing system" OR "mock interview system"
Agents	"Avatar" OR "Virtual agent"
Skills Assessed	"Soft skills" OR "Communication"
Interview Types	"Job interview" OR "Mock interview" OR "Coding interview" OR "Soft skill interview"
Feedback Mechanism	"Personalized feedback" OR "Adaptive feedback"
Immersive Technology	"AR" OR "Augmented Reality" OR "VR" OR "Virtual Reality"
Skills Emphasis	"Communication" OR "Soft skills" OR "Coding skill" OR "Technical skill"

B. Inclusion and Exclusion Criteria

The inclusion criteria were designed to prioritize peer-reviewed research studies relevant to the development of Job Interview systems. However, high-quality preprints and early-stage publications were also considered when they contributed substantially to emerging trends or recent technological advancements. These were critically appraised for relevance and quality before inclusion. We included studies that:

(1) Presented or evaluated Job Interview platforms.
(2) Presented or evaluated Job Interview platforms focusing on Augmented Reality, Virtual Reality, and Gamification.
(3) Integrated avatar-based or embodied conversational agents.
(4) Offered mock coding interviews or simulated roleplay interactions.
(5) Provided personalized or adaptive feedback mechanisms and studies were excluded if they:
(a) Not written in English.
(b) Focused solely on general e-learning.

C. Selection Process

The literature selection process followed a systematic four-stage PRISMA approach (Mutter et al., 2021), as illustrated in the flowchart (Figure 1).

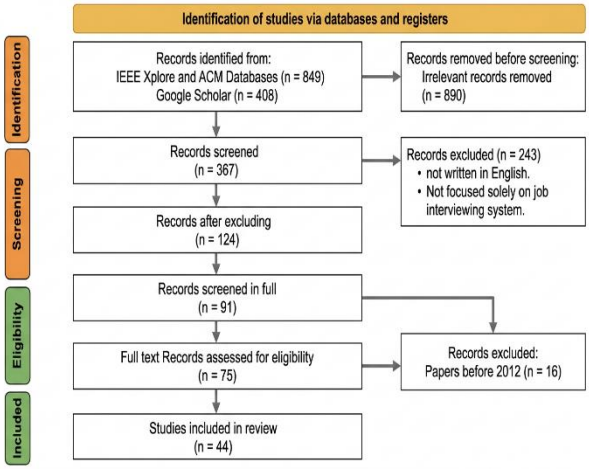


Figure 1: Paper Selection Process

Identification Phase: An initial pool of 1257 records were retrieved through comprehensive database searches across IEEE Xplore and ACM Digital Library (n = 849) and Google Scholar (n = 408) using a predefined set of related keywords as specified before. Prior to formal screening, 890 records were excluded for irrelevance based on manual inspection of titles and abstracts. This review relies on IEEE Xplore, ACM Digital Library, and Google Scholar as primary databases. While this ensured strong coverage of computing and AI-focused publications, the exclusion of broader indexing services such as Scopus and Web of Science may have omitted some interdisciplinary studies. However, the inclusion of Google Scholar helped mitigate this limitation by retrieving relevant works.

Screening Phase: A total of 367 articles were subjected to title and abstract screening. Of these, 243 records were excluded for failing to meet the inclusion criteria. This resulted in 124 potentially relevant records progressing to the next stage.

Eligibility Assessment: Full texts of 91 studies were reviewed in detail to determine their suitability for inclusion. At this stage, an additional 16 papers were excluded due to being published prior to 2012, as they did not reflect the recent advances in AI, natural language processing, and immersive technologies. The remaining 75 studies were evaluated against the full inclusion criteria. This threshold ensured that it included works aligned with the technological advancements most relevant to current and future interview training systems.

Inclusion Phase: Following a rigorous assessment, 44 studies were determined to meet all predefined inclusion criteria and were selected for detailed analysis in this systematic review. This assessment went beyond basic inclusion criteria and considered factors such as methodological soundness (clarity of design, evaluation approach), technological relevance (use of AI, VR/AR, or feedback mechanisms), and contribution to the research questions. Preprints were included only if they demonstrated clear methodological transparency and novel contributions.

Having established the methodological foundation and selection criteria, the following section is organized along three dimensions technological approaches, skills assessed, and interaction modalities to provide a comprehensive understanding of the current landscape of AI-enhanced interviewing systems.

3. REVIEW OF INTERVIEW TRAINING SYSTEMS

This review first categorizes interview training systems by (1) "Technological Approaches" (2) "Skills Assessed", and (3) "Modality".

.

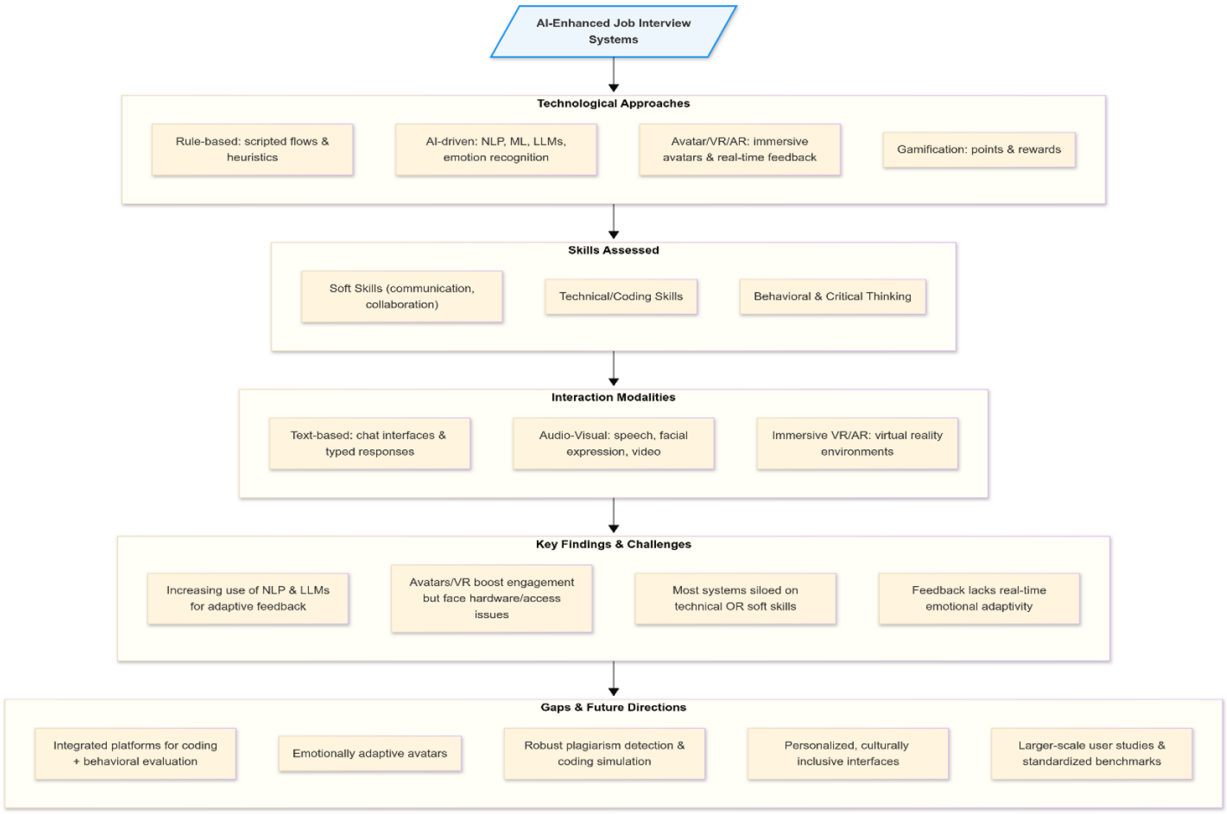


Figure 2: Overview of AI-Enhanced Job Interview Systems: Taxonomy and Future Directions

A high-level taxonomy of AI-enhanced interview systems is shown in Figure 2. This diagram summarizes how existing systems are categorized and where research gaps emerge. It provides a taxonomy of AI-enhanced job interview systems across five dimensions: technologies, skills, modalities, challenges, and future directions. It identifies four main technological approaches rule-based, AI-driven (NLP/LLMs), immersive VR/AR, and gamification which were used to assess soft, technical, and behavioral skills. Here, interaction modalities include text-based, audio-visual, and immersive environments.

Key challenges include limited real-time emotional feedback, skill-specific system silos, and accessibility issues in VR. Future directions emphasize the need for integrated platforms, emotionally adaptive avatars, plagiarism detection, culturally inclusive interfaces, and large-scale evaluations with standardized benchmarks

A. Technological Approaches

i. Rule-based Systems: Early efforts in virtual interview training often relied on rule-based or scripted approaches. Rule-based systems employ scripted dialogue and decision trees to ensure predictable interactions. Early examples include ERICA, which used keyword detection for follow-up questions but lacked adaptability (Kawahara et al., 2021), and semi-automated multi-agent interviewing with partial ML support (Kawai et al., 2022). Other efforts applied rule-based templates with NLP for question generation (Pandey et al., 2023), conversational chatbots with state-transition models (Boudjani et al., 2023), finite state machines for social cue detection (Baur et al., 2013), and hybrid Theory of Mind models (Belkaid et al., 2014). These systems provide structure but remain rigid, non-adaptive, and limited in scalability.

While rule-based systems laid the foundation for automated interviewing, their limited flexibility led to the development of more adaptive AI-driven approaches that support dynamic interaction and personalized feedback.

ii. Intelligent Systems (AI-driven): The advent of Artificial Intelligence (AI) has significantly reshaped job interview preparation, enabling systems that integrate natural language processing (NLP), computer vision, and large language models (LLMs) for dynamic evaluation. Early platforms such as Big Interview (Fulk et al., 2022) and NexInterview (S et al., 2025) illustrate how NLP and generative AI can support structured mock interviews. Building on these foundations, systems like CIRVR (Adiani et al., 2022) and ITEM (Nofal et al., 2025) incorporate virtual reality and real-time feedback to deliver personalized, adaptive training.

A central trend is the shift from static, rule-based approaches to adaptive agents powered by LLMs and multimodal cues. For example, GPT-4o has been applied in technical interviews with high realism (Gomez et al., 2025), yet these implementations often neglect emotion awareness and plagiarism detection. Similarly, Gemini-based adaptive questioning (Rai, 2025) advances interactivity but lacks affective sensing.

Fairness and transparency have also emerged as critical concerns. Pathak et al. (2024) demonstrate that asynchronous, LLM-driven video interviews can enhance demographic fairness, though delayed feedback and reliance on a single modality limit effectiveness. Complementary approaches integrate inclusivity tools such as gaze-tracking in VR (Adjani et al., 2022), yet hardware demands and limited feedback remain barriers. Likewise, video simulation with analytics has supported language training (Jarvis et al., 2024), though it remains domain specific. Nofal et al. (2025) further combines VR, LLMs, and bias-testing frameworks to create bias-aware practice, though fidelity is constrained by the absence of gaze or head tracking.

Expanding beyond evaluation, several systems employ predictive and multimodal models. NLP–CNN pipelines have been explored for soft-skill and personality prediction, albeit with generic feedback (Rao et al., 2025). Voice-first, role-specific simulations extend Gemini applications (S et al., 2025), yet real-time analytics and expressiveness are limited. Other innovations include vision–speech fusion for emotion recognition (Golande et al., 2025), graph-based skill-targeted questioning (Qin et al., 2024), and speaker-willingness recognition for adaptive questioning (Nagasawa et al., 2023), though each faces constraints in validation, flexibility, or scope.

Beyond the systems reviewed above, recent advances in multimodal foundation models such as GPT-4o's native audio-visual capabilities and Google's Gemini architecture enable simultaneous processing of text, speech, and visual input within a single model, reducing integration complexity and latency. Agentic AI

frameworks, where LLMs autonomously plan and execute multi-step interview workflows, represent an emerging paradigm that could enable fully autonomous adaptive interviewers capable of adjusting difficulty, topic, and questioning style in real time based on candidate performance. While these technologies are still maturing and have limited representation in the peer-reviewed literature captured by our review window, they signal important directions for future system design.

Collectively, these systems affirm the potential of AI to enhance confidence, self-assessment, and accessibility. However, persistent limitations including limited personalization, underdeveloped affective sensing, and the absence of standardized benchmarking underscore the need for larger-scale evaluations and rigorous validation.

iii. Avatar-based / Immersive VR/AR-based Systems: AI-powered interview simulators are increasingly adopting avatars and extended reality to deliver immersive and interactive experiences. By moving beyond static interfaces and text-based exchanges, these systems aim to provide dynamic, lifelike simulations that enhance user engagement and skill transfer (Nofal et al., 2025; Sahani et al., 2025).

To illustrate, Nofal et al. (2025) introduced a VR-based platform built with Oculus Quest and Unity, where animated avatars and ChatGPT-generated questions were used to assess communication, leadership, and domain knowledge through sentiment-driven scoring and bias analysis. Although the system proved effective in controlled studies, it remains constrained by its dependence on high-end hardware, the absence of gaze tracking, and the lack of large-scale validation. Building on this trajectory, Sahani et al. (2025) presented a scalable web-based platform with a 3D avatar created using React.js and Three.js. While it demonstrated high question accuracy (~95%), robust speech-to-text performance (>90%), and measurable gains in user confidence (80%), its limited realism, absence of multilingual capacity, and lack of expert-curated content reduce its applicability in broader contexts.

In parallel, several systems have sought to broaden accessibility. AIVATAR (Bachhav et al., 2023) integrates 3D avatars with aptitude testing and instant textual feedback to reduce interview anxiety in low-stakes practice environments. Yet, despite its promise, the platform remains conceptual, with no empirical validation and limited non-verbal cue integration. Extending this line of research, Hassan et al. (2023) developed a multimodal platform for investigative interviews across VR, desktop, and audio formats. Their findings indicate that VR fosters stronger presence and realism compared to other modalities; however, binary feedback, reliance on proprietary APIs, and synthesized voices limit its utility. Similarly, Røed et al. (2023) used a fine-tuned GPT-3 to simulate conversations with a child avatar, showing that personalized feedback significantly improved questioning strategies skills directly transferable to interview contexts.

At the same time, efforts have also focused on enhancing realism. Ashrafi et al. (2024) employed Unreal Engine, MetaHuman, and Convai to develop avatars across VR, AR, and desktop environments. While the system captures physiological signals to detect anxiety, it still lacks real-time coaching, and its AR features remain incomplete. Likewise, Hasan et al. (2023) introduced SAPIEN, a demo platform with emotionally expressive 3D agents powered by LLMs and multilingual speech technologies. Despite its potential for communication training, its short sessions, limited conversational memory, and lack of empirical evaluation restrict practical adoption.

Taken together, Avatar-based and immersive VR and AR systems improve realism and personalization, but accessibility, nonverbal analysis, and validation remain key challenges.

iv. Gamification-based Systems: Gamification has emerged as a powerful strategy in the design of modern job interview systems. By incorporating game elements such as points, achievements, interactive simulations, and immersive environments, these systems aim to enhance candidate engagement, reduce interview anxiety, and deliver more meaningful assessments. This review explores recent developments in gamification-based job interview systems, summarizing the technological approaches and key contributions of several published works. Table 1 in Appendix A depicts an overview of gamification-based systems.

The reviewed systems collectively highlight the diverse application of gamification in job interview settings. Some platforms, like the Metahuman-based VR system (Ashrafi et al., 2024) and the agent-based VR training, prioritize immersion and realism to build

candidate confidence, while others, like Conversate (Daryanto et al., 2025) and the cognitive assessment tool, integrate adaptive simulations and machine learning to personalize feedback and scoring. Despite the varied approaches, all these systems converge on the goal of making interview preparation more interactive, insightful, and equitable.

B. Skills Assessed

i. Soft Skills: Table 2 in Appendix A below presents a comparative overview of prominent AI-powered systems designed for job interview preparation, with a particular emphasis on soft skills and competency-based assessment. Each system is evaluated across multiple dimensions, including its primary skill focus, underlying AI technologies, input modalities, feedback mechanisms, and notable strengths and limitations. This structured comparison offers insights into how different platforms approach interview readiness through multimodal interaction, adaptive feedback, and targeted soft skill development, while also highlighting existing gaps in non-verbal analysis, empirical validation, and immersive realism.

ii. Technical and Coding Skills: Over the past few years, a growing body of research has explored how AI-based systems can simulate, assess, and enhance technical interviews, particularly for roles requiring coding and problem-solving skills. This review in Table 3 in Appendix A depicts recent academic work on AI-driven interview systems designed for technical and coding roles, summarizing the innovations, methodologies, and implications of each.

The reviewed systems collectively demonstrate how AI enhances technical interviews through adaptive questioning and multimodal evaluation.

iii. Behavioral and Critical Thinking Skills: Some systems extend beyond communication by targeting behavioral readiness and critical thinking. For instance, Conversate, developed by Daryanto et al. (2025), employed dialogic reflection, while STAR-based evaluations (Siswanto et al., 2022) scaffold structured behavioral responses. Adaptive questioning (Nagasawa et al., 2023) and bias-testing frameworks (Nofal et al., 2025) also prompt reasoning beyond surface-level answers. However, explicit support for these skills remains limited compared to technical and soft-skill training.

C. Modality

Text-based and Multimodal Interviewing Systems: The reviewed AI-based job interview systems in Figure 3 were categorized based on their primary interaction modalities: Text-based systems primarily rely on typed input and output, enabling question delivery and response evaluation via chat-like interfaces. Textual systems include Anaza et al. (2023), Ashrafi et al. (2024), Bachhav et al. (2023), Chou et al. (2022), Hasan et al. (2023), Mishra et al. (2024), Namratha et al. (2024), Nofal et al. (2025), Rao et al. (2025), Sahani et al. (2025), Senthilkumar et al. (2025), and Wilkie and Rosendale (2024).

Audio-visual systems integrate spoken input and/or output, often enhancing realism by incorporating speech recognition (e.g., Google STT, DeepSpeech) and text-to-speech (TTS) engines (e.g., Amazon Polly, Google TTS). Audio-based systems have been presented by Anaza et al. (2023), Ashrafi et al. (2024), Bachhav et al. (2023), Chou et al. (2022), Hasan et al. (2023), Namratha et al. (2024), Nofal et al. (2025), Rao et al. (2025), Røed et al. (2023), Sahani et al. (2025), Senthilkumar et al. (2025), Siswanto et al. (2022), and Wilkie and Rosendale (2024).

Visual modalities are represented by Anaza et al. (2023), Ashrafi et al. (2024), Bachhav et al. (2023), Chou et al. (2022), Hasan et al. (2023), Namratha et al. (2024), Nofal et al. (2025), Røed et al. (2023), Sahani et al. (2025), Senthilkumar et al. (2025), and Wilkie and Rosendale (2024).

Text-based systems are efficient but limited. Audio improves verbal interaction but depends on reliable speech processing. Video increases realism at higher computational and privacy costs. Together, they show a tradeoff between accessibility and realism.

