

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Volume 19, No. 2
June 2026
ISSN: 1946-1836

In this issue:

- 4. Immersive Virtual Reality to Support Blood Donation: A Pilot Study on Engagement and Anxiety Reduction**
Yi Li, Kennesaw State University
Maria Valero, Kennesaw State University
Brooke Berman, Kennesaw State University
Junhong Min, Kennesaw State University
Lakeeya Wiggins, Kennesaw State University
Lin Li, Kennesaw State University
Afekwo Mary Ukuku, Grady Memorial Hospital
Robert Keyser, Kennesaw State University
- 11. A Study of Chatbots Performance with Large Language Models in Scenario of a Higher Education Institution**
April Villeda Roblero, University of North Carolina Wilmington
Yang Song, University of North Carolina Wilmington
Judith Gebauer, University of North Carolina Wilmington
Yao Shi, University of North Carolina Wilmington
- 25. Examining Artificial Intelligence Job Descriptions: An Analysis of Current Job Postings**
Alan Peslak, Penn State University
Wendy Ceccucci, Quinnipiac University
Kiku Jones, Quinnipiac University
Margaret Wanjiku Gachau, Quinnipiac University
- 36. Retrieval Augmented Generation (RAG) Based Question and Answer System**
Emmanuel Balogun, Georgia Southern University
Hayden Wimmer, Georgia Southern University
Carl Michael, University of San Diego
- 50. Improving AI-Driven Stroke Prediction Models: A Comparative Evaluation of SMOTE and Undersampling Methods**
Dara Tourt, Metropolitan State University Minnesota
Queen Booker, Metropolitan State University Minnesota
- 71. What Agile Skills Are Needed for IT Jobs?**
Stephanie Totty, Middle Tennessee State University
Carol Clark, Middle Tennessee State University
Amy Harris, Middle Tennessee State University

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.

2026 ISCAP Board of Directors

Amy Connolly
James Madison University
President

Michael Smith
Georgia Institute of Technology
Vice President

Jeff Cummings
Univ of NC Wilmington
Past President

David Firth
University of Montana
Director

Mark Frydenberg
Bentley University
Director/Secretary

Leigh Mutchler
James Madison University
Director

RJ Podeschi
Millikin University
Director/Treasurer

Bryan Reinicke
Rochester Institute of
Technology / Director

Jeffrey Babb
West Texas A&M University
Director/Curricular Matters

Eric Breimer
Siena University
Director/2026 Conf Chair

Tom Janicki
Univ of NC Wilmington
Director/Meeting Planner

Xihui "Paul" Zhang
University of North Alabama
Director/JISE Editor

Copyright © 2026 by Information Systems and Computing Academic Professionals (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to Scott Hunsinger, Editor, editor@jisara.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH AND ANALYTICS

Editors

Scott Hunsinger
Senior Editor
Appalachian State University

Thomas Janicki
Publisher
University of North Carolina Wilmington

2026 JISARA Editorial Board

Biju Bajracharya
East Tennessee State University

Jason Price
Nichols College

Queen Booker
Metro State

Bryan Reinicke
Rochester Institute of Technology

Wendy Ceccucci
Quinnipiac University

Asish Satpathy
Arizona State University

Biswadip Ghosh
Metro State University

Dana Schwieger
Southeast Missouri State University

Russell Haines
Appalachian State University

Jeff Strain
Brigham Young University - Hawaii

Melinda Korzaan
Middle Tennessee State University

Katarzyna Toskin
Southern Connecticut University

Will Ledbetter
Perdue University

Karthikeyan Umapathy
University of North Florida

Li-Jen Lester
Sam Houston State University

Hayden Wimmer
Georgia Southern University

Muhammed Miah
Tennessee State University

David Woods
University of Miami Regionals

Alan Peslak
Penn State University

David Yates
Bentley University

Immersive Virtual Reality to Support Blood Donation: A Pilot Study on Engagement and Anxiety Reduction

Yi Li
joy.li@kennesaw.edu

Maria Valero
mvalero2@kennesaw.edu

Brooke Berman
bberman6@students.kennesaw.edu

Junhong Min
jmin7@students.kennesaw.edu

Lakeeya Wiggins
lwiggi14@students.kennesaw.edu

Lin Li
lli19@kennesaw.edu

Afekwo Mary Ukuku
amukuku@gmh.edu

Robert Keyser
rkeyser@kennesaw.edu

Kennesaw State University
Marietta, GA 30060

Abstract

Blood supply shortages continue to challenge healthcare systems, with particularly low participation rates among young adults. This paper reports on the development and pilot evaluation of an immersive virtual reality (VR) system designed to improve donor comfort and strengthen motivation to donate. Informed by the Theory of Planned Behavior and an extensive review of prior literature, the system integrates multiple virtual environments that combine education, relaxation, and interactive engagement, including a media space, information hub, calming room, game-based activities, and an exploratory gallery. Twenty-two participants between the ages of 18 and 39 completed pre- and post-experience measures assessing satisfaction, perceived anxiety, usability, and donation intent. The results indicate high overall acceptance of the VR experience, with most participants reporting strong satisfaction and an increased willingness to donate. Interactive and visually engaging components were rated most favorably, while reports of cybersickness highlighted important usability and technical limitations. Together, these findings demonstrate the promise of immersive VR as a tool for addressing psychological and experiential barriers to blood donation among younger populations. Ongoing work will focus on improving system stability, expanding content, and examining longer-term impacts on donor behavior through larger-scale studies.

Keywords: Virtual reality, blood donation, donor recruitment, immersive technology, anxiety reduction

Recommended Citation: Li, Y., Valero, M., Berman, B., Min, J., Wiggins, L., Li, L., Ukuku, A., Keyser, R., (2026). Evaluating Immersive Virtual Reality for Supporting Blood Donor Experience and Engagement. *Journal of Information Systems Applied Research and Analytics*, v19(n2) pp 4-10. DOI# <https://doi.org/10.62273/BSIW6059>

Immersive Virtual Reality to Support Blood Donation: A Pilot Study on Engagement and Anxiety Reduction

*Yi Li, Maria Valero, Brooke Berman, Junhong Min, Lakeeya Wiggins, Lin Li
Afekwo Mary Ukuku and Robert Keyser*

1. INTRODUCTION

Blood donation remains a cornerstone of modern healthcare, providing essential support for surgeries, trauma care, cancer treatments, and chronic illnesses. Despite this critical need, blood supplies in many regions, including the United States, have faced persistent shortages. According to recent estimates, although approximately 62% of Americans are eligible to donate blood, only about 3% do so regularly (Keyser et al., 2025). This gap between eligibility and participation poses significant challenges for maintaining a stable blood supply.

The problem is especially pronounced among younger adults. From 2019 to 2021, donations among individuals aged 16–24 dropped sharply, driven in part by the COVID-19 pandemic's impact on donation events and broader shifts in attitudes and lifestyle. Common barriers identified in prior research include fear of needles, discomfort with medical environments, lack of time, and limited awareness of the need for donation (Armitage & Conner, 2001; Lemmens et al., 2009).

VR is an emerging tool to address these challenges. Prior studies show that VR can reduce anxiety and increase engagement in healthcare, from surgical preparation to phobia treatment. In the blood donation context, VR may combine education, distraction, and relaxation to make the process less intimidating. It can also create positive associations for younger donors who are comfortable with interactive media. Yet, its use for blood donation has been studied little.

This study presents the design and pilot evaluation of a VR application developed to support donor education and reduce anxiety. The intervention was informed by the Theory of Planned Behavior and prior research on donor motivations and technology-enabled engagement. The goal was to assess the feasibility, acceptability, and perceived impact of the VR experience on participants' attitudes and willingness to donate blood.

2. LITERATURE REVIEW

The persistent shortage of blood donations remains a critical public health concern in the United States and worldwide. According to Keyser et al. (2025), while approximately 62% of the U.S. population is eligible to donate blood, only about 3% donate each year, underscoring an urgent need to attract and retain new donors. Declines have been most severe among younger age cohorts: from 2019 to 2021, donations fell by 62.5% among individuals aged 16–18 and by 34.9% among those aged 19–24, partly due to the COVID-19 pandemic's impact on school-based drives and partly due to broader shifts in attitudes and habits among young adults. Additional factors, including fear of needles, lack of time, perceived inconvenience, and distrust in healthcare institutions, contribute to this trend (Keyser et al., 2025).

To address these barriers, more research has explored emerging and disruptive technologies such as mobile applications, gamified engagement, mixed reality (MR), and VR. Ghouri et al. (2023) advocate the integration of technologies such as blockchain, AI, and neural networks to improve the transparency, safety, and efficiency of donation systems.

Evidence supporting the use of VR in blood donation contexts is robust. A pilot study published in *Transfusion* found that 68.4% of donors who used MR during blood collection reported a measurable reduction in pre-donation anxiety (Pandey et al., 2024). Williams et al. (2024) demonstrated that immersive VR interventions elicit positive emotions and can reassure donors, contributing to improved willingness to return. Research showed VR interventions can improve mood and significantly reduce tension and fear of needles in first-time donors (Tarrant et al., 2019). Additionally, Deighan et al. (2024) reported that while healthcare professionals recognize the promise of VR in clinical workflows, they note the importance of addressing challenges such as motion sickness and usability constraints in confined donation spaces.

The ACM Symposium on Virtual Reality Software and Technology presented further evidence of VR's motivational potential, leveraging the Proteus effect—where users' avatars progressively became more energetic and positive as they engaged with a donation simulation—to foster positive associations with the donation experience among young adults. Mäkinen et al. (2020) concluded from a systematic review that VR supports enhanced learning and engagement in healthcare training, suggesting similar potential to educate and motivate donors.

To bridge knowledge gaps and address behavioral barriers, our team's prior work (Keyser et al., 2025) synthesized 166 studies and recommended integrated interventions combining accurate education, interactive components, and opportunities for self-directed exploration. This recommendation aligns with the Theory of Planned Behavior (TPB), which identifies perceived behavioral control, subjective norms, and attitudes as the primary predictors of donation intention (Ajzen, 1991). Studies applying TPB in blood donation contexts (Armitage & Conner, 2001; Lemmens et al., 2009) have underscored the importance of providing clear information and reinforcing perceptions of control and self-efficacy.

These insights directly informed the design of our VR system. The prototype combines a suite of selectable environments: (1) a virtual theater offering high-quality informational videos on the donation process and its lifesaving impact; (2) a library where users can read articles and complete interactive quizzes to reinforce learning; (3) a meditation room with calming natural scenery and guided breathing exercises to reduce anxiety; (4) an arcade with simple, low-effort games designed to distract and occupy the donor during the procedure; and (5) an art gallery featuring curated works with optional descriptive captions. This approach aims to build confidence, alleviate fear, and strengthen positive associations with donation, especially in the critical 18–39 age group.

3. METHODOLOGY

This project followed a multi-phase, mixed-methods research design combining literature review, user-centered design, prototype development, and pilot evaluation.

Phase 1: Literature Review and Needs Assessment

We conducted a systematic literature review

informed by PRISMA guidelines, examining 166 peer-reviewed studies published between 1970 and 2024 (Keyser et al., 2025). Search queries included "blood donation barriers," "emerging technology in donor recruitment," and "VR in healthcare education." Major themes were synthesized and mapped to the Theory of Planned Behavior framework to identify determinants of donation intention.

Phase 2: Prototype Design

Drawing on research showing that VR can help reduce anxiety and improve engagement, we developed the VR prototype in Unity 3D and tailored it for Oculus Quest 2 and 3 headsets. Table 1 provides an overview of the main features and the reasons behind their inclusion. The design focused on four key priorities:

Educational Relevance: Delivering clear, factual information about the donation process, its benefits, and overall safety.

Anxiety Reduction: Offering calming environments, guided breathing exercises, and options for distraction, whether passive or interactive.

User Agency: Using a simple navigation model so donors could easily choose the environment that suited them best.

Accessibility: Keeping interaction straightforward and comfortable for people who would have a needle in one arm during the experience. The one-handed control setup allowed users to explore and interact without needing both hands.

| Feature | Description | Rationale |
|---------------------|---|--|
| Educational Theater | Informational videos with donation process and its impact | Increase knowledge; correct misconceptions |
| Library w/ Trivia | Articles; interactive questions | Reinforced learning; confidence building |
| Meditation scenery | Calming visuals; guided breathing | Reduce anxiety & stress |
| Arcade Games | Simple engaging activities | Distraction & fun |
| Art Gallery | Curated artworks with captions | Additional calming content |

Table 1: VR Prototype Design Overview

Phase 3: Iterative Development and Pilot Testing

During the pilot testing sessions, participants were invited to a controlled lab environment where they received an orientation on the VR equipment and study purpose. Each participant completed the pre-survey before being fitted with the VR headset. They were guided through the scene calibration and shown how to navigate the scene selection menu using their unoccupied hand. Participants then explored each of the five VR scenarios at their own pace, spending approximately 20–30 minutes in total. Research staff observed interactions unobtrusively, recording notes on usability issues, observed discomfort, and participant questions. After completing the experience, participants immediately completed the post-survey and were encouraged to elaborate on their ratings and suggestions in open-ended responses. Data were collected anonymously and linked using participant IDs for paired pre/post analysis.

The prototype underwent iterative refinement with feedback from stakeholders, including healthcare professionals and prospective donors aged 18–39. Usability issues—such as headset fit, navigation clarity, and cybersickness—were systematically documented and addressed.

Phase 4: Evaluation

Participants completed a pre-survey capturing demographic data (age, gender, prior donation history, prior VR experience) and attitudes about blood donation and VR. The post-survey included ratings of overall satisfaction, experiences of cybersickness, perceived anxiety reduction, favorite scenarios, interaction effectiveness, helpfulness of hints, suggestions for improvement, and willingness to recommend the program. Participants also reported whether the VR experience increased their inclination to donate and provided open-ended feedback describing their feelings during specific activities. Quantitative items were summarized with descriptive statistics, while qualitative responses were thematically coded to identify patterns in perceived effectiveness, usability, and emotional impact.

Ethics

This study protocol was reviewed and approved by the relevant institutional ethics board. All participants provided informed consent.

4. RESULTS

Participants were recruited through campus flyers, classroom announcements, and word-of-

mouth among college students. The final sample consisted of 22 participants aged 18 to 26. The demographics information is shown in Table 2. Almost half (46%) were between 18 and 20 years old, with another 46% between 21 and 23, and one participant aged 26. The majority identified as female (59%), with 36% identifying as male and 5% as another gender.

| Variable | Category | N | % |
|----------------------------|-----------------|----|-----|
| Gender | Female | 13 | 59% |
| | Male | 8 | 36% |
| | Other | 1 | 5% |
| Age Group | 18–20 | 10 | 46% |
| | 21–23 | 10 | 46% |
| | 24–26 | 1 | 5% |
| Education Level | College | 21 | 95% |
| | Graduate School | 1 | 5% |
| Prior Donation | Yes | 7 | 32% |
| | No | 15 | 68% |
| Prior VR Experience | Yes | 11 | 50% |
| | No | 11 | 50% |

Table 2: Demographics information

Education levels were diverse, with 32% of participants in their final year of college, 41% enrolled in college without senior standing, and smaller groups representing juniors, freshmen, and graduate students. Approximately one-third of participants (32%) reported having previous experience in blood donation, while two-thirds never donated blood. Prior VR experience was evenly distributed, with 50% indicating they had used VR before and 50% reporting no prior experience. This demographic diversity provided a relevant perspective on the usability and perceived impact of the VR intervention targeting young adults.

All participants completed the pre- and post-surveys following the VR experience. Overall satisfaction with the program was high, with 82% (18 participants) rating their experience as “Very Satisfied” or “Extremely Satisfied,” and 18% (4 participants) indicating a neutral response. No participants reported dissatisfaction. The most frequently cited positive aspects included the variety of interactive environments and the informative content, particularly in the arcade and art gallery scenarios.

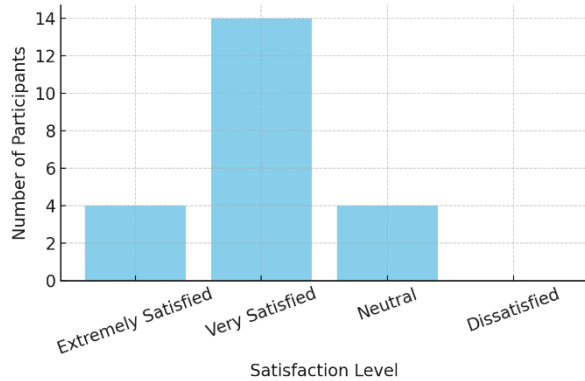


Figure 1: Overall satisfaction with VR prototype

Cybersickness was reported by 9 participants (41%), with symptoms ranging from mild dizziness and nausea to eye discomfort. Notably, several participants who experienced cybersickness still expressed overall satisfaction with the program.

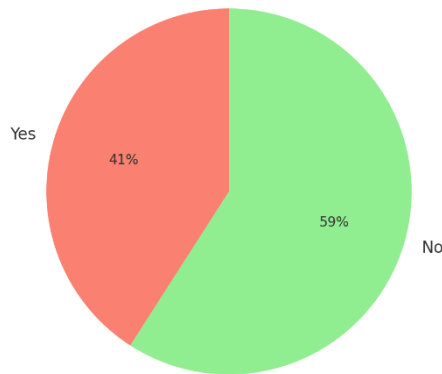


Figure 2: Participants reporting cybersickness

Half of the participants (11) indicated that the VR activities helped alleviate their concerns about blood donation. Among the factors contributing to reduced anxiety were the immersive distractions (e.g., arcade games, art gallery) and the clear, structured information provided in the library and theater environments. For some participants, the calming visuals and guided breathing in the meditation scene also played a role in reducing stress.

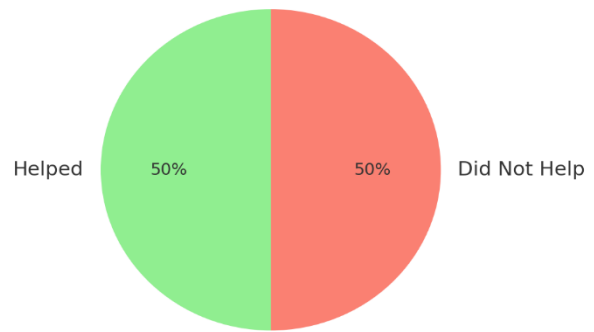


Figure 3: Concern alleviation.

The arcade and art gallery were the most popular activities, selected as favorites by 16 and 13 participants respectively. The library was endorsed by 9 participants for its educational value, though it was also among the most commonly cited least favorite experiences, along with the meditation room, due to overstimulation or confusion about purpose. Several participants noted that more instructions and clearer navigation cues would improve usability.

Regarding interaction effectiveness, 11 participants reported being able to interact effectively with the VR environment, while 10 indicated they were somewhat effective, suggesting moderate learning curves and opportunities for refinement. Hints provided within the program were rated as helpful or very helpful by most respondents (86%).

Importantly, 64% of participants (14) stated that they were more inclined to donate blood after completing the VR experience. Reported reasons included increased understanding of the process, reduced anxiety, and the appeal of engaging distractions during donation. Five participants indicated they were less inclined, primarily due to persistent needle anxiety or discomfort with VR itself.

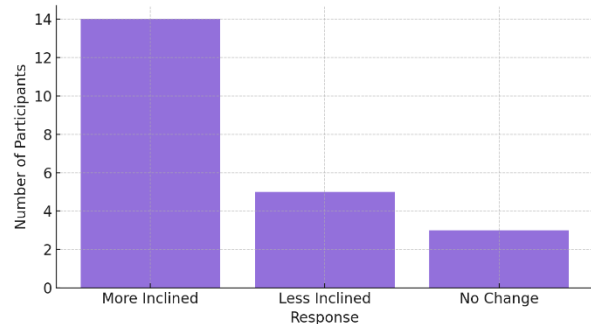


Figure 4: Change in willingness to donate blood

On a scale from 1 to 10, the likelihood of recommending the VR program to others averaged approximately 8.3, with scores ranging from 3 to 10. The most frequent ratings were 8 and 10, reflecting positive overall perceptions.

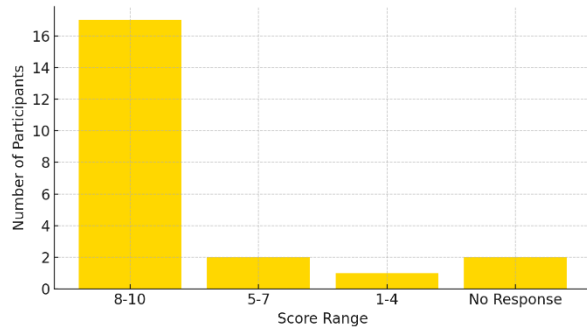


Figure 5: Recommendation likelihood (1-10 scale).

5. DISCUSSIONS

Overall, the findings indicate that VR interventions hold considerable promise for enhancing donor education, reducing anxiety, and increasing willingness to donate.

Participants' satisfaction ratings were largely positive, with 82% reporting being "Very Satisfied" or "Extremely Satisfied." This aligns with prior research showing that immersive technologies can elicit positive emotional responses and improve perceptions of medical procedures (Williams et al., 2024; Pandey et al., 2024). Notably, 64% of participants reported that the VR experience increased their inclination to donate blood, suggesting the approach is effective in addressing attitudinal barriers.

The most frequently endorsed features were the arcade and art gallery, underscoring the value of interactive and visually engaging environments for distraction. This supports the design rationale that combining educational components with appealing, low-pressure activities can maintain engagement and reduce perceived stress. However, the mixed feedback regarding the meditation and library scenes highlights the need for careful calibration of content complexity and sensory input to avoid overstimulation or confusion.

Cybersickness emerged as a substantial limitation, with 41% of participants experiencing some discomfort. This proportion is consistent with other VR studies in healthcare contexts and reinforces the importance of optimizing frame

rates, headset ergonomics, and motion design to mitigate adverse effects (Deighan et al., 2024). Despite these issues, most participants still rated their experience positively, indicating that the benefits of immersion outweighed the drawbacks for many users.

Interaction effectiveness ratings and feedback also pointed to opportunities for improvement. While the majority of participants could navigate and interact with the scenes successfully, a subset indicated only partial success, often related to unfamiliarity with VR controls or headset adjustments. Enhancing onboarding instructions, simplifying navigation gestures, and providing real-time guidance may help address these usability barriers.