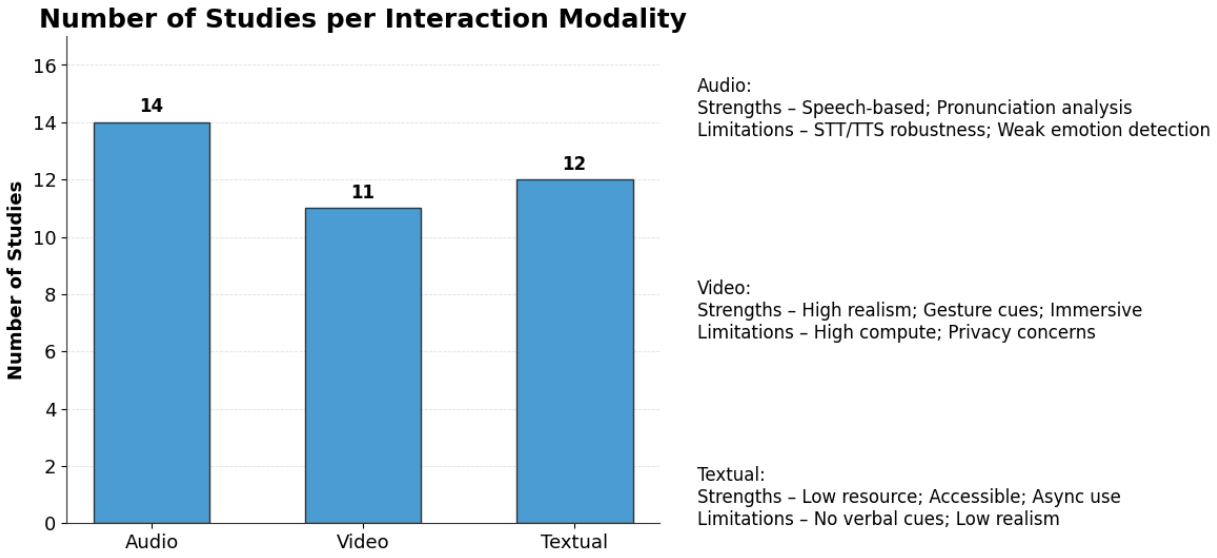


Figure 3: Summary of Job Interview System based on their primary interaction modalities

This section focused on text-based and multimodal systems, while immersive VR/AR treated as a distinct modality was discussed earlier under Technological Approaches (Section 3.A.iii).

By synthesizing existing interviewing systems, this systematic review makes significant contributions. Compared to previous reviews that are either technically narrow e.g., Barpute et al. (2024) or ethically focused like Hunkenschroer and Luetge (2022), or have insufficient system diversity and technical depth e.g., Abedi (2022) which lacks breadth in reviewed platforms and ignores cutting-edge LLM/VR-based tools, our review takes a comprehensive approach by integrating the evaluation of both coding and soft skills, feedback mechanisms, and immersive interaction modalities (VR/AR/avatar), reflecting the multidimensional demands of real-world job interviews. Furthermore, our review bridges the gap between educational training and industry expectations, an area largely overlooked by prior surveys.

Notably, in this process the authors discovered several critical research gaps, including the absence of comprehensive systems capable of concurrently evaluating technical coding skills and behavioral competencies, inadequate real-time emotionally adaptive feedback, limitations in avatar-driven interactions, insufficient adaptability to varying professional and user contexts, lack of specialized support for realistic coding interview simulations, insufficient attention to plagiarism detection and limited personalized interviewing experiences. This paper identifies several key future research directions essential for advancing interview preparation systems. Specifically, progress should focus on:

- Integrating technical and behavioral assessments, enabling holistic evaluations that capture both domain expertise and interpersonal competencies.
- Enhancing avatar realism and emotional responsiveness, thereby fostering more authentic and immersive candidate–interviewer interactions.
- Providing real-time adaptive feedback, ensuring that learners receive personalized guidance aligned with their evolving performance.
- Developing robust coding simulations with plagiarism detection, to uphold integrity and rigor in technical skill assessment.
- Designing authentic, participatory mock interviews, mirroring real-world hiring practices and supporting experiential learning.

Collectively, these directions guide the development of intelligent, inclusive platforms that replicate real-world interview complexity while enhancing candidate readiness through personalized and adaptive

training. The prospective architecture transformation in AI-Driven Interview Systems from the present context is illustrated in Figure 4.

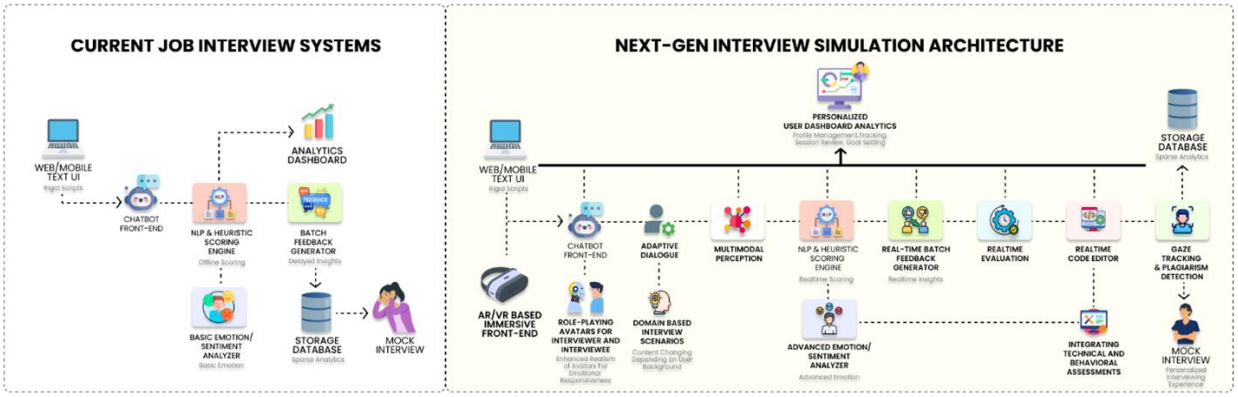


Figure 4: Architectural trajectory of AI-driven interview systems

Here, this figure 4 illustrates the evolution from conventional job interview systems to proposed next-generation AI-driven interview simulation architecture. The left side depicts current systems that primarily rely on web/mobile text-based user interfaces, and chatbot front ends powered by offline NLP and heuristic scoring. These systems often provide delayed batch feedback, rely on basic sentiment analysis, and support only limited user analytics and mock interview functionality. In contrast, the right side presents a next-generation architecture that integrates AR/VR-based immersive interfaces, emotionally responsive role-playing avatars, domain-based adaptive dialogue, and multimodal perception. Real-time scoring, evaluation, and feedback are supported by advanced NLP models and sentiment analyzers. Technical assessments are conducted through an embedded real-time code editor with integrated gaze tracking and plagiarism detection. A personalized user dashboard enables analytics-driven session review, goal tracking, and profile management. Collectively, these enhancements enable more engaging, adaptive, and realistic interview simulations that holistically assess both technical and behavioral competencies.

The following section interprets the reviewed interviewing systems through the framework of the five research questions.

4. ANALYSES OF RESEARCH QUESTIONS

A. RQ1: What are the key features, technologies, and pedagogical strategies employed in existing job interviewing systems?

Section 3 outlines the technological evolution of job interview systems, classifying them into four primary categories: rule-based, AI-driven, avatar-based/immersive, and gamification-enhanced platforms. Rule-based systems rely on scripted logic and predefined keyword triggers, offering reliable and consistent interactions but lacking adaptability and personalization. In contrast, AI-driven systems leverage advanced techniques such as natural language processing, computer vision, and LLMs to enable real-time, multimodal evaluation. These platforms provide detailed, context-aware feedback on candidate performance, encompassing verbal, nonverbal, and emotional dimensions, thereby supporting a more comprehensive assessment of interview readiness.

Each category offers distinct affordances but exhibit varied levels of pedagogical integration. Rule-based systems align with behaviorist pedagogy, emphasizing scripted interactions and fixed feedback, though they lack adaptability and depth. Intelligent systems employing NLP and LLMs enable dynamic, real-time feedback aligned with formative and adaptive learning principles, but few consistently incorporate structured pedagogical scaffolding. Immersive platforms support experiential learning through realistic simulations, reflecting constructivist ideals; however, they often neglect structured reflection and personalized guidance. Gamified systems enhance engagement but typically lack instructional depth, with feedback and learning pathways remaining underdeveloped.

So, technically robust but pedagogically limited, these systems require adaptive, learner-centered strategies to enable effective and transferable learning.

B. RQ2: To what extent do existing job interview systems in the computing field bridge the gap between academic training and real-world hiring expectations?
The review finds a moderate alignment between job interview systems and real-world expectations. Some systems offer realistic whiteboard style coding simulations and behavioral analysis, bridging academic exercises with practical hiring practices. However, a significant portion of platforms still focus narrowly on either behavioral or technical aspects, failing to present the integrative complexity of actual job interviews.

A few systems have made strides toward realism and engagement and integrate dialogic feedback and transcript annotation, encouraging reflective learning aligned with real-world communication tasks. Nonetheless, many reviewed systems lack authentic, industry-driven evaluation models and employer-aligned performance metrics.

C. RQ3: How do these systems support different aspects of professional development, including communication skills, critical thinking, and behavioral readiness?
Soft skills development is increasingly embedded in intelligent and immersive systems. AI-enhanced platforms like *SAPIEN* and *InterviewPal* incorporate sentiment analysis, speech modulation, and facial emotion recognition to deliver multimodal feedback (Hasan et al., 2023; Namratha et al., 2024). This allows candidates to reflect not only on what they say, but also how they say it, a key component of behavioral readiness.

Critical thinking is indirectly supported through scenario-based questioning, adaptive follow-up prompts, and STAR model-based evaluations. Yet, explicit support for reflective learning and metacognitive feedback is present in only a few systems (e.g., *Conversate*), indicating an underexplored opportunity for systems to scaffold users' self-regulated learning processes (Daryanto et al., 2025).

Overall, systems integrating emotion-aware AI and adaptive feedback mechanisms are more likely to foster deep skill development and sustained user engagement.

D. RQ4: What gaps exist in the current systems that future research must address to build intelligent, context-aware, and career-aligned interview preparation platforms?

Our analysis highlights several underexplored directions in the development of job interview simulation systems. Notably, there is a lack of platforms that integrate mock coding and behavioral interviews in a unified environment, despite the prevalence of hybrid formats in real-world hiring processes particularly for computing students. Furthermore, current systems show limited alignment with computing education, missing opportunities to embed interview training into curriculum-relevant activities like code reviews or plagiarism detection.

Real-time coding environments remain largely neglected, with most systems offering non-interactive assessments that fail to simulate live technical interviews. The importance of real-time mock coding interview lies in their continued relevance for assessing core technical competencies. Major technology employers still rely on live coding interviews to assess core competencies such as algorithmic reasoning, problem decomposition, and decision making under ambiguity. At present, software engineering roles are shifting toward higher level responsibilities, including system design, debugging, code review, all of which still depend on strong foundational coding ability. From an educational perspective, computing curricula must prepare students for both current hiring practices and industry expectations, which further strengthens the value of real-time, proctored coding environments in interviewing systems.

Although avatars are increasingly used in interview training, their contribution to soft skill development remains limited by weak emotional expressiveness, adaptability, and feedback quality. Emotionally adaptive feedback is also uncommon, as few systems use multimodal signals to generate dynamic, personalized responses. In addition, most platforms lack content adaptation based on user interests or prior performance, reducing relevance and long-term engagement. These limitations highlight important opportunities for future research and innovation.

E. RQ5: What are the major challenges and limitations (technical, pedagogical, ethical) faced by these systems in achieving sustained learning impact and user trust?
From a technical perspective, many systems face challenges such as latency in real-time multimodal processing, limited scalability, and unstable integration across core components, including speech-to-text, natural language processing, and computer vision. Pedagogically,

feedback mechanisms are often generic, lacking the granularity and adaptability required for personalized learning paths or formative assessment. Ethically, major concerns persist around potential privacy risks associated with the collection of sensitive multimodal user data. Together, these challenges hinder the effectiveness, fairness, and trustworthiness of current systems, posing significant barriers to their widespread adoption and long-term impact in educational and professional settings.

5. CONCLUSION

This article presented a systematic review of AI-driven interview systems, synthesizing research across four technological approaches rule-based, intelligent, avatar-based immersive, and gamification. By examining assessed skills and interaction modalities, the review provided a holistic account of how current systems simulate interviews, deliver feedback, and support professional skill development. A central contribution of this work is its integration of technical and behavioral perspectives, which prior reviews have largely treated in isolation. In doing so, the study establishes a comprehensive foundation for aligning interview preparation platforms with both academic training and employer expectations. This review is the first to systematically integrate technical, behavioral, and immersive perspectives in interview systems.

The analysis shows that advances in NLP, LLMs, and immersive VR and AR have improved realism and adaptivity, but the field remains fragmented. Key limitations include weak personalization, limited affective and multimodal feedback, scarce empirical validation, and a lack of standardized benchmarks. For Information Systems and Computer Information Systems education, these findings highlight both the promise and the current limits of intelligent interview simulations for career readiness. Future systems should integrate real-time coding tasks, plagiarism-aware assessment, affective feedback, culturally inclusive multilingual design, and rigorous benchmarking to support fairness and transparency. A notable limitation of the existing literature is the limited empirical validation across the systems. Many platforms remain at the prototype or demonstration stage, with evaluations based on small sample sizes, short-duration studies, and non-standardized metrics. The field would benefit significantly from large-scale, longitudinal studies that employ standardized evaluation frameworks to establish robust evidence of system effectiveness on learning outcomes, interview performance, and career readiness.

By synthesizing the existing literature, identifying critical gaps, and outlining research directions, this review establishes a roadmap for the next generation of intelligent, inclusive, and context-aware interview preparation systems. Such systems are essential for bridging the gap between academic preparation and professional hiring demands in an increasingly competitive global employment landscape.

6. REFERENCES

Adiani, D., Itzkovitz, A., Bian, D., Katz, H., Breen, M., Hunt, S., … & Sarkar, N. (2022). Career interview readiness in virtual reality (CIRVR): A platform for simulated interview training for autistic individuals and their employers. ACM Transactions on Accessible Computing (TACCESS), 15(1), 1–28. https://doi.org/10.1145/3505560

Anaza, E., Mabrey, P., Sato, M., Miller, O., & Thompson, J. (2023). Improving student interview preparation through collaborative multimodal mock-interview assignments. Sport Management Education Journal, 17(2), 164–176. https://doi.org/10.1123/smej.2021-0021

Ashrafi, N., Vona, F., Ringsdorf, C., Hertel, C., Toni, L., Kailer, S., … & Voigt-Antons, J. N. (2024, October). Enhancing job interview preparation through immersive experiences using photorealistic, AI-powered Metahuman avatars. In 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (pp. 345–346). IEEE. https://doi.org/10.1109/ISMAR-Adjunct64951.2024.00083

Bachhav, P. K., Khandale, O. S., & Karnavat, S. R. (2023, November). AIVATAR: Scrutinizing interview readiness platform with intelligent assessment and aptitude. International Research Journal of Modernization in Engineering Technology and Science, 5(11), 1900–1903. https://www.irjmets.com/uploadedfiles/paper//issue_11_november_2023/46469/final/fin_irjmets1700472402.pdf

Barpute, J. V., Wattamwar, O., Pakjade, S., & Diwate, S. (2024, December). A survey of AI-driven mock interviews using GenAI and machine learning (InterviewX). In 2024 4th

International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS) (pp. 217–224). IEEE. https://doi.org/10.1109/ICUIS64676.2024.10866631

Abedi, M. (2022). Towards a reference architecture of AI-based job interview systems. Ecole Polytechnique de Montréal. https://publications.polymtl.ca/10215/

Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., & André, E. (2013, September). A job interview simulation: Social cue-based interaction with a virtual character. In 2013 International Conference on Social Computing (pp. 220–227). IEEE. https://doi.org/10.1109/SocialCom.2013.39

Belkaid, M., & Sabouret, N. (2014). A logical model of theory of mind for virtual agents in the context of job interview simulation. arXiv preprint arXiv:1402.5043. https://doi.org/10.48550/arXiv.1402.5043

Boudjani, N., Colas, V., Joubert, C., & Amor, D. B. (2023, May). AI chatbot for a job interview. In 2023 46th MIPRO ICT and Electronics Convention (MIPRO) (pp. 1155–1160). IEEE. https://doi.org/10.23919/MIPRO57284.2023.10159831

Chou, Y. C., Wongso, F. R., Chao, C. Y., & Yu, H. Y. (2022, April). An AI mock-interview platform for interview performance analysis. In 2022 10th International Conference on Information and Education Technology (ICIET) (pp. 37–41). IEEE. https://doi.org/10.1109/ICIET54416.2022.9753837

Daryanto, T., Ding, X., Wilhelm, L. T., Stil, S., Knutsen, K. M., & Rho, E. H. (2025). Conversate: Supporting reflective learning in interview practice through interactive simulation and dialogic feedback. Proceedings of the ACM on Human-Computer Interaction, 9(1), 1–32. https://doi.org/10.1145/3701188

Dascalescu, S., Dumitran, A. M., & Vasiluta, M. A. (2025). Leveraging generative AI for enhancing automated assessment in programming education contests. arXiv preprint arXiv:2506.05990. https://doi.org/10.48550/arXiv.2506.05990

Dougherty, Q., & Mehta, R. (2025, May). Proving the coding interview: A benchmark for formally verified code generation. In 2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code) (pp. 72–79). IEEE. https://doi.org/10.1109/LLM4Code66737.2025.00014

Fulk, H. K., Dent, H. L., Kapakos, W. A., & White, B. J. (2022). Doing more with less: Using AI-based Big Interview to combine exam preparation and interview practice. Issues in Information Systems, 23(4), 118–127. https://doi.org/10.48009/4_iis_2022_118

Geng, W., Zhou, C., & Bian, Y. (2024). Change gently: An agent-based virtual interview training for college students with great shyness. Virtual Reality, 29(1), 12. https://doi.org/10.1007/s10055-024-01076-y

Gomez, N., Batham, S. S., Volonte, M., & Do, T. D. (2025). Virtual interviewers, real results: Exploring AI-driven mock technical interviews on student readiness and confidence. arXiv preprint arXiv:2506.16542. https://doi.org/10.48550/arXiv.2506.16542

Hasan, M., Ozel, C., Potter, S., & Hoque, E. (2023, September). SAPIEN: Affective virtual agents powered by large language models. In 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1–3). IEEE. https://doi.org/10.1109/ACIIW59127.2023.10388188

Hassan, S. Z., Sabet, S. S., Riegler, M. A., Baugerud, G. A., Ko, H., Salehi, P., … & Halvorsen, P. (2023). Enhancing investigative interview training using a child avatar system: A comparative study of interactive environments. Scientific Reports, 13(1), 20403. https://doi.org/10.1038/s41598-023-47368-2

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. Journal of Business Ethics, 178(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Jarvis, A., Ho, A., & Lim, G. (2024). Impressing artificial intelligence: Automated job interview training in professional English subjects. RELC Journal. Advance online publication. https://doi.org/10.1177/00336882241245449

Kawahara, T., Inoue, K., & Lala, D. (2021). Intelligent conversational android ERICA applied to attentive listening and job interview. arXiv preprint arXiv:2105.00403. https://doi.org/10.48550/arXiv.2105.00403

Kawai, H., Muraki, Y., Yamamoto, K., Lala, D., Inoue, K., & Kawahara, T. (2022, September). Simultaneous job interview system using multiple semi-autonomous agents. In Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL) (pp. 107–110). ACL. https://doi.org/10.18653/v1/2022.sigdial-1.12

Lee, J., Kim, H., Kim, K. H., Jung, D., Jowsey, T., & Webster, C. S. (2020). Effective virtual patient simulators for medical communication training: A systematic review. Medical Education, 54(9), 786–795. https://doi.org/10.1111/medu.14152

Leutner, F., Codreanu, S. C., Brink, S., & Bitsakis, T. (2023). Game-based assessments of cognitive ability in recruitment: Validity, fairness and test-taking experience. Frontiers in Psychology, 13, 942662. https://doi.org/10.3389/fpsyg.2022.942662

Luo, X., Wang, Y., Lee, L. H., Xing, Z., Jin, S., Dong, B., … & Hui, P. (2024). Using a virtual reality interview simulator to explore factors influencing people's behavior. Virtual Reality, 28(1), 56. https://doi.org/10.1007/s10055-023-00934-5