Future work will focus on several areas. First, expanding the study to larger and more diverse participant samples will help validate these findings and explore subgroup differences in response to VR interventions. Second, incorporating longitudinal follow-up can assess whether the positive impacts on willingness to donate translate into actual donation behavior over time. Finally, iterative development will continue to refine content, interaction design, and technical performance to maximize both engagement and comfort.

6. CONCLUSIONS

This pilot study demonstrates that a VR approach can be a promising tool for engaging and educating prospective blood donors. The majority of participants expressed high satisfaction with the experience, and many reported that it reduced their anxiety and increased their willingness to donate. Interactive and visually rich environments, particularly the arcade and art gallery, were especially well received. These findings are consistent with prior studies that have highlighted VR's capacity to improve mood and support health-related behavior change.

In parallel, the prevalence of cybersickness among participants underscores the need for careful attention to technical optimization and ergonomic design. Usability feedback also indicated opportunities to improve navigation clarity and onboarding instructions to ensure a comfortable and accessible experience for first-time VR users.

In summary, the proposed work contributes to the growing evidence base supporting the role of immersive technologies in healthcare

engagement and highlights the potential for VR to become a valuable component of blood donation recruitment and retention strategies.

7. REFERENCES

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.
- Armitage, C. J., & Conner, M. (2001). Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *British Journal of Social Psychology*, 40(4), 471–499. <https://doi.org/10.1348/014466601164939>
- Deighan, M.T., Simpson, D., Brigden, A. et al. Perspectives of healthcare professionals on the use of immersive virtual reality in teenage and young adult oncology: a qualitative interview study. *BMC Digit Health* 2, 22 (2024). <https://doi.org/10.1186/s44247-024-00071-2>
- Keyser, R. S, Li, D. L., Kimber, E., Berman, B., Kerbage, G., Valero, M., Li, Y. Joy and Ukuku, A. Mary (2025). Exploring the blood donation problem in the U.S. and the use of emergent and disruptive technologies to attract new donors. *International Journal of Research in Industrial Engineering*, 14(2), 256-280. doi: 10.22105/rirej.2024.472560.1466
- Lemmens, K. P., Abraham, C., Ruitter, R. A., Veldhuizen, I. J., Dehing, C. J., Bos, A. E., & Schaalma, H. P. (2009). Modelling antecedents of blood donation motivation among non-donors of varying age and education. *British Journal of psychology*, 100(1), 71-90. <https://doi.org/10.1348/000712608X310237>
- Mäkinen, H., Haavisto, E., Havola, S., & Koivisto, J. M. (2020). User experiences of virtual reality technologies for healthcare in learning: an integrative review. *Behaviour & Information Technology*, 41(1), 1–17. <https://doi.org/10.1080/0144929X.2020.1788162>
- Pandey, S., Goel, R., Kapral, J., Kieffer, T., Kang, J., Shaffer, H., Hermelin, D., & Hartwell, B. (2024). The use of mixed reality technology within the donor collection experience. *Transfusion*, 64(2), 315–324. <https://doi.org/10.1111/trf.17712>
- Tarrant, J., Abrams, J. S., & Jackson, R. (2019). The impact of virtual reality on mood states prior to blood donation. *Depression*, 1, 0-7. <https://dx.doi.org/10.32474/SJPBS.2019.02.000150>
- Ghouri, A. M., Khan, H. R., Mani, V., ul Haq, M. A., & de Sousa Jabbour, A. B. L. (2023). An Artificial-Intelligence-Based omnichannel blood supply chain: A pathway for sustainable development. *Journal of Business Research*, 164, 113980. <https://doi.org/10.1016/j.jbusres.2023.113980>
- Williams, L.A., Tzelios, K., Masser, B. et al. A virtual reality paradigm simulating blood donation serves as a platform to test interventions to promote donation. *Sci Rep* 14, 10334 (2024). <https://doi.org/10.1038/s41598-024-60578-6>

A Study of Chatbots Performance with Large Language Models in Scenario of a Higher Education Institution

April Villeda Roblero
av9401@uncw.edu

Yang Song
songy@uncw.edu

Judith Gebauer
gebauerj@uncw.edu

Yao Shi
shiy@uncw.edu

University of North Carolina Wilmington
Wilmington, NC 28405

Abstract

Large language models (LLMs) and the use of artificial intelligence (AI) are becoming increasingly integrated into everyday life. As the models continue their training and improvement, leading technology companies are competing to create the most advanced artificial assistant. This competition has produced a variety of large language models, each with different capabilities. While LLMs like ChatGPT, Claude, and Gemini have proven effective at handling everyday tasks, they often lack the domain-specific expertise necessary for more specialized consultations. This limitation makes it essential to integrate a targeted knowledge base when developing chatbots for specific domains. Therefore, this study aims to address the question: Is there a model that clearly outperforms the others, considering correctness and comprehensiveness? In the study, we conducted an experiment to investigate the performance of chatbots utilizing LLMs within the context of a custom, mid-scale knowledge base designed for a university located in the southeastern United States. Using a web-crawled knowledge base, we created chatbots across multiple platforms and LLMs, testing them against a set of predefined questions to evaluate correctness and comprehensiveness. The findings highlight the disparity in the capability of LLMs and offer practical guidance for their effective use.

Keywords: Chatbot, Large Language Models, LLMs, AI, Higher Education

Recommended Citation: Villeda Roblero, A., Song, Y., Gebauer, J., Shi, Y., (2026). A Study of Chatbots Performance with Large Language Models in Scenario of a Higher Education Institution. *Journal of Information Systems Applied Research and Analytics*, v19(n2) pp 11-26. DOI# <https://doi.org/10.62273/SIVA7683>

A Study of Chatbots Performance with Large Language Models in Scenario of a Higher Education Institution

April Villeda Roblero, Yang Song, Judith Gebauer and Yao Shi

1. INTRODUCTION

Generative artificial intelligence has developed rapidly in recent years. Large Language Models (LLMs) represent a breakthrough in AI technology. These models are changing how people acquire, understand, and use information. Popular examples include ChatGPT (Firat, 2023), Claude (Liu et al., 2024), and Gemini (Islam & Ahmed, 2024). LLMs possess powerful natural language processing and knowledge generation capabilities. This enhances the efficiency and convenience of obtaining information through the internet. In fields that require domain knowledge and understanding of the context/scenario, such as consultation, LLMs have demonstrated "remarkable potential" (Song et al., 2023). In education settings, LLMs are also "promising tools for open education" as they can provide customized and interactive assistance to students and thereby improve the independence and autonomy of the learners (Firat, 2023). "Domain-specific" LLMs (Zhang et al., 2024) can respond quickly to diverse student needs. They provide personalized advice and information based on massive datasets. This capability reduces the burden on traditional consultation, support, and feedback methods.

However, LLMs face significant reliability and accuracy challenges in practical applications. The training data for these models is complex, diverse, and potentially biased. This leads to several problems during the understanding and inference processes. LLMs may misunderstand queries or engage in erroneous reasoning. They can even generate fabricated information, a phenomenon sometimes called "hallucination" (Chang et al., 2024). These issues highlight the gap between LLMs' demonstrated potential and their current limitations in real-world settings.

This uncertainty creates potential risks for users. Many users rely on LLM recommendations when making important decisions. Poor model accuracy could lead to harmful outcomes. Therefore, researchers must examine the accuracy and reliability of LLMs in consultation contexts. Such exploration is essential for

promoting safe applications across different domains. Understanding these limitations will help ensure healthy deployment of LLM technology.

This study systematically evaluates the performance of major LLMs in answering a common set of questions. We used web crawling technology to collect extensive information about an R2 university in the southeastern United States. This information served as our knowledge base. The evaluation included two types of questions. **Objective questions** focused on new student enrollment topics such as tuition fees, accommodation, and food services. **Subjective questions** used specific scenarios and personas to assess contextual understanding. We input this data into multiple LLM platforms to create chatbots. The platforms included both commercial and open-source models: ChatGPT, Claude, Gemini, Copilot, Llama, and DeepSeek. These chatbots were designed to simulate real-life question-answering scenarios using the university knowledge base.

We compared and analyzed answers from different LLMs to achieve three main objectives. First, we demonstrate our evaluation framework for LLMs. This framework uses a knowledge base and question set that includes both objective and subjective items. Second, we explore the impact of reliable knowledge bases versus online search capabilities in building domain-specific chatbots. We analyze how these components improve answer quality and identify their limitations. Third, we propose strategies for improving LLM applications in educational consulting based on our research results. These recommendations serve as a reference for practitioners and researchers in related fields.

2. BACKGROUND

Chatbots represent an important research area within natural language processing (NLP), which is a subfield of AI that uses machine learning to help computers interpret, manipulate, and comprehend human language. Early machine translation research, a subdomain of NLP, relied

on specific evaluation metrics. BLEU (Papineni et al., 2002) and METEOR (Banerjee & Lavie, 2005) were two common approaches. BLEU evaluates translation accuracy by calculating similarity between machine-generated translations and reference translations. METEOR improves upon BLEU by introducing an alignment algorithm. This algorithm better handles synonyms and word order differences when evaluating similarity between generated and reference translations. However, these metrics are ineffective for evaluating chatbots.

Mehri and Eskenazi (2020) proposed the FED measurement metric to address this limitation. FED measures fine-grained dialogue quality at two levels. It evaluates individual dialogue turns, defined as "a dialog context and a system response (from chatbot)" (Mehri & Eskenazi, 2020), and entire dialogues. FED achieved moderate to strong correlation with human judgments at both levels.

Chatbot technology has advanced significantly with the introduction of LLMs. End users now have higher requirements for chatbot performance. User focus has shifted from basic fluency to multi-dimensional evaluations. Current assessments examine reading comprehension, reasoning capability, mathematical skills, and other technical measures (AI et al., 2024).

Higher expectations for chatbot performance bring attention to the Turing Test. Alan Turing, often called the "father of computer science," first introduced this concept in 1950. The test was originally known as the imitation game (Oppy & Dowe, 2021). The game involves behavioral evaluations that assess whether a machine can imitate human conversation. The key question is whether machine responses become indistinguishable from human responses. Turing argued that if a computer's response seems indistinguishable from a human response, we should consider whether it qualifies as a thinking entity. This question remains relevant today. Some research claims that modern LLMs have "passed" the Turing Test. This study does not conduct the Turing Test directly. However, we draw from similar principles by analyzing chatbot responses. Our focus is on their ability to provide accurate and contextually appropriate answers.

Researchers use specific datasets to evaluate dialogue response quality. These evaluations test fluency, naturalness, and other aspects to determine whether chatbots reach human-level

performance. Mendonça et al. introduced a new evaluation benchmark called SODA-EVAL (Mendonça et al., 2024). This dataset used more than 120,000 turn-level assessments for training across 10,000 conversations. The researchers conducted human validation and annotation tasks to confirm automatic annotation quality. The evaluation system used a rating scale from 1 to 5 to assess dialogue response quality. Several factors determined these ratings. Evaluators examined whether responses contradicted dialogue history information. They assessed whether models covered all relevant conversation information. The system also measured fluency and naturalness of dialogue responses. This included evaluating common sense, participation levels, and repetition in conversations. The research revealed important findings about GPT-4 performance. Although GPT-4 performed well in many aspects, it showed room for improvement in coherence and common-sense reasoning. This indicates that large language models have made significant progress in generating fluent and relevant responses. However, they still face challenges in simulating the complexity of human dialogue.

Researchers have successfully created domain-specific chatbots that support domain experts in reading and decision-making. This progress has stimulated research in domain-specific chatbot evaluation methods. Song et al. (2023) developed an evaluation framework for medical applications. They created a questionnaire with 21 questions and two clinical scenarios related to urolithiasis. The researchers tested four LLMs: Claude, Bard, ChatGPT-4, and Bing. Domain experts evaluated the model responses using multiple criteria. These included accuracy, comprehensiveness, legibility, human care, and clinical case analysis ability. Evaluations used a 5-point Likert scale for systematic assessment. The study found that Claude and GPT-4 were the top-performing LLMs across their evaluation metrics. However, the research had a limited scope. The evaluation focused only on clinical urolithiasis-related dialogues, which represent a relatively narrow domain.

Educational applications of LLMs have also received research attention. Hwang et al. (2023) used an AI-driven approach to create and assess multiple-choice questions in chemistry and biology. They evaluated question quality using Item Writing Flaw (IWF) criteria (Breakall et al., 2019). The researchers combined machine learning models with human assessments to verify question alignment with Bloom's Taxonomy. The study used the RoBERTa model

to validate 120 generated questions. A domain expert with over 28 years of STEM education experience assessed 57 of these questions. The research found that GPT-3.5 can generate questions aligned with Bloom's Taxonomy levels. However, notable differences emerged between human and machine quality assessments. These findings suggest a discrepancy between machine learning models and human evaluations when assessing AI-generated content. The results indicate that carefully considered evaluation standards are necessary for assessing AI-generated materials. This highlights ongoing challenges in developing reliable automated evaluation methods for educational content.

3. EXPERIMENT DESIGN

The scope of the experiment involves building chatbots on different platforms using data related to UNC Wilmington, a university we selected for this research, which is also the home institution of the authors. The data is crawled mainly from the website of the university under the uncw.edu domain. We also compare the performance of the chatbots when they answer evaluation questions with or without restricted internet access. In the data acquisition process, a web crawler was used to gather content from multiple pages within the domain from a URL such as <https://library.uncw.edu/>, <https://uncw.edu/seahawk-life/dining-housing/housing/>, or <https://uncw.edu/research/>. Those are the URLs that students may need to visit frequently, especially in their freshman year. The web crawling process ensures that the pages cover comprehensive information about this university, including details on housing, transportation, safety reports, meal plans, tuition fees, scholarships, and student life. The web crawler organized content into 20 topic-specific files. These files were saved in .docx, .pdf, or .csv formats. The collected data generated more than 148,000 tokens, which formed the foundation for the chatbot knowledge base. We conducted a preliminary comparison to determine the optimal file format. Two chatbots used the same knowledge base content, but one received a .pdf file while the other received a .docx file. The chatbot that used the .docx knowledge base performed better than its counterpart. Based on this finding, all subsequent chatbots in this study were trained using .docx files as their knowledge base unless otherwise noted.

We explored various open-source LLM and commercial AI chatbot frameworks to identify

suitable tools for testing and comparison. Six platforms were selected for this project: GPT-4o, Gemini 1.5, Claude, Copilot Studio, Llama, and DeepSeek. We created multiple chatbots using these platforms and provided our knowledge base to each for testing. We included a baseline comparison by testing GPT-4o without any knowledge base. We also analyzed the impact of allowing chatbots to access the internet for searches in addition to using the knowledge base. For Llama, we tested different model sizes to assess performance variations. These included Llama 2 13B and Llama 2 70B models.

We focused exclusively on independent models and model providers when selecting commercial LLMs. Therefore, the "secondary" AI platforms that use those LLMs (such as BoodleBox) were excluded. **Table 4** in Appendix A lists the chatbots we have tested. When testing these LLMs, to minimize the impact of irrelevant factors, we provided the same prompt for the chatbots on each platform and conducted the tests using identical questions.

4. EVALUATION FRAMEWORK

We developed a structured framework to assess chatbot performance using two sets of testing questions. The first set contained 20 objective questions designed for single-turn interactions. In these tests, users asked one question and received one response without follow-up or context retention. This approach allowed us to evaluate basic chatbot performance. The second set introduced three fictional personas to test more complex interactions. Each persona engaged with chatbots through 13 subjective questions in multi-turn dialogues. These conversations involved multiple exchanges between users and chatbots. The format allowed for context retention and follow-up questions within coherent conversations. **Table 3** in Appendix B lists the objective and subjective questions we created.

We created two test sets covering various aspects of university life. The first set contained 22 **objective** questions that focused on factual information. These questions addressed topics such as tuition and fees, housing options, and campus safety policies. The second set included three fictional personas and 13 **subjective** questions related to personal preferences. Example questions included "What food options that align with my taste does the on-campus dining offer?" and "Are there any on-campus housing options that will fit my lifestyle and preferences?" We developed detailed profiles for

three virtual incoming college students to create these personas. Each profile specified preferred majors, food preferences, personal hobbies, and other individualized characteristics. This approach allowed us to test how well chatbots could provide personalized responses based on specific user needs and preferences.

We used different testing approaches for each question set. For the first set of questions, we conducted single-turn conversations to obtain and evaluate responses directly. This provided straightforward assessment of factual question handling. For the second set, we employed a multi-turn dialogue approach. We first provided each fictional persona's background information to establish context for the interaction. We then asked subjective questions from the test set. We expected chatbots to respond using both the knowledge base and specific persona details. **Table 3** in Appendix B lists all the assessment questions we have created for our experiments.

We evaluated chatbot dialogues using two key aspects: *correctness* and *comprehensiveness*. Each chatbot response received ratings on both dimensions using a 3-point Likert scale. Two researchers conducted the rating process independently. The inter-rater reliability achieved a Kappa index of 0.68, indicating substantial agreement between evaluators. The Likert scales were defined as follows in **Table 3**.

Please note that we used a 3-point Likert scale instead of the more commonly used 5-point version. We adopted simplicity and faster training of raters for the project timeline.

| Correctness | |
|-------------------|---|
| 0 (Poor) | Wrong information (hallucination) or refusal to answer when the knowledge base includes the information needed. |
| 1 (Fair) | Partially wrong/inaccurate or misleading information provided, or refusal to answer when the knowledge base does not provide adequate information to answer. |
| 2 (Good) | Accurate information, or the information is very close to the information provided (considering some numbers, like student population, will change every year). |
| Comprehensiveness | |
| 0 (Poor) | Not enough information is provided to cover all the aspects of the question. |
| 1 (Fair) | Some information related to the |

| | |
|----------|---|
| | question but inadequate, or too much additional and unrelated information. |
| 2 (Good) | Good coverage of all the aspects, no or limited unrelated information provided. |

Table 3: Evaluation Scales
5. EVALUATION RESULTS

Objective Questions

We created three ChatGPT-based chatbots with different configurations: GPT-4o without a knowledge base (GPT-4o), GPT-4o with knowledge base only (GPT-4o+KB), and GPT-4o with knowledge base plus search capabilities (GPT-4o+KB+S). The baseline GPT-4o relied on built-in knowledge and real-time internet searches. We found that GPT-4o was able to produce reasonably accurate and comprehensive answers that are comparable to GPT-4o+KB and better than Gemini, Copilot Studio, Llama 13B, Llama 70B, and DeepSeek. However, GPT-4o may make mistakes because the information searched may not be from the university website/domain. The ChatGPT-based model that uses the given knowledge base but no real-time searches, GPT-4o+KB, could produce a similar level of correctness, but we observed that its performance of comprehensiveness was slightly weaker. The ChatGPT-based model that uses both the given knowledge base and online searches (the prompt required it to use the knowledge base as the *primary* information source), GPT4o+KB+S, produced the highest correctness and comprehensiveness among all the models. The results are listed in Appendix C (**Table 6, Table 7**)

Claude achieved the highest accuracy despite being an offline LLM without real-time search capabilities. Claude's correctness scores were phenomenal (tied for highest among all models). However, Claude frequently provided additional unrelated information, resulting in the fifth-highest comprehensiveness score due to irrelevant content. Claude's knowledge base capacity limitations required us to split our files and create four separate chatbots to answer all test questions.

Our Gemini-based chatbot missed important information from the knowledge base and often provided incomplete answers. For example, when asked about tuition, it provided only in-state rates while omitting out-of-state costs. Additionally, Gemini could not process tabular data such as Excel or CSV files. This limitation prevented it from accessing major lists and core course requirements, resulting in several "n/a"

responses in our results.

The Copilot Studio-based chatbot received the most zero scores, indicating frequent inability to provide answers even when information existed in the knowledge base. Copilot provided shorter, less comprehensive responses compared to other chatbots. Despite allowing searches across four hyperlink domains, it performed weakest in both correctness and comprehensiveness. However, Copilot Studio offered the most extensive customization features, including conversational flow management, suggesting potential for handling complex tasks in future applications.

We tested two different Llama chatbots: Llama 2 13B and Llama 2 70B, both tested on LM Studio. Although Llama 13B is significantly smaller in size, it outperformed Llama 70B in both correctness and comprehensiveness. This suggests that larger model size does not necessarily lead to better performance, especially in domain-specific tasks. Llama70B struggled more with hallucinations but generally stayed on topic. In addition, both models outperformed Copilot Studio.

DeepSeek 7B performed surprisingly well despite being one of the smallest models in the study. It received second place for comprehensiveness and third place for accuracy. DeepSeek frequently produced well-rounded responses, but it primarily suffered from providing partially incorrect information. Unlike Claude, DeepSeek was thorough enough without going into too much detail or overburdening the response. It is worth noting that DeepSeek is a relatively new model and was added towards the end of the study. Testing was restricted to a shorter period of time after access was made available via LM Studio. Despite these challenges, DeepSeek still performed well considering it's the smallest model tested in this study. **Tables 4 and 5** list the scores we have given to our chatbot on correctness and comprehensiveness.

Subjective Questions

We tested chatbot performance using three fictional personas with different university fit levels. The first persona was a poor fit based on intended major, location preferences, and city setting. The second was clearly a good fit considering academic plans and personal hobbies. The third represented a "borderline fit" for academic goals. These personas tested chatbot ability to answer subjective questions tailored to specific needs and interests. The results are listed in Appendix C (**Table 8, Table**

9, Table 10).

ChatGPT-based models (GPT-4o+KB and GPT-4o+KB+S) showed good but unstable performance across the three personas. Zero ratings primarily resulted from hallucinations where models fabricated information not in the knowledge base. For example, GPT-4o+KB+S incorrectly suggested a "Bachelor of Science in Civil Engineering" degree that wasn't on the provided major list. GPT-4o+KB, which relied solely on the knowledge base, showed fewer hallucinations.

Claude was most straightforward in advising "this university is not a good fit" for the first and third personas, making it the top performer for the first persona. However, Claude sometimes couldn't provide personalized suggestions due to limited information access (no real-time search capability). When lacking specific knowledge base information, it offered general responses instead. For example, when asked about off-campus apartment rentals aligned with personal needs, Claude provided only general information about housing availability and costs rather than personalized recommendations.

Gemini often ignored specific persona attributes when answering questions. For example, it calculated student loans using in-state tuition rates for an out-of-state student. Gemini provided lengthy responses with factual information that wasn't necessarily relevant to individual needs. However, it performed better than GPT-4o and Claude for the third persona. Gemini's inability to process tabular data prevented us from using all test questions, resulting in "n/a" entries in **Tables 6-8**.

Copilot Studio provided answers consistent with the knowledge base but often ignored fictional persona needs. For example, when the first persona preferred "her own apartment/suite with a private bedroom and bathroom," Copilot still suggested "double-occupancy, pod-style rooms." Copilot's responses were also less comprehensive than other chatbots. When asked to estimate student loans, most chatbots provided calculations and numbers, while Copilot offered only generic advice: "You can estimate your total student loan payment by multiplying the annual loan amount by the number of years and adding any interest accrued." While technically correct, this response wasn't helpful.

Both versions of LLama performed moderately; LLama 13B slightly outperformed LLama 70B in both correctness and comprehensiveness. LLama

13B tended to produce more grounded and relevant responses, particularly when dealing with questions related to housing, campus life, and tuition estimates. In contrast, LLama 70B occasionally hallucinated program names or exaggerated student life details that were not included in the knowledge base. However, both models did better than expected, especially considering they operated entirely offline without real-time search. Llama 13B and Llama 70B hallucinated, but their responses were often aligned with the persona's background. For example, both models sometimes offered generic recommendations without fully incorporating the unique hobbies or goals of the student persona. Nonetheless, their mostly accurate responses allowed them to outperform Copilot Studio and, in some cases, Gemini.

DeepSeek 7B proved to be both precise and efficient in addressing the subjective needs of the personas. Deepseek responded with relevant answers that aligned well with the persona's background and avoided unrelated content. The third persona was notably difficult for DeepSeek to handle. DeepSeek's responses lacked context of this persona, and it frequently overexplained or hallucinated information such as degrees or on-campus student services. Additionally, DeepSeek occasionally lacked deeper personalization when compared to models like Claude or GPT-4o+KB+S. Overall, DeepSeek's performance was solid and commendable.

6. CONCLUSION

In this paper, we crawled available information from a university in the southeastern United States, UNC Wilmington. The data covered multiple aspects, including academics, admission, student life, research, and university athletics. The volume of this unstructured knowledge base was comparable to that of a small-to-medium corporation (Jordão & Novas, 2024). We also included the university's annual security report as a PDF file. Additionally, we incorporated a major list and core course list that were crawled as Excel files. We created a testing framework to assess chatbot capabilities using the given knowledge base. The framework targeted chatbot performance in answering both objective and subjective questions. Objective questions focused on factual information while subjective questions addressed personal preferences. All questions were designed from the perspective of potential incoming students.

Among the commercial LLM models tested, ChatGPT-4o-based models performed strongest

in both correctness and comprehensiveness for objective questions. This finding aligns with Rydzewski's research (Rydzewski et al., 2024). However, no single model consistently achieved the highest performance on subjective questions across different fictional personas. ChatGPT-4o, Claude, DeepSeek, and Gemini each demonstrated superior capabilities for different fictional personas. The following section presents our qualitative findings for different commercial LLM models.

ChatGPT-4o: Providing a customized knowledge base and utilizing online search options tends to improve responses for objective questions. However, hallucination remains a significant concern with this configuration. Maintaining the knowledge base while disabling online search can reduce hallucination occurrences.

Claude: This model does not offer real-time search capabilities. The responses are generally more "faithful" to the provided knowledge base. However, Claude's allowed knowledge base size is smaller than ChatGPT's capacity. This limitation makes it challenging to build domain-specific models on Claude when the knowledge base exceeds the textual data volume of an R2 university.

Gemini: This model does not support real-time searches. It typically excels at providing general answers and considering personal needs. However, the current version does not support structured input formats such as Excel files as part of the knowledge base.

Copilot Studio: This platform currently shows weakness in both correctness and comprehensiveness as of the first quarter of 2025. It is particularly weak in answering subjective questions related to specific persona needs. However, the developer's ambitious approach is evident. We remain hopeful for improved performance in future versions.

Llama: Both open-source models do not support real-time searches. The responses are stronger in relating back to the persona, but hallucination is of significant concern, especially with the 70B model. Despite the model being considered "open source," the limitations of licensing and regulations made the model difficult to use locally. The models performed relatively well overall and performed adequately on the subjective scenarios.

DeepSeek: The new model has generated significant buzz since its release, often compared

to ChatGPT and its ability to compete. Although we tested a small 7B version, DeepSeek scored third place in the objective questions. Similarly to ChatGPT, it struggled with hallucination to a moderate extent. However, as the smallest model tested in this study, DeepSeek performed especially well in comprehensiveness. There is strong potential in this model as it continues to grow.

Limitations

The knowledge base we have created for this paper is for one single university, and yet this knowledge base cannot be processed by some commercialized LLM models (such as Claude). For even large knowledge bases, our approach will not be applicable.

This study aimed to create an assessment framework and evaluate current state-of-the-art LLM models. However, the rapid pace of innovation in this field will inevitably limit the generalizability of our results and observations. Our assessment framework has longer-lasting impact compared to our specific findings. The experiment was conducted between May 2024 and March 2025. The new OpenAI o1 model was not included in the scope of this project.

Institutional or corporate regulations may prevent knowledge bases from being uploaded or shared online in certain cases. The technical approaches discussed in this paper will not be applicable under these circumstances. For researchers and working professionals facing such constraints, the optimal solution involves creating an in-house LLM server. This approach would utilize open-source LLM models such as Llama or DeepSeek.

Future work

For future work, instead of general daily tasks (like a college-related information chatbot in this paper), the authors are developing similar testing frameworks on more domain-specific chatbots with an unstructured or semi-structured knowledge base, like network system risk analysis.

7. REFERENCES

- AI, et al. (2024). Yi: Open foundation models by 01.AI.
- Banerjee, S., & Lavie, A. (2005). An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (pp. 65–72).
- Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*, 20(2), 369–382. DOI: <https://doi.org/10.1039/C8RP00262B>.
- Chang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45. DOI: <https://doi.org/10.1145/3641289>.
- Firat, M. (2023). How Chat GPT can transform autodidactic experiences and open education? OSF. DOI: <https://doi.org/10.31219/osf.io/9ge8m>.
- Hwang, K., Challagundla, S., Alomair, M., Chen, L. K., & Choa, F. S. (2023). Towards AI-assisted multiple choice question generation and quality evaluation at scale: Aligning with Bloom's taxonomy. In Workshop on Generative AI for Education.
- Islam, R., & Ahmed, I. (2024). Gemini—the most powerful LLM: Myth or truth. In 2024 5th Information Communication Technologies Conference (ICTC) (pp. 303–308). IEEE. DOI:10.36227/techrxiv.171177477.70151414/v1.
- Jordão, R. V. D., & Novas, J. C. (2024). Information and knowledge management, intellectual capital, and sustainable growth in networked small and medium enterprises. *Journal of the Knowledge Economy*, 15(1), 563–595. <https://doi.org/10.1007/s13132-022-01043-5>
- Liu, X., et al. (2024). Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: Comparative performance analysis. *JMIR Medical Informatics*, 12, e59273. DOI: 10.2196/59273.
- Mehri, S., & Eskenazi, M. (2020). Unsupervised evaluation of interactive dialog with DialoGPT. DOI: 10.18653/v1/2020.sigdial-1.28.
- Mendonça, J., Trancoso, I., & Lavie, A. (2024). Soda-Eval: Open-domain dialogue evaluation in the age of LLMs. DOI: 10.18653/v1/2024.findings-emnlp.684.
- Oppy, G., & Dowe, D. (2021). The Turing test. *Stanford Encyclopedia of Philosophy*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). DOI: 10.3115/1073083.1073135.

Rydzewski, N. R., et al. (2024). Comparative evaluation of LLMs in clinical oncology. NEJM AI, 1(5). <https://doi.org/10.1056/AIoa2300151>

Song, H., et al. (2023). Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. Journal of Medical Systems, 47(1), 125. <https://doi.org/10.1007/s10916-023-02021-3>.

Zhang, Q., et al. (2024). A critical review of large language model on software engineering: An example from ChatGPT and automated program repair. DOI:10.48550/arXiv.2310.08879.

APPENDIX A
LLMs Studied in the Project

| Chatbot | Domain Based Chatbot Function with Knowledge Base Support | Cost | Search Online in Real-Time? | Additional Constraints |
|----------------|--|-------------------------|------------------------------------|--|
| ChatGPT | My GPTs | \$20/month | yes | Up to 20 files as the knowledge base |
| Gemini | Vertex AI | Free (Gemini 1.5 flash) | no | All the input knowledge base should be in one category: (1) PDF or doc, (2) JSON, or (3) CSV; 10 files, 1-million-line window size |
| Claude | Claude Pro projects | \$20/month | no | Input knowledge base total line limit (20,000-40,000) because of Claude's maximum context window size |
| Copilot | Copilot studio | \$200/month | Yes (but only from 4 links) | Can only support four URLs, two Dataverses (up to 15 tables), supports file uploads |
| Llama | LM studio | free | no | Supports file-based knowledge base |
| DeepSeek | LM studio | free | no | Supports file-based knowledge base |

Table 4: LLMs Studied in the Project

APPENDIX B
Questions for Chatbots

| Objective Questions | |
|----------------------------|---|
| Q1 | What is the size of the school? |
| Q2 | What is the location of the school, rural or urban? |
| Q3 | What is the student-to-faculty ratio, and how large are the class sizes? |
| Q4 | What is the tuition cost, and what financial aid options are available? |
| Q5 | Does the school offer a computer science major? |
| Q6 | What is the college's retention rate, and how many students complete their degrees? |
| Q7 | How many students are there? |
| Q8 | What is the university's overall ranking among national universities? |
| Q9 | How many residence halls or dorms does the school have? |
| Q10 | How many dining halls or cafeterias are available for students on campus? |
| Q11 | What percentage of incoming freshmen receive scholarships? |
| Q12 | What is the reported campus crime rate per year by the school's police department? |
| Q13 | What is the average starting salary for students who graduate with a Bachelor's degree? |
| Q14 | How many student organizations and clubs does the school have? |
| Q15 | How to get access to the college's wifi? |
| Q16 | What is the most recent year's acceptance rate for incoming freshmen? |
| Q17 | What is the volume of the university's library collection? |
| Q18 | What is the proportion of Asian students in the total number of students? |
| Q19 | What are the school's wireless network coverage and stability metrics? |
| Q20 | How many free or low-cost campus transportation options does the school offer? |

| | |
|-----------------------------|---|
| Q21 | Which majors of this school are better known? |
| Q22 | Does the school offer any job connections with any company? [you may not find a good answer] |
| Subjective Questions | |
| Q1 | What is my intended major, and does the college offer a strong program for it? |
| Q2 | Considering my preference of the campus size, does this university's campus size a good fit for me? |
| Q3 | Will I enjoy the location of this university considering my personal preferences? |
| Q4 | Are there any on-campus housing options that will fit my lifestyle and preferences? |
| Q5 | What on-campus clubs or activities can I join during my college years? |
| Q6 | Would the tuition be too expensive for me considering the tuition range I am okay with? And considering my academic level, if there are possible scholarships for me from this university? |
| Q7 | Could you compare my intended major at this university with other university at similar level? |
| Q8 | How is the safety of the school compared to the average of the universities in the U.S.? |
| Q9 | What food options that aligns to my state does the on-campus dining offer? |
| Q10 | Is there any famous scenery/place to visit outside campus that are aligned to my personal preference? |
| Q11 | What kind of sports am I likely to watch on campus at this university? |
| Q12 | If I am taking a full student loan and using a 10-year payment plan, could you give me an estimate about how much I'm paying after college? And is this too high for the average salary of the students who graduate with my major? |
| Q13 | Is it easy to rent an off-campus apartment near the school that is aligned to my personal need, and do I need to buy a car if I live off campus? |

Table 5: Questions for Chatbots

APPENDIX C
Performance of Chatbots

| Question | GPT-4o | GPT4o+ KB | GPT4o+ KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
|------------------|-------------|-------------|-------------|-------------|------------------|----------------|-------------|-------------|-------------|
| Q1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 0 | 2 |
| Q2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 2 | 2 |
| Q3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 |
| Q4 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| Q5 | 2 | 2 | 2 | 2 | n/a | 2 | 2 | 1 | 2 |
| Q6 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q8 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q9 | 0 | 1 | 2 | 2 | 1 | 0 | 1 | 1 | 0 |
| Q10 | 2 | 2 | 1 | 2 | 2 | 0 | 1 | 0 | 1 |
| Q11 | 2 | 2 | 2 | 2 | 0 | 0 | 1 | 1 | 1 |
| Q12 | 1 | 1 | 2 | 1 | 0 | 1 | 2 | 1 | 1 |
| Q13 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 0 |
| Q14 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q15 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| Q16 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q17 | 2 | 2 | 2 | 2 | 0 | 2 | 0 | 2 | 1 |
| Q18 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| Q19 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q20 | 2 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 |
| Q21 | 2 | 1 | 2 | 1 | 0 | 2 | 2 | 1 | 2 |
| AVG. | 1.67 | 1.67 | 1.90 | 1.90 | 1.50 | 1.33 | 1.48 | 1.38 | 1.52 |
| Std. Dev. | 0.58 | 0.58 | 0.30 | 0.30 | 0.83 | 0.86 | 0.68 | 0.67 | 0.68 |
| CV. | 0.35 | 0.35 | 0.16 | 0.16 | 0.55 | 0.64 | 0.46 | 0.48 | 0.45 |

Table 6: Correctness of Chatbots on Objective Questions

| Questions | GPT-4o | GPT4o+ KB | GPT4o+ KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
|-----------|--------|-----------|-------------|--------|------------------|----------------|-----------|-----------|-------------|
| Q1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 0 | 2 |
| Q2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 2 |
| Q3 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 |
| Q4 | 0 | 0 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q5 | 2 | 2 | 2 | 1 | n/a | 2 | 2 | 2 | 2 |
| Q6 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 |
| Q7 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 |
| Q8 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q9 | 0 | 1 | 2 | 2 | 1 | 0 | 1 | 1 | 1 |
| Q10 | 1 | 2 | 1 | 2 | 2 | 0 | 1 | 0 | 1 |
| Q11 | 2 | 2 | 2 | 1 | 0 | 0 | 2 | 2 | 2 |
| Q12 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| Q13 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| Q14 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 |
| Q15 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 |
| Q16 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q17 | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 2 | 2 |
| Q18 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 |
| Q19 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q20 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |

| | | | | | | | | | |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Q21 | 1 | 1 | 1 | 0 | 0 | 1 | 2 | 1 | 1 |
| AVG. | 1.52 | 1.48 | 1.86 | 1.52 | 1.30 | 1.33 | 1.67 | 1.57 | 1.76 |
| Std. Dev. | 0.68 | 0.68 | 0.36 | 0.60 | 0.73 | 0.73 | 0.48 | 0.68 | 0.44 |
| CV. | 0.45 | 0.46 | 0.19 | 0.39 | 0.56 | 0.55 | 0.29 | 0.43 | 0.25 |

Table 7: Comprehensiveness of Chatbots on Objective Questions

| Questions | Correctness | | | | | | | | Comprehensiveness | | | | | | | |
|------------------|-------------|-------------|-------------|------------------|----------------|-------------|-------------|-------------|-------------------|-------------|-------------|------------------|----------------|-------------|-------------|-------------|
| | GPT4o+KB | GPT4o+KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B | GPT4o+KB | GPT4o+KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
| Q1 | 2 | 2 | 2 | n/a | 0 | 0 | 0 | 1 | 2 | 2 | 1 | n/a | 0 | 1 | 1 | 1 |
| Q2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 0 | 2 |
| Q3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q4 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q5 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 |
| Q6 | 0 | 0 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 2 | 1 | 1 | 2 | 0 | 2 |
| Q7 | 0 | 0 | 2 | N/A | 0 | 0 | 0 | 0 | 1 | 1 | 2 | N/A | 1 | 1 | 1 | 1 |
| Q8 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q9 | 1 | 2 | 0 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q10 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 |
| Q11 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| Q12 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q13 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 0 | 2 | 2 |
| Avg. | 1.38 | 1.54 | 1.85 | 1.82 | 1.23 | 1.23 | 1.15 | 1.54 | 1.46 | 1.69 | 1.69 | 1.45 | 1.15 | 1.54 | 1.46 | 1.77 |
| Std. Dev. | 0.77 | 0.78 | 0.55 | 0.40 | 0.73 | 0.73 | 0.69 | 0.66 | 0.66 | 0.48 | 0.48 | 0.52 | 0.55 | 0.66 | 0.78 | 0.44 |
| CV. | 0.55 | 0.50 | 0.30 | 0.22 | 0.59 | 0.59 | 0.60 | 0.43 | 0.45 | 0.28 | 0.28 | 0.36 | 0.48 | 0.43 | 0.53 | 0.25 |

Table 8: Performance of Chatbots on Subjective Questions (First Persona)

| Questions | Correctness | | | | | | | | Comprehensiveness | | | | | | | |
|-----------|-------------|------------|--------|------------------|----------------|-----------|-----------|-------------|-------------------|------------|--------|------------------|----------------|-----------|-----------|-------------|
| | GPT4o+KB | GPT4o+KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B | GPT4o+KB | GPT4o+KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
| Q1 | 2 | 2 | 2 | n/a | 2 | 2 | 2 | 1 | 2 | 2 | 2 | n/a | 1 | 2 | 2 | 2 |
| Q2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| Q3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Q4 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Q5 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 2 |
| Q6 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| Q7 | 1 | 2 | 1 | n/a | 1 | 1 | 1 | 1 | 0 | 2 | 2 | n/a | n/a | 1 | 2 | 1 |
| Q8 | 1 | 2 | 2 | 1 | n/a | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 |
| Q9 | 2 | 2 | 2 | 1 | 2 | n/a | 1 | 2 | 2 | 2 | 2 | 2 | 2 | n/a | 0 | 2 |

| | | | | | | | | | | | | | | | | |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Q10 | 2 | 2 | 2 | 2 | n/a | 2 | 1 | 1 | 2 | 1 | 2 | 2 | n/a | 2 | 0 | 2 |
| Q11 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 2 |
| Q12 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q13 | 2 | 2 | 2 | n/a | 1 | 2 | 2 | 2 | 2 | 2 | 1 | n/a | 1 | 2 | 2 | 1 |
| Avg. | 1.69 | 2 | 1.69 | 1.5 | 1.27 | 1.75 | 1.62 | 1.69 | 1.54 | 1.77 | 1.77 | 1.67 | 1.09 | 1.83 | 1.54 | 1.85 |
| Std. Dev. | 0.48 | 0.00 | 0.48 | 0.53 | 0.79 | 0.45 | 0.51 | 0.48 | 0.66 | 0.44 | 0.44 | 0.48 | 0.70 | 0.39 | 0.78 | 0.38 |
| CV. | 0.28 | 0.00 | 0.28 | 0.35 | 0.62 | 0.26 | 0.31 | 0.28 | 0.43 | 0.25 | 0.25 | 0.28 | 0.64 | 0.21 | 0.50 | 0.20 |

Table 9: Performance of Chatbots on Subjective Questions (Second Persona)

| Questions | Correctness | | | | | | | | Comprehensiveness | | | | | | | |
|------------------|-------------|-------------|-------------|------------------|----------------|-------------|-------------|-------------|-------------------|-------------|-------------|------------------|----------------|-------------|-------------|-------------|
| | GPT4o+KB | GPT4o+KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B | GPT4o+KB | GPT4o+KB+S | Claude | Gemini-1.5-flash | Copilot Studio | Llama 13B | Llama 70B | DeepSeek 7B |
| Q1 | 1 | 1 | 2 | n/a | 0 | 0 | 0 | 0 | 2 | 2 | 2 | n/a | 1 | 1 | 1 | 0 |
| Q2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 2 | 2 |
| Q3 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 0 | 1 | 2 | 2 |
| Q4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 |
| Q5 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| Q6 | 0 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 2 |
| Q7 | 2 | 0 | 1 | n/a | 0 | 0 | 0 | 0 | 2 | 1 | 2 | n/a | 0 | 1 | 1 | 1 |
| Q8 | 2 | 2 | 1 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |
| Q9 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| Q10 | 1 | 2 | 1 | 2 | 0 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 |
| Q11 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 |
| Q12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| Q13 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 2 |
| Avg. | 1.54 | 1.46 | 1.54 | 1.72 | 1 | 1.15 | 1.15 | 1.30 | 1.85 | 1.62 | 1.62 | 1.73 | 0.85 | 1.38 | 1.62 | 1.54 |
| Std. Dev. | 0.66 | 0.66 | 0.52 | 0.47 | 0.82 | 0.69 | 0.69 | 0.95 | 0.38 | 0.51 | 0.51 | 0.47 | 0.55 | 0.65 | 0.65 | 0.66 |
| CV. | 0.43 | 0.45 | 0.34 | 0.27 | 0.82 | 0.60 | 0.60 | 0.72 | 0.20 | 0.31 | 0.31 | 0.27 | 0.66 | 0.47 | 0.40 | 0.43 |

Table 10: Performance of Chatbots on Subjective Questions (Third Persona)

Examining Artificial Intelligence Job Descriptions: An Analysis of Current Job Postings

Alan Peslak
arp14@psu.edu
Penn State University
University Park, PA 16802

Wendy Ceccucci
Wendy.Ceccucci@qu.edu
Quinnipiac University
Hamden, CT 06518

Kiku Jones
Kiku.Jones@qu.edu
Quinnipiac University
Hamden, CT 06518

Margaret Wanjiku Gachau
margaret.gachau@qu.edu
Quinnipiac University
Hamden, CT 06518

Abstract

Artificial intelligence continues to transform the workforce. Educators, job seekers, and policymakers need to understand the changing needs of AI-related job roles. This study examines 435 AI job postings on PostJobsFree.com utilizing text mining techniques that include word frequency analysis, TF-IDF, Named Entity Recognition (NER), and Linguistic Inquiry and Word Count (LIWC). The analysis shows that the job posts are requesting a well-rounded skill set that includes both technical skills - such as Data Science and Natural Language Processing - with durable skills such as teamwork, communication, problem solving, and interpersonal skills. The most used words are AI, Data, and Team. TF-IDF displays more domain-specific words, such as Generative AI and AI Models. NER indicates the importance of tools (such as Python and AWS) and security clearances for DOD personnel. LIWC shows that the descriptions are extremely analytical and achievement oriented. These findings suggest that companies continue to value core technical proficiency while also looking for applicants who can work in complex, interdisciplinary environments. This study adds to the literature on AI labor market trends and provides insights into the AI industry's dynamic demands. In addition, this study can be used by both job seekers aiming for AI positions and employers seeking professionals in the AI field. The results can assist job seekers in identifying the necessary technical skills they need to develop to obtain specific AI jobs. Employers can use the result to have a better understanding of the essential skill sets to look for in candidates applying for AI related positions.