Mishra, P. K., Arulappan, A. K., Ra, I. H., Rose, G., & Lee, Y. S. (2024, November). AI-driven virtual mock interview development. In 2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS & ISIS) (pp. 1–4). IEEE. https://doi.org/10.1109/SCISISIS61014.2024.10760210

Mutter, B. A., Tetzlaff, J., Moher, D., & Altman, D. G. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Nagasawa, F., Okada, S., Ishihara, T., & Nitta, K. (2023). Adaptive interview strategy based on interviewees' speaking willingness recognition for interview robots. IEEE Transactions on Affective Computing. Advance online publication. https://doi.org/10.1109/TAFFC.2023.3309640

Namratha, M., Lokesh, R., Bhat, P., Srikanth, N., & Gagan, M. (2024, April). InterviewPal—Elevating interview automation with deep learning and natural language processing perspectives. In 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICETCS61022.2024.10543368

Nofal, A. B., Ali, H., Hadi, M., Ahmad, A., Qayyum, A., Johri, A., … & Qadir, J. (2025). AI-enhanced interview simulation in the metaverse: Transforming professional skills training through VR and generative conversational AI. Computers and Education: Artificial Intelligence, 8, 100347. https://doi.org/10.1016/j.caeai.2024.100347

Pandey, R., Chaudhari, D., Bhawani, S., Pawar, O., & Barve, S. (2023, March). Interview bot with automatic question generation and answer evaluation. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1279–1286). IEEE. https://doi.org/10.1109/ICACCS57279.2023.10112918

Pathak, G. (2024, June). Asynchronous AI interviews for technical roles: Improving candidate experience and reducing interview fatigue. International Journal of Novel Research and Development, 9(6), f140–f149. http://ijnrd.org/papers/IJNRD2406504.pdf

Qin, C., Zhu, H., Shen, D., Sun, Y., Yao, K., Wang, P., & Xiong, H. (2023). Automatic skill-oriented question generation and recommendation for intelligent job interviews. ACM Transactions on Information

Systems, 42(1), 1–32. https://doi.org/10.1145/3604552

Qin, C., Zhu, H., Zhu, C., Xu, T., Zhuang, F., Ma, C., … & Xiong, H. (2019, July). DuerQuiz: A personalized question recommender system for intelligent job interview. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2165–2173). ACM. https://doi.org/10.1145/3292500.3330706

Prajwal, R., Divya, D., Gowda, H. K., & Hemalatha, K. L. (2023). AI interview agent for predicting communication skills and personality traits. International Journal of Engineering Research & Technology (IJERT), 11(8), RTCSIT–2023. https://doi.org/10.17577/IJERTCONV11IS08018

Rädel-Ablass, K., Schliz, K., Schlick, C., Meindl, B., Pahr-Hosbach, S., Schwendemann, H., … & Miersch, C. (2025). Teaching opportunities for anamnesis interviews through AI-based teaching role plays: A survey with online learning students from health study programs. BMC Medical Education, 25(1), 259. https://doi.org/10.1186/s12909-025-06756-0

Rai, A. N. (2025, March 17). AI mock interview chatbot using Gen AI. International Journal of Science, Engineering and Technology, 13(2), 189–197. https://www.ijset.in/wp-content/uploads/IJSET_V13_issue2_242.pdf

Rao, G. S., Jaiganesh, M., & Parida, P. K. (2025, April). AI-powered virtual job interview simulator using natural language processing. In 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1413–1421). IEEE. https://doi.org/10.1109/ICOEI65986.2025.11013504

Røed, R. K., Baugerud, G. A., Hassan, S. Z., Sabet, S. S., Salehi, P., Powell, M. B., … & Johnson, M. S. (2023). Enhancing questioning skills through child avatar chatbot training with feedback. Frontiers in Psychology, 14, 1198235. https://doi.org/10.3389/fpsyg.2023.1198235

S, P., Siranjeevi, K., Kumar, N. V., Pathmesh, G., & Ponnarasu, A. (2025, May). NexInterview—AI-driven mock interview preparation platform. International Journal of Advanced Research in Science, Communication and Technology, 5(7), 26835. https://www.ijarsct.co.in/Paper26835.pdf

Sahani, K. K., Khan, M. S., Khatwani, S., Gupta, S., & Dubey, A. (2025). A smart interview simulator using AI avatars and real-time feedback mechanisms (AI avatar for interview preparation). International Journal of Engineering Technologies and Management Research, 12(5), 66–75. https://doi.org/10.29121/ijetmr.v12.i5.2025.1618

Sahu, A., Khare, R. K., Singh, Y. R., & Sahu, Y. (2025, June). AI interviewer using generative AI. In International Conference on Advances and Applications in Artificial Intelligence (ICAAAI 2025) (pp. 1231–1240). Atlantis Press. https://doi.org/10.2991/978-94-6463-738-0_94

Senthilkumar, K., Ranjith, S., Sivasakthi, M., & Ramvignesh, R. (2025, March). AI-based mock interview system using natural language processing. In 2025 International Conference on Advanced Computing Technologies (ICoACT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICoACT63339.2025.11005032

Siswanto, J., Suakanto, S., Andriani, M., Hardiyanti, M., & Kusumasari, T. F. (2022). Interview bot development with natural language processing and machine learning. International Journal of Technology, 13(2), 274–285. https://doi.org/10.14716/ijtech.v13i2.5018

Takeuchi, N., & Koda, T. (2021, September). Job interview training system using multimodal behavior analysis. In 2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1–3). IEEE. https://doi.org/10.1109/ACIIW52867.2021.9666270

Golande, S. V., Dandage, P., Jadhav, A., Mohite, P., & Shahane, A. (2025). Mock interview evaluator powered by AI. Excel Journal: Technology, Engineering, & Management Research, 12(2), 1–9. https://www.researchgate.net/publication/391976964_MOCK_INTERVIEW_EVALUATOR_POWERED_BY_AI

Vardarlıer, P. (2023). A system that allows users to have a job interview experience. Sustainability, 15(22), 16031. https://doi.org/10.3390/su152216031

Wilkie, L., & Rosendale, J. (2024). Efficacy and benefits of virtual mock interviews: Analysing student perceptions of digital employment preparations. Journal of University Teaching and Learning Practice, 21(1), 177–195. https://doi.org/10.53761/rvtxt659

APPENDIX A
Tables

Papers	Gamification	Technology	Strengths	Limitations
(Ashrafi et al., 2024)	Immersive Metahuman avatars in VR/AR/desktop	Unreal Engine (Metahuman), Convai (Chatbot), Meta Quest 3, Empatica Embrace+ (biosensor), TTS/STT modules	Photorealism, emotional impact tracking, comparative analysis	No real-time feedback, lacks adaptive coaching, AR in progress
(Daryanto et al., 2025)	Adaptive LLM-based simulation with dialogic feedback	Web app with GPT-3.5/4 and transcript annotation	Realistic, interactive simulation, dialogical feedback	Lacks multimodal feedback, not domain-specific
(Vardarlıer et al, 2023)	Gamified scoring through points and achievements, immersive VR-based simulations	Web-based system, NLP, emotion and gesture analysis, VR headset, chatbot, sensors	Gamified scoring, immersive UX, NLP & emotion tracking	Feedback general, no validation study
(Leutner et al., 2023)	Cognitive ability assessment via games (Shapedance, Numerosity)	Machine learning (Ridge Regression with Bias Penalization), ICAR, CRT, HireVue platform	Valid and fair cognitive scoring, positive user feedback	Concerns over face validity, narrow task types
(Geng et al., 2024)	Immersive VR agents and biofeedback	VR headset with EEG/ECG-enabled virtual agents	Reduced anxiety, improved performance, multimodal	Requires biofeedback hardware, small sample

Table 1: Summary of Recent Advancements in Gamification-Based Job Interview Systems

Papers	Soft Skill Focus	Key AI Technologies Used	Strengths	Limitations
(Rao et al., 2025)	Technical, behavioral, situational questions; response relevance, semantic coherence, sentiment, keywords.	GPT/BERT, semantic similarity, sentiment analysis, and Google Speech-to-Text (STT) for speech-to-text.	Scalable, adaptive QG, heatmap feedback, role-based scoring	No emotion recognition, limited non-verbal feedback
(Senthilkumar et al., 2025)	Fluency, coherence, and knowledge are assessed through verbal communication and behavioral observation.	Speech-to-Text (Mozilla DeepSpeech), Text-to-Speech (Google TTS, Amazon Polly), Natural Language Processing (GPT, T5, BERT, RoBERTa), and Computer Vision (OpenCV, MediaPipe).	Multimodal feedback (verbal and behavioral), real-time evaluation	Limited scenario variation

Papers	Soft Skill Focus	Key AI Technologies Used	Strengths	Limitations
(Nofal et al., 2025)	Leadership, communication, and domain knowledge; quantitative reliability metrics; and bias analysis	Unity 3D + OpenXR for the VR environment, ChatGPT for question generation and feedback, Wit. AI for Text-to-Speech, Whisper. AI for Speech-to-Text, BART/topic modeling & Distil-BERT for semantic similarity, and RoBERTa for the scorer model.	Bias analysis, consistent scoring, VR immersion, real-time feedback	Hardware intensive, no gaze detection, lab-only testing
(Sahani et al., 2025)	Key qualities for effective communication and professional aptitude, emphasizing clear expression, topical relevance, and proficiency in both technical and interpersonal skills.	GPT-3, Whisper, Google STT/TTS, React.js	Immersive, scalable 3D avatar interface with real-time, multi-modal feedback. High STT/TTS accuracy, low latency (<2.5s)	No non-verbal analysis lacks multilingual support
(Bachhav et al., 2023)	Common hiring assessments include general job-skill Q&A, soft-skill simulations, and aptitude tests (covering logical and verbal reasoning).	STT, TTS, web-based layered architecture	Aptitude test, realistic Q&A, anxiety reduction	No validation study, no emotional feedback, no coding tasks
(Chou et al., 2022)	Intrinsic and DISC personality traits, along with facial emotions, head poses, speaking rate, amplitude, and pitch, contribute to interview performance.	Linear regression for scoring, Gamma distribution, and Automatic Relevance Determination (ARD).	Personality/behavioral modeling, asynchronous simulations	No live interaction, no adaptive feedback
(Namratha et al., 2024)	Factors assessed include content accuracy, response delivery, emotional state (via facial expressions), confidence levels (through voice analysis), and sentiment.	NLP, along with Transformer-based models for STT conversion, and Convolutional Neural Networks (CNNs) for image analysis	Emotion & voice analysis, adaptive feedback	Generic scoring models, unclear personalization logic
(Siswanto et al., 2022)	Evaluation of competency levels using the STAR model (Behavioral Event Interview) based on predefined categories.	NLP techniques, such as tokenization, stop-word removal, stemming, and part-of-speech tagging. Machine Learning methodologies, including Bayesian inference and TF-based weighting	Real-time competency evaluation, scalable	Behavior only via text, lacks audiovisual feedback

Papers	Soft Skill Focus	Key AI Technologies Used	Strengths	Limitations
(Hasan et al., 2023)	Language learning, mental health, public speaking, social skill development, and emotionally expressive communication	Large Language Models (LLMs), STT, TTS, emotion modeling, and avatars (within a 3D game engine).	Multilingual, emotion-aware avatar, personalized coaching	Short duration, no memory persistence, demo stage

Table 2: Summary of Interview Systems Focused on Soft Skills

Papers	Main Features	Technologies Used	Strengths	Limitations
(Sahu et al., 2025)	Generates coding questions from resumes; cheat detection; live editor	LLMs (GPT), facial/voice analysis, behavioral metrics	Context-aware question generation, cheat detection, interactive coding	Facial/voice model performance unclear, lacks soft skill evaluation
(Gomez et al., 2025)	Whiteboard-style technical interviews, code and voice analysis	Multimodal NLP, whiteboarding tools, and feedback systems	Realistic simulation, high user engagement, multimodal analysis	No emotion detection, lacks plagiarism control
(Qin et al., 2024)	Generates skill-aligned technical questions using deep learning and graph-based recommendation	Skill entity mining, question generation, neural ranking models	High relevance of Questions, user skill adaptability, modular design	Limited soft skill integration, rule-based rigidity in places
(Dougherty et al., 2025)	Coding interview benchmark with formal verification	Formal methods, Lean 4, coding benchmark creation	Verified test cases, benchmarking standard for interviews	No user interaction features, lacks AI/NLP analysis
(Dascalescu et al., 2025)	Generates edge case test cases for coding contests	LLMs, test case generation, code evaluation	Complement human test design, increases grading accuracy	Not a traditional interview system, no behavioral component
(Chou et al., 2022)	Mock interview platform for tech and behavioral evaluation	AI evaluation models (unspecified), behavioral scoring	Dual focus on technical and behavioral, asynchronous simulation	No live feedback, lacks scenario customization

Table 3: Summary of Recent Advancements in Technical Job Interview Systems

AI-Powered Study Assistant for Exams: QuizAI

Thi Hong Anh Nguyen
honganh3179@gmail.com
City University of Seattle
Seattle, WA 98121

Sam Chung
chungsam@cityu.edu
City University of Seattle
Seattle, WA 98121

Abstract

The advancement of Artificial Intelligence (AI) has created numerous opportunities to enhance the self-learning experience. Automated quiz generation has attracted attention as a way to support self-learning, particularly in digital education environments. However, existing systems often face key limitations: they frequently lack references, fail to provide meaningful feedback, and do not adapt content to individual needs. These shortcomings hinder the systems' effectiveness in promoting deep understanding and engagement. This paper introduces QuizAI, a novel AI-powered quiz generation system designed to address these issues. QuizAI can generate multiple-choice questions (MCQs) and open-ended questions based on user-provided sources, including PDF files and web pages. By utilizing Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs), the system ensures that questions are grounded in their source material and accompanied by supportive feedback. While QuizAI demonstrates comparable latency in generating dynamic content, it has produced a 50% duplication rate in five test documents, particularly when processing a larger number of questions per request. Despite these limitations, QuizAI achieves moderate to strong performance in terms of response time and question diversity, effectively complementing existing systems in the field. The proposed system enhances learning outcomes, supports recollection, and helps reduce learners' anxiety in self-paced educational settings.

Keywords: Quiz Generator, AI-Quiz Maker, RAG, Retrieval-Augmented Generator, Personalized Practice, LLM-Powered Learning,

GitHub: https://github.com/hongeinh/quizai.git

Recommended Citation: Nguyen, T., Chung, S., (2026). AI-Powered Study Assistant for Exams: QuizAI. *Journal of Information Systems Applied Research and Analytics*, v19(n3) pp 68-81. DOI# https://doi.org/10.62273/FHRO6233

AI-Powered Study Assistant for Exams: QuizAI

Thi Hong Anh Nguyen and Sam Chung

1. INTRODUCTION

Continuous learning is essential for success, especially in today's rapidly changing digital landscape (Hennekam, 2015). London and Smither (1999) emphasize that empowered self-development and continuous learning are crucial for individuals to adapt to new environments, enhance their skills, and achieve personal growth. With the rise of technology and the availability of vast amounts of data, numerous learning platforms and applications have emerged to help learners acquire new skills and specific knowledge, often without the need for formal education. This allows learners to study at their own pace while following a structured pathway.

However, despite the abundance of educational tools, many are primarily focused on subjects such as linguistics and mathematics. In contrast, the increasing demand for tech-related skills, particularly in fields like computer science, has exposed a gap in the availability of learning resources for technology-focused areas. As the industry continues to evolve, a growing number of learners are seeking to build or reinforce their technical knowledge to remain competitive in the job market.

Laguna et al. (2021) indicate that there is growing interest in skill enhancement within the tech industry. This increasing demand is driven by individuals who are looking to improve their employability or address skill gaps through industry certifications and post-secondary credentials.

Supporting this trend, Ehlinger and Stephany (2023) conducted a large-scale analysis of over eleven million job postings across the UK from 2018 to mid-2024. Their findings reveal that employers are placing greater emphasis on specific, demonstrable skills rather than traditional academic degrees, especially in the field of Artificial Intelligence (AI).

Many professionals are motivated to pursue certifications, but they often face significant challenges when preparing for exams. One major issue is the sheer volume of content, which includes textbooks, online articles, and technical documentation. Learners must sift through complex materials to identify key concepts within tight time constraints, which can be overwhelming. This challenge is particularly tough for individuals with weaker academic backgrounds, who may struggle to succeed in virtual learning environments due to a lack of interaction and engagement. As noted by Baum and McPherson (2019), students who are less academically prepared are especially vulnerable in fully online courses.

While resources such as practice tests and study groups are available, they often do not provide the personalized and adaptive revision that many learners require. Most current tools lack intelligent, on-demand question generation and do not offer actionable feedback. According to Patterson et al. (2024), most digital assessment platforms provide limited feedback and insufficient support for open-ended questions, which are crucial for evaluating deep understanding.

Additionally, popular platforms like Duolingo, Socratic, and ALEKS are effective in their respective fields, but they do not meet the specific needs of learners preparing for technical certifications. These tools primarily focus on general education and do not cover the foundational knowledge required in areas such as software development, data science, or cybersecurity.

To foster better engagement, Hirulkar and Athawale (2024) introduced a gamified multiple-choice quiz generator, promoting a competitive and collaborative learning environment. While effective in boosting participation, the system lacks essential features for personalized revision and alternative question types beyond multiple-choice questions (MCQs).

To overcome these limitations, this paper introduces an AI-powered quiz generator that dynamically creates personalized quizzes based on the learner's study materials, particularly PDFs and websites. By extracting key concepts and topics from the content, the system generates relevant multiple-choice questions that adapt in real time based on the learner's performance.

As users engage with the quizzes, the system adapts the difficulty of the questions and targets specific areas based on the frequency of correct and incorrect answers. This feature allows for focused revision by emphasizing the concepts that the learner struggles with the most. Additionally, the tool provides immediate and detailed feedback, including contextual explanations and direct references to the source material. This approach not only encourages active and targeted learning but also empowers users to track their progress and effectively address any knowledge gaps. By transforming static content into interactive, personalized assessments, the system aims to enhance understanding and retention.

2. BACKGROUND

Assessment is a vital component of a student's learning journey, helping to evaluate knowledge and reinforce learning through active recall and feedback. Among the various assessment methods, quizzes prove to be an effective tool for enhancing memory retention and engagement. They create a low-pressure, interactive environment that encourages students to reflect on their understanding of the material. However, effective learning relies not only on the format of the assessment but also on the quality of the feedback provided. Supportive feedback has been shown to positively influence students' mental preparation and self-assessment (Patterson et al., 2024). Incorporating mechanisms that simulate human interaction can enhance students' learning experiences and progress.

Recent advancements in Artificial Intelligence (AI), particularly in Large Language Models (LLMs), have facilitated the generation of questions. Several studies have developed systems capable of creating multiple-choice and open-ended questions from text inputs (Kurdi et al., 2019; Mulla & Gharpure, 2023; Das et al., 2021; Hirulkar & Athawale, 2024). Despite these advancements, existing systems often lack effective feedback mechanisms and tend to generate generic questions due to the limited availability of datasets.

3. RELATED WORK

The necessity for upskilling has become increasingly critical, particularly in the tech and AI industries. Laguna et al. (2021) emphasize that hiring managers prioritize technical skills when evaluating candidates, making certifications an essential complement to project experience and work history. Similarly, Ehlinger and Stephany (2023) note a shift towards skill-based hiring in the AI and green technology sectors, where digital certifications serve as important proof of an individual's technical abilities. Together, these studies highlight the growing demand for skill development through certifications.