Keywords: AI jobs, Job applications, text mining, job market, AI, Artificial Intelligence

Recommended Citation: Peslak, A., Ceccucci, W., Jones, K., Gachau, M.W., (2026). Examining Artificial Intelligence Job Descriptions: An Analysis of Current Job Postings. *Journal of Information Systems Applied Research and Analytics* v19(n2) pp 25-35. DOI# <https://doi.org/10.62273/CCJX2208>

Examining Artificial Intelligence Job Descriptions: An Analysis of Current Job Postings

Alan Peslak, Wendy Ceccucci, Kiku Jones and Margaret Wanjiku Gachau

1. INTRODUCTION

Artificial intelligence (AI) has become a transformative force across industries, reshaping workflows, decision-making processes, and organizational priorities. As companies increasingly seek to integrate AI technologies into their operations, the demand for skilled professionals in this field has surged. This demand is reflected in the job postings across various online platforms, where employers highlight the skills, qualifications, and roles required for AI-related positions.

PostJobsFree.com, a widely used platform for job advertisements, provides a unique opportunity to analyze trends in AI job postings. By examining these listings, it is possible to gain insights into the evolving requirements of the AI workforce, including the technical proficiencies and domain expertise that employers prioritize.

This paper aims to systematically review AI job postings on PostJobsFree.com to identify key trends and patterns in the field. The study examines job titles, required qualifications, preferred skill sets, and job descriptions to uncover the nuances of the AI labor market. Through this analysis, the paper seeks to provide actionable insights for educators, policymakers, and job seekers, enabling them to align their strategies with the dynamic demands of the AI industry.

2. LITERATURE REVIEW

Job Market Trends

The AI job market is experiencing unprecedented growth, with demand surging across multiple sectors including technology, healthcare, finance, and manufacturing. Recent data from UMD-LinkUp (2025) and the University of Maryland reveals a remarkable 68 percent increase in AI job postings in the United States, rising from 29,509 in 2022 to 49,577 by the end of 2024.

The World Economic Forum projects a significant reshaping of the global job landscape by 2030. Technological advances are expected to create

170 million new jobs while displacing 92 million, resulting in a net gain of 78 million jobs. This transformation underscores the critical need for comprehensive workforce training and education reforms to prepare workers for emerging roles.

AI is fundamentally reshaping the workplace, presenting a complex landscape of both opportunities and challenges. Technology has the potential to simultaneously automate routine tasks and create new roles requiring specialized AI expertise.

Generative AI offers promising opportunities for workforce empowerment. It could enable less specialized employees to perform more complex tasks, expanding the functional capabilities of roles such as accounting clerks, nurses, and teaching assistants. For skilled professionals like electricians, doctors, and engineers, AI can provide access to cutting-edge knowledge, dramatically improving problem-solving efficiency.

Researchers like Shen (2024) emphasize that the most beneficial technological developments will focus on enhancing human capabilities rather than simply substituting human labor.

The integration of AI into the workforce is creating a polarized employment landscape. While generating numerous new jobs, AI is also causing significant disruption, particularly in low-cost, labor-intensive sectors.

Key observations include:

- Increasing employment thresholds
- Widening income disparities
- Higher compensation for AI managers and experts compared to manual workers
- Automation technologies replacing repetitive and basic skilled roles

As noted by Shen and Zhang (2024), these technological shifts are fundamentally altering traditional labor market dynamics, presenting both challenges and opportunities for workers across various industries.

In the healthcare sector, AI has revolutionized diagnostics, drug discovery, personalized medicine through technologies like image recognition for disease detection, and the development of AI-powered prosthetics (Kolakowski, 2025).

In Finance, AI has transformed financial services through applications such as fraud detection, algorithmic trading, and personalized financial advice (Shen, 2024). AI is now being used in traditional financial activities such as customer collection, identification, risk management, investment advisory and customer service.

In Manufacturing, AI-powered robotics and automation are streamlining production processes, improving efficiency, and enhancing product quality (Shen, 2024).

From the perspective of the application of artificial intelligence technology, there will be many jobs set around AI in the future, which will produce more new occupations and new opportunities that are difficult to evaluate (Jepperson, 2025). The proportion of blue-collar workers will gradually decrease while the proportion of white-collar workers in the new knowledge economy will expand. Comunale and Manera (2024) conclude that occupations with high task exposure will face higher displacement, and exposed workers will see reduced employment opportunities.

Asim and Ding (2025) used Structural Equation Modeling with 500 professionals from various industry sectors, and found that artificial intelligence adoption creates measurable improvements in both workforce engagement and job performance outcomes. Additionally, their research identified job complexity levels and individual AI expertise as powerful moderating variables that can either strengthen or diminish these positive effects, indicating that both personal capabilities and work environment characteristics play decisive roles in determining how effectively organizations can harness artificial intelligence benefits.

Han, Chen Wand, and Xu (2025) conducted an online survey of 519 Chinese employees with AI experience. Using regression analysis they found that the use of AI positively enhanced innovative workplace behaviors through its ability to enable job crafting activities, with this effect being more pronounced among employees who possessed greater confidence in their creative abilities. The study also found that workplace environments emphasizing employee strengths served as a

positive moderating factor, amplifying the connection between job crafting practices and innovative behavioral outcomes.

The evolving technological landscape presents significant challenges for workers with limited skill sets. Professionals who possess only basic operational skills face heightened risks of:

- Potential unemployment
- Reduced income opportunities
- Displacement by intelligent engineering systems

As intelligent technologies systematically replace traditional employment roles, enterprises' human resource demands are expected to contract, creating substantial pressure on the job market. This transformation underscores the critical importance of continuous learning, skill development, and technological adaptability.

Skills and Qualifications in AI Job Postings

Generative AI technologies like ChatGPT are fundamentally reshaping creative industries, catalyzing the emergence of innovative professional roles that demand sophisticated technological understanding. This technological shift is creating unprecedented opportunities for professionals who can strategically leverage AI for content creation, design, and complex problem-solving across diverse sectors.

Employers are increasingly seeking hybrid talent with a skill set that combines:

- Technical capabilities such as data manipulation and training data development
- Strategic competencies including advanced problem-solving and strategic decision-making

According to research by Jepperson (2025), the AI revolution is driving substantial demand for professionals with expertise in:

1. Data analytics
2. STEM disciplines
3. Information technology

Organizations integrating AI are rapidly transitioning toward more educated and specialized workforces, with a particular emphasis on technical domains. Hazan, Madgavkar, Chui, Smit, Maor, Dandona and Huyghues-Despointes (2024) further highlight this trend, noting a significant increase in job postings requiring specialized skills like:

- Robotics engineering
- Advanced data analysis
- Complex system integration

The AI professional landscape requires a sophisticated and multifaceted skill set that exceeds the traditional technological boundaries. Technical competencies form the foundational framework for success in this rapidly evolving field, with programming languages like Python and R serving as critical tools for innovation.

Machine learning frameworks such as TensorFlow and PyTorch represent essential platforms through which professionals can develop and implement advanced computational solutions.

Data engineering knowledge has become increasingly important, reflecting the growing complexity of AI systems. Professionals must demonstrate advanced capabilities in data manipulation, analysis, and interpretation. This requires not only technical proficiency but also the ability to extract meaningful insights from intricate datasets using advanced visualization tools like Power BI and Tableau.

Cloud computing platforms including Azure and AWS have emerged as crucial infrastructure for AI development, necessitating strong familiarity and strategic implementation skills. These platforms enable scalable, flexible technological solutions that are fundamental to modern AI research and application.

Beyond technical skills, employers are placing increased emphasis on complementary professional attributes. The World Economic Forum's (2025) research highlights the critical importance of soft skills such as strategic problem-solving, advanced critical thinking, professional communication, and effective collaboration. These capabilities enable AI professionals to translate complex technological concepts into actionable business strategies.

Educational pathways for AI professionals have become increasingly specialized. Typical credentials include bachelor's degrees in computer science, master's degrees in AI-related disciplines, and doctoral research focusing on advanced technological applications. Emerging specialized roles like Generative AI Engineer, Computer Vision Engineer, and Remote AI Training Specialist underscore the field's growing complexity and segmentation.

Researchers like Shen (2024) emphasize that success in the AI ecosystem requires continuous learning and adaptive professional development. Professionals must consistently update their knowledge frameworks, cultivate innovative thinking, and maintain a dynamic approach to technological engagement. This involves not only mastering existing technologies but also developing the capacity to anticipate and integrate emerging computational paradigms.

The rapidly transforming labor market demands that AI professionals remain agile, proactively updating their skills to match evolving technological demands. This requires a holistic approach that balances deep technical expertise with broader strategic understanding, positioning professionals to navigate the complex and dynamic landscape of artificial intelligence.

3. THEORETICAL FRAMEWORKS

Labor market analysis theories and computational linguistics frameworks provide the foundation for this study. The application of text mining techniques in employment research is an emerging area of interest.

Text Analysis in Job Postings & Theoretical Framework

Prior research has used NLP techniques to analyze job postings for various fields. Text mining, sentiment analysis, and keyword extraction have been employed to understand employment trends and required skills.

Based on our goal of understanding the current AI job requirements, we have developed the following research questions.

Research Questions

1. What are the most frequently occurring words and phrases in AI job postings?
2. What insights can Term Frequency-Inverse Document Frequent analysis provide about distinguishing terms in AI job descriptions?
3. What key entities (e.g., organizations, technologies, skills) appear most frequently in AI job postings?
4. What psychological, emotional, and linguistic attributes are present in AI job postings as analyzed by LIWC?

4. METHODOLOGY

Data Collection

AI job postings were collected from PostJobfree.com. This source was used since it does not have restrictions on web scraping whereas other job boards such as Indeed and Ziprecruiter have specific prohibitions against web scraping in their terms of service. Tools used for data analysis included Wordstat 8, LIWC, and Voyant Tools.

Data Preprocessing

Text preprocessing steps included removing stop words, special characters, and irrelevant data. Tokenization and lemmatization were applied to standardize the text.

Analytical Methods

Five different methods were used to analyze the job descriptions.

1. Word Clouds

Word clouds are visual representations of frequently occurring words. In a word cloud, the size of each word corresponds to its frequency: the more often a word appears, the larger it is displayed. This visualization makes it easy to identify the most commonly used words.

2. Word and Phrase Frequency

Analyzing the frequency of the most common words and two- or three-word phrases helps to identify which terms and expressions appear most frequently in job descriptions. This method provides a clear understanding of the prevalent skills and requirements requested by employers.

3. TF-IDF (Term Frequency-Inverse Document Frequency)

The term frequency (TF) measures how often a word appears in a document. The Inverse Document Frequency (IDF) measures how rare or less common a word is. It is a term frequency measure which gives a larger weight to terms which are less common in the document set (Nguyen, 2014). This method "allows for common words that appear in many documents (like 'the' or 'and') to be assigned less weight, while more distinctive words receive higher weights, making TF-IDF a useful tool for identifying keywords and important themes within documents." (Fan & Qin, 2018). The result is a determination of the most distinctive words in AI job descriptions compared to other job postings.

4. Named Entity Recognition (NER)

"Named entity recognition is an important basic tool for information extraction, question

answering system, syntactic analysis, machine translation and other application fields, and plays an important role in natural language processing technology" (Chang, Zhang, Lv, Zhou, & Bai, 2022). As a result it has become an integral tool in text analytics research.

5. LIWC (Linguistic Inquiry and Word Count) Analysis

LIWC is a tool used by many researchers for understanding sentiment and many other linguistic characteristics of documents (Bahgat, Wilson, & Magdy, 2022).

It is used to assess and measure psychological, emotional, and linguistic dimensions, including tone, formality, and cognitive processes in AI job descriptions. LIWC is the software tool developed by Pennebaker.

5. RESULTS

1. Word Cloud Analysis

Figure 1 illustrates the 100 most frequently occurring individual words in the job descriptions through a word cloud. The prominence of AI and Data is evident, as they are the most commonly used terms. Additionally, the word cloud highlights other frequently mentioned words such as Team, Development, Learning, Business, and Support.

Figure 2 the most common phrases found in job descriptions through a word cloud. Artificial Intelligence and Machine Learning stand out as the most frequently used phrases, appearing significantly larger than the others. The next most prominent phrases include AI ML, and Computer Science which are noticeably smaller in comparison.



Figure 1: Top 100 Individual Words



Figure 2: Top 50 Phrases

2. Word and Phrase Frequency Analysis

As anticipated, AI is the most frequently occurring word, appearing 2094 times. Following closely is Data, with 1438 instances, and Team, with 982 instances. Table 1 lists the most common words that appear at least 500 times. The complete table of 100 individual words and their frequencies can be found in Appendix A. The phrases Artificial Intelligence and Machine Learning are the most frequently used, with 599 and 515 instances respectively. The third most common phrase is AI ML, which appears 265 times, notably much less often. The full table of most common phrases and their frequencies is available in Appendix B. Note that common EEO phrases were excluded in this table.

| Individual Words | Frequency |
|------------------|-----------|
| AI | 2094 |
| Data | 1438 |
| Team | 982 |
| Development | 838 |
| Learning | 815 |
| Support | 671 |
| Business | 641 |
| Ability | 620 |
| Technology | 611 |
| Systems | 605 |
| Information | 587 |
| Including | 582 |
| Years | 581 |
| Solutions | 578 |
| Research | 578 |
| Job | 573 |
| Technical | 572 |
| Machine | 548 |
| Position | 546 |
| Technologies | 513 |
| Requirements | 507 |
| Science | 504 |
| Environment | 502 |

Table 1: Frequency of Individual Words With at Least 500 Instances

3. TF-IDF Results

The TF-IDF analysis (Figure 3) surfaces domain-specific terms that are distinctively emphasized in AI job postings. High scores for "Generative AI" (139.5) and "AI ML" (201) underscore the increased demand for cutting-edge skills following the 2022–2023 surge in public awareness and use of tools like ChatGPT. This finding directly supports the narrative of “dynamic demands” within the AI industry, reflecting how quickly generative technologies have moved from research into practice.

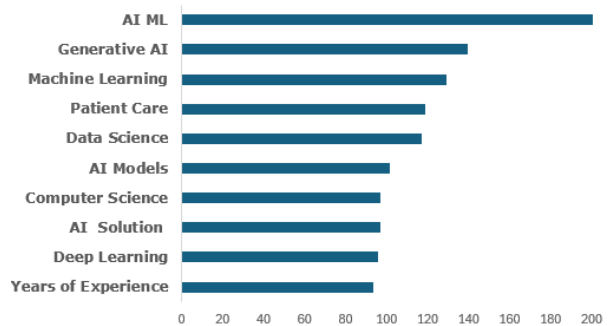


Figure 3. Top AI Job Description Terms by TF-IDF Score

Additionally, the presence of terms like "Patient Care" (119) and "Data Science" (117) suggests a significant overlap between AI and healthcare, as well as data-centric applications. Technical terms such as "AI Models", "Deep Learning", and "Computer Science" further emphasize the skillsets and knowledge areas critical to the field. Overall, TF-IDF scoring is effective for identifying terms that are both highly frequent in a specific context and relatively rare elsewhere, making them strong indicators of domain-specific language—in this case, uniquely associated with the AI workforce.

4. Named Entity Recognition (NER) Findings

NER analysis (Figure 4) reinforces these trends by highlighting specific technologies and organizational affiliations. Python, TensorFlow, and PyTorch feature prominently—common across machine learning stacks—while Azure and AWS suggest a strong orientation toward cloud-native AI deployment. This supports the view that AI job roles now demand “full-stack” proficiency: from programming and model training to cloud integration and deployment.

Furthermore, the frequent mention of "DoD", "SCI", and "Security Clearance" reflects the unique demand for AI professionals in defense and government sectors, particularly for roles

requiring advanced clearance. This implies that national security applications of AI are a growing employment segment, which carries implications for both technical preparedness and citizenship eligibility.

Together, the TF-IDF and NER results confirm that AI hiring reflects a dynamic and integrated skills economy—one in which deep technical fluency must coexist with security compliance, interdisciplinary understanding, and deployment readiness.

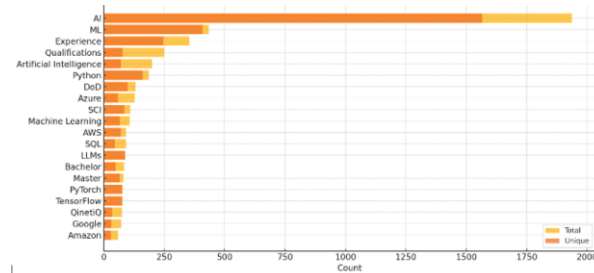


Figure 4. Top 20 Entities by Total and Unique Mentions

5. LIWC Analysis

The composite LIWC analysis of 436 AI job descriptions reveals a communication style that is highly structured, confident, and performance-driven. The **Analytic** score is exceptionally high (mean = 93.44), indicating logically organized, formal text, while the **Clout** score (75.49) reflects a commanding and authoritative voice. Despite this assertiveness, the **Authentic** score is low (19.28), suggesting that the language is impersonal and detached—typical of standardized, professional postings. A high **Tone** score (84.23) signals a generally positive emotional valence, likely used to attract applicants while maintaining a formal tone. Linguistically, the descriptions are complex, featuring high usage of six-letter words (40.82%) and elevated words-per-sentence (26.98), reflecting a sophisticated and technical vocabulary. The psychological dimension of these postings emphasizes cognition and achievement: scores are high for **Cognitive Processes** (9.94), especially **Insight** (3.81) and **Cause** (2.55), and for **Achievement** (5.51) and **Work** (13.14), underscoring the intellectual and goal-oriented nature of AI roles. In contrast, **social** references (7.16) and **personal pronouns**—especially first-person “I” (0.03)—are used sparingly, indicating a depersonalized, objective style with limited interpersonal or community language. Emotional expressiveness is tightly restrained, with minimal **negative emotion** (0.56), **anxiety** (0.13), or **anger** (0.09), reinforcing a composed and positive

narrative. Temporal focus centers on the **present** (5.10), with relatively little emphasis on **past** (0.46) or **future** (1.07), further suggesting an emphasis on immediate performance rather than historical experience or long-term vision. Overall, AI job postings construct a professional persona that is analytical, upbeat, intellectually demanding, and achievement-focused, while downplaying emotional tone, personal expression, and social connectedness—strategically aligning with the technical rigor and high expectations of roles in artificial intelligence.

6. DISCUSSION

Interpretation of Findings

Analyzing the most frequently occurring words and phrases, along with their frequencies, reveals trends in the expected and required technical and durable skills for job applicants. While it is unsurprising to see terms like AI and Data frequently mentioned, other technical skill words such as Development (838 instances), Engineering (484 instances), Models (474 instances), Program (312 instances), and Cloud (245 instances) highlight the technical expertise employers are seeking. Additionally, common phrases provide further insight into the specific technical skills required: Computer Science (227 instances), Data Science (172 instances), Generative AI (162 instances), Deep Learning (124 instances), and Natural Language Processing (79 instances, with Natural Language appearing uniquely 14 times).

Beyond technical terms, other words like Team (982 instances, with Teams appearing 340 times), Business (641 instances), Research (578 instances), Management (485 instances), and Analysis (303 instances) paint a broader picture of the qualities expected from applicants. These words suggest that employers are looking for candidates with a well-rounded skill set. This is further supported by phrases such as Communication Skills (124 instances), Problem Solving (111 instances), Cross Functional (107 instances), Interpersonal Skills (83 instances), and Solving Complex Problems (81 instances).

Implications for Researchers

The linguistic analysis of AI job descriptions has several important implications for both researchers and practitioners. For researchers, the consistent use of highly analytical, impersonal, and achievement-focused language raises questions about how such communication styles might influence perceptions of AI roles and contribute to patterns in applicant self-

selection. The minimal use of personal pronouns, emotional expression, and social references suggests a depersonalized tone that could deter candidates who prioritize collaboration, inclusivity, or emotional intelligence. This opens opportunities for further research into the framing effects of job language on different demographic groups, and whether such language reinforces existing gender or diversity imbalances in the tech workforce. Researchers may also explore how linguistic markers like cognitive complexity and goal orientation reflect broader cultural norms in AI-related fields and how they compare to other domains.

Implications for Practitioners

These findings can serve as a valuable resource for both job seekers aiming for AI positions and employers seeking AI professionals. Job seekers can leverage these findings to gain a clearer understanding of the qualities and skills employers are currently prioritizing. With this insight, they can identify which technical skills they need to develop before beginning their job search. Additionally, recognizing that employers value more than just technical expertise will help job seekers highlight their durable skills, presenting themselves as well-rounded candidates.

For employers, reviewing these findings can provide a comprehensive understanding of the essential skills required for AI positions. This ensures they do not overlook any critical components of the skill set necessary for success in the role. These findings will help employers avoid the consequences of focusing solely on technical skills, enabling them to seek candidates with a balanced mix of technical and durable skills.

Additionally, these findings point to the need for more balanced job descriptions. While the current emphasis on confidence, technical skill, and achievement is aligned with performance expectations, it may unintentionally limit the appeal of these roles to a narrow segment of applicants. Incorporating more authentic, inclusive, and people-oriented language could broaden the talent pool and better reflect collaborative and human-centered workplace values. Additionally, re-evaluating how work intensity and expectations are framed can help ensure job ads don't inadvertently suggest high-pressure or unsupportive environments. Ultimately, aligning the tone and content of job descriptions with both organizational values and evolving workforce expectations can support more effective and inclusive recruitment in the AI field.

7. Limitations and Conclusions

Limitations

The major limitation of this study is the relatively small sample set. This was due to download restrictions from postjobsfree.com. Further analysis with larger datasets is suggested to confirm our findings.

Future Research Directions

Future researchers should aim to expand this analysis by collecting job descriptions from global job markets, including those in different languages. Expanding the scope of the analysis could allow for the integration of data to identify the most critical skills worldwide. Additionally, comparing job descriptions from various countries might reveal whether similar priorities are shared globally.

Incorporating sentiment analysis into future research could provide insights into the tone of job postings, offering a deeper understanding of how job descriptions are framed. This could help employers as they are creating their job descriptions to convey their intended message rather than risk misunderstandings in what they are putting in the descriptions.

Lastly, employing deep learning models for advanced classification of job descriptions would enhance the understanding of different AI position segments and the specific skill sets required for success. This knowledge would help job seekers identify the roles best suited to their skills and enable employers to craft job descriptions that more accurately reflect their candidate requirements.