Although online courses are accessible, Baum and McPherson (2019) argue that human interaction and engagement are essential for achieving positive learning outcomes. To improve the learning experience, they recommend incorporating feedback, competition, and collaboration as key features of these platforms. Supporting this idea, Patterson et al. (2024) demonstrate that providing real-time, supportive feedback during tests can help reduce anxiety and enhance students' self-assessment. These studies suggest that learning platforms should integrate encouraging feedback to maintain students' motivation and improve learning efficiency.

One effective way to enhance student engagement is by using quizzes. Yang et al. (2021) found that testing, including quizzes, significantly boosts academic achievement in various educational settings. Additionally, El-Hashash (2022) demonstrated that weekly quizzes help to improve both attendance and engagement among students. These findings indicate that incorporating quizzes can be a successful strategy for promoting active learning.

Recent advancements in AI have facilitated the automatic generation of questions. Comprehensive reviews by Kurdi et al. (2019) and Mulla and Gharpure (2023) highlight the latest methodologies in question generation, focusing on the integration of feedback, control of question difficulty, and the use of neural architecture such as sequence-to-sequence (seq2seq) models. Das et al. (2021) introduced a system that generates and evaluates subjective questions, promoting deeper understanding rather than mere recall from students. Additionally, Hirulkar and Athawale (2024) developed an AI-based quiz generator that creates multiple-choice questions based on selected topics and difficulty levels. However, the broad selection of topics may limit the effectiveness of targeted learning.

Recent studies are exploring the adjustment of question difficulty in educational settings. Alkhuzaey et al. (2023) conduct a systematic

review of methods for predicting the difficulty of generated questions and advocate for personalized quiz adaptations that are based on students' performance. Tomikawa et al. (2024) utilize transformer models along with Item Response Theory (IRT) to dynamically manage question difficulty. Additionally, Fu et al. (2025) introduce the ConQuer framework, which uses external knowledge to ensure that question generation is grounded, though the system is limited by predefined concepts.

Effective feedback mechanisms are crucial for online educational tools. Tobler (2023) discusses a generative AI-based system for automated grading, addressing challenges such as LLM hallucinations and the need for verified references. Additionally, Quizzio, Junior (n.d.) offers personalized feedback and tracks user progress, although it is limited by the length of user-submitted text. Together, these works emphasize that feedback should be immediate, relevant, and accurate to enhance learning outcomes.

Numerous educational resources exist in unstructured formats, such as PDFs, which necessitate effective methods for content extraction. Siegler (2025) introduces LlamaParse, a tool designed to parse complex documents into usable formats suitable for Retrieval-Augmented Generation (RAG) systems. Additionally, Google Cloud (n.d.) outlines strategies for semantic search and content extraction within its RAG pipeline for PDFs.

Despite significant advancements, current AI-powered educational tools still face notable limitations. These systems often rely on predefined content, lack interactivity, and fail to offer personalized experiences. Furthermore, the evaluation of answers is not yet perfect, as pointed out by Tomikawa et al. (2024), due to limitations in the datasets used. Future systems must accurately analyze diverse content sources and generate meaningful assessments while providing real-time, positive feedback to enhance learning engagement.

4. APPROACH

Figure 1 outlines the main use cases available to users of QuizAI. Users will be authenticated and authorized to upload documents in either PDF or URL format. They can generate quizzes based on the uploaded documents and access previously created documents and quizzes.

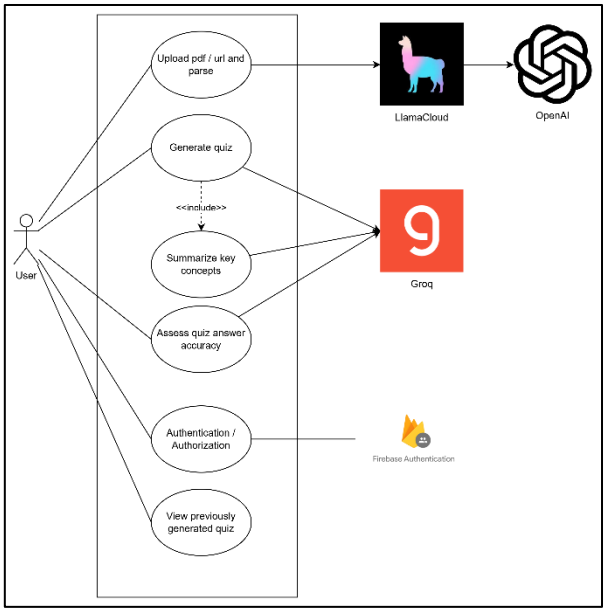


Figure 1: QuizAI Use Case Diagram

4.1. User Requirements

QuizAI requires users to authenticate before accessing its core functionalities. This authentication process protects users' privacy and digital footprints by ensuring that only authorized individuals can access their study materials and activity history. Additionally, it enhances data management and system integrity by preventing the misuse of resources.

Authentication is managed through Firebase Authentication, a service that provides identity management and access control for web and mobile applications. Firebase is a Backend-as-a-Service (BaaS) platform developed by Google, offering tools and services for building web and mobile applications. In QuizAI, users can create a new account using their email and password, or sign in using supported federated identity providers.

4.2. Design

The activity flow is outlined in Figure 2 in the Appendix. Users can either log in or register to access the core features. Once authenticated, users are able to upload a document in either PDF or URL format.

This input is then processed using LlamaParse (via LlamaCloud), a parsing service created to extract structured text and metadata from unstructured sources. Each page or section is extracted and saved to Firebase.

The parsed content is organized into smaller, semantically meaningful chunks using

LlamaIndex (via LlamaCloud). This framework links these chunks to their corresponding embeddings. The embeddings are then stored in a vector database based on FAISS (Facebook AI Similarity Search). FAISS is an open-source library developed by Meta that enables efficient similarity searches for embeddings of multimedia documents. By creating searchable indices for these embeddings, FAISS allows the system to quickly retrieve semantically relevant chunks.

Once embeddings are stored, the system uses Retrieval-Augmented Generation (RAG), which integrates document retrieval with language generation to improve the process. RAG facilitates accurate and grounded question generation from the uploaded and parsed documents.

After the user has chosen the question type and number of questions, the system will retrieve relevant document chunks and generate questions. For open-ended questions, user answers are assessed to evaluate correctness and quality. The user receives feedback with direct citation to the original page or section.

Figure 3, the Class Diagram located in the Appendix, outlines the main entities stored in the database. The User collection holds the user's metadata, while authentication information is securely managed by Firebase authentication and is therefore not included in the User collection. The Document collection contains metadata for documents and ensures that only the file owners can access the content. Lastly, the Question collection stores the questions generated by QuizAI, allowing users to revisit previously generated questions if they wish to attempt them again.

4.3. Implementation

The core REST API endpoints are described below:

- POST /pdf/upload - Accepts PDF input and stores the structured result.
- POST /html/upload – Accepts a webpage URL and stores the structured result.
- POST /process - Processes parsed content and indexes using LlamaIndex.
- GET /chunks/retrieve - Retrieves the most relevant chunk.
- GET /pdf/page - Returns the specific page or section from the original document.
- POST /quiz/generate - Generates multiple-choice or open-ended quizzes based on the document.
- POST /answer/evaluate - Evaluates user-provided answers for open-ended questions.

These APIs enable clear separation of concerns and make the backend extensible for any future additions.

4.4. Technologies Used

Table 1 lists the technologies used for implementation.

Frontend	ReactJS, TailwindCSS
Backend	FastAPI (Python)
Authentication & database	Firebase authentication & Firestore
AI/ML	LlamaCloud
LLM	Groq, model llama-3.1-8b-instant

Table 1: Technologies Used

4.5. Deployment architecture

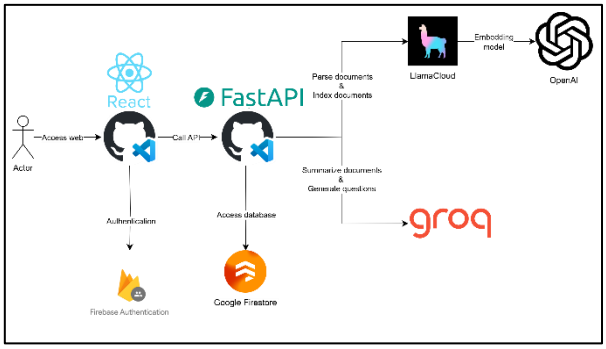


Figure 4: Deployment Diagram

The frontend and backend of the project can be deployed using GitHub Codespaces, a cloud-based environment that provides access to the project's code and ready-to-use coding environments, as illustrated in Figure 4. This setup incorporates several tools and technologies.

For user authentication, Firebase Authentication is employed, supporting both the Email/Password method and third-party Google accounts. User data is stored using Firestore, a service from Google Cloud, which handles internal information, including user profiles, documents, and quizzes.

Quizzes are generated in collaboration with LlamaCloud for RAG, OpenAI for embeddings, and Groq for Large Language Model support. All external tools can be accessed via the API provided and secret keys.

5. DATA COLLECTION

Figure 5 in the Appendix illustrates the Input-Process-Output framework necessary for QuizAI.

Input:
Users are prompted to upload PDFs or provide URL links from which they want to generate quizzes.

Process:
Both PDF files and webpage links must be processed before use. Uploading a PDF will trigger LlamaParse to extract text, while URLs will require BeautifulSoup, a Python library used for parsing HTML and XML documents into plain text.

The extracted text is then cleaned by removing unnecessary content, such as headers, footers, references, and copyright information. SpaCy, a library for tokenization and named entity recognition, is employed to extract key concepts from the document, aiding in the cohesive generation of quizzes.

Output:
The cleaned text and the associated concept metadata are then input into LlamaIndex to create index objects, which are essential for future retrieval steps.

6. DATA ANALYSIS

To assess the performance and reliability of QuizAI, we use both qualitative and quantitative analyses, each focusing on different aspects of system reliability, efficiency, and learning effectiveness. Data is gathered by testing FastAPI endpoints through both automated and manual methods.

Quantitative Analysis:
The quantitative analysis aims to evaluate the efficiency of the system using the following metrics:

API Response Time: The response times for key endpoints—specifically, document processing, quiz generation using a large language model (LLM), and answer evaluation using an LLM—are recorded to calculate the average response time. Figure 6 in the Appendix illustrates the methodology used to calculate the API endpoint response time, while Figure 7 in the Appendix presents the actual response times for each API endpoint.

Question uniqueness refers to the ratio of duplicated to unique quiz questions within a document, which helps evaluate the diversity and relevance of the generated quizzes (see Figure 8 in the Appendix). On average, the repetition rate exceeds 50% for each generation. This issue is illustrated in Figure 9 of the Appendix, where the output of the LLM may contain several duplicated questions. There are two main reasons for this duplication:

1. The prompt given to the LLM does not require strictly unique questions and answer generation.
2. The summarization of the document limited the context available for the LLM, resulting in fewer unique questions being produced.

The quantitative analysis offers insights into system responsiveness and content diversity, which are essential for keeping users engaged.

Qualitative Analysis:
The qualitative analysis focuses on evaluating the accuracy and consistency of outputs generated by the LLM based on the following metrics:

Document Reference Accuracy: The generated questions are compared to the original document to assess their correctness. The Section ID provided by the endpoints (/quiz/generate) can be verified using both manual and automated methods, as illustrated in Figure 10 of the Appendix.

Scoring Consistency: Figure 11 in the Appendix shows how the LLM evaluates and scores a user's response to an open-ended question. Semantically similar user responses to the same question can be used to determine the reliability of the answer evaluation mechanism. Consistency is measured using the standard deviation of scores, as demonstrated in Figure 12 of the Appendix. Through testing, the system consistently assesses answers with an average score of 60 and delivers the same feedback for the same question and answer.
The quantitative analysis offers insights into the factual alignment, coherence, and reliability of the system's outputs.

7. FINDINGS

QuizAI demonstrates efficient processing across its core endpoints. The average response times for each endpoint are shown in Figure 7 in the Appendix. The results reveal that the system experiences moderate latency, with slightly longer processing times when handling large

documents. However, it remains effective in maintaining user engagement, especially when compared to QuizMasterAI by Hirulkar and Athawale (2024), which has longer processed periods. In comparison to Quizzio (Junior), QuizAI maintains similar latency while generating dynamic content.

The quality of question generation is assessed based on the frequency of duplicate questions created in response to user requests. In testing five documents, the system produced a duplication rate of 50%, particularly when a higher number of questions were requested. This limitation arises primarily from two factors: (1) the current prompt does not ensure strict uniqueness, and (2) document summarization is truncated, limiting the context available for question generation. With a shorter summarization length, the process lacks enough information to generate a diverse set of questions.

In comparison, ConQuer by Fu et al. (2025) faces challenges with redundancy control and is constrained by predefined concepts. On the other hand, V-Doc by Ding et al. (2022) does not effectively address question diversity or system performance.

The current system is limited to PDFs and web URLs, excluding other commonly used formats such as Word documents, PowerPoint presentations, and various multimedia sources. Additionally, while answer evaluation is consistently scored for responses that are semantically similar, the current mechanism may not accurately capture subtle nuances.

Despite these limitations, QuizAI demonstrates moderate to strong performance in both response time and question diversity, making it a valuable complement to existing systems in the field. As a result, QuizAI serves as an effective tool for personalized learning and assessment.

8. CONCLUSION

This paper introduces QuizAI, an AI-powered quiz generation system aimed at helping students and learners master content through personalized assessments. By utilizing RAG and LLMs, QuizAI converts PDFs and webpage URLs into interactive quizzes for personal assessment.

QuizAI addresses several key challenges in self-study environments, such as content overload and the lack of direct engagement and feedback. Through both quantitative and qualitative analyses, the system demonstrates its ability to generate useful quizzes for recall and effectively assess users' answers. However, one challenge remains: question duplication, which is influenced by current prompt engineering and context truncation. Additionally, the research scope—specifically, focusing on an AI-powered study assistant for exams—prevented us from tackling pedagogical and ethical considerations.

Future efforts will concentrate on improving question duplication, refining answer evaluation scoring, and supporting a wider variety of content formats. Broader testing across diverse domains and user groups is necessary. We are also considering a refined user interface and user experience (UI/UX) along with an interactive game mode that allows users to compete against each other for enhanced engagement. Furthermore, we need to explore the pedagogical implications of QuizAI: How will it influence learners' critical thinking, engagement, and exam performance? What ethical considerations, such as AI bias and fairness in scoring, must we address?

9. REFERENCES

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2023). Text-based question difficulty Prediction: A Systematic review of automatic approaches. International Journal of Artificial Intelligence in Education, 34(3), 862–914. https://doi.org/10.1007/s40593-023-00362-1

Baum, S., & McPherson, M. (2019). The Human Factor: The Promise & Limits of Online Education. Daedalus, 148(4), 235–254. https://doi.org/10.1162/daed_a_01769

Das, B., Majumder, M., Sekh, A. A., & Phadikar, S. (2021). Automatic question generation and answer assessment for subjective examination. Cognitive Systems Research, 72, 14–22. https://doi.org/10.1016/j.cogsys.2021.11.002

Ding, Y., Huang, Z., Wang, R., Zhang, Y., Chen, X., Ma, Y., Chung, H., & Han, S. C. (2022). V-Doc : Visual questions answers with Documents. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21460–21466. https://doi.org/10.1109/cvpr52688.2022.02083

Ehlinger, E. G., & Stephany, F. (2023). Skills or degree? The rise of Skill-Based hiring for AI and green jobs. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2312.11942

El-Hashash, A. (2022). Weekly Quizzes Reinforce Student Learning Outcomes and Performance in Biomedical Sciences in-course Assessments. Open Journal of Educational Research, 2(4), 168–178. Retrieved from https://www.scipublications.com/journal/index.php/ojer/article/view/273

Firebase Authentication. (n.d.). Firebase. https://firebase.google.com/docs/auth

Fu, Y., Wang, Z., Yang, L., Huo, M., & Dai, Z. (2025). ConQuer: A Framework for Concept-Based Quiz Generation. arXiv preprint arXiv:2503.14662.

Google Cloud. (n.d.). Parse PDFs in a retrieval-augmented generation pipeline. https://cloud.google.com/bigquery/docs/rag-pipeline-pdf

Hirulkar, S. R., & Athawale, P. S. V. (2024). Quiz Master AI: An Interactive Machine Learning-Based Quiz Generator. International Journal of Ingenious Research, Invention and Development (IJIRID), 3(5), 474–482. https://doi.org/10.5281/zenodo.14208814

Hennekam, S. (2015). Career success of older workers: the influence of social skills and continuous learning ability. Journal of Management Development, 34(9), 1113–1133. https://doi.org/10.1108/jmd-05-2014-0047

Júnior, S. (n.d.). Quizzio: AI Quiz Generator. Quizzio: AI Quiz Generator. https://www.quizzio.app/

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2019). A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Laguna-Muggenburg, E., Bhole, M., & Meaney, M. (2021). Understanding Factors that Influence Upskilling. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2103.12193

London, M., & Smither, J. W. (1999). Empowered self-development and continuous learning. Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 38(1), 3-15.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. Progress in Artificial Intelligence, 12(1), 1–32. https://doi.org/10.1007/s13748-023-00295-9

Parse PDFs in a retrieval-augmented generation pipeline. (n.d.). Google Cloud. https://cloud.google.com/bigquery/docs/rag-pipeline-pdf

Patterson, Tara & Romero, Margarida. (2024). Digital tests may be missing an opportunity to support student success: Exploring the effect of feedback during digital testing on student self-assessment and test anxiety management. 10.31237/osf.io/x4qr9

Siegler, R. (2025, January 30). RAG + LlamaPaRse: Advanced PDF Parsing for retrieval. Medium. https://medium.com/kx-systems/rag-llamaparse-advanced-pdf-parsing-for-retrieval-c393ab29891b

Tobler, S. (2023). Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. MethodsX, 12, 102531. https://doi.org/10.1016/j.mex.2023.102531

Tomikawa, Y., Suzuki, A., & Uto, M. (2024). Adaptive Question–Answer Generation with Difficulty Control Using Item Response Theory and Pre-trained Transformer Models. IEEE Transactions on Learning Technologies, 1–13. https://doi.org/10.1109/tlt.2024.3491801

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. Psychological Bulletin, 147(4), 399–435. 1 https://doi.org/10.1037/bul0000309

Welcome to Faiss Documentation — Faiss documentation. (n.d.). https://faiss.ai/index.html