Conclusion

This study provides key insights into AI job postings using text analysis techniques. The findings contribute to AI labor market research and computational linguistics, highlighting the evolving nature of AI job descriptions. Generative AI being highly weighted aligns with the post-ChatGPT surge in interest for that skillset and the prominence of cloud platforms and ML frameworks in entities confirms how AI roles now demand full-stack implementation abilities.

Disclaimer

The authors generated parts of this paper including in the introduction and literature review with assistance from GPT-4, OpenAI's large-scale language-generation model. Upon generating draft language, the author reviewed, edited, and revised the language to their own

liking and takes ultimate responsibility for the content of this publication (OpenAI Publication Policy, 2022)

8. REFERENCES

- Asim, M. & Ding, W. (2025). AI Usage, Employee Engagement, and Work Performance: Examining the Roles of Job Complexity and AI Knowledge *Proceedings of the 2025 4th International Conference on Big Data and Digital Management*, 191, 294 – 303. [pphttps://doi.org/10.2991/978-94-6463-710-6_34](https://doi.org/10.2991/978-94-6463-710-6_34)
- Bahgat, M., Wilson, S., & Magdy, W. (2022). LIWC-UD: classifying online slang terms into LIWC categories. 14th ACM Web Science Conference 2022, 422-432. <https://doi.org/10.1145/3501247.3531572>
- Chang, L., Zhang, R., Lv, J., Zhou, W., & Bai, Y. (2022). A review of biomedical named entity recognition. *Journal of Computational Methods in Sciences and Engineering*, 22(3), 893-900. <https://doi.org/10.3233/jcm-225952>
- Comunale, M. & Manera, A. (2024). The Economic Impacts and the Regulation of AI: A Review of the Academic Literature and Policy Actions. (n.p.): International Monetary Fund.
- Fan, H. & Qin, Y. (2018). Research on text classification based on improved TF-IDF algorithm. *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering* <https://doi.org/10.2991/ncce-18.2018.79>
- Han, X., Chen, F., Wang, H., & Xu, S. (2025). Unlocking innovation: Artificial intelligence usage and innovative behavior in the workplace. *Social Behavior and Personality: An international Journal* 53(3). <https://doi.org/10.224.sbp.13851>
- Hazan, E., Madgavkar, A., Chui, M., Smit, S., Maor, D., Dandona, G., & Huyghues-Despointes, R. (2024). A new future of work: The race to deploy AI and raise skills in Europe and beyond. McKinsey Global Institute. Retrieved March 27, 2025 from <https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond>.
- Jepperson, T. (2025). How Is Artificial Intelligence Impacting the Job Market? LINQTO. Retrieved March 27, 2025 from <https://www.linqto.com/blog/how-is-artificial-intelligence-impacting-the-job-market/>
- Kolakowski, N. (2025). AI Job Market 2025: Trends and Opportunities Across the US. Dice. Retrieved March 27, 2025 from <https://www.dice.com/career-advice/artificial-intelligence-a-i-job-market-small-but-growing-in-many-states>
- Nguyen, E. (2014). Chapter 4 – Text Mining and Network Analysis of Digital Libraries in R, Editors Zhao, Y. & Cen, Y., Data Mining Applications with R, Academic Press <https://doi.org/10.1016/B978-0-12-411511-8.00004-9>.
- Shen, Y. (2024). Artificial Intelligence and the Economy . The Impact of Artificial Intelligence on the Job Market, *Advances in Economics Management and Political Sciences*, 92(1) 71-74. <https://doi.org/10.54254/2754-1169/92/20231275>
- Shen, Y. & Zhang, X. (2024). The impact of artificial intelligence on employment: the role of virtual agglomeration. *Humanities Social Sciences Communications*, 11(122). <https://doi.org/10.1057/s41599-024-02647-9>
- UMD-LinkUp AIMaps (2025). Tracking where AI jobs are being created. University of Maryland. Retrieved March 27, 2025 from <https://www.aimaps.ai/>
- World Economic Forum. (2025). Future of Jobs Report 2025. Retrieved March 27, 2025 from https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf?_gl=1*1gzrggi*_up*MQ..*_gs*MQ..&clid=Cj0KCQjw-e6-BhDmARIsAOxxlxVEsX6hUGqzyeDxVx0-3zQdyIcGqsNwLW9c7e3T599mAJi2YRZfxFQaAoDpEALw_wcB

Appendix A: Top 100 Most Common Individual Words

| Individual Words | Frequency |
|------------------|-----------|
| AI | 2094 |
| Data | 1438 |
| Team | 982 |
| Development | 838 |
| Learning | 815 |
| Support | 671 |
| Business | 641 |
| Ability | 620 |
| Technology | 611 |
| Systems | 605 |
| Information | 587 |
| Including | 582 |
| Years | 581 |
| Solutions | 578 |
| Research | 578 |
| Job | 573 |
| Technical | 572 |
| Machine | 548 |
| Position | 546 |
| Technologies | 513 |
| Requirements | 507 |
| Science | 504 |
| Environment | 502 |
| Engineering | 494 |
| Provide | 490 |
| Required | 488 |
| Related | 486 |
| Management | 485 |
| Models | 474 |
| Design | 468 |
| Company | 467 |
| Knowledge | 453 |
| Role | 442 |

| Individual Words | Frequency |
|------------------|-----------|
| ML | 441 |
| Employment | 436 |
| Status | 422 |
| Software | 422 |
| Qualifications | 418 |
| Strong | 414 |
| Security | 409 |
| Opportunity | 407 |
| Time | 402 |
| Develop | 394 |
| Computer | 384 |
| World | 377 |
| Process | 376 |
| High | 371 |
| Training | 369 |
| Degree | 364 |
| Quality | 360 |
| Tools | 354 |
| Professional | 352 |
| Complex | 349 |
| Working | 344 |
| Teams | 340 |
| Opportunities | 335 |
| Services | 330 |
| Patients | 327 |
| Ensure | 325 |
| Disability | 324 |
| Based | 322 |
| National | 317 |
| Across | 314 |
| Applications | 314 |
| Program | 312 |
| Range | 310 |

| Individual Words | Frequency |
|------------------|-----------|
| Patient | 309 |
| Expertise | 308 |
| Benefits | 305 |
| Analysis | 303 |
| Applicants | 301 |
| Field | 291 |
| Performance | 283 |
| Full | 278 |
| Equal | 278 |
| Products | 278 |
| Industry | 277 |
| Application | 277 |
| Mission | 275 |
| Education | 270 |
| Product | 268 |
| Gender | 266 |
| Problems | 265 |
| Level | 264 |
| Success | 260 |
| Employees | 260 |
| Deliver | 255 |
| Include | 255 |
| Stakeholders | 254 |
| Understanding | 252 |
| Marketing | 250 |
| Part | 247 |
| Operations | 245 |
| Cloud | 245 |
| Projects | 241 |
| State | 240 |
| Relevant | 238 |
| Customers | 237 |
| Language | 235 |

Appendix B: Most Common Phrases

| Phrase | Frequency | Phrase | Frequency |
|----------------------------|-----------|-----------------------------------|-----------|
| Artificial Intelligence | 599 | Job Description | 109 |
| Machine Learning | 515 | AI Models | 108 |
| AI ML | 265 | Cross Functional | 107 |
| Computer Science | 227 | Team Members | 106 |
| Data Science | 172 | AI Solutions | 104 |
| Generative AI | 162 | Team Of Professionals Who Deliver | 95 |
| Years Of Experience | 155 | Natural Language | 93 |
| Veteran Status | 153 | Preferred Qualifications | 91 |
| Patient Care | 151 | Fast Paced Environment | 86 |
| High Quality | 150 | Relevant Experience | 84 |
| Full Time | 148 | Interpersonal Skills | 83 |
| Related Field | 141 | Cutting Edge | 82 |
| Fast Paced | 130 | Ability To Work | 81 |
| Artificial Intelligence Ai | 129 | Masters Degree | 81 |
| Communication Skills | 124 | Solving Complex Problems | 81 |
| Deep Learning | 124 | TS SCI | 81 |
| Qualified Applicants | 122 | Security Clearance | 80 |
| Bachelors Degree | 118 | Degree In Computer Science | 79 |
| Complex Problems | 114 | Natural Language Processing | 79 |
| Problem Solving | 111 | | |

Retrieval Augmented Generation (RAG) Based Question and Answer System

Emmanuel Balogun
eb20284@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

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

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

Abstract

This study focuses on improving LLM contextual understanding through external context sourcing RAG. In this approach, documents containing text, images, and tables, are transformed into high-dimensional vectors via LLM-generated embeddings, then stored in a vector database. When a user submits a query, the LLM retrieves the most relevant documents from this vector store, enabling a deeper understanding of the query and improving response quality. A comparative analysis—supported by a t-test—indicates that this RAG-based model outperforms typical LLM, demonstrating greater efficiency and accuracy. These research studies are combined in this thesis to unlock the hidden potentials of LLMs by providing a comprehensive understanding of improving Artificial Intelligence in contextualization for accurate answer generation and engagement. This solution eradicates LLMs' hallucinations by providing them with the right context and putting machines' reasoning on the same page as humans. The thesis also creates a solution to the long challenges of AI and the battle with the spread.

Keywords: RAG, Retrieval Augmented Generation, LLM, Vector Store, Query, Prompt.

Recommended Citation: Balogun, E., Wimmer, H., Rebman Jr., C.M., (2026). A Comparative Analysis of an LLM and RAG (Retrieval Augmented Generation) in a Question and Answer System. *Journal of Information Systems Applied Research and Analytics*, v19(n2) pp 36-49. DOI# <https://doi.org/10.62273/RJM2996>

Retrieval Augmented Generation (RAG) Based Question and Answer System

Emmanuel Balogun, Hayden Wimmer and Carl Michael

1. INTRODUCTION

The integration of artificial intelligence (AI) into healthcare has introduced transformative tools, with AI-powered chatbots emerging as a key technology for patient engagement and information dissemination (Laranjo, 2018). In a field where the precision and reliability of information can have profound consequences, the demand for accurate and contextually aware conversational agents is paramount. However, the efficacy of many current chatbot systems is often undermined by their reliance on generalist language models (Kung et al., 2023). These models, trained on vast but non-specialized datasets, are prone to significant shortcomings in high-stakes environments like healthcare, frequently producing responses plagued by factual inaccuracies, hallucinations, and a general lack of domain-specific knowledge, thereby eroding user trust and limiting their clinical utility (Jiang et al., 2023; Thirunavukarasu et al., 2023).

While various approaches have been developed to ground AI models in specific knowledge domains, a persistent challenge lies in effectively parsing the complex and varied document formats endemic to medical literature, such as PDFs containing a mix of unstructured text and structured tabular data (Lin et al., 2019). Existing systems often struggle to extract and synthesize information from these heterogeneous sources, failing to capture the complete context required for precise query resolution (Wang et al., 2018). This gap highlights the need for a sophisticated data processing pipeline capable of interpreting both textual and tabular content to build a comprehensive and reliable knowledge base for specialized applications.

To address these challenges, this paper introduces a novel Retrieval-Augmented Generation (RAG) framework specifically designed to enhance the accuracy of healthcare chatbots for answering frequently asked questions (FAQs). The model implements a specialized ingestion process that not only extracts textual content from PDF documents but also transforms embedded tabular data into a markdown text format, a method shown to

improve multimodal data handling (Lewis et al., 2020). By converting both content types into a unified embedding representation, the system creates a robust, domain-specific knowledge base, building on prior work in the joint understanding of textual and tabular data (Yin et al., 2020).

This research argues that by systematically processing the entirety of specialized documents, RAG-based approach significantly mitigates the risks of misinformation and hallucination, leading to a more reliable and effective AI-powered conversational tool for the healthcare sector (Shuster et al., 2021).

2. LITERATURE REVIEW

The fast rising and adoption of Large Language Models (LLMs) has created new opportunities for patient engagement, with AI-powered chatbots being explored for tasks ranging from answering patient queries to medical education (Laranjo et al., 2018). However, the high-stakes nature of healthcare demands a high level of precision and reliability that general-purpose models often fail to provide. Studies evaluating models like ChatGPT on medical licensing exams have shown promising results, but also highlight significant limitations (Brucks & von Bayern, 2020). These models, trained on broad internet corpora, are likely to generating factually incorrect information, a phenomenon often termed "hallucination" [3] (Thirunavukarasu et al., 2023). In a clinical context, such inaccuracies can erode user trust and pose significant risks, underscoring the critical need for systems grounded in domain-specific, verified knowledge.

RAG has emerged as a leading architectural model to address the limitations of standalone LLMs. First proposed by (Lewis et al., 2020), the RAG framework enhances generative models by enabling them to retrieve relevant information from an external knowledge base before generating a response. This approach directly grounds the model's output in factual data, which has been shown to significantly reduce the incidence of hallucination in conversational AI (Shuster et al., 2021). In the medical field, RAG offers a pathway to create more reliable AI systems by connecting them to curated sources of medical literature and data. Recent work by

(Yang et al., 2024) demonstrates the application of RAG in healthcare to mitigate biases and overcome the static knowledge limitations of generative models, thereby improving the reliability of AI-driven responses.

While RAG provides a robust framework, its effectiveness is fundamentally dependent on the quality of its knowledge base. In healthcare, a significant portion of specialized knowledge is encapsulated in PDF documents, which often feature complex layouts containing a mixture of unstructured text, tables, and figures. Traditional methods for clinical information extraction have long grappled with the complexities of parsing these varied formats (Lin et al., 2019; Wang et al., 2018). This challenge persists in modern RAG systems. (Lin, 2024) highlights that conventional PDF parsing methods often fail to accurately recognize document structures, leading to flawed information retrieval. To address this, the introduction of deep learning-based parsers like ChatDOC has been shown to significantly improve the chunking and extraction of content from complex layouts, including tables and multi-column text, thereby enhancing the accuracy of downstream RAG tasks.

Beyond simple text extraction, a key challenge is creating a unified understanding of both the textual and tabular data within medical documents. Foundational work (Yin et al., 2020) with TaBERT demonstrated the feasibility of pretraining models to jointly comprehend information from natural language text and structured tables. This is crucial for healthcare applications where clinical data, lab results, and treatment guidelines are often presented in tabular form.

Building on this, recent advancements in RAG have focused on optimizing the retrieval process itself. For instance, the Blended RAG framework proposed by (Sawarkar et al., 2024) combines keyword-based, dense vector, and semantic search to form hybrid queries, improving retrieval relevance. Similarly, (Ghali et al., 2024; Koo et al., 2024) introduced a query optimization method that refines user queries to enhance retrieval accuracy, particularly in scientific contexts.

These advanced retrieval strategies, when combined with sophisticated parsing techniques, provide a foundation for building a comprehensive and reliable knowledge base from the heterogeneous data found in medical literature. This review reveals a clear need for an integrated RAG framework that specifically

addresses the end-to-end challenge of ingesting complex PDF documents, transforming both textual and tabular data into a unified representation, and leveraging this for accurate, domain-specific answer generation in healthcare.

3. METHODOLOGY

The methodology described herein outlines the development of a RAG based question-answering (QA) system. The system's primary goal is to utilize document retrieval to enhance the relevance and accuracy of LLM outputs when answering questions based on knowledge retrieved from PDF documents. While we use PDF documents as case study, modern RAG can accept a myriad of document types. This section delves into all RAG process stages, from data preprocessing, embedding, and indexing strategies, retrieval techniques, and the final question-answering mechanism, thoroughly exploring each component of the system.

We utilized a series of highly optimized Python libraries like **Unstructured**, **LangChain**, **PDFplumber**, **Natural Language Processing (NLP)**, **Numpy**, and **pandas** to handle the various data types and formats present in PDF documents. By partnering with Institute for Health Logistics & Analytics (IHLA), case study documents which contains the One health joint plan of action (2022–2026) and The US investigating the origins of covid-19 hearing. These documents were parsed using UnstructuredPDFLoader, a component of langchain library built to process unstructured text data from diverse document formats, making it well-suited for extracting both textual and tabular data from complex PDF documents. Additionally, the PDFplumber was employed to extract tabular data from document. This choice was guided by the PDFplumber's ability to accurately detect table boundaries and parse them into structured data formats. Using multiple parsing libraries was critical to ensure data integrity and maximize the quality of the extracted content.

Document Preprocessing

Document processing involved reading the PDF input document using the UnstructuredPDFLoader. The text contained within the document was extracted and prepared for subsequent operations. This was done to ensure that the model received a clean and consistent input across all document sections.

One of the significant challenges when handling PDFs is the extraction of tabular data. Tables, which often contain structured numerical or

categorical information, cannot be processed in their native format by LLMs. To address this, PDFplumber was used to extract tables, which were then converted into a Markdown format. Markdown allows for the structured representation of tabular data using simple text symbols, which the downstream LLM can easily understand during the question-answering phase.

Figures 1 and 2 as shown below illustrates how tabular data is converted into markdown format to help LLMs understand the concept of tables.

| Column 1 | Column 2 | Column 3 |
|--------------|--------------|--------------|
| Row 1, Col 1 | Row 1, Col 2 | Row 1, Col 3 |
| Row 2, Col 1 | Row 2, Col 2 | Row 2, Col 3 |
| Row 3, Col 1 | Row 3, Col 2 | Row 3, Col 3 |

Figure 1: Initial Table

```
| Column 1 | Column 2 | Column 3 |
|-----|-----|-----|
| Row 1, Col 1 | Row 1, Col 2 | Row 1, Col 3 |
| Row 2, Col 1 | Row 2, Col 2 | Row 2, Col 3 |
| Row 3, Col 1 | Row 3, Col 2 | Row 3, Col 3 |
```

Figure 2: Markdown text format for LLM

This process of converting tables into Markdown offers two key advantages:

- **Text-based Representation:** The LLM can now process the tables as text, enabling it to understand tabular relationships linearly, which is vital for generating answers that consider tabular data.
- **Seamless Integration:** Markdown tables integrate smoothly with other text, ensuring that the context of both text and tabular data is preserved during the chunking and embedding stages.

Once the text and tables were extracted and cleaned, the next step involved dividing the document into smaller chunks for embedding. The document was split into text blocks of approximately 4,000 characters each using the RecursiveCharacterTextSplitter, a LangChain Python library tool. This recursive text splitter was selected due to its ability to maintain context while chunking large volumes of text.

Chunking is an essential step in building retrieval-based systems because large documents must be broken down into more manageable pieces. In this context, a balance was struck between chunk size and computational efficiency:

- **Chunk Size:** A 4,000-word chunk was chosen because it fits well with the input token limits of 4,000 OpenAI's language models input token and ensures that the context within a chunk remains coherent.
- **Recursive Splitting:** The recursive splitting strategy allows the algorithm to adjust chunk boundaries intelligently, ensuring that the content isn't arbitrarily truncated in a way that may lose important context (e.g., between paragraphs or sections).

Text Embedding and Indexing

After chunking, each text block was passed through the OpenAI embedding model to generate a dense vector representation. Embeddings are critical in enabling the system to perform semantic search, as they convert textual information into numerical vectors that capture the meaning of the text.

Mathematical Representation

Let:

T_i represent the i -th text block in the set of text blocks.

$f(T_i) = v_i$, Where v_i is the dense vector representation (embedding) of dimension k generated by the embedding model $f(\cdot)$

Thus, the embedding generation process can be expressed as:

$$v_i = f(T_i), \text{ for each } i \in \{1, 2, \dots, n\}$$

Here, $f(T_i)$ maps the text block T_i to its corresponding embedding vector v_i which enables tasks like semantic search by comparing vectors in the embedding space.

We used OpenAI's embedding model to generate vector representations of each text chunk (Figure 3). Embeddings are a core component of most modern NLP models, as they encode the semantic content of text in a continuous vector space. The advantage of OpenAI embeddings lies in their pre-training on diverse datasets, enabling robust and generalizable representations of a wide variety of text types.

By encoding the text as vectors, we could transform the text retrieval problem into a vector similarity search problem. This transformation enables the model to retrieve chunks based on the semantic similarity between the user query and the embedded text.

Once the document chunks were embedded, they were stored in a vector index for fast and efficient retrieval. The FAISS (Facebook AI

Similarity Search) library was chosen to build this vector store. FAISS is a widely used library for fast approximate nearest neighbor (ANN) search, particularly when working with large sets of dense vectors. FAISS supports both exact and approximate searches, with its approximate search algorithms being ideal for large-scale systems where retrieval speed is critical. In this system, FAISS provided the following benefits:

Efficiency: FAISS offers highly efficient indexing and search capabilities, even on datasets containing millions of vectors.

Scalability: The system can easily be scaled up to handle much larger or multiple documents without a significant increase in query latency.

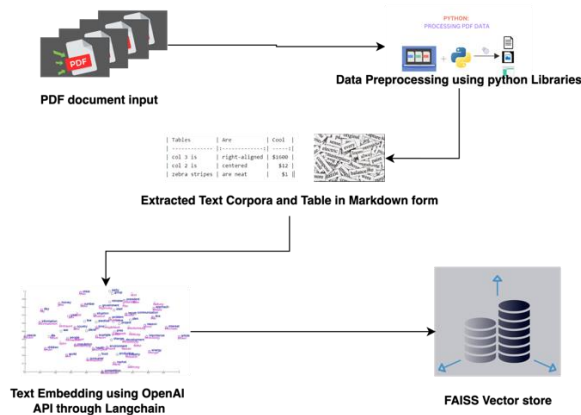


Figure 3: Document Embedding using FAISS Vector store.

Retrieval and Answer Generation

Each document chunk's embedding was added to the FAISS index. This allowed the system to perform a similarity search across the document when the user posed a query. The index was built using FAISS's IndexFlatL2 algorithm, which performs an exact search based on Euclidean distance between vectors. FAISS was chosen for its ability to handle large amounts of vector data in memory, providing a balance between accuracy and speed. By storing the embedded document chunks in this index, we enabled the model to perform quick lookups of relevant sections of the document based on a user's query.

The core of the retrieval process is based on cosine similarity, which measures the angular distance between vectors. Cosine similarity was chosen for this application because it has been shown to be highly effective in semantic search tasks, where the goal is to retrieve content

based on meaning rather than exact keyword matches.

Given two vectors v_q and v_d representing the query embedding and document embedding, respectively, the cosine similarity $\cos(v_q, v_d)$ between the query and a document chunk can be computed using the following formula:

$$\cos(v_q, v_d) = \frac{v_q \cdot v_d}{|v_q| |v_d|}$$

Where:

$v_q \cdot v_d$ is the dot product of the query vector and the document vector.