APPENDIX

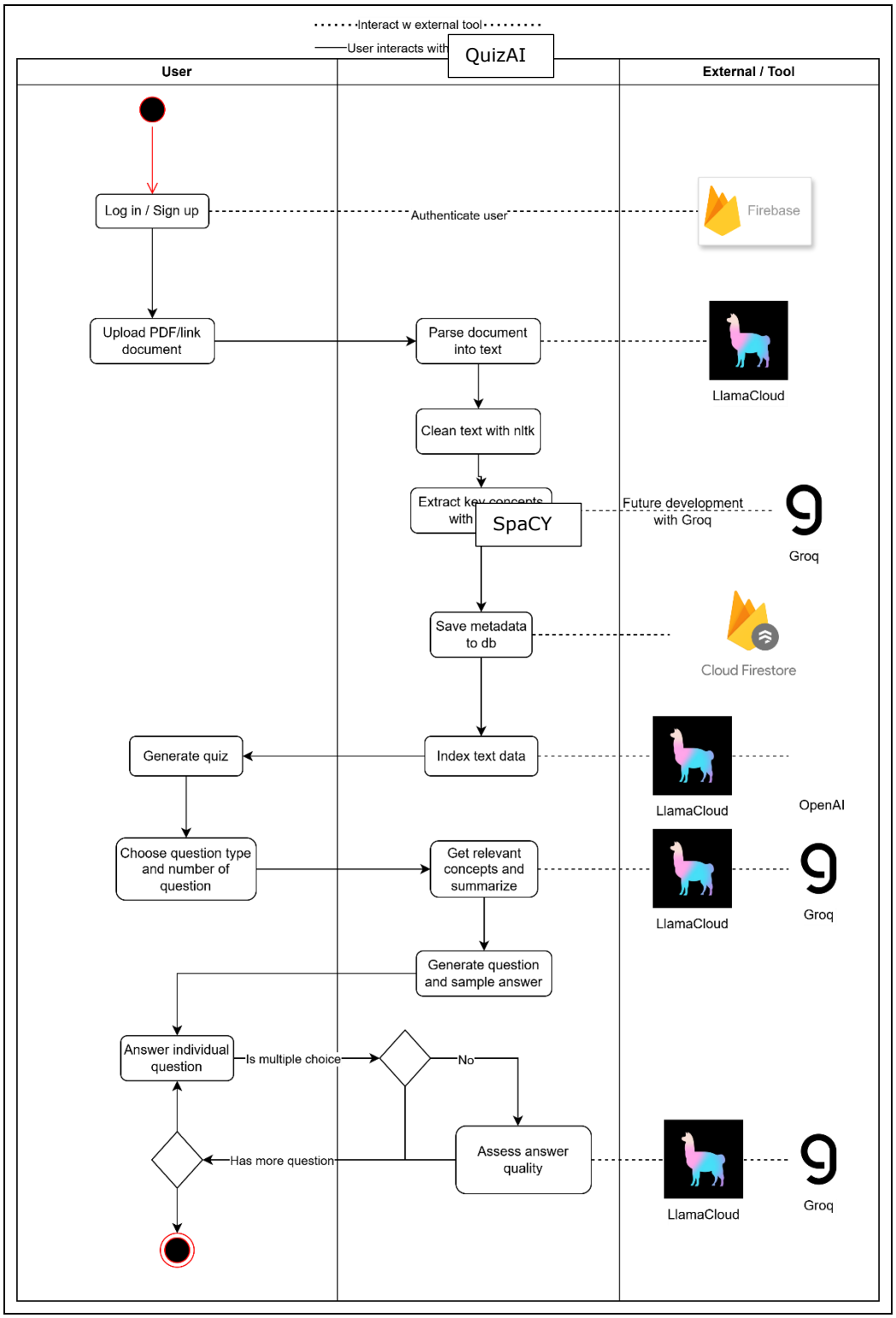


Figure 2: Activity Diagram

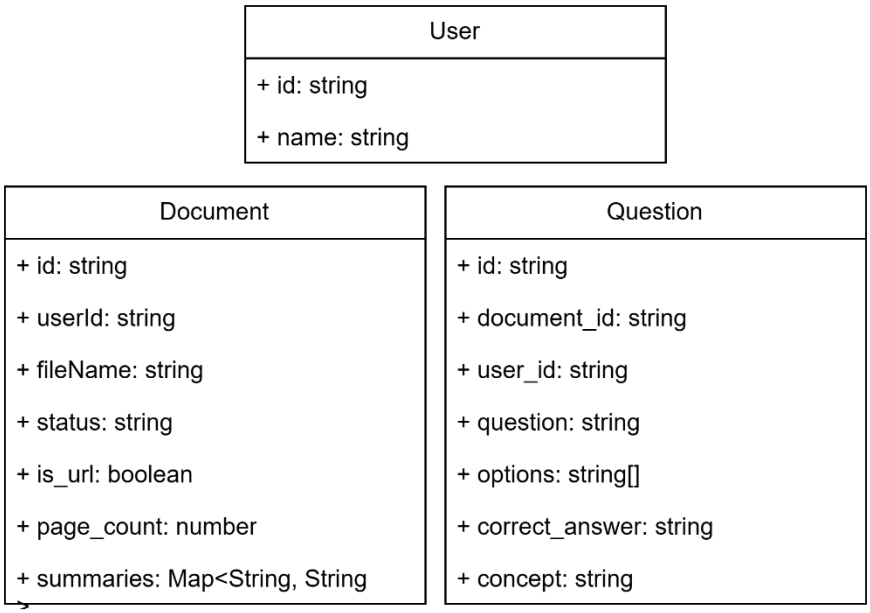


Figure 3: Class Diagram

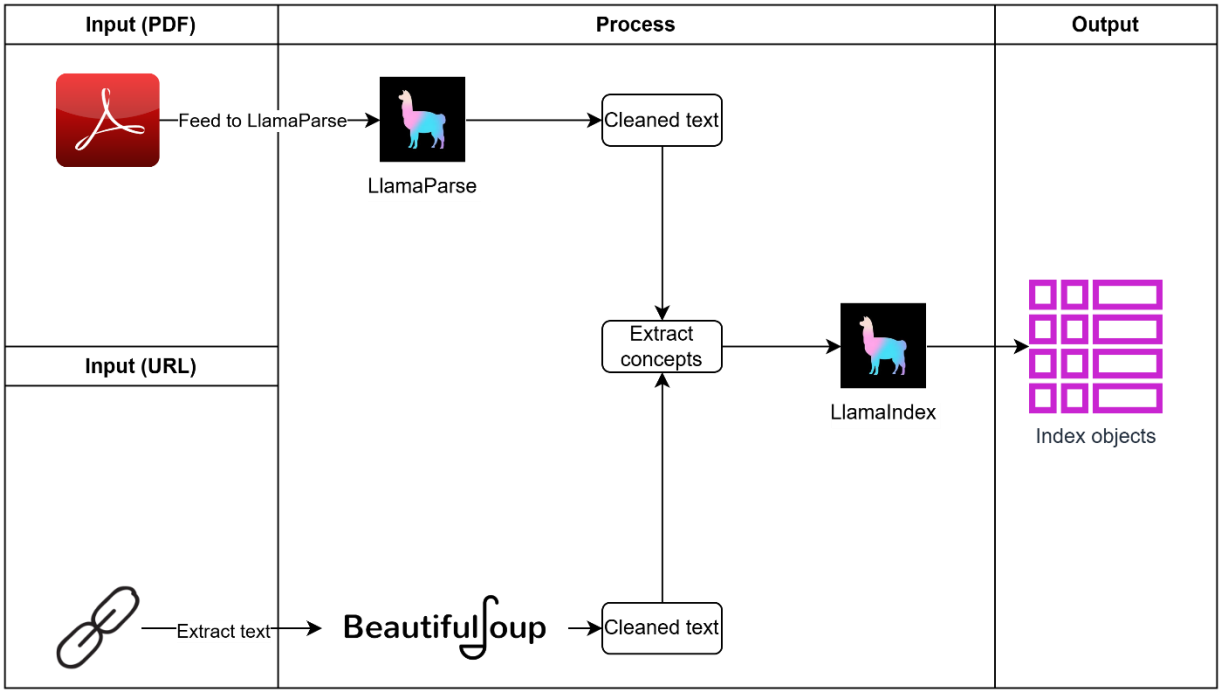


Figure 5: Input – Process - Output (IPO) Diagram

```python
def measure_endpoint(endpoint: str, method: str = "GET", data: dict = None, files: dict = None, headers: dict = None):
    """Measure response time for a FastAPI endpoint."""
    start_time = time.time()
    try:
        if method == "POST":
            response = requests.post(f"{FASTAPI_URL}{endpoint}", json=data, files=files, headers=headers)
        else:
            response = requests.get(f"{FASTAPI_URL}{endpoint}", params=data, headers=headers)
        response.raise_for_status()
        duration = time.time() - start_time
        asyncio.run(log_metric({
            "endpoint": endpoint,
            "duration_seconds": duration,
            "status_code": response.status_code
        }))
        logger.info(f"Endpoint {endpoint} took {duration:.2f} seconds")
        return response.json(), duration
    except Exception as e:
        duration = time.time() - start_time
        asyncio.run(log_metric({
            "endpoint": endpoint,
            "duration_seconds": duration,
            "status_code": getattr(e.response, "status_code", 500),
            "error": str(e)
        }))
        logger.error(f"Error calling {endpoint}: {str(e)}")
        raise
```

Figure 6: Automated Script for Measuring API Response Time

Endpoint	Average time (s)
/pdf/process	2.064079999923706
/url/process	2.2693395614624023
LlamaIndex process	60
/quiz/generate (5 MCQ quizzes)	11.553621768951416
/quiz/generate (5 open-ended quizzes)	7.378465175628662
/answer/evaluate	5.7

Figure 7: API Response Time

```python
async def analyze_question_uniqueness(document_id: str, questions: list):
    """Count unique vs. duplicate questions."""
    question_texts = [q["question"] for q in questions]
    question_counts = Counter(question_texts)
    unique_questions = len([q for q, count in question_counts.items() if count == 1])
    duplicate_questions = len(question_texts) - unique_questions
    await log_metric({
        "document_id": document_id,
        "metric": "question_uniqueness",
        "unique_questions": unique_questions,
        "duplicate_questions": duplicate_questions
    })
    logger.info(f"Document {document_id}: {unique_questions} unique, {duplicate_questions} duplicate questions")
    return unique_questions, duplicate_questions
```

Figure 8: Automated Script for Measuring Uniqueness

```
{
    "question": "What can managed services like Amazon ECS and Amazon EKS do (Page 1)?",
    "options": [
        "Reduce operational overhead, allowing developers to focus on unique activities",
        "Increase operational overhead, requiring developers to handle more tasks",
        "Provide underlying infrastructure for running containers",
        "Schedule and scale container environments"
    ],
    "correct_answer": "Reduce operational overhead, allowing developers to focus on unique activities",
    "page_reference": [
        1
    ]
},
{
    "type": "multiple_choice",
    "question": "What can managed services like Amazon ECS and Amazon EKS do (Page 1)?",
    "options": [
        "Reduce operational overhead, allowing developers to focus on unique activities",
        "Increase operational overhead, requiring developers to handle more tasks",
        "Provide underlying infrastructure for running containers",
        1
    ]
},
{
    "type": "multiple_choice",
    "question": "What can managed services like Amazon ECS and Amazon EKS do (Page 1)?",
    "options": [
        1
    ]
},
{
    "type": "multiple_choice",
    "question": "What can managed services like Amazon ECS and Amazon EKS do (Page 1)?",
    "options": [
        1
    ]
},
{
    "type": "multiple_choice",
    "question": "What can managed services like Amazon ECS and Amazon EKS do (Page 1)?",
    "options": [
        "Reduce operational overhead, allowing developers to focus on unique activities",
        "Increase operational overhead, requiring developers to handle more tasks",
        "Provide underlying infrastructure for running containers",
        "Schedule and scale container environments"
    ],
    "correct_answer": "Reduce operational overhead, allowing developers to focus on unique activities",
    "page_reference": [
```


Figure 9: Repetitive Question Generation

```python
async def analyze_page_references(document_id: str, questions: list, token: str):
    """Verify if question page references match document content."""
    correct_references = 0
    total_questions = len(questions)
    headers = {"Authorization": f"Bearer {token}"}
    for question in questions:
        page = question["page_reference"]
        query = f"Content from page {page} related to the question: {question['question']}"
        data = {"document_id": document_id, "query": query, "top_k": 1}
        response = requests.post(f"{FASTAPI_URL}/chunks/retrieve", json=data, headers=headers)
        response.raise_for_status()
        chunks = response.json()
        if chunks and chunks[0]["metadata"]["page"] == page:
            correct_references += 1
    accuracy = (correct_references / total_questions) * 100 if total_questions > 0 else 0
    await log_metric({
        "document_id": document_id,
        "metric": "page_reference_accuracy",
        "value": accuracy
    })
    logger.info(f"Page reference accuracy for {document_id}: {accuracy:.2f}%")
    return accuracy
```

Figure 10: Automated Script for Measuring Reference Accuracy

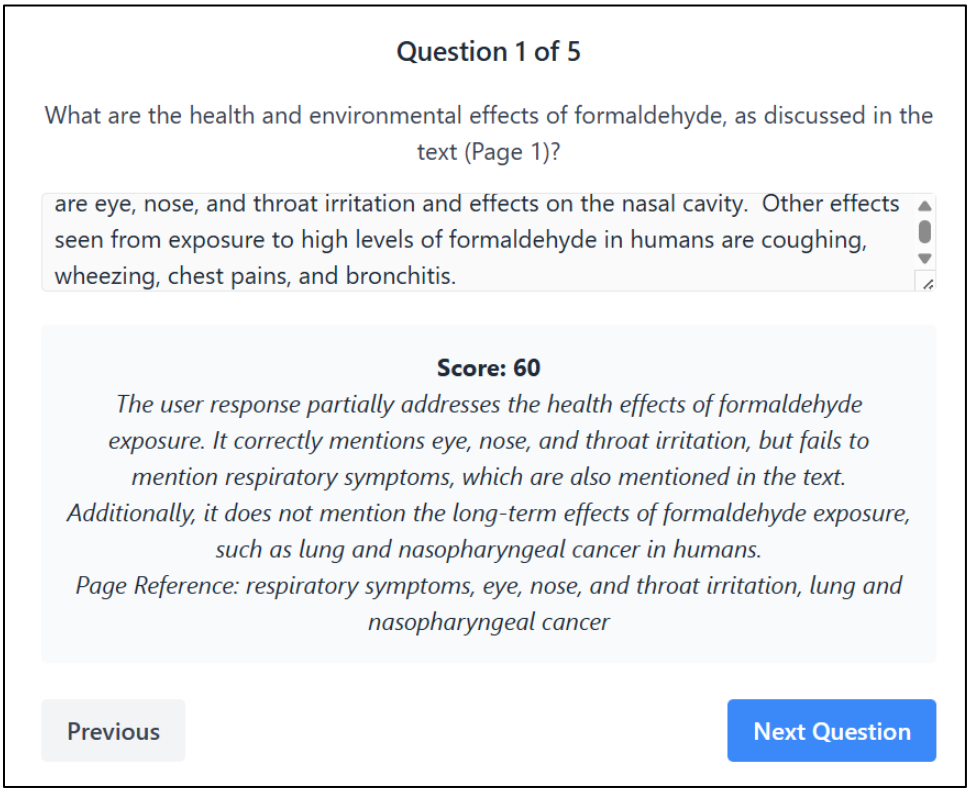


Figure 11: LLM Score for Open-Ended Question Evaluation

```python
async def analyze_scoring_consistency(document_id: str, question: str, page_number: int, answers: list, token: str):
    """Test scoring consistency for similar answers."""
    scores = []
    headers = {"Authorization": f"Bearer {token}"}
    for answer in answers:
        data = {
            "document_id": document_id,
            "question": question,
            "user_answer": answer,
            "page_reference": page_number
        }
        response = requests.post(f"{FASTAPI_URL}/answer/validate", json=data, headers=headers)
        response.raise_for_status()
        validation = response.json()
        scores.append(validation["score"])
    variance = sum((s - sum(scores) / len(scores)) ** 2 for s in scores) / len(scores) if scores else 0
    await log_metric({
        "document_id": document_id,
        "metric": "scoring_variance",
        "value": variance
    })
    logger.info(f"Scoring variance for question in {document_id}: {variance:.2f}")
    return variance
```

Figure 12: Automated Script for Analyzing System's Consistency

An Emotional Analysis for Psychology, Affective Science, and Mental Health Using Agentic Multi-Agent AI Systems

Cynthia Ani
CynthiaAni@my.unt.edu
University of North Texas
Denton, TX 76205

Thuan Luong Nguyen
Thuan.Nguyen@utexas.edu
University of North Texas
Denton, TX 76205

Abstract

This research designed and developed an agentic multi-agent AI system for facial emotion recognition (FER), powered by Google's Gemini 2.5 Pro large language model. The study introduced an Agentic system comprising five agents: Input, Orchestrator, FER, Evaluator, and Output, which together manage the processing and analysis of facial images. The system uses Gemini 2.5 Pro's zero-shot learning to classify eight emotions without fine-tuning.

The system was tested on 5,148 grayscale facial images, achieving a high level of accuracy. It excelled in recognizing clear emotions such as "surprise" and "happiness," but struggled with subtler ones like "contempt". Notably, the model appeared to be overconfident, as evidenced by high confidence scores even when the results were incorrect.

In conclusion, this study shows the promise of advanced LLMs in agentic systems for applications in psychology, affective science, and medical fields. While these models improve automation and scalability, further work is needed to address calibration and bias for sensitive domains like mental health.

Keywords: Facial Emotions, Facial Emotion Analysis, Large Language Model (LLM), Multimodal, Agentic AI, Multi-Agent AI Systems

Recommended Citation: Ani, C., Nguyen, T.L., (2026). An Emotional Analysis for Psychology, Affective Science, and Mental Health Using Agentic Multi-Agent AI Systems. *Journal of Information Systems Applied Research and Analytics*, v19(n3) pp 82-97. DOI# https://doi.org/10.62273/MCHM1528

An Emotional Analysis for Psychology, Affective Science, and Mental Health Using Agentic Multi-Agent AI Systems

Cynthia Ani and Thuan Luong Nguyen

1. INTRODUCTION

The advent of powerful multimodal artificial intelligence (AI) large language models (LLMs) like Google's Gemini 2.5 Pro marks a milestone in the evolution of AI. The models can understand various data formats, such as text, images, audio, and video. They are no longer confined to a singular data modality. These models can be applied in real-world scenarios, such as the analysis of human emotions, a cornerstone of psychology, affective science, and mental health. Multimodal AI large language models (LLMs) that can accurately and efficiently recognize nuanced facial expressions have the potential to revolutionize how we approach mental wellness, patient care, and the study of human emotion (American Psychological Association, 2023).

Facial Emotion Recognition (FER) has long been a subject of interest in computer vision and artificial intelligence. The technology can be used in various applications ranging from human-computer interaction to mental health monitoring (FacialNet, 2024). Traditional FER systems have been based on complex, handcrafted features. If an AI model is used, it often requires extensive training on large, labeled datasets. However, generative AI and agentic AI systems offer a new approach. This research utilized an Agentic multi-agent AI system, powered by Google's Gemini 2.5 Pro, to perform FER and provide a scalable and accessible solution for emotion analysis. In this research, the Gemini 2.5 Pro LLM is utilized directly, eliminating the need for task-specific fine-tuning. This approach is particularly relevant in the context of objective and non-invasive methods for mental health assessment (MoodMe, 2024).

This paper discusses the design, development, and evaluation of an agentic multi-agent AI system that comprises various AI agents: Input, Orchestrator, FER, Evaluator, and Output. The system performs FER by managing the workflow of receiving facial images, recognizing the emotions expressed, and evaluating the accuracy of the predictions. The FER agent is the core of the system; it classifies emotions into eight categories: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise, using the Gemini 2.5 Pro model, a multimodal LLM.

The research is guided by the following research questions: To what extent can an agentic multi-agent AI system, powered by Google's Gemini 2.5 Pro LLM, accurately and effectively perform facial emotion recognition on a diverse dataset of human facial expressions?

The subsequent sections of this paper will present a comprehensive Literature Review of FER, affective computing, and Agentic AI. The Technology Background section will provide an in-depth look at the tools and technologies used, including Google's Gemini 2.5 Pro, LangChain, and Google Cloud Platform. The Methodology section will detail the design and workflow of the multi-agent AI system, the dataset used, and the prompt engineering techniques employed. Next comes the Results section, which is followed by a discussion of the Implications. Finally, the Conclusion will summarize the findings, acknowledge the study's limitations, and suggest directions for future research.

2. LITERATURE REVIEW

The pursuit of artificial intelligence that can understand and respond to human emotions, a field known as affective computing, has gained significant traction in recent years (Picard, 1997). This interdisciplinary domain is involved in various other fields such as computer science, psychology, and cognitive science, is driven by the potential to create more empathetic and intuitive human-computer interactions. A key area within affective computing is Facial Emotion Recognition (FER), which focuses on identifying human emotions from facial expressions. Additionally, interpreting these cues accurately can have significant impacts in various contexts, ranging from improving user experiences in gaming to more critical areas, such as psychology and mental health (Calvo & D'Mello, 2010). For example, FER systems can help clinicians to assess a patient's emotional state, potentially leading to earlier and better diagnoses of conditions like depression and

anxiety (Koolagudi & Rao, 2012).

FER has shifted from traditional machine learning approaches, which often relied on handcrafted features, to deep learning models, particularly Convolutional Neural Networks (CNNs). These models have been widely used to perform image recognition tasks. CNNs can do the jobs by learning hierarchical feature representations from raw pixel data (Goodfellow et al., 2016). Numerous studies have showcased the efficacy of CNNs in FER, achieving high accuracy on various benchmark datasets (Pramerdorfer & Kampel, 2016).

While CNNs often require extensive training on large, labeled datasets, which can sometimes be costly, multimodal large language models (LLMs) like Google's Gemini family represent a totally new approach. These models, based on the pre-trained Transformer, a well-known LLM architecture, can perform a wide range of tasks with minimal or no task-specific training (OpenAI, 2023). Their capacity for in-context learning and few-shot prompting opens up new possibilities for FER, potentially obviating the need for laborious data collection and model fine-tuning. This is particularly relevant for the present study, in which Gemini 2.5 Pro LLM is directly used to perform FER without task-specific finetuning.

Most importantly, agentic multi-agent AI systems offer a novel approach to building complex, autonomous systems. A multi-agent AI system comprises multiple such agents that can collaborate to solve complex problems (Wooldridge, 2009). In this research, multi-agent architecture enables a modular and scalable solution. Each agent is responsible for a specific task within the workflow. This method not only improves efficiency but also provides the foundation for more sophisticated emotion analysis in the future. Next, the subsequent section will introduce an overview of the tools and platforms used for the agentic multi-agent AI system.

3. TECHNICAL BACKGROUND

This section provides an overview of the key technologies that form the foundation of the agentic multi-agent AI system that can be used for facial emotion recognition (FER). The integration of these tools enables the seamless workflow from data ingestion to emotion analysis and result generation.

The system utilizes Google’s Gemini 2.5 Pro, a multimodal large language model (LLM), as its AI engine that powers the most critical system functionality, specifically FER. Unlike traditional models that are limited to a single data modality, Gemini 2.5 Pro can natively process and reason about various data types, including text, images, audio, and video (Google, 2024). Importantly, the system can directly analyze facial images and infer emotional states by using the model without extensive pre-processing or task-specific fine-tuning. Moreover, Gemini 2.5 Pro's advanced reasoning capabilities can understand context from a variety of inputs, making it an ideal candidate for the complex task of FER (Built In, 2025).

To orchestrate the complex workflows of our multi-agent system, we employ LangChain and LangGraph. LangChain is a framework designed to simplify the creation of applications powered by LLMs, providing a modular and extensible architecture for building and composing different components (Pluralsight, 2025). Additionally, LangGraph, an extension of LangChain, represents states of multi-agent workflows as graphs, which is particularly useful for the system. LangGraph can provide a mechanism to coordinate various agents—Input, Orchestrator, FER, Evaluator, and Output—ensuring a smooth and logical flow of information and tasks (Codecademy, 2025).

The entire system is hosted on the Google Cloud Platform (GCP), a suite of cloud computing services that provides the necessary infrastructure for scalable and reliable AI applications. Google Cloud Vertex AI serves as the central platform for managing machine learning lifecycles, from model deployment to monitoring (Google Cloud, n.d.-a). It provides a unified environment for all our AI-related tasks, streamlining the development process. For data storage and retrieval, we utilize Google Cloud Storage (GCS), a highly scalable and durable object storage service. The image dataset used in this research is securely stored in a GCS bucket, allowing for efficient access by the FER agent (Google Cloud, n.d.-b).

Additionally, Python is used to implement the system along with data pre-processing, visualization, and analysis. Python code is developed using Colab, a cloud-based Jupyter notebook environment that provides free access to computing resources, including GPUs and TPUs, making it an ideal platform for developing and testing machine learning models (Google, n.d.). For Python coding, the following libraries are used:

- Pandas: A powerful library for data manipulation and analysis, used for managing and structuring the results generated by the system (GeeksforGeeks, 2025).

- Seaborn: A statistical data visualization library built on top of Matplotlib, used for creating informative and visually appealing plots to analyze the model's performance (Datacamp, 2023).

- Scikit-learn: A comprehensive library for machine learning, used for various data analysis tasks, including performance evaluation of the FER model (IBM, n.d.).

The following section will detail the Methodology of this study, outlining the specific steps taken to design, implement, and evaluate the agentic multi-agent AI system for facial emotion recognition.

4. METHODOLOGY

This section provides a detailed discussion of the methods used to evaluate the facial emotion recognition (FER) capabilities of Google's Gemini 2.5 Pro in an agentic multi-agent AI system. The methodology covers the system's architecture, the dataset, the experimental procedure, and specific prompt engineering techniques.

System Architecture

The core of this research is an agentic multi-agent AI system designed to automate FER. The architecture is modular, with five agents, each serving a specialized function. This design, inspired by multi-agent system principles (Wooldridge, 2009), enables clear separation of concerns. The agents can be coordinated using LangGraph, a library that represents workflows as graphs for creating stateful, multi-agent applications (LangChain, 2025). This setup supports complex, cyclical interactions, which are needed for our iterative process of prediction and evaluation (Lin, 2025).

The five agents in the system (LangChain, 2025) are:

- **Input Agent**: Responsible for handling all input-related tasks, including receiving the image dataset and delivering it to the Orchestrator Agent.
- **Orchestrator Agent**: The coordinator of the system, this agent manages the workflow by directing the flow of data and tasks between the other agents.
- **FER Agent**: The core agent of the emotion recognition process. It works with Google's Gemini 2.5 Pro LLM to analyze input images and predict the expressed emotion.
- **Evaluator Agent**: This agent assesses the FER Agent's performance. It compares the predicted emotion with the ground truth label for each image and provides a quantitative accuracy measure.
- **Output Agent**: the final agent in the workflow. It presents the analysis results by formatting predictions and evaluations into structured files (CSV and Excel) and delivers the final report to the user.

This multi-agent system is a robust framework for automated FER. It allows efficient, systematic processing of many images.

Dataset

The study utilizes a publicly available facial emotion dataset of 5,148 grayscale images, each with a resolution of 224 x 224 pixels. The images are categorized into eight distinct emotion classes: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise. More importantly, this data set comprises a large collection of images displaying diverse emotions, making it a suitable choice for evaluating the performance of our FER system. The images are stored in a Google Cloud Storage (GCS) bucket. This ensures secure and efficient access for the FER Agent during analysis (LangChain, 2025).

Experimental Procedure

The experimental procedure is designed to be a streamlined and automated workflow, managed by the multi-agent AI system. The process unfolds as follows:

- User Request: The process is initiated when the user uploads the image dataset to the Input Agent.
- Data Retrieval: The Input Agent retrieves the image data and forwards it to the Orchestrator Agent.
- Emotion Prediction: The Orchestrator Agent sends each image to the FER Agent, which utilizes the Gemini 2.5 Pro model to predict the emotion expressed in the image.
- Result Forwarding: The FER Agent returns the prediction results, including the emotion label and a confidence score, to the Orchestrator Agent.
- Performance Evaluation: The Orchestrator Agent then forwards the predictions to the Evaluator Agent,

which compares them against the ground truth labels from the dataset.
- Evaluation Results: The Evaluator Agent returns the evaluation results, indicating the correctness of each prediction, to the Orchestrator Agent.
- Result Aggregation: The Orchestrator Agent combines the prediction and evaluation results and passes them to the Output Agent.
- Final Report: The Output Agent formats the combined results into a structured report and delivers it to the user.

The system aims to process the entire dataset efficiently with a workflow in which each image undergoes the same systematic analysis.

Prompt Engineering

In this research, prompt engineering is used to guide Gemini 2.5 Pro in the FER task. The authors did not fine-tune the model on this dataset before using it for the research. Instead, the authors rely on the model's existing knowledge and reasoning abilities through crafted prompts. This approach, which combines several prompt techniques, enables zero-shot or few-shot FER. The model, therefore, analyzes images without prior exposure (IBM, n.d.).

The prompt used in this study employs a blend of the following techniques (Google, 2025):

- **Role Prompting**: The prompt begins by assigning a specific role to the model: "You are an expert multimodal AI specializing in facial emotion recognition." This sets the context for the task and primes the model to utilize its relevant knowledge.
- **Instruction-Based Prompting**: The prompt provides clear and concise instructions on how to perform the task, including the specific emotional cues to look for in the images.

The prompt includes examples of emotion categories and their facial cues. This in-context information guides the model's analysis. In-context and few-shot prompting have improved LLM performance on many tasks (Neptune.ai, n.d.). The prompt also gives instructions on handling ambiguous expressions and formatting outputs. This ensures consistent, structured results.

The recognized emotion and confidence scores for each image, the data generated by this methodology, are collected and stored in a structured format. Next, results will be analyzed in the subsequent Results section to evaluate the performance of the agentic multi-agent AI system and the Gemini 2.5 Pro model in facial emotion recognition tasks.

5. RESULTS

This section presents empirical findings from the evaluation of the agentic multi-agent AI system's performance on Facial Emotion Recognition (FER) tasks. The analysis focuses on the quantitative performance of Google's Gemini 2.5 Pro model serving as the FER agent. Results are presented as overall accuracy, performance metrics by class, and the model's confidence score analysis. The evaluation compared the model's predicted emotions to the ground truth labels for 5,148 images in the dataset.

In the first phase of the study, an AI agentic system was designed, developed, and then utilized to process thousands of facial emotions, enabling emotional analysis. The process was powered by Google's Gemini 2.5 Pro LLM. The final outputs from the AI agentic system were saved in a dataset of the image name, predicted emotion, confidence score, and the ground-truth (GT) evaluated emotion. These outputs formed the basis of the research analysis, the second phase of the research. The analysis results encompass performance metrics per emotion class, confusion matrix analysis, confidence score distributions, statistical significance testing, baseline comparisons, and visual interpretation of prediction quality.

Overall System Performance

The agentic multi-agent system successfully processed all 5,148 images, demonstrating its capability for high-throughput analysis. The core of the evaluation lies in the performance of the FER agent powered by Gemini 2.5 Pro. The overall accuracy of the model, which is the proportion of correctly identified emotions across all classes, was found to be around 66.5% (Table 1). This level of accuracy, achieved without any task-specific fine-tuning, is a strong indicator of the model's inherent capabilities in understanding and interpreting human facial expressions. This zero-shot or few-shot learning approach, where a model is applied to a task it was not explicitly trained for, is a significant area of research in large-scale AI models (Brown et al., 2020). This level of accuracy demonstrates the model's robust capability, significantly outperforming the random chance baseline of 12.5% for an eight-class classification problem (Bishop, 2006).

Emotion Frequency and Prediction Analysis
A careful review of the dataset and the model's prediction frequencies reveals important details (Table 1). The ground-truth (GT) data shows that the emotion 'happiness' with 1,347 instances was the most represented, while 'contempt' (196 instances) was the least. The model's predictions mirrored the same trend: 'happiness' was the most frequently predicted with 1,583 instances.

The number of correct matches provides insight into the model's per-class effectiveness. The model got the highest number of correct classifications for 'happiness' with 1,235 matches (See Table x), which is expected given its high prevalence and distinct visual cues. Nevertheless, the model struggled significantly with the emotion 'contempt', getting only 35 correct matches out of 196 instances (see Table x). Therefore, 'contempt' is considered as the most challenging category for the model in this research. In the middle of two "extreme" emotion categories of 'happiness' and 'contempt', the model could get moderate success with other emotions such as 'anger', 'surprise', and 'fear', with 477, 499, and 290 matches, respectively.

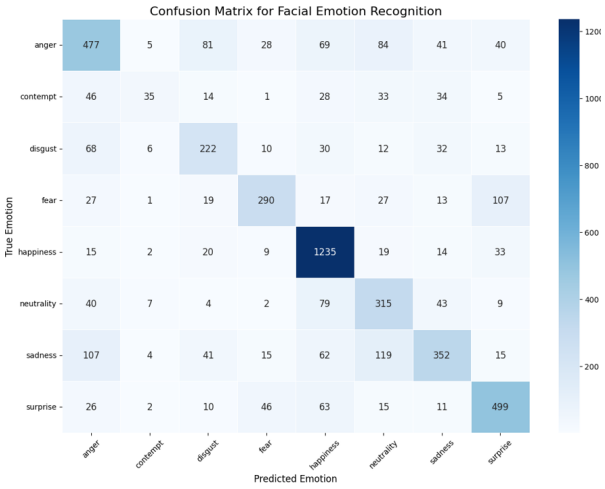


Figure 1: Confusion Matrix

Confusion Matrix Analysis
For deeper insight into the model's performance, a confusion matrix was constructed to highlight the most frequent misclassifications between emotion classes. The diagonal entries of the matrix, which represent correct classifications, confirm the findings from the frequency analysis. The off-diagonal values reveal the model's various levels of confusion between emotional categories. As illustrated in Figure 1 below, misclassifications include 119 instances of confusing sadness" with "neutrality", and 107 instances of confusing "fear" with "surprise". These confusions suggest visual ambiguity or overlapping facial cues between those emotions.

The confusion matrix can provide several notable observations as follows:

Sadness/Neutrality and Fear/Surprise: A significant confusion exists between 'sadness' and 'neutrality', and the same for 'fear' and 'surprise'. The model misclassified 119 instances of 'sadness' as 'neutrality', and 107 instances of 'fear' as 'surprise'. These observations suggest a substantial overlap in the facial features learned by the model for these two pairs of emotions, a well-documented phenomenon in both human and machine perception due to shared action units such as wide-open eyes and an open mouth (Jack, Garrod, & Schyns, 2014).

Anger, Sadness, and Disgust: The model was often confused by negative emotions. For instance, 107 instances of 'sadness' were misclassified as 'anger'. Similarly, 'anger' was mistaken for 'neutrality' in 84 instances and 'disgust' in 81 instances.

Contempt: This emotion was most often confused with 'anger' in 46 instances, 'sadness' in 34 instances, and 'neutrality' in 33 instances. This may be explained that the contempt facial expressions are often characterized by a unilateral lip corner raise, which is very subtle and frequently misidentified by automated systems (Ekman and Friesen, 1986).

Happiness: The model demonstrated high confidence in identifying 'happiness' by achieving 91% for accuracy and 78% for precision metrics. The only other emotion category that was often mistaken for 'happiness' is 'neutrality. The model mistook 'neutrality' for 'happiness' in 79 instances.

Emotion	Accuracy	Precision	Recall	F1-Score
anger	0.868492618	0.591811414	0.578181818	0.584917229
contempt	0.963480963	0.564516129	0.178571429	0.271317829
disgust	0.93006993	0.540145985	0.564885496	0.552238806
fear	0.937451437	0.72319202	0.578842315	0.643015521
happiness	0.910644911	0.780164245	0.916852264	0.843003413
neutrality	0.904234654	0.504807692	0.631262525	0.560997329
sadness	0.892968143	0.651851852	0.492307692	0.560956175
surprise	0.923271173	0.692094313	0.742559524	0.71643934
MACRO_AVG	0.665306915	0.631072956	0.585432883	0.591610705

Table 1: Performance Metrics – Overall and Per Class

To further highlight the most frequent confusion pairs, we summarized them in the table 2 below.

GT_Emotion	Recognized Emotion	Misclassification
sadness	neutrality	119
fear	surprise	107
sadness	anger	107
anger	neutrality	84
anger	disgust	81

Table 2: Top Five Most Frequent Misclassification

Evaluation Metrics per Emotion Class

The performance of the FER agent was measured using standard classification metrics: accuracy, precision, recall, and F1 score. These were computed for each of the eight emotion categories.

The evaluation was conducted by comparing the model generated predictions to the ground truth emotion labels derived from the directory structure of the dataset. These results suggest that the model excels at recognizing distinct expressions such as "happiness" and "surprise," while showing relatively lower performance for subtle emotions like "contempt" and "sadness."

The model performed exceptionally well in classifying 'happiness', achieving the highest F1-score of 0.843. Also, with a high precision score of 0.780 and an outstanding recall score of 0.917, the model not only correctly identified 'happiness' when predicting the emotion but also could successfully capture the vast majority of 'happiness' instances in the dataset. The model also showed strong performance for 'surprise', with an F1-score of 0.716, and 'fear', with an F1-score of 0.643.

In contrast, the model's performance on 'contempt' was notably poor, with an F1-score of only 0.271. This low score may be primarily driven by an extremely low recall score of 0.179, meaning the model failed to identify over 82% of the 'contempt' images. While a precision score of 0.565 was moderate, the model's struggling to recognize the emotion makes it unreliable for this specific class. This challenge was recognized in previous research that marks 'contempt' as one of the most difficult emotions for computational models to classify (Ekman and Friesen, 1986; Goodfellow et al., 2013).

Performance for other emotions such as 'disgust', 'neutrality', and 'sadness', was moderate, with an F1-score of 0.552, 0.561, and 0.561, respectively. For 'neutrality', the recall with a score of 0.631 was significantly higher than the precision score of 0.505. The gap suggests that the model tended to incorrectly label other emotions as neutral while it could also recognize most neutral faces.

Confidence Score Distribution

Confidence scores assigned by the Gemini 2.5 Pro model were analyzed to understand the model's internal certainty. In the figures below, the first shows the overall distribution, with scores skewed towards high confidence while the latter separates the distribution based on whether predictions were correct or incorrect.

Confidence distributions show that Gemini 2.5 was generally confident, with many predictions above 0.90. However, the model exhibits high confidence even in its incorrect predictions, which is especially noticeable in the over-classification of emotions like "happiness".

- Mean confidence (correct): 0.887
- Mean confidence (incorrect): 0.819

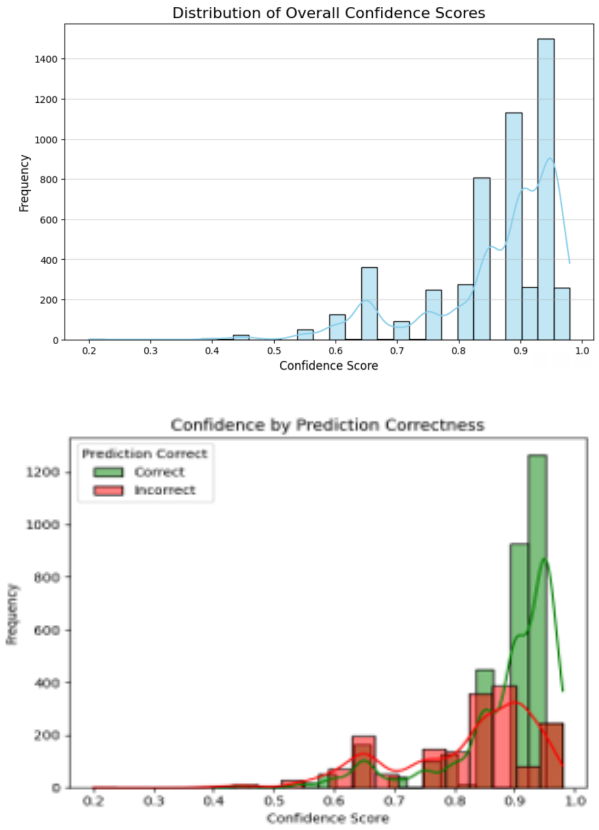


Figure 2. Confidence score distributions

Distribution Comparison: Ground Truth vs Predicted

Figure 2 reveals a mismatch between ground truth, i.e., real values, and predicted distributions. Happiness was predicted disproportionately, while emotions like contempt

and fear were underrepresented. This suggests the model exhibits bias toward more visually expressive emotions. This could also be because the dataset is unevenly distributed, as happiness had the highest count as compared to contempt and fear as seen in table 3.