$|v_q|$ is the Euclidean norm (magnitude) of the query vector.

$|v_d|$ is the Euclidean norm (magnitude) of the document vector.

The cosine similarity measures the cosine of the angle between these two vectors. A higher value (closer to 1) indicates higher semantic similarity between the query and the document chunk, while a value closer to 0 indicates less similarity.

Therefore, in the retrieval process:

- For each document chunk d in the FAISS index, compute its embedding v_d
- For a given query, compute its embedding v_q using the same embedding model.
- Cosine similarity between v_q and v_d .
- Rank the document chunks based on their cosine similarity scores and retrieve the most relevant chunks (those with the highest similarity).

When the user asks a question (query), the system first generates an embedding for the query using the OpenAI embedding. This query embedding will then be compared against the document embeddings stored in the FAISS index using cosine similarity to determine which chunks from the document are semantically closely related to the query.

Once the most relevant document chunks were retrieved, they were passed along with the user's question into the final question-answering stage. The retrieval process ensured that the model had access to only the most pertinent sections of the document, reducing noise and improving the quality of the generated response.

The system as shown in Figure 4 used LangChain, a framework designed to integrate document retrieval with LLMs to generate an

answer from the retrieved document chunks. LangChain was selected due to its modular architecture, which facilitates easy integration of various components, such as document retrieval, question answering, and generation workflows.

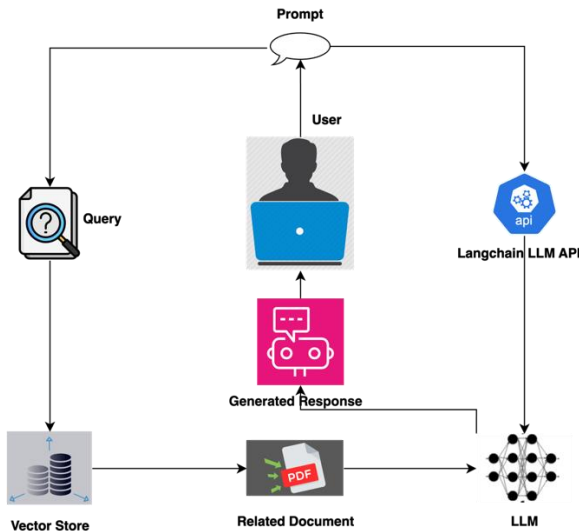


Figure 4: RAG Question and Answer working system.

Once LangChain combined the retrieved document chunks and the user's query, it was passed into OpenAI's language model for answer generation. The language model used the retrieved context to generate accurate, context-aware answers to the user's question. This hybrid approach, leveraging both retrieval and generation, ensured that the system could answer complex queries grounded in the document's content while benefiting from the broad linguistic capabilities of OpenAI's language models.

4. RESULTS

The evaluation of LangChain and LLM involves assessing the semantic accuracy of their answers compared to a ground truth answer. These answers were evaluated based on Bidirectional Encoder Representations from Transformers score (BERT) similarity, providing a detailed comparison of the effectiveness of each model in understanding and generating responses.

Model Performance on Various Questions

Various questions from the document were passed on to the models; their answers were evaluated against the ground truth using BERT. BERT scores quantify the semantic similarity between two pieces of text, and in this case, it

was used to compare the answers generated by the models to the ground truth. A detailed breakdown of the questions, model answers, and their corresponding BERT scores is shown in the Appendix.

Each model demonstrated distinct strengths across different types of questions. The LLM was generally more consistent in providing concise and factual responses, particularly when the question directly aligned with specific document content. However, with more BERT score values, the RAG often outperformed the LLM in generating more precise and contextually richer responses that offered more detailed insights, especially when dealing with more complex or interpretive queries.

Research Result Hypothesis

Answers generated by RAG and LLM seem to show a significant difference in efficiency. To further demonstrate these differences, we conducted a human survey through Institutional Review Board (IRB) approved health professionals recruited from a prominent One Health conference. All subjects are verified experts in the OneHealth model. A total of 17 respondents participated, each rated 15 sets of questions and answers. There is a null hypothesis (H0) and an alternative hypothesis (H1).

H0 (null hypothesis): No statistically significant difference exists between standard LLM and RAG.

H1 (research hypothesis): RAG significantly outperforms LLM.

We used Qualtrics, an online survey platform, to conduct our survey. Out of 20 RAG questions, 14 were randomly shown to participants. The answers from the model and RAG were also randomized, ensuring that the order of the answer switches between the LLM and RAG responses. Participants were asked to rate the accuracy of the answers based on how closely they matched the ground truth, using a 1 to 7 scale. On this scale, a lower score indicated more substantial agreement with the answer, while a higher score indicated stronger disagreement. Data from the survey was collected and analyzed, and the hypothesis testing was followed by applying a t-test evaluation to assess the differences between the models. The results of the analysis are presented in the *Table 1* below.

| | LLM | RAG |
|-------------------------------------|--------|-------|
| Mean | 3.919 | 2.784 |
| Variance | 3.993 | 2.583 |
| Observations | 74 | 74 |
| Pearson Correlation | 0.225 | |
| Hypothesized Mean Difference | 0 | |
| df | 73 | |
| t Stat | 4.31 | |
| P(T<=t) one-tail | 0.0000 | |
| t Critical one-tail | 1.666 | |
| P(T<=t) two-tail | 0 | |
| t Critical two-tail | 1.993 | |

Table 1: T-test for Research Result Hypothesis.

The mean score for RAG (2.784) is lower than that for LLM (3.919), indicating that participants rated RAG answers as more accurate and closely aligned with the ground truth. The t-statistic of 4.31 exceeds the critical t-value for both one-tail (1.666) and two-tail (1.993) tests, indicating that the difference is statistically significant at the 5% level. The p-value for the one-tailed test is approximately 0.000000, which is well below the significance threshold (0.05). Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference between LLM and RAG.

Since RAG has a lower mean score (indicating higher accuracy based on the rating scale), the results support the alternative hypothesis (H1) that RAG with BERT similarity improves robustness and significantly outperforms LLM model. Through human evaluation, findings shows that RAG significantly outperforms LLM regarding accuracy and alignment with the ground truth.

5. CONCLUSION

A comparative analysis was performed to evaluate the performance of a RAG model against a standard LLM in a question-and-answer system. The study aimed to assess the accuracy and alignment of responses generated by both models through a structured human survey. Participants rated the accuracy of the

answers based on their agreement with the ground truth using a 1 to 7 scale, where lower scores indicated more substantial agreement and higher scores indicated stronger disagreement. A paired t-test was conducted to determine whether this difference was statistically significant. The t-statistic value of 4.31 exceeded the critical t-values for both one-tailed (1.666) and two-tailed (1.993) tests at the 5% significance level.

The p-value for the one-tailed test was approximately 0.000000, which is well below the significance threshold of 0.05. As a result, the null hypothesis (H_0) was rejected, and findings support the alternative hypothesis (H_1) that RAG significantly outperforms LLM in terms of accuracy based on human evaluation. The superior performance of RAG is attributed to its ability to combine external knowledge retrieval with language generation.

Unlike standard LLMs relying solely on pre-trained knowledge, RAG systems enhance accuracy by retrieving information from authoritative sources, ensuring that generated answers are contextually grounded and accurate. This hybrid approach enables RAG to provide more detailed and reliable answers, particularly for complex or fact-based queries.

These results underscore the potential of RAG-based systems to improve the accuracy and trustworthiness of AI-generated content, especially in real-time information validation and fact-checking applications. This research has several key limitations that will be addressed in future work, including a limited case study dataset, the exclusion of image content retrieval, and the use of a single LLM model for comparison.

In conclusion, the findings demonstrate that RAG significantly outperforms LLM regarding accuracy and alignment with the ground truth, highlighting the value of retrieval-augmented frameworks in delivering more reliable and context-aware responses.

6. REFERENCES

Brucks, D., & von Bayern, A. M. P. (2020). Parrots Voluntarily Help Each Other to Obtain Food Rewards. *Current Biology*. <https://doi.org/10.1016/j.cub.2019.11.030>

Ghali, J. P. E., Shima, K., Moriyama, K., Mutoh, A., & Inuzuka, N. (2024). Enhancing Retrieval Processes for Language Generation with Augmented Queries. *arXiv preprint*

- arXiv:2402.16874*.
<https://doi.org/10.48550/arXiv.2402.16874>
- Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.
<https://doi.org/10.18653/v1/2023.emnlp-main.495>
- Koo, H., Kim, M., & Hwang, S. J. (2024). Optimizing Query Generation for Enhanced Document Retrieval in RAG. *arXiv preprint arXiv:2407.12325*.
<https://doi.org/10.48550/arXiv.2407.12325>
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., & Maningo, J. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2), e0000198.
<https://doi.org/10.1371/journal.pdig.0000198>
- Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., & Lau, A. Y. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248-1258.
<https://doi.org/10.1093/jamia/ocy072>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
<https://doi.org/10.48550/arXiv.2005.11401>
- Lin, C., Miller, T., Dligach, D., Bethard, S., & Savova, G. (2019). A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. Proceedings of the 2nd clinical natural language processing workshop,
<https://doi.org/10.18653/v1/W19-1908>
- Lin, D. (2024). Revolutionizing retrieval-augmented generation with enhanced PDF structure recognition. *arXiv preprint arXiv:2401.12599*.
<https://doi.org/10.48550/arXiv.2401.12599>
- Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. *arXiv preprint arXiv:2404.07220*.
<https://doi.org/10.1109/MIPR62202.2024.00031>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.
<https://doi.org/10.48550/arXiv.2104.07567>
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8), 1930-1940.
<https://doi.org/10.1038/s41591-023-02448-8>
- Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., & Sohn, S. (2018). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77, 34-49.
<https://doi.org/10.1016/j.jbi.2017.11.011>
- Yang, R., Ning, Y., Keppo, E., Liu, M., Hong, C., Bitterman, D. S., Ong, J. C. L., Ting, D. S. W., & Liu, N. (2024). Retrieval-Augmented Generation for Generative Artificial Intelligence in Medicine. *arXiv preprint arXiv:2406.12449*.
<https://doi.org/10.48550/arXiv.2406.12449>
- Yin, P., Neubig, G., Yih, W.-t., & Riedel, S. (2020). TaBERT: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*.
<https://doi.org/10.48550/arXiv.2005.08314>

Appendix

Detailed breakdown of the questions, model answers, and their corresponding BERT scores

QUESTION 1 *What organizations make up the quadripartite?*

LLM ANSWER

The Quadripartite is composed of four international organizations: the World Health Organization (WHO), the Food and Agriculture Organization (FAO), the World Organisation for Animal Health (WOAH, formerly OIE), and the United Nations Environment Programme (UNEP). These organizations collaborate to address global health challenges at the human-animal-environment interface, focusing on areas like zoonotic diseases, food safety, and antimicrobial resistance. Their combined efforts aim to promote a unified approach to global health, known as the One Health initiative.

LANGCHAIN-ANSWER

The quadripartite is made up of the Food and Agriculture Organization of the United Nations (FAO), the United Nations Environment Programme (UNEP), the World Organisation for Animal Health (WOAH, founded as OIE), and the World Health Organization (WHO). These organizations collaborate to address health challenges at the human-animal-plant-environment interface. They have developed the One Health Joint Plan of Action (OH JPA) to promote sustainable health solutions.

LANGCHAIN-BERT-score 0.9366606

LLM-BERT-score 0.90872216

QUESTION 2 *Why is a joint Plan of Action needed?*

LLM ANSWER

A joint Plan of Action is needed to ensure coordinated efforts and alignment among stakeholders toward achieving a common goal. It helps in pooling resources, avoiding duplication of efforts, and maximizing the impact of collective actions. Moreover, it establishes clear roles and responsibilities, which enhances accountability and efficiency in addressing the targeted issue.

LANGCHAIN-ANSWER

A joint Plan of Action is needed to address the interconnected health threats to humans, animals, plants, and the environment in an integrated manner, promoting sustainable development. It aims to improve collaboration, coordination, and capacity-building across sectors to effectively prevent, predict, detect, and respond to health threats. Additionally, it seeks to leverage resources and expertise from various stakeholders to achieve comprehensive One Health outcomes.

LANGCHAIN-BERT-score 0.8954361

LLM-BERT-score 0.87486625

QUESTION 3 *What is the definition of One Health?*

LLM ANSWER

One Health is an integrated, unifying approach that aims to optimize the health of people, animals, and ecosystems by recognizing their interconnection. It involves collaboration across various disciplines, including human medicine, veterinary medicine, environmental science, and public health, to address complex health challenges. The One Health approach is critical for effectively managing zoonotic diseases, food safety, and environmental hazards that impact global health.

LANGCHAIN-ANSWER

One Health is an integrated, unifying approach that aims to sustainably balance and optimize the health of humans, animals, plants, and ecosystems. It recognizes that the health of humans, domestic and wild animals, plants, and the wider environment are closely linked and interdependent. The approach mobilizes multiple sectors, disciplines, and communities at varying levels of society to work together to foster well-being and tackle threats to health and ecosystems.

LANGCHAIN-BERT-score 0.9097894

LLM-BERT-score 0.91218096

QUESTION 4 *List the key objectives of the Joint Plan of Action.*

LLM ANSWER

The Joint Plan of Action (JPA) aimed to ensure that Iran's nuclear program would be exclusively peaceful by limiting its nuclear activities. It sought to provide Iran with limited, temporary, and

reversible sanctions relief in exchange for verifiable actions that curtail its nuclear program. Additionally, the JPA established a framework for future negotiations towards a comprehensive and long-term solution to Iran's nuclear issue.

LANGCHAIN-ANSWER

The key objectives of the One Health Joint Plan of Action (OH JPA) include providing a framework for collective and coordinated action to mainstream the One Health approach at various levels, offering policy and legislative advice and technical assistance to set national targets and priorities, and promoting collaboration, learning, and exchange among nations, sectors, and disciplines. Additionally, the OH JPA aims to take stock of existing cross-sectoral initiatives, identify synergies and overlaps, and support coordination. It also focuses on mobilizing and optimizing the use of resources across sectors, disciplines, and stakeholders.

LANGCHAIN-BERT-score 0.84955204

LLM-BERT-score 0.8311661

QUESTION 5 *How can the environmental sector contribute to the initiative?*

LLM ANSWER

The environmental sector can contribute to the initiative by promoting sustainable practices that reduce carbon emissions and conserve natural resources, ensuring long-term environmental health. By integrating eco-friendly technologies and supporting renewable energy solutions, it can drive innovation in reducing environmental impact. Additionally, collaboration with other sectors can help raise awareness and implement policies that prioritize sustainability in both local and global contexts.

LANGCHAIN-ANSWER

The environmental sector can contribute to the initiative by developing and implementing advocacy training and tools for decision-makers, conducting national environment sector needs assessments, and creating interoperable One Health training courses for professionals. Additionally, the sector can ensure the inclusion of environmental considerations in medical, public health, and veterinary training, and integrate systems thinking into core academic modules. These efforts will help influence One Health policies and ensure that environmental priorities are adequately represented.

LANGCHAIN-BERT-score 0.88010305

LLM-BERT-score 0.8791212

QUESTION 6 *What is the Joint Plan of Action?*

LLM ANSWER

The Joint Plan of Action (JPOA) is an interim agreement reached in 2013 between Iran and the P5+1 (the U.S., UK, France, Russia, China, and Germany) concerning Iran's nuclear program. It aimed to limit Iran's nuclear activities in exchange for temporary relief from economic sanctions while a more comprehensive agreement was negotiated. The JPOA served as the foundation for the subsequent 2015 Joint Comprehensive Plan of Action (JCPOA), commonly known as the Iran nuclear deal.

LANGCHAIN-ANSWER

The One Health Joint Plan of Action (OH JPA) is a strategic framework developed by the Quadripartite organizations (FAO, WHO, WOA, and UNEP) to address health threats at the human-animal-environment interface. It aims to prevent, predict, detect, and respond to health threats while promoting sustainable development. The OH JPA outlines specific actions, deliverables, and timelines to achieve long-term and medium-term health outcomes through a coordinated One Health approach.

LANGCHAIN-BERT-score 0.8350722

LLM-BERT-score 0.83486396

QUESTION 7 *Describe the theoretical basis of the plan*

LLM ANSWER

The theoretical basis of the AI-Driven Agile Risk Management Tool lies in leveraging machine learning algorithms to predict risks by analyzing historical project data, team dynamics, and project complexity. This approach integrates principles from Agile methodologies, emphasizing adaptability and iterative improvements in risk management through real-time assessments. By utilizing predictive analytics, the tool provides proactive risk mitigation strategies, enabling teams to make informed decisions and maintain project momentum amidst uncertainties.

LANGCHAIN-ANSWER

The theoretical basis of the One Health Joint Plan of Action (OH JPA) is built on the premise that an integrated, multisectoral, holistic, and transdisciplinary approach can effectively address pressing

health challenges. The theory of change is supported by three pathways: policy, legislation, advocacy, and financing; organizational development, implementation, and sectoral integration; and data, evidence, and knowledge. These pathways collectively aim to achieve the medium- and long-term outcomes of improved health for humans, animals, plants, and the environment, while reducing health threats at the human-animal-plant-environment interface.

LANGCHAIN-BERT-score 0.8406563
LLM-BERT-score 0.81940776

QUESTION 9 *What are the six action tracks?*

LLM ANSWER

The six action tracks refer to key focus areas outlined in various global frameworks for addressing critical challenges. These tracks typically include areas like health, climate action, poverty reduction, education, economic development, and technology innovation. Each track represents a targeted approach to solving specific issues within these broad domains, often fostering collaborative efforts between governments, organizations, and communities.

LANGCHAIN-ANSWER

"The six action tracks are:

1. Enhancing One Health capacities to strengthen health systems.
2. Reducing the risks from emerging and re-emerging zoonotic epidemics and pandemics.
3. Controlling and eliminating endemic zoonotic, neglected tropical, and vector-borne diseases.
4. Strengthening the assessment, management, and communication of food safety risks.
5. Curbing the silent pandemic of antimicrobial resistance.
6. Integrating the environment into One Health."

LANGCHAIN-BERT-score 0.83532774
LLM-BERT-score 0.84236777

QUESTION 9 *How does the plan of action build upon other initiatives?*

LLM ANSWER

The plan of action builds upon other initiatives by leveraging past successful strategies and incorporating lessons learned to enhance current objectives. It aligns with ongoing projects to create synergy, ensuring that resources and expertise are shared effectively. Additionally, it introduces innovative approaches to address gaps identified in previous efforts, aiming for more comprehensive and sustainable outcomes.

LANGCHAIN-ANSWER

The One Health Joint Plan of Action (OH JPA) builds upon other initiatives by complementing and supporting their implementation at various levels, including global, regional, and national. It strategically aligns with relevant initiatives such as the Global Action Plan on AMR, the Road Map for Neglected Tropical Diseases, and the WHO global strategy on health, environment, and climate change. The OH JPA offers a coordinated framework that integrates existing guidance, best practices, and tools from these initiatives to advance One Health objectives comprehensively.

LANGCHAIN-BERT-score 0.87298435
LLM-BERT-score 0.87678677

QUESTION 10 *How do the activities in 6.1 differ from the activities in 6.3?*

LLM ANSWER

The activities in 6.1 focus on preventive measures and surveillance systems to monitor public health risks, emphasizing early detection and proactive interventions. In contrast, the activities in 6.3 are centered on response and recovery, addressing the management of health crises and mitigation of their impact on affected populations. While 6.1 is more focused on preparedness, 6.3 deals with reactive measures and long-term recovery strategies.

LANGCHAIN-ANSWER

The activities in 6.1 focus on identifying and mitigating anthropogenic factors leading to environmental degradation, promoting sustainable resource management, and supporting policies for Indigenous Peoples and wildlife economies. In contrast, the activities in 6.3 are not detailed in the provided context, so specific differences cannot be fully outlined. However, 6.1 emphasizes practical measures for environmental conservation and policy support, while 6.3 likely involves different aspects of integrating the environment into One Health.

LANGCHAIN-BERT-score 0.89234847
LLM-BERT-score 0.89168185

QUESTION 11 *What action track has the highest number of deliverables?*

LLM ANSWER

The action track with the highest number of deliverables is typically the one focused on capacity building and stakeholder engagement. This track often includes numerous initiatives aimed at training, partnership development, and resource allocation. As a result, it generates a larger number of tangible outputs compared to more specialized tracks

LANGCHAIN-ANSWER

Action track 6, "Integrating the environment into One Health," has the highest number of deliverables. This track includes multiple activities aimed at promoting sustainable management of nature, enhancing ecosystem integrity, and integrating environmental knowledge into One Health decision-making. The deliverables span various timelines and involve extensive collaboration across sectors.

LANGCHAIN-BERT-score 0.88623744

LLM-BERT-score 0.90800357

QUESTION 12 *Summarize the recommendations for enhancing One Health capacities to strengthen health systems.*

LLM ANSWER

The recommendations for enhancing One Health capacities to strengthen health systems focus on improving cross-sectoral collaboration, particularly between human, animal, and environmental health sectors. They emphasize the need for stronger surveillance systems, better data sharing, and integrated response mechanisms to detect and manage health threats more effectively. Additionally, capacity-building initiatives, such as training and resource allocation, are recommended to ensure sustained coordination and preparedness across all sectors.

LANGCHAIN-ANSWER

The recommendations for enhancing One Health capacities include developing methodologies and tools for monitoring and evaluation, establishing financial needs and mechanisms for One Health investments, and advocating for political prioritization through economic case studies and national strategies. Additionally, it emphasizes the creation of a competent One Health workforce through joint processes, capacity building, and public engagement. These efforts aim to improve coordination, communication, and the integration of health systems at the human-animal-plant-environment interface.

LANGCHAIN-BERT-score 0.89351654

LLM-BERT-score 0.8894139

QUESTION 13 *In what ways does gender equality serve as a guiding principle in the Joint Plan of Action?*

LLM ANSWER

Gender equality serves as a guiding principle in the Joint Plan of Action by ensuring that health initiatives consider the specific needs and vulnerabilities of all genders. It promotes equal access to resources, healthcare, and decision-making opportunities, fostering inclusive policies that address disparities. Additionally, it emphasizes the importance of empowering women and marginalized groups in the implementation and leadership of health interventions.

LANGCHAIN-ANSWER

Gender equality in the Joint Plan of Action (OH JPA) supports gender equity and women's empowerment, taking a gender-sensitive perspective that considers vulnerabilities specific to each national context. This principle is consistent with the 2030 Agenda for Sustainable Development. The OH JPA emphasizes inclusiveness and equity, ensuring active participation and communication with communities, and recognizing the importance of local and traditional knowledge alongside scientific research.

LANGCHAIN-BERT-score 0.8834198

LLM-BERT-score 0.8693475

QUESTION 14 *What is the vision?*

LLM ANSWER

The vision is to create a healthier, more resilient world by fostering collaboration across human, animal, and environmental health sectors under the One Health approach. It aims to prevent, detect,

and respond to health threats more effectively by integrating diverse expertise and resources. Ultimately, the vision aspires to build sustainable systems that safeguard global health and well-being for current and future generations.

LANGCHAIN-ANSWER

The vision (impact) of the One Health Joint Plan of Action (OH JPA) is to create a world better able to prevent, predict, detect, and respond to health threats while improving the health of humans, animals, plants, and the environment. This vision aims to contribute to sustainable development. The desired impact is expected to be achieved within a 15-20-year timeline.

LANGCHAIN-BERT-score 0.89629817

LLM-BERT-score 0.86143655

QUESTION 15 What factors influence the emergence of antimicrobial resistance?

LLM ANSWER

The emergence of antimicrobial resistance (AMR) is influenced by the overuse and misuse of antibiotics in both human medicine and agriculture, which accelerates the adaptation of microbes. Poor infection prevention and control practices in healthcare settings and insufficient sanitation in communities also contribute to the spread of resistant strains. Additionally, the lack of new antibiotic development and the global movement of people and animals further exacerbate the spread of AMR.

LANGCHAIN-ANSWER

The emergence of antimicrobial resistance (AMR) is influenced by the overuse and misuse of antimicrobials in human, animal, and plant sectors. Social and environmental factors such as insufficient access to health services, inadequate production and housing, lack of clean water, poor sanitation, waste management, and hygiene also accelerate the spread of resistant genes and pathogens. Additionally, insufficient regulatory frameworks and a lack of awareness and education about the risks of AMR and the appropriate use of antimicrobials contribute to the problem.

LANGCHAIN-BERT-score 0.918555

LLM-BERT-score 0.9042939

QUESTION 16 What timeline is covered in the Joint Plan of Action?

LLM ANSWER

The Joint Plan of Action typically covers a timeline of five years, providing a structured framework for addressing health threats through coordinated efforts. This period allows for phased implementation, monitoring, and evaluation of key activities aimed at strengthening One Health capacities. It also ensures flexibility to adapt to emerging health challenges within the specified timeline.

LANGCHAIN-ANSWER

The Joint Plan of Action covers a timeline of 1 to 5 years. Specific activities and deliverables are planned across this period, with some actions scheduled for completion within 1 year, others within 2-3 years, and the remainder within 4-5 years. The plan aims to achieve medium-term outcomes by 2026.

LANGCHAIN-BERT-score 0.8803032

LLM-BERT-score 0.8429157

QUESTION 17 Who provided input into the Joint Plan of Action?

LLM ANSWER

Input into the Joint Plan of Action was provided by a wide range of stakeholders, including representatives from human, animal, and environmental health sectors. Key contributors included international organizations like the World Health Organization (WHO), the Food and Agriculture Organization (FAO), and the World Organisation for Animal Health (WOAH). Additionally, national governments, academic institutions, civil society, and local communities played vital roles in shaping the plan through consultations and collaborative efforts.

LANGCHAIN-ANSWER

The Joint Plan of Action (OH JPA) was informed by evidence, best practices, and existing guidance from the Quadripartite. It also incorporated input from countries, international partners, and non-state actors such as civil-society organizations, professional associations, academia, and research institutions. The implementation of proposed actions at the national level involved consultation with Members, Member States, and State Parties.

LANGCHAIN-BERT-score 0.86545044

LLM-BERT-score 0.83925253

QUESTION 18 *What are common barriers observed in enhancing One Health capacities?*

LLM ANSWER

Common barriers to enhancing One Health capacities include a lack of coordinated communication and collaboration between human, animal, and environmental health sectors. Insufficient funding and resources often hinder the implementation of integrated health initiatives, while gaps in surveillance systems limit the early detection of health threats. Additionally, varying political priorities and regulatory frameworks across sectors and countries create challenges for unified action and sustained progress.

LANGCHAIN-ANSWER

Common barriers observed in enhancing One Health capacities include a lack of cooperation between internal and external stakeholders, limited engagement with the environmental sector, and professional segregation. Additionally, there is limited standardization around One Health curricula and competency-based frameworks. Commercial, academic, reputational, and profit motives often supersede collaborative efforts, further hindering progress.

LANGCHAIN-BERT-score 0.89567107

LLM-BERT-score 0.89926213

QUESTION 19 *List all the activities that need to happen in Year 1.*

LLM ANSWER

In Year 1, key activities include establishing cross-sectoral collaboration frameworks and setting up integrated surveillance systems for early detection of health threats. Capacity-building programs must be initiated to train personnel across human, animal, and environmental health sectors. Additionally, initial assessments and resource mapping are required to identify gaps in current health systems and allocate necessary resources for effective implementation.

LANGCHAIN-ANSWER

In Year 1, the activities that need to happen include developing frameworks and mechanisms for public participation, promoting One Health cross-sectoral collaboration and partnerships, developing operational tools to support science-based One Health coordinated strategic technical actions, and providing guidance on the appropriate use of integrated One Health information, surveillance, and emergency response systems. Additionally, guidance and tools for transparent and trusted One Health collaborative governance, effective communication structures, and mechanisms for joint funding and resource mobilization need to be provided. Other activities include promoting One Health task forces and working groups, identifying and quantifying anthropogenic factors leading to environmental degradation, enhancing private-sector and NGO engagement in sustainable natural resource management, and promoting the transition towards sustainable, climate-smart, agroecological approaches.

LANGCHAIN-BERT-score 0.88191336

LLM-BERT-score 0.8496687

QUESTION 20 *How does natural resource management contribute to the One Health approach?*

LLM ANSWER

Natural resource management contributes to the One Health approach by ensuring the sustainable use of ecosystems, which helps prevent the emergence and spread of zoonotic diseases. By maintaining healthy environments, it reduces human and animal exposure to pathogens that thrive in degraded or disrupted ecosystems. Additionally, responsible management of natural resources promotes biodiversity, which is critical for ecosystem resilience and the overall health of both people and animals.

LANGCHAIN-ANSWER

Natural resource management contributes to the One Health approach by preserving biodiversity and maintaining ecosystem health, which helps limit the emergence and transmission of diseases across the animal-human-plant interface. It also addresses environmental hazards such as pollution and climate change, which pose significant health threats to humans, animals, and plants. Integrating environmental considerations into One Health policies and interventions enhances disease prevention, preparedness, and promotes sustainable health solutions.

LANGCHAIN-BERT-score 0.9136984

LLM-BERT-score 0.9048505

Improving AI-Driven Stroke Prediction Models: A Comparative Evaluation of SMOTE and Undersampling Methods

Dara Tourt
dara.tourt@my.metrostate.edu
Metropolitan State University Minnesota
St Paul, MN 55106

Queen Booker
queen.booker@metrostate.edu
Metropolitan State University Minnesota
St Paul, MN 55106

Abstract

Artificial intelligence (AI) is improving the field of predictive healthcare by enabling data-driven decision-making through advanced machine learning (ML) algorithms. Stroke prediction is challenging due to highly imbalanced clinical datasets, where positive cases are rare. This study investigates the impact of data-level resampling methods on the performance of AI-driven predictive models. Four widely used classifiers—Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting (GB)—were applied to a highly imbalanced stroke dataset. Models were evaluated across key AI performance metrics. Paired t-tests assessed the statistical significance of observed differences. This comparative analysis offers critical insights into how data balancing techniques impact the reliability of AI models. The findings support the development of more effective and ethically responsible AI systems for early stroke detection.

Keywords: Stroke Prediction, Class Imbalance, SMOTE, Undersampling, Machine Learning

Recommended Citation: Tourt, D., Booker, Q., Jin, S.S., (2026). Improving AI-Driven Stroke Prediction Models: A Comparative Evaluation of SMOTE and Undersampling Methods. *Journal of Information Systems Applied Research and Analytics*, v19(n2) pp 50-70. DOI# <https://doi.org/10.62273/MWHS5422>

Improving AI-Driven Stroke Prediction Models: A Comparative Evaluation of SMOTE and Undersampling Methods

Dara Tourt and Queen Booker

1. INTRODUCTION

Artificial intelligence (AI) is increasingly important in healthcare, supporting diagnosis, prediction, and management of complex conditions. Stroke prediction is a particularly high-impact application given the sudden onset and severe consequences of stroke. Machine learning (ML) models show promise for identifying risk by analyzing large-scale electronic health records, but their effectiveness is often limited by classification imbalance: stroke-positive cases represent only a small fraction of the data, making accurate detection difficult. In such scenarios, models tend to favor the majority class and overlook minority cases, raising the risk of false negatives—an unacceptable outcome in clinical settings where early detection is critical.

To address this, researchers apply resampling techniques that adjust class distribution in training data. Oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) create synthetic minority samples, while undersampling approaches (e.g., Tomek Links, Edited Nearest Neighbors, and NearMiss) reduce majority samples to balance the data. Despite their widespread use, few studies provide systematic, side-by-side comparisons of these methods across multiple ML classifiers in stroke prediction.

This study fills that gap by evaluating the impact of SMOTE and several undersampling techniques on four common classifiers: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting (GB). We assess performance using multiple evaluation metrics and statistical testing to identify trade-offs and practical implications.

Our research is guided by two questions:

- **RQ1:** How do SMOTE and selected undersampling methods compare in improving model performance for stroke prediction with highly imbalanced datasets?
- **RQ2:** What trade-offs arise between predictive performance when using SMOTE versus undersampling techniques?

By addressing these questions, this study contributes to building more accurate and clinically relevant AI models for early stroke detection.

2. LITERATURE REVIEW

This section reviews key literature on stroke prediction by focusing on classification imbalance, oversampling methods such as SMOTE, undersampling techniques, and the use of machine learning models. It concludes by identifying current research gaps that this study aims to address.

Class Imbalance in Stroke Prediction

Class imbalance is a well-documented challenge in healthcare datasets, where stroke-positive cases are far fewer than non-stroke cases. This imbalance biases models toward majority classifications, leading to poor sensitivity in detecting actual stroke cases and an elevated risk of false negatives (Salmi et al., 2024; Chen et al., 2024; Lin et al., 2024). In clinical contexts, missed diagnoses have serious consequences, underscoring the importance of addressing imbalance in predictive modeling (Aish et al., 2024).

Over-sampling Techniques: SMOTE

Over-sampling increases the representation of minority cases in training data. The Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely adopted methods, generating synthetic samples by interpolating between existing minority cases (Chawla et al., 2002). Studies show that SMOTE improves sensitivity and F1-scores in medical predictions (Salmi et al., 2024), though it can also create overlapping regions or introduce noise, increasing overfitting risk (Elreedy et al., 2024; Fernández et al., 2018a, 2018b).

Undersampling Techniques

Undersampling reduces imbalance by removing majority-class cases. While effective in improving minority detection, it risks discarding valuable information in smaller datasets.

- **Random Undersampling (RU):** Efficient and often improves recall but may remove informative samples (He & Garcia, 2009).

- **Tomek Links:** Identifies neighboring pairs from different classes and removes majority instances to clean decision boundaries. It can enhance separability but may discard useful borderline cases (Tomek, 1976; Batista et al., 2004).
- **Edited Nearest Neighbors (ENN):** Removes samples that disagree with most neighbors, reducing noise but sometimes overly eliminating data and increasing computational cost (Wilson, 1972; Laurikkala, 2001).
- **NearMiss:** Retains majority cases based on distance to minority samples, emphasizing boundary representation (Mani & Zhang, 2003). Variants differ in focus: NearMiss-1 enhances sensitivity but raises false positives; NearMiss-2 reduces overlap but may miss borderline patterns (Yen & Lee, 2009); NearMiss-3 sharpens boundaries but can preserve noisy examples.

Machine Learning and Sampling Methods

Machine learning classifiers such as Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB) are widely used in healthcare prediction. Logistic Regression remains a strong baseline due to its interpretability, particularly when imbalance is corrected through resampling (Sáez et al., 2015). Ensemble models like RF and GB capture nonlinear patterns and perform well with SMOTE or hybrid approaches (Chawla et al., 2002; Fernández et al., 2018). XGB, leveraging gradient boosting, demonstrates strong predictive power and is enhanced by resampling techniques (Chen & Guestrin, 2016; Haixiang et al., 2017). Combining these classifiers with oversampling or undersampling consistently improves minority-class detection in medical datasets (Douzas & Bacao, 2018).

3. METHODOLOGY

This study follows the research methodology used by Kamiri and Mariga (2021), which includes Data Collection, Data Pre-processing, Model Training, Model Testing, and Model Evaluation. A process model outlining the methodology’s steps is shown in Appendix F.

We aim to look at how different resampling techniques can help improve stroke prediction when the data is heavily imbalanced. We compare SMOTE, a popular oversampling method, with several undersampling approaches like Random Undersampling, Tomek Links, Edited Nearest Neighbors (ENN), and the three

versions of NearMiss. To see how these methods impact results, we test them across four commonly used machine learning models: Logistic Regression, Random Forest, XGBoost, and Gradient Boosting, all using a real-world stroke dataset where positive cases are rare.

Gaps in Existing Research

While SMOTE and undersampling techniques have been studied independently, there is limited research that compares a broad range of these methods within the context of stroke prediction across multiple classifiers. Many existing studies also lack statistical validation of their findings, which limits the reproducibility and reliability of their conclusions. Furthermore, few works evaluate performance using a comprehensive set of metrics, such as precision, recall, F1-score, ROC-AUC, and PR-AUC, which are essential for assessing models trained on imbalanced data.

This study addresses these limitations by performing a side-by-side comparison of several resampling techniques across four classifiers, applying consistent evaluation criteria and statistical significance testing. The findings offer practical insights into how different resampling methods affect stroke prediction performance and provide guidance for building more reliable machine learning models in healthcare.

| Feature | Feature description |
|-------------------|--|
| id | Unique identifier |
| gender | Male, Female, or Other |
| age | Age of the patient |
| hypertension | 0 if the patient doesn't have hypertension, 1 if the patient has hypertension |
| heart_disease | 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease |
| ever_married | No or Yes |
| work_type | Children, Govt_job, Never_worked, Private, or Self-employed |
| Residence_type | Rural or Urban |
| avg_glucose_level | Average glucose level in the blood |
| bmi | Body mass index |
| smoking_status | Formerly smoked, Never smoked, Smokes, or Unknown |
| stroke | 1 if the patient had a stroke, 0 if not |

Table 11: Description of Stroke Dataset

The following section provides a detailed explanation of the research methodology adopted in this study.

Dataset

The dataset used in this study contains 5,110 instances, each representing a patient, and is publicly available on Kaggle (Fedesoriano, n.d.). The dataset includes a variety of attributes relevant to predicting the occurrence of a stroke, with detailed descriptions provided in Table 1 below.

Among the patients, 2,994 were female, 2,115 were male, and 1 was categorized as other. The average age was 43 years, with a range from 18 to 82 years. Additionally, 498 patients had hypertension, and 276 were diagnosed with heart disease. The variable `smoking_status` represents the patient's self-reported smoking behavior. It is a categorical feature with four possible values:

- formerly smoked – the individual has smoked in the past but is no longer a smoker.
- never smoked – the individual has never smoked.
- smokes – the individual is a current smoker.
- Unknown – the smoking history of the individual is not recorded (i.e., missing or unavailable information).

The variable `work_type` describes a patient's type of employment or occupational status. It is a categorical variable with five distinct values:

- Private – employed in the private sector.
- Self-employed – working independently or running their own business.
- Govt_job – employed in government service.
- Never_worked – individuals who have never been employed.
- Children – individuals classified as a dependent and had not entered into the workforce.

The dataset exhibits a significant classification imbalance, with the majority of cases being non-stroke. Specifically, 95.1% (4,861 cases) are non-stroke, while 4.9% (249) represent stroke cases, as shown in Figure 1. This imbalance mirrors real-world scenarios, where stroke events are less frequent but have significant clinical implications.

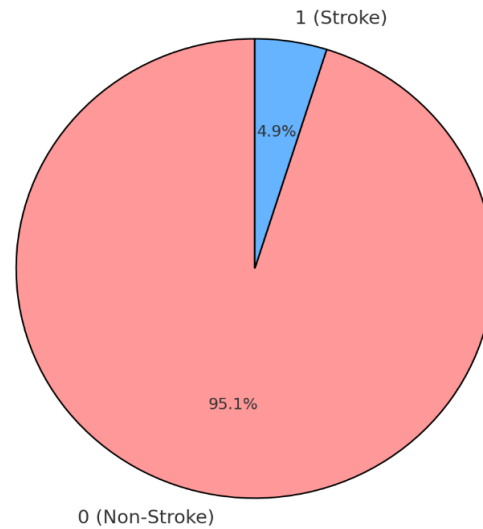


Figure 1: Classification Distribution in Stroke Dataset

Data Preprocessing

To prepare the dataset for model development, several preprocessing steps were implemented. Missing BMI values, which accounted for approximately 4% of the data, were imputed using the mean to maintain consistency. Irrelevant features, such as patient ID, were excluded, while categorical features (e.g., gender, ever_married, work_type, Residence_type, and smoking_status) were label-encoded into binary values for compatibility with the model.

Numerical features were standardized to address discrepancies in magnitude and units, ensuring fair evaluation and preventing data leakage. For example, average glucose levels are in the hundreds, while BMI values are typically in the tens.

Selected Features

Significant independent variables and independent variables deemed significant to stroke prediction but determined not highly correlated to stroke were included in our analysis. Based on the correlation between the target variable (stroke) and the independent variables, as shown in Figure 2 and Figure 4 in Appendix D, we excluded variables such as `work_type`, `Residence_type`, and `gender` from model development because their correlations with the target variable were negligible ($|r| \leq 0.03$; Cohen, 1988) and were not determined to not be significant to the model.

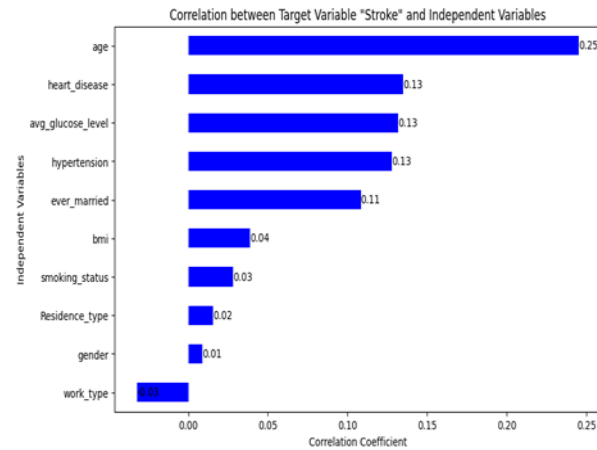


Figure 2: Correlation between Target Variable (Stroke) and Independent Variables

Although the simple correlations between BMI, smoking status, and stroke were negligible in our dataset, we retained these variables in model development for both theoretical and methodological reasons. First, BMI and smoking are widely recognized in the epidemiological literature as important risk factors for stroke (Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration, 2014; Pan et al., 2019), and excluding them could undermine the clinical relevance of our findings. Second, correlation with the outcome alone does not capture the potential contribution of these variables in a multivariate framework, where nonlinear associations or interactions with other predictors may enhance predictive performance (Molnar, 2022). Finally, including BMI and smoking status supports comparability with prior stroke prediction studies, ensuring that our results can be interpreted within the broader body of research.

Model Selection and Development

To evaluate the impact of different resampling strategies on stroke prediction, we implemented a pipeline-based approach that combined resampling, standardization, and classification. Four machine learning classifiers were selected based on their widespread use and demonstrated effectiveness in binary classification tasks. These included Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting Classifier (GB). Together, these models represent a balanced mix of linear and ensemble-based learning algorithms commonly used in healthcare data analysis.

The dataset was split into 70 percent for training

and 30 percent for testing. All resampling techniques and machine learning models were implemented using their default settings, with the random state parameter set to 42 to ensure reproducibility. To ensure reliable performance estimates, we used 5-fold stratified cross-validation during model training. In this approach, the dataset was divided into five equal parts (folds) while preserving the proportion of stroke and non-stroke cases in each fold. For each iteration, four folds were used to train the model and the remaining fold was used for testing. This process was repeated five times, with each fold serving once as the test set. The results from all five iterations were then averaged to produce a more stable and generalizable estimate of model performance.

Cross-validation is particularly important in imbalanced datasets such as stroke prediction, because it prevents performance results from being overly influenced by a single train-test split. Without this procedure, the distribution of minority cases (stroke events) in the test set could vary widely, leading to poor sensitivity and a higher risk of false negatives. By averaging across folds, cross-validation provides a more accurate picture of how often the model is likely to produce false negatives (missed stroke cases, which carry high clinical cost) and false positives (incorrectly flagged non-stroke cases, which increase system burden and unnecessary interventions). This ensures that the evaluation reflects not only statistical performance but also the potential clinical and operational costs associated with deploying the models in practice.

Each classifier was evaluated using a range of data balancing techniques. These included one oversampling method, Synthetic Minority Over-sampling Technique (SMOTE), and several undersampling methods, such as Random Undersampling, Tomek Links, Edited Nearest Neighbors (ENN), and the three NearMiss variants (NearMiss-1, NearMiss-2, and NearMiss-3). Model performance was also evaluated using the original imbalanced dataset, referred to as the "None" configuration, to serve as a baseline.

For every combination of classifier and resampling method, a machine learning pipeline was constructed. The pipeline began with a resampling step (where applicable), followed by feature standardization using the StandardScaler, and concluded with the selected classifier. To ensure a fair and robust assessment of model performance, five-fold stratified cross-validation was employed. This approach maintained the original classification

distribution in each fold, which is particularly important when dealing with imbalanced datasets.

Model performance was evaluated using six metrics: accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. These metrics provided a well-rounded evaluation framework, especially suitable for assessing models trained on imbalanced data. To avoid data leakage, only the test scores from each fold were collected during cross-validation.

After cross-validation, each model pipeline was trained on the full training set and then evaluated on the held-out test set. A confusion matrix was generated for each configuration to capture the counts of true positives, false positives, false negatives, and true negatives. These results were recorded for detailed error analysis.