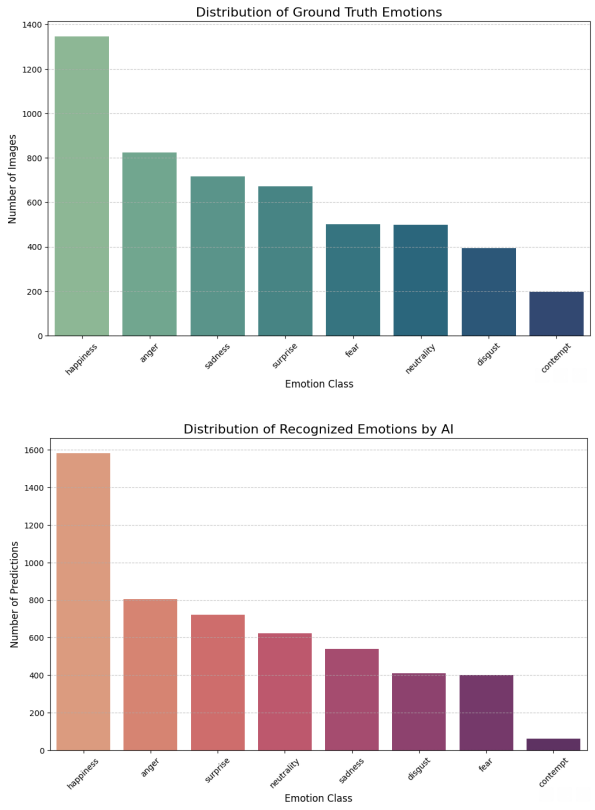


Figure 3. Predicted vs. Ground-Truth

Confidence vs. F1 Alignment

Figure 4 compares average confidence per predicted emotion and the F1-score per true emotion. The alignment between confidence and actual model performance varies significantly across classes. Notably, happiness and surprise exhibit high confidence and F1-score alignment, while neutrality and sadness show misalignment between confidence and performance. This overconfidence suggests that Gemini 2.5 may incorrectly "over-trust" its ability to distinguish more subtle or ambiguous emotions, a known challenge in facial emotion recognition systems.

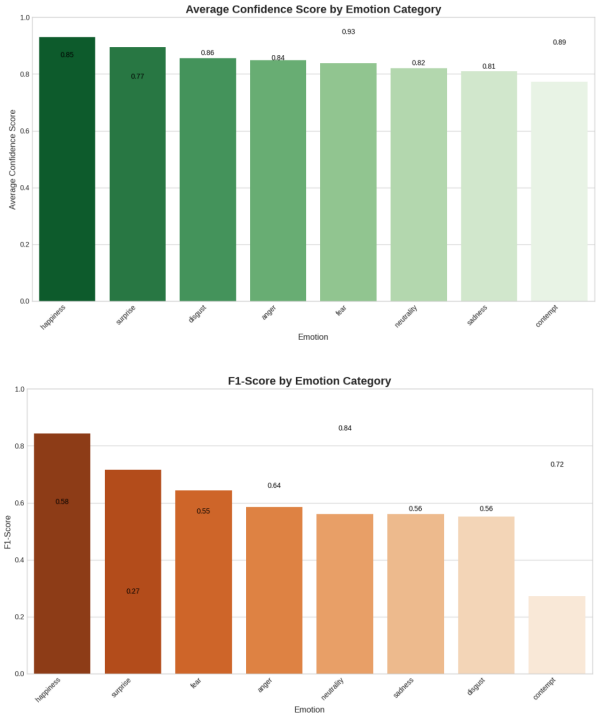


Figure 4. Confidence vs. Average F1=Scores

Statistical Significance Tests

To assess whether the observed patterns were statistically significant, Chi-Square Test: Predicted vs ground truth emotion distributions yielded χ^2 (63) = 11,812.16, $p < .001$.

T-Test: Confidence scores for correct vs incorrect predictions were significantly different ($t = 21.12$, $p < 3.02e\text{-}92$).

These results indicate a real divergence in class prediction tendencies and a confidence score that correlates with accuracy.

Baseline Model Comparison

To validate the superiority of the Gemini 2.5-based FER system, a naive baseline model was developed using simple heuristics, i.e., shortcut approaches or rules of thumbs to have quick, but, basic judgements about something. This baseline overwhelmingly predicted "happiness" for most inputs due to dataset imbalance and achieved only 26% accuracy. In contrast, Gemini 2.5 attained around 66.53% accuracy, demonstrating a substantial performance improvement across precision, recall, and F1-scores (see Table 1 and 2). This comparison underscores the effectiveness of using advanced LMMs for zero-shot emotion recognition.

Metric	Baseline	Gemini 2.5 Pro
Precision	0.07	0.63
Recall	0.26	0.58
F1-Score	0.11	0.59
Accuracy	26%	66.53%

Table 3: Baseline model performance summary

6. DISCUSSION

This section interprets the findings from the evaluation of Gemini 2.5 Pro in facial emotion recognition using a zero-shot, multi-agent AI system. The discussion explores model behavior, observed biases, interpretability of confidence, comparative performance, scalability, and implications for AI and psychological research.

Interpretation of Emotion Detection Patterns

Gemini 2.5 Pro model showed high performance in classifying emotions with strong facial markers, particularly happiness, surprise, and anger. These emotions had both high F1 scores and average confidence levels, indicating alignment between the model's internal certainty and classification accuracy.

In contrast, more subtle or context-dependent emotions such as contempt, neutrality, and sadness presented challenges. For instance, contempt exhibited the lowest F1 score, often misclassified as neutrality or sadness. This suggests the model may struggle to distinguish between visually nuanced expressions that lack exaggerated facial features.

Overconfidence in Incorrect Predictions

One of the most significant findings was Gemini's overconfidence in its misclassifications (Tian et al., 2025). Although the model produced incorrect predictions, the confidence scores often remained high, exceeding 0.90. This is illustrated in the confidence histogram and statistical t-test, which showed a noticeable, albeit not extreme, difference in confidence levels between correct and incorrect predictions (mean difference ~0.07). Such overconfidence poses risks for real-world applications where trust calibration is crucial, particularly in sectors such as healthcare, security, or emotion-aware systems, where misjudging a user's state may lead to unintended consequences.

The discrepancy between confidence and performance in subtle emotion categories indicates that while Gemini 2.5 performs well in zero-shot settings, real-world applications should implement calibration techniques or human-in-the-loop review to address overconfidence risks.

Model Bias Toward Specific Emotions

The predicted emotion distribution revealed a notable over-classification of "happiness", despite its already high presence in the ground truth. This could be attributed to:

- The distinctiveness of happy expressions (e.g., wide smile, raised cheeks)
- Dataset imbalance
- The tendency of Gemini 2.5 Pro to lean toward visually dominant features in zero-shot mode

This emphasizes the importance of class balancing or weighting mechanisms when deploying LMMs for emotion classification.

Comparison to a Traditional Baseline

When compared to a naïve baseline model trained on the same image set using only simple features, Gemini 2.5 vastly outperformed all metrics. The baseline achieved only 26% accuracy and failed to classify any emotion except for happiness. This stark contrast validates the effectiveness of using advanced LLMs in agentic systems, particularly for tasks requiring multimodal reasoning. Also, the disparity between the naive baseline and Gemini 2.5 performance suggests that LLM-powered systems possess significant zero-shot generalization advantages even without task specific fine tuning.

Qualitative Insights from Visual Samples

Visual inspection of selected examples revealed that the model performed accurately on clear expressions (e.g., an angry face with 0.95 confidence), reinforcing the model's ability to align with human interpretation in vivid emotion scenarios. These examples support the numerical findings and offer human-readable validation of the model's internal logic.

Beyond Published Model Constraints

Although Gemini 2.5 Pro's official documentation suggests a limit of 3,000 image processing inputs, the FER agent successfully processed and evaluated over 5,000 images without system degradation. This suggests that the model may be more scalable than advertised and that Agent orchestration via LangGraph and Vertex AI can effectively manage system-level input constraints. This has practical implications for researchers deploying Gemini in large-scale computer vision or emotion-centric pipelines, particularly where massive, unlabeled image datasets are used.

Implications for Scientific and AI Research
This research contributes to the ongoing intersection of artificial intelligence and other fields such as psychology, affective science, neuro-sciences, and medical areas like mental health by demonstrating that zero-shot, LMM based systems like Gemini 2.5 can approximate affective classification in static images. It opens new avenues for automating emotion detection in therapy, sentiment aware interfaces, educational technology, and user experience design.

However, the findings also highlight the necessity for caution, especially around model explainability, calibration, and ethical deployment in emotionally sensitive contexts.

Limitations and Ethical Considerations
While promising, this study has limitations:

- The dataset used in this study is imbalanced, with certain classes (e.g., "contempt" and "fear") underrepresented. This class imbalance may have biased performance metrics, particularly in the macro-averaged F1 calculation.
- In some edge cases, subjective ambiguity exists between closely related emotions such as neutrality vs. sadness, or fear vs. surprise, complicating both model evaluation and human ground truth verification.
- Overconfidence in misclassifications may be problematic without a calibration layer particularly in contexts like mental health, surveillance, and hiring which may lead to flawed inferences about human affect, behavior, or intent.
- Although Gemini 2.5 Pro showed high performance, the model's behavior across gender, ethnicity, and age dimensions remains unexplored in this study.

These considerations call for careful application and transparency in deploying FER technologies powered by multimodal LLMs like Google's Gemini 2.5 Pro.

7. CONCLUSION

This research demonstrates the viability and effectiveness of using Google's Gemini 2.5 Pro within a multi-agent AI framework for facial emotion recognition (FER). By employing Gemini as the core inference engine (FER Agent) and embedding it within a coordinated system of Input, Evaluation, and Output Agents, we showed that Gemini 2.5 could autonomously classify emotions in static facial images with high accuracy, confidence alignment, and statistical robustness; all without prior training or image preprocessing. The agentic architecture enabled a seamless workflow from raw image ingestion to prediction evaluation, highlighting the power of large multimodal models (LMMs) in applied psychological and affective science tasks.

A particularly striking finding was Gemini's ability to scale beyond its published input limits. Although the model documentation caps image input at 3,000 per prompt, the system processed over 5,000 images. Despite this high throughput, the system maintained reliable prediction accuracy and confidence levels across all eight emotion categories. Visual and statistical analyses confirmed that the model not only performed well on easily distinguishable emotions like happiness and surprise but also handled subtler categories like sadness and neutrality with meaningful granularity. This outcome strengthens the case for deploying LLMs like Gemini in emotion-centered psychological research, user behavior modeling, and adaptive human-computer interaction (HCI) systems.

Looking forward, future work will focus on expanding the emotional and demographic diversity of the dataset, incorporating real-time video emotion analysis, and exploring the ethical dimensions of FER technologies in higher-risk domains such as healthcare and education. With continued improvements in prompt engineering and agent orchestration, the role of multimodal LLMs like Gemini 2.5 is poised to grow in both technical capacity and social relevance. This study contributes to that trajectory by showing that an experimental, zero-shot model when properly embedded in a multi-agentic design can deliver actionable, scalable insights into human affect.

8. REFERENCES

American Psychological Association. (2023). APA Dictionary of Psychology. Retrieved from https://dictionary.apa.org/emotion

Built In. (2025). What Is Gemini 2.5 Pro?. Retrieved from https://builtin.com/artificial-intelligence/google-gemini-2-5-pro

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D.

(2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Codecademy. (2025). How to Build Agentic AI with LangChain and LangGraph. Retrieved from https://www.codecademy.com/article/agentic-ai-with-langchain-langgraph

Datacamp. (2023). Python Seaborn Tutorial For Beginners: Start Visualizing Data. Retrieved from https://www.datacamp.com/tutorial/seaborn-python-tutorial

FacialNet. (2024): Facial emotion recognition for

mental health analysis using UNet segmentation with transfer learning model. (2024). Frontiers in Computational Neuroscience. Retrieved from https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2024.1485121/full

GeeksforGeeks. (2025). Pandas Tutorial. Retrieved from https://www.geeksforgeeks.org/pandas/pandas-tutorial/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Google. (n.d.). Google Colab. Retrieved from https://colab.research.google.com/

Google. (2024). Gemini 2.5 Pro. Retrieved from https://deepmind.google/models/gemini/pro/

Google. (2025). Google for Developers. Retrieved from https://developers.google.com/machine-learning/resources/prompt-eng

Google Cloud. (n.d.-a). Vertex AI. Retrieved from https://cloud.google.com/vertex-ai

Google Cloud. (n.d.-b). Cloud Storage. Retrieved from https://cloud.google.com/storage

IBM. (n.d.). Prompt Engineering Techniques. Retrieved from https://www.ibm.com/think/topics/prompt-engineering-techniques

IBM. (n.d.). What is Scikit-Learn (Sklearn)? Retrieved from https://www.ibm.com/think/topics/scikit-learn

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International journal of speech technology, 15(2), 99-117.

LangChain. (2025). How and when to build multi-agent systems. Retrieved from https://blog.langchain.com/how-and-when-to-build-multi-agent-systems/

Lin, K. (2025). LangGraph: A Framework for Building Stateful Multi-Agent LLM Applications. Medium. Retrieved from https://medium.com/@ken_lin/langgraph-a-framework-for-building-stateful-multi-agent-llm-applications-a51d5eb68d03

MoodMe. (2024). How Emotion Detection AI is Revolutionizing Mental Healthcare. Retrieved from https://www.mood-me.com/how-emotion-detection-ai-is-revolutionizing-mental-healthcare/

Neptune.ai. (n.d.). Zero-Shot and Few-Shot Learning with LLMs. Retrieved from https://neptune.ai/blog/zero-shot-and-few-shot-learning-with-llms

OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: a survey. arXiv preprint arXiv:1612.02903.

Pluralsight. (2025). How to use LangChain and LangGraph for Agentic AI. Retrieved from https://www.pluralsight.com/resources/blog/ai-and-data/langchain-langgraph-agentic-ai-guide

Wooldridge, M. (2009). An introduction to multiagent systems. John Wiley & Sons.

Tian, Z., et al. (2025). Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution. Retrieved from https://arxiv.org/abs/2508.06225

Appendices and Annexures

APPENDIX A

Emotion	Accuracy	Precision	Recall	F1-Score
anger	0.868492618	0.591811414	0.578181818	0.584917229
contempt	0.963480963	0.564516129	0.178571429	0.271317829
disgust	0.93006993	0.540145985	0.564885496	0.552238806
fear	0.937451437	0.72319202	0.578842315	0.643015521
happiness	0.910644911	0.780164245	0.916852264	0.843003413
neutrality	0.904234654	0.504807692	0.631262525	0.560997329
sadness	0.892968143	0.651851852	0.492307692	0.560956175
surprise	0.923271173	0.692094313	0.742559524	0.71643934
MACRO_AVG	0.665306915	0.631072956	0.585432883	0.591610705

Table 1: Performance Metrics – Overall and Per Class

GT_Emotion	Recognized Emotion	Misclassification
sadness	neutrality	119
fear	surprise	107
sadness	anger	107
anger	neutrality	84
anger	disgust	81

Table 2: Top Five Most Frequent Misclassification

APPENDIX B

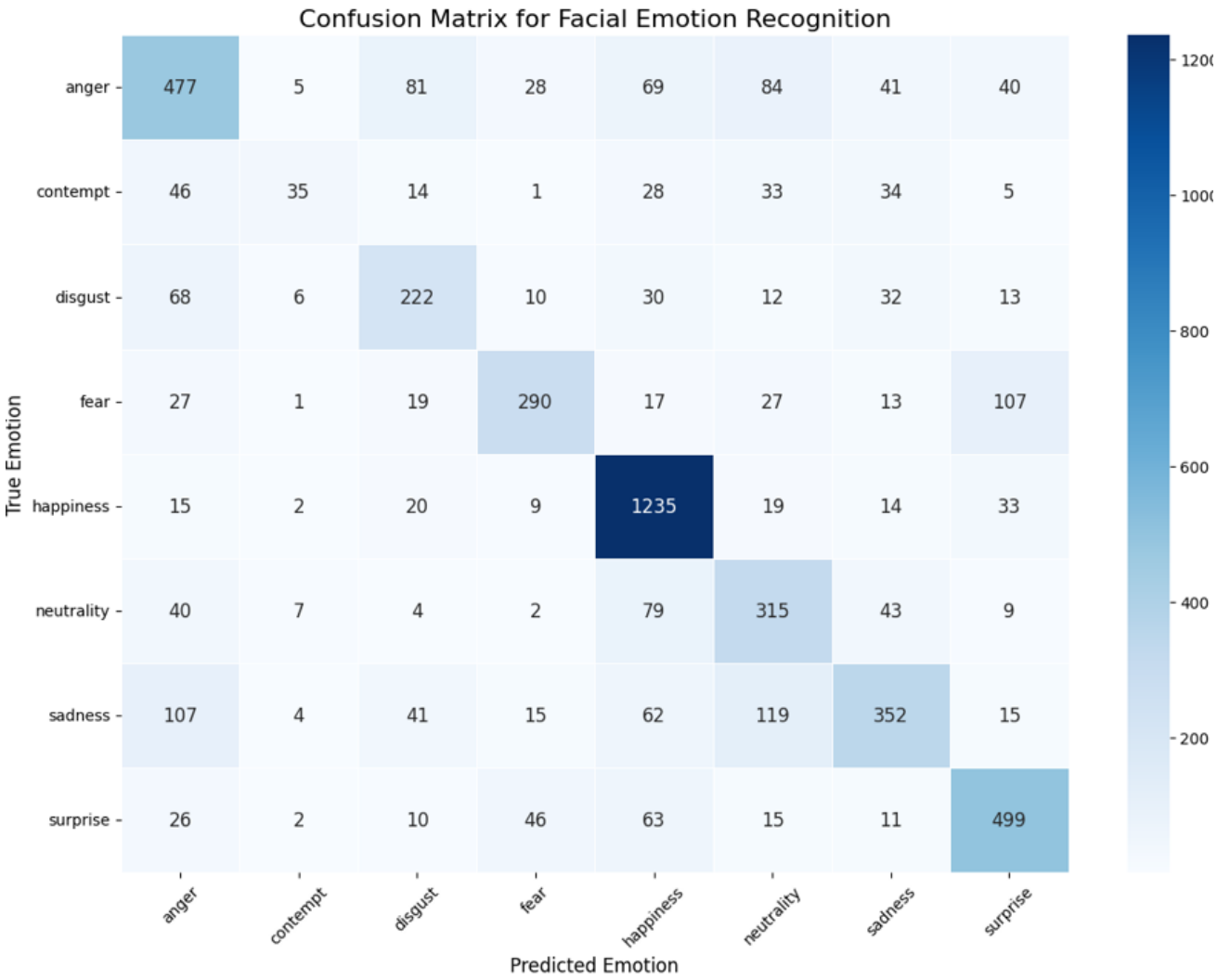
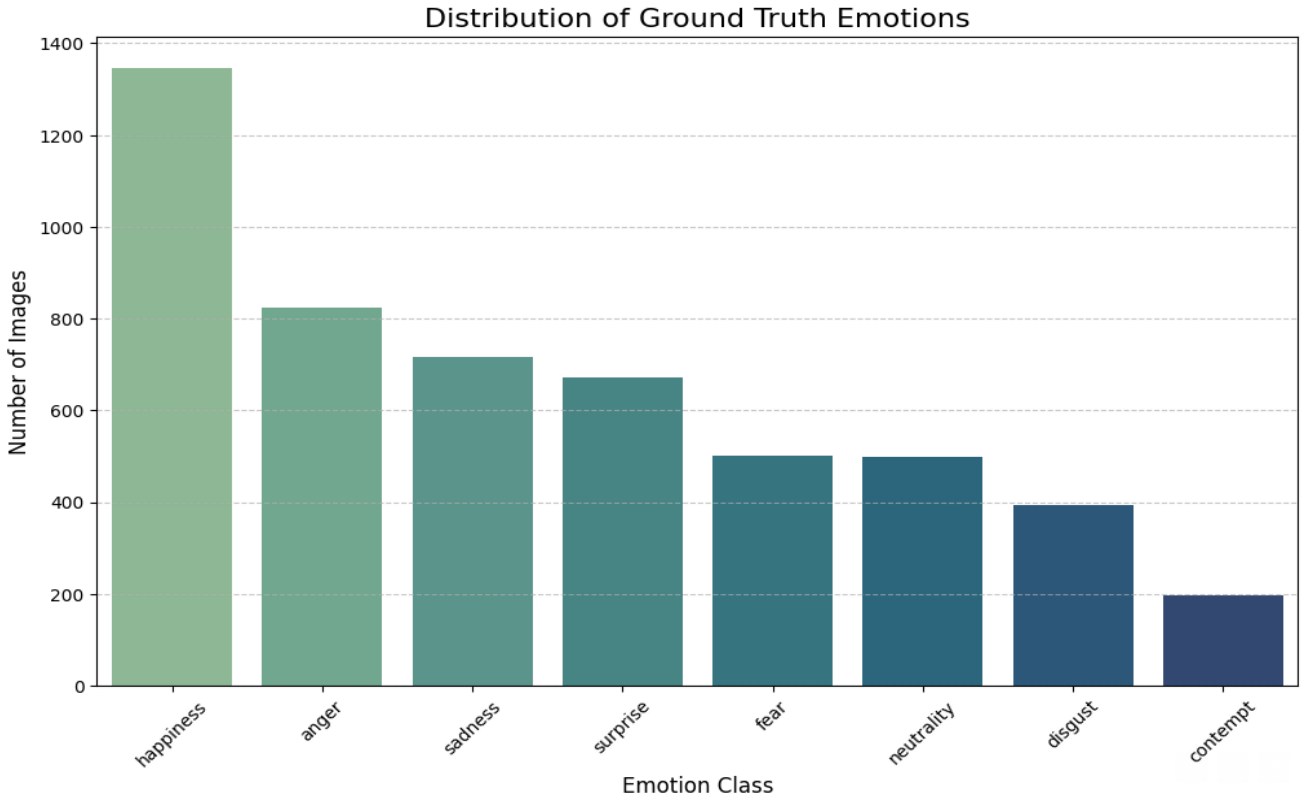
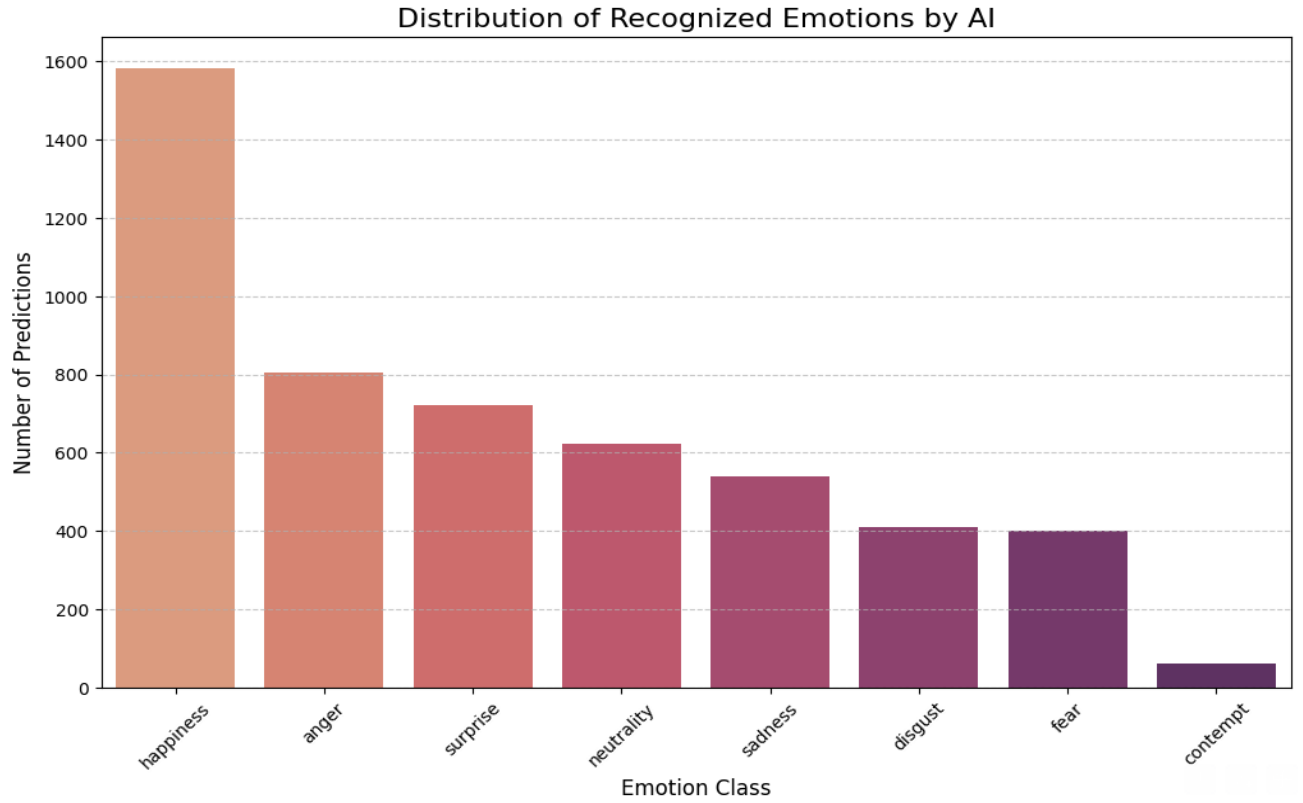
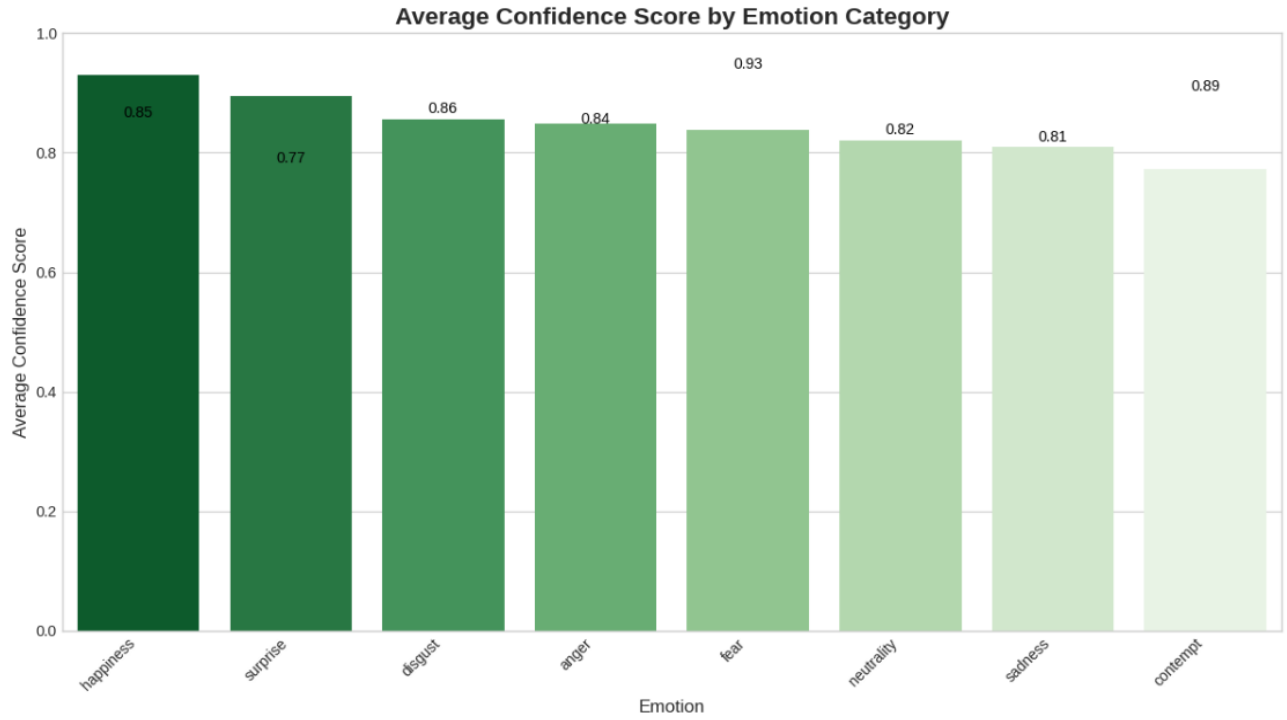
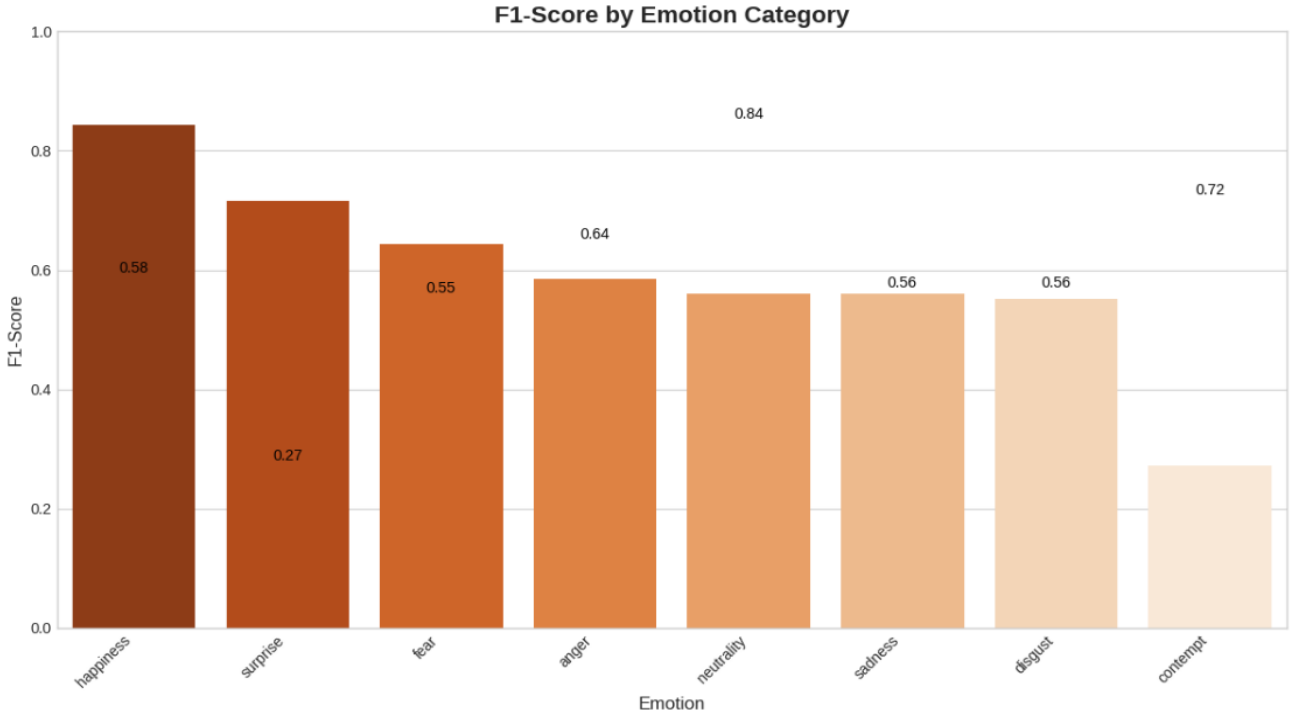


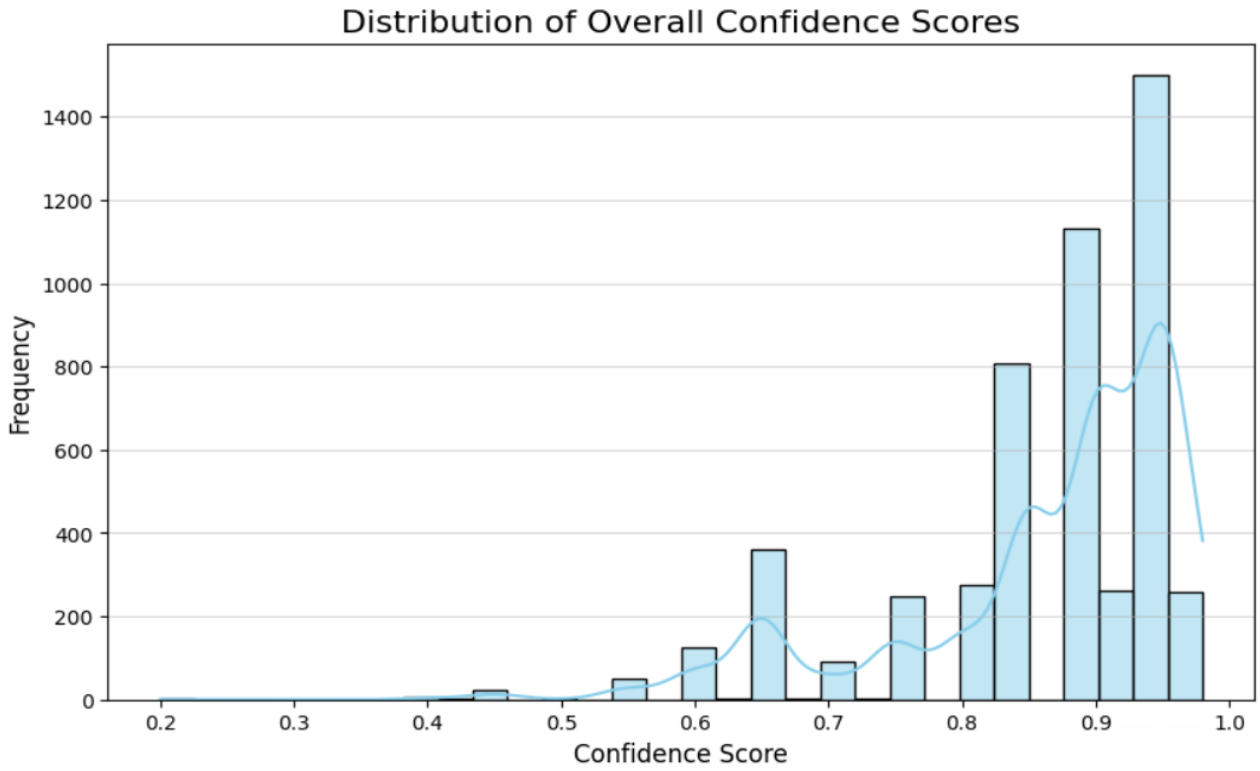
Figure 1: Confusion Matrix

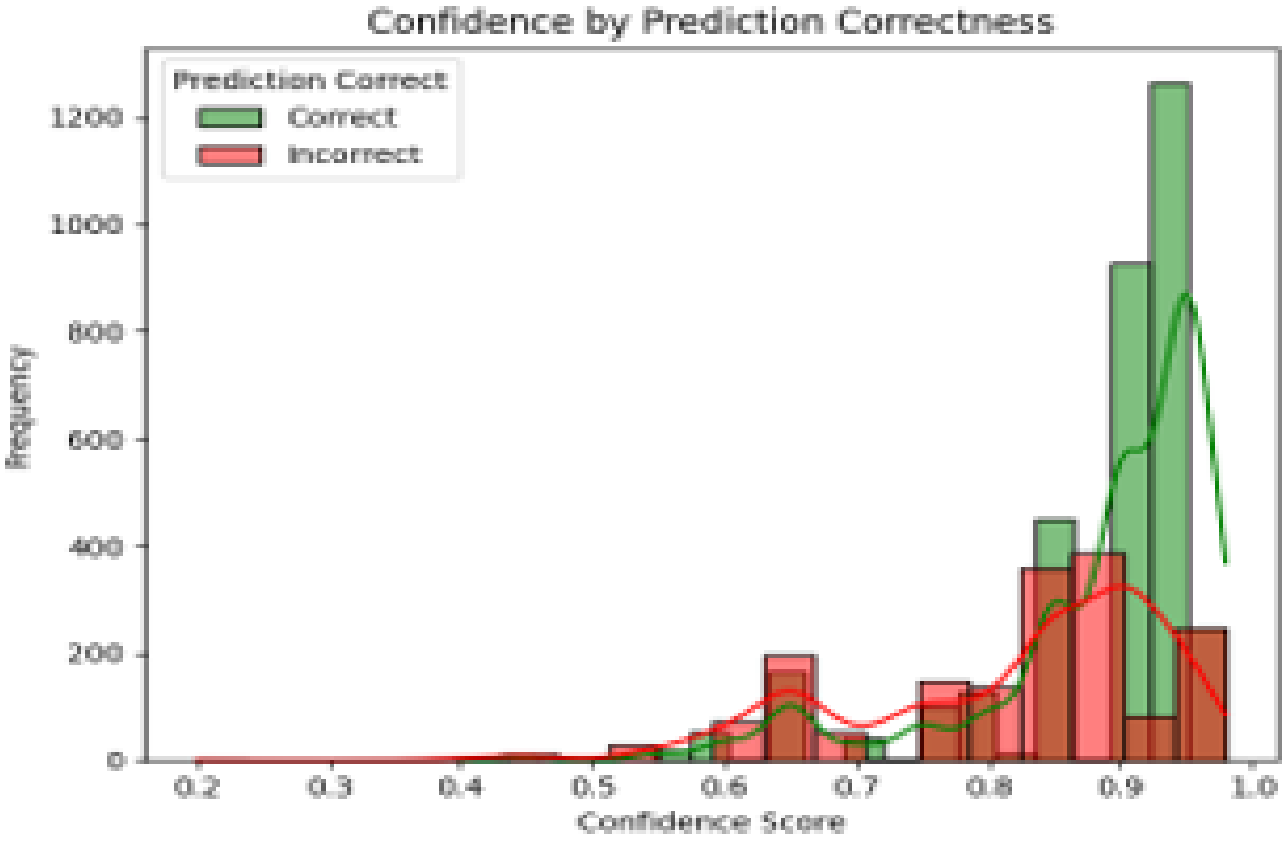

Distribution of Ground Truth Emotions
Number of Images
0
200
400
600
800
1000
1200
1400
happiness
anger
sadness
surprise
fear
neutrality
disgust
contempt
Emotion Class


Distribution of Recognized Emotions by AI
Number of Predictions
0
200
400
600
800
1000
1200
1400
1600
happiness
anger
surprise
neutrality
sadness
disgust
fear
contempt
Emotion Class


Average Confidence Score by Emotion Category
Average Confidence Score
1.0
0.8
0.6
0.4
0.2
0.0
0.85
0.77
0.86
0.84
0.93
0.82
0.81
0.89
happiness
surprise
disgust
anger
fear
neutrality
sadness
contempt
Emotion


F1-Score by Emotion Category
F1-Score
1.0
0.8
0.6
0.4
0.2
0.0
0.58
0.27
0.55
0.64
0.84
0.56
0.56
0.72
happiness
surprise
fear
anger
neutrality
sadness
disgust
contempt
Emotion


Distribution of Overall Confidence Scores
Frequency
Confidence Score
0
200
400
600
800
1000
1200
1400
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


Confidence by Prediction Correctness
Prediction Correct
Correct
Incorrect
Frequency
Confidence Score
0
200
400
600
800
1000
1200
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0