Finally, predictions from each model-resampling configuration were saved for further statistical analysis and visualizations. Confusion matrices were also plotted to offer a visual understanding of how well each model performed under different resampling scenarios.

This experimental setup provided a consistent and reproducible framework for evaluating how various resampling methods influenced the classification performance of different machine learning models.

Evaluation Metrics

To assess model performance under different resampling strategies, several well-established evaluation metrics were used. These metrics help capture both overall accuracy and the model's ability to correctly identify minority classification instances in an imbalanced dataset (He & Garcia, 2009).

Accuracy

Accuracy measures the proportion of correctly classified instances. Although commonly used, it can be misleading in imbalanced datasets since a model may achieve high accuracy by always predicting the majority classification (Jeni et al., 2013).

Precision (Positive Predictive Value)

Precision calculates the proportion of true stroke cases among all cases predicted as stroke. High precision reflects fewer false positives, which is essential to avoid unnecessary medical interventions (Sokolova & Lapalme, 2009).

Recall (Sensitivity or True Positive Rate)

Also known as sensitivity, recall measures the ability to correctly identify all actual stroke cases. It is critical in healthcare to minimize false negatives, which may result in missed diagnoses (Davis & Goadrich, 2006).

A low recall score indicates that many stroke patients are incorrectly classified as non-stroke, increasing the risk of undiagnosed cases.

F1-Score

The F1-score is the harmonic mean of precision and recall, offering a balanced view when both false positives and false negatives matter. It is particularly useful in imbalanced classification settings (Fernández et al., 2018a).

A high F1-score means that the model effectively balances precision and recall, making it a more informative metric for evaluating stroke prediction performance.

Confusion Matrix

The confusion matrix summarizes predictions into four categories: true positives, true negatives, false positives, and false negatives. It provides a clear view of how the model performs on each classification (Tharwat, 2020; Swaminathan & Tantri, 2024), especially in minimizing false negatives.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) indicates how well the model distinguishes between stroke and non-stroke cases across various threshold settings. It evaluates the trade-off between true positive rate and false positive rate and is widely used in binary classification tasks (Fawcett, 2006; Choi et al., 2024).

Area Under the Precision-Recall Curve (PR-AUC)

The Area Under the Precision-Recall Curve (PR-AUC) focuses on the model's ability to correctly identify stroke cases among all positive predictions. It is particularly useful when dealing with highly imbalanced datasets, where the number of actual positive cases is small (Saito & Rehmsmeier, 2015; Sofaer et al., 2019).

Statistical Significance: t-test

To determine whether the differences in model performance was statistically significant, a paired t-test was conducted to confirm or reject our hypotheses. The t-test evaluates whether the observed performance variations are

attributable to the resampling techniques or occur by chance (Demšar, 2006).

After obtaining the prediction results from each model across different sampling methods, we manually performed a paired t-test in Excel, using the two-sample test assuming equal variances.

The statistical significance level was set at a p-value of less than 0.05, indicating that differences between models are considered statistically significant if the p-value falls below this threshold.

The next section details the results and discussion of this study.

4. RESULTS AND DISCUSSION

This section presents the results of the model evaluations using several key performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. The findings, summarized in Table 2 in Appendix A, highlight how each resampling method affected the performance of the machine learning models on the imbalanced stroke dataset.

Accuracy

Models trained on the original imbalanced dataset (None) with Logistic Regression (0.951), the dataset processed with TomekLinks using Logistic Regression (0.951), and the original imbalanced dataset with Gradient Boosting Classifier (0.951) all achieved the highest accuracy. However, this metric was misleading, since models achieved near-perfect accuracy by predicting the majority classification (non-stroke) while entirely failing to detect stroke cases (Recall = 0.000). This highlights the critical limitation of accuracy as a performance measure in highly imbalanced datasets.

Precision

Precision was highest at 0.413 with Gradient Boosting and no resampling. However, once resampling techniques were introduced, precision dropped significantly. For example, when using NearMiss2 with Random Forest, precision fell to just 0.049. This reflects a common trade-off in resampling: as recall improves, precision tends to decline (Saito & Rehmsmeier, 2015). No single method achieved strong results for both.

Recall

As shown in Table 2, without resampling, recall scores were nearly zero across all models, which

means they failed to identify most stroke cases. The highest recall, 0.905, was achieved by NearMiss2 combined with Random Forest. While this result shows that almost all stroke cases were caught, it came at the cost of many false alarms.

RandomUnder combined with Logistic Regression offered a more balanced approach, reaching a recall of 0.784 and a more moderate precision of 0.131.

F1-Score

F1-scores were generally low across all combinations. The best result came from RandomUnder with Logistic Regression, reaching 0.224. Although SMOTE increased recall for Gradient Boosting up to 0.447, its F1-score remained lower at 0.201. This suggests that RandomUnder offered a better balance between recall and precision in this context.

ROC-AUC

Several models without resampling showed high ROC-AUC values. For example, Gradient Boosting reached 0.839. However, this did not reflect meaningful performance, since the models failed to detect stroke cases. Among the resampled methods, ENN combined with Gradient Boosting achieved the highest ROC-AUC at 0.846. Still, its recall remained low, which reinforces the idea that ROC-AUC can be misleading when working with imbalanced datasets.

PR-AUC

Precision-recall area under the curve scores were low across the board. The highest score was 0.213, recorded by TomekLinks with Gradient Boosting. These low values show how difficult it is to achieve both strong precision and recall when stroke cases are rare.

Precision-Recall Trade-off

Each resampling method showed a clear trade-off between precision and recall. NearMiss2 reached very high recall, such as 0.905 with Random Forest, but suffered a major loss in precision. Non-resampled models had high precision but almost no recall.

RandomUnder combined with Logistic Regression stood out as a reasonable compromise. It offered a recall of 0.784 and a precision of 0.131. SMOTE with Gradient Boosting also performed well in recall at 0.447, though its precision was lower.

Confusion Matrix of a Well-Balanced Model: RandomUnder with Logistic Regression

Table 3 of Appendix B shows the confusion matrix for the Random Undersampling + Logistic Regression configuration, which demonstrated one of the best trade-offs between precision and recall.

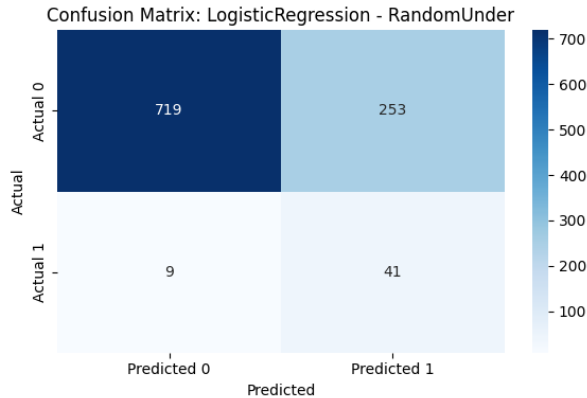


Figure 3: Confusion Matrix of Logistic Regression with Random Undersampling Method

Figure 3 above shows the model correctly identified 41 stroke cases (true positives) while minimizing false negatives (only 9 missed cases). Although 253 non-stroke cases were incorrectly classified as strokes (false positives), the model achieved a strong recall of 0.784, making it a practical choice for screening scenarios where detecting true stroke cases is critical.

Notable Exceptions (p-value > 0.05):

Logistic Regression (LR):

| Model + Method | Model + Method | p-value |
|----------------|----------------|---------|
| LR + None | LR + TL | 1.00 |
| | LR + ENN | 0.76 |
| LR + TL | LR + ENN | 0.76 |
| LR + NM2 | LR + NM3 | 0.78 |

Random Forest (RF):

| Model + Method | Model + Method | p-value |
|----------------|----------------|---------|
| RF + None | RF + TL | 0.84 |
| | RF + ENN | 0.77 |
| | LR + None | 0.84 |
| | LR + TL | 0.84 |
| | LR + ENN | 0.92 |
| RF + RU | LR + RU | 0.21 |
| RF + TL | LR + None | 0.69 |
| | LR + TL | 0.69 |
| | RF + ENN | 0.92 |
| RF + ENN | LR + None | 0.62 |
| | LR + TL | 0.62 |
| | RF + ENN | 0.84 |
| RF + NM3 | LR + NM3 | 0.08 |

Statistical Significance: t-test

Tables 4(a) and 4(b) of Appendix C present the p-values obtained from paired t-tests comparing the accuracy of various machine learning models under different resampling methods for stroke prediction. The majority of the p-values are approximately 0.00, suggesting that differences in model performance are statistically significant across most resampling techniques. However, several exceptions with higher p-values were observed, indicating no statistically significant difference in those specific comparisons.

Extreme Gradient Boosting (XGB):

| Model + Method | Model + Method | p-value |
|----------------|----------------|---------|
| XGB + None | XGB + TL | 0.84 |
| | XGB + ENN | 0.12 |
| | LR + None | 0.55 |
| | LR + TL | 0.55 |
| | LR + ENN | 0.77 |
| XGB + RU | LR + RU | 0.05 |
| | LR + NM2 | 0.15 |
| | LR + NM3 | 0.09 |
| XGB + TL | LR + None | 0.69 |
| | LR + TL | 0.69 |
| | LR + ENN | 0.92 |
| | XGB + ENN | 0.08 |

Gradient Boosting (GB):

| Model + Method | Model + Method | p-value |
|----------------|----------------|---------|
| GB + None | GB + TL | 1.00 |
| | GB + ENN | 0.39 |
| | LR + None | 0.76 |
| | LR + TL | 0.76 |
| GB + RU | LR + ENN | 1.00 |
| | GB + NM3 | 0.92 |
| | LR + NM2 | 0.81 |
| | LR + NM3 | 0.60 |
| GB + TL | LR + None | 0.76 |
| | LR + TL | 0.76 |
| | LR + ENN | 1.00 |
| GB + ENN | GB + ENN | 0.39 |
| | LR + None | 0.24 |
| | LR + TL | 0.24 |
| GB + NM3 | LR + ENN | 0.39 |
| | LR + NM2 | 0.74 |
| | LR + NM3 | 0.54 |

These findings highlight that some resampling strategies yield similar classification performance, especially when applied to models that share similar decision boundary behavior or sensitivity to classification imbalance (e.g., Tomek Links and ENN). These statistically non-significant results offer insight into which

combinations may provide equivalent predictive performance, allowing for flexibility in method selection.

Answering Research Questions

After evaluating individual performance metrics, this study aimed to answer two guiding research questions related to the effectiveness of resampling strategies for stroke prediction.

RQ1: How do SMOTE and selected undersampling techniques compare in improving the performance of machine learning models for stroke prediction using highly imbalanced datasets?

The results across all metrics and models indicate that both SMOTE and undersampling techniques significantly outperformed models trained without resampling. SMOTE was especially effective in boosting recall for ensemble models like Gradient Boosting, while RandomUnder with Logistic Regression achieved the best balance between recall (0.784) and F1-score (0.224). NearMiss2 achieved the highest recall overall (0.905 with Random Forest), but at the cost of extremely low precision. These findings confirm that resampling methods are essential for improving minority-class detection and overall model effectiveness.

RQ2: What are the trade-offs between predictive performance when using SMOTE versus undersampling techniques in highly imbalanced stroke prediction models?

The analysis showed a consistent trade-off between recall and precision across all resampling methods. Techniques like SMOTE and NearMiss2 greatly improved recall but significantly reduced precision, leading to more false positives. Among the resampling techniques, RandomUndersampling offered the best compromise, demonstrating that carefully chosen undersampling methods can improve recall without overwhelming the system with false positives.

The next section provides concluding insights based on the findings and details of the limitations of this study.

5. CONCLUSION AND LIMITATIONS

This study conducted a comprehensive comparative analysis of SMOTE and various undersampling techniques for addressing classification imbalance in stroke prediction using four artificial intelligence (AI)-driven machine learning models: Logistic Regression

(LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting (GB). By evaluating model performance across key AI evaluation metrics—including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC—this study provides actionable insights into how data-level resampling methods influence the reliability and fairness of AI systems in healthcare.

The results demonstrate that all resampling strategies significantly improved the ability of AI models to detect minority stroke cases compared to models trained on imbalanced data. Notably, SMOTE and NearMiss2 achieved high recall rates, enhancing the AI system's sensitivity to stroke events, while Random Undersampling combined with Logistic Regression achieved the most balanced performance. These findings affirm the critical role of resampling in enhancing the trustworthiness and effectiveness of AI-based stroke prediction tools. The statistical significance of most performance differences, confirmed through paired t-tests, further validates the robustness of these observations.

The trade-off between recall and precision observed in the study has important cost implications for stroke prediction. High recall models, such as NearMiss2 with Random Forest, successfully identified nearly all stroke cases but did so at the expense of extremely low precision, resulting in a large number of false positives. In clinical practice, false positives generate unnecessary diagnostic procedures, increase healthcare expenditures, and contribute to patient anxiety. They may also create "alert fatigue" for clinicians, reducing trust in AI systems and limiting adoption. Conversely, models with high precision but low recall risk producing false negatives, where actual stroke cases are missed. This outcome carries an even higher patient cost, as it delays treatment and increases the likelihood of long-term disability or mortality.

From a methodological perspective, this trade-off demonstrates that optimizing for recall alone is insufficient. Instead, model evaluation must consider both types of errors and their asymmetric consequences. Future research should incorporate cost-sensitive learning frameworks that explicitly weight false negatives more heavily, while controlling the operational burden of false positives (Khan et al., 2017). Such approaches would allow stroke prediction models to better align with the realities of

clinical decision-making, where both medical safety and resource efficiency are critical

Despite its contributions, this study has several limitations. First, the stroke dataset provided by Fedesoriano on Kaggle is a publicly available collection of patient-level records designed for predictive modeling of stroke risk. While the dataset offers a useful benchmark for developing and testing machine learning models, it has notable data integrity considerations. For example, there are missing values in the BMI attribute that require imputation before analysis. Furthermore, since the dataset is aggregated and anonymized, there is limited information on its clinical provenance, meaning that while it is suitable for methodological exploration and comparative studies, caution should be exercised in generalizing findings to real-world clinical populations.

Next, the analysis is confined to a single publicly available stroke dataset, which may affect the generalizability of results to other medical conditions or populations. All models and resampling methods were applied using default hyperparameters, suggesting that further tuning could yield even stronger results. Third, while traditional classification metrics were used, the study did not explicitly incorporate fairness, interpretability, or cost-sensitive evaluation—critical considerations for the responsible deployment of AI in clinical settings. Finally, the use of t-tests assuming equal variances may not fully account for dependencies introduced through resampling.

The following section outlines the future research directions based on the findings of this study.

6. FUTURE RESEARCH DIRECTIONS

Future AI-driven research should expand this work by exploring hybrid resampling strategies, integrating fairness-aware and interpretable AI models, tuning hyperparameters, and validating findings across multiple datasets. These enhancements will support the development of more accurate, equitable, and clinically actionable AI systems for early stroke detection and other high-impact healthcare applications.

Specifically, the following avenues should be considered and implemented below:

- **Exploring Hybrid and Ensemble**

Resampling Methods: Combining oversampling and undersampling strategies (e.g., SMOTE-ENN, SMOTE-Tomek) or integrating resampling within ensemble frameworks (e.g., BalancedBagging) may further improve performance.

- **Model Tuning and Optimization:** Future studies should investigate the impact of hyperparameter tuning on both classifiers and resampling methods to optimize performance.
- **Fairness and Interpretability:** Incorporating fairness-aware algorithms and interpretable models is critical, especially when deploying in high-stakes domains like healthcare.
- **Cross-Dataset Evaluation:** To ensure generalizability, testing on multiple stroke or related healthcare datasets from diverse populations would strengthen the findings.
- **Cost-Sensitive Learning:** Integrating cost-sensitive learning approaches could help reduce false negatives while accounting for the asymmetric costs of misclassification in clinical decision-making.

7. REFERENCES

- Aish, M. A., Ghafoor, A. A., Nasim, F., Ali, K. I., Akhter, S., & Azeem, S. (2024). Improving stroke prediction accuracy through machine learning and synthetic minority over-sampling. *Journal of Computing & Biomedical Informatics*, 7(2), 566-0702. <https://doi.org/10.56979/702/2024>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29. <https://doi.org/10.1145/1007730.1007735>
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2), Article 31, 1-50. <https://doi.org/10.1145/2907070>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*,

- 16, 321-357.
<https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
<https://doi.org/10.1145/2939672.2939785>
- Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*, 57, 137. <https://doi.org/10.1007/s10462-024-10759-6>
- Choi, Y. J., Kim, J. H., Lee, S. H., & Park, M. J. (2024). Explainable artificial intelligence for stroke prediction through deep learning and machine learning models. *Scientific Reports*, 14, Article 82931.
<https://doi.org/10.1038/s41598-024-82931-5>
- Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24, 136-158.
<https://doi.org/10.1007/s10618-011-0222-1>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 233-240.
<https://doi.org/10.1145/1143844.1143874>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1-30.
<https://dl.acm.org/doi/10.5555/1248547.1248548>
- Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91, 464-471.
<https://doi.org/10.1016/j.eswa.2017.09.030>
- Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 113(12), 4903-4923.
<https://doi.org/10.1007/s10994-022-06296-4>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.
<https://doi.org/10.1016/j.patrec.2005.10.010>
- Fedesoriano. (n.d.). *Stroke Prediction Dataset*. Kaggle. Retrieved April 21, 2025, from <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- Fernández, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018a). *Learning from Imbalanced Data Sets*. Springer.
- Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018b). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905.
<https://doi.org/10.1613/jair.1.11192>
- Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration. (2014). Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: A pooled analysis of 97 prospective cohorts with 1.8 million participants. *The Lancet*, 383(9921), 970-983. [https://doi.org/10.1016/S0140-6736\(13\)61836-X](https://doi.org/10.1016/S0140-6736(13)61836-X)
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications*, 73, 220-239.
<https://doi.org/10.1016/j.eswa.2016.12.035>
- Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing*, 3644, 878-887. Springer.
https://doi.org/10.1007/11538059_91
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.
<https://doi.org/10.1109/TKDE.2008.239>
- Hoens, T. R., Qian, Q., Chawla, N. V., & Zhou, Z. H. (2012). Building decision trees for the multi-class imbalance problem. In *Pacific-Asia conference on knowledge discovery and data mining*, 122-134. Berlin, Heidelberg: Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-30217-6_11

- Hoens, T. R., & Chawla, N. V. (2013). Imbalanced datasets: from sampling to classifiers. *Imbalanced learning: foundations, algorithms, and applications*, 43-59. <https://doi.org/10.1002/9781118646106.ch3>
- Japkowicz, N. (2000a). The classification imbalance problem: significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence*, 56, 111-117.
- Japkowicz, N. (2000b). Learning from imbalanced data sets: a comparison of various strategies. In *Proceedings of the AAAI 2000 Workshop on Learning from Imbalanced Data Sets*. AAAI Press.
- Japkowicz, N., & Stephen, S. (2002). The classification imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429-449. <https://doi.org/10.3233/IDA-2002-6504>
- Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data—recommendations for the use of performance metrics. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245-251. <https://doi.org/10.1109/ACII.2013.47>
- Kamiri, J., & Mariga, G. W. (2021). Research methods in machine learning: a content analysis. *International Journal of Computer and Information Technology*, 10(2), 78-84. <https://doi.org/10.24203/ijcit.v10i2.79>
- Khan, S. H., Hayat, M., Bennamoun, M., Soheli, F. A., & Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8), 3573-3587. <https://doi.org/10.1109/TNNLS.2017.2732482>
- Laurikkala, J. (2001). Improving identification of difficult small classes by balancing classification distribution. In *Proceedings of the 8th Conference on AI in Medicine in Europe*, 2101, 63-66. Springer. https://doi.org/10.1007/3-540-48229-6_9
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17), 1-5. <http://jmlr.org/papers/v18/16-365.html>
- Lin, CH., Chen, YA., Jeng, JS., Sun, Y., Wei, CY., Yeh, PY., Chang, WL., Fann, YC., Hsu, KC., Lee, JT., & Taiwan Stroke Registry Investigators. (2024). Predicting ischemic stroke patients' prognosis changes using machine learning in a nationwide stroke registry. *Medical & Biological Engineering & Computing*, 62, 2343-2354. <https://doi.org/10.1007/s11517-024-03073-4>
- López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250, 113-141. <https://doi.org/10.1016/j.ins.2013.07.007>
- Mani, I., & Zhang, I. (2003). kNN approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*, 126, 1-7. ICML.
- Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Pan, B., Jin, X., Jun, L., Qiu, S., Zheng, Q., & Pan, M. (2019). The relationship between smoking and stroke: A meta-analysis. *Medicine*, 98(12), e14872. <https://doi.org/10.1097/MD.00000000000014872>
- Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291, 184-203. <https://doi.org/10.1016/j.ins.2014.08.051>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57, 273. <https://doi.org/10.1007/s10462-024-10884-2>

- Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4), 565-577. <https://doi.org/10.1111/2041-210X.13140>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: a review. *International journal of pattern recognition and artificial intelligence*, 23(04), 687-719. <https://doi.org/10.1142/S0218001409007326>
- Swaminathan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, 27(4s), 4023-4031. <https://doi.org/10.53555/AJBR.v27i4S.4345>
- Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168-192. <https://doi.org/10.1016/j.aci.2018.08.003>
- Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(11), 769-772. <http://dx.doi.org/10.1109/TSMC.1976.4309452>
- Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3), 408-421. <https://doi.org/10.1109/TSMC.1972.4309137>
- Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3), Part 1, 5718-5727. <https://doi.org/10.1016/j.eswa.2008.06.108>

APPENDIX A

| Method | Model | *Accuracy | *Precision | *Recall | *F1 | *ROC-AUC | *PR-AUC |
|----------------------|---------|--------------|--------------|--------------|--------------|--------------|--------------|
| No Resampling | LR | 0.951 | 0.000 | 0.000 | 0.000 | 0.838 | 0.187 |
| | RF | 0.950 | 0.040 | 0.005 | 0.009 | 0.813 | 0.171 |
| | XGBoost | 0.941 | 0.217 | 0.075 | 0.111 | 0.808 | 0.173 |
| | GB | 0.951 | 0.413 | 0.025 | 0.046 | 0.839 | 0.201 |
| SMOTE | LR | 0.779 | 0.122 | 0.568 | 0.200 | 0.777 | 0.138 |
| | RF | 0.899 | 0.147 | 0.221 | 0.176 | 0.791 | 0.146 |
| | XGBoost | 0.899 | 0.134 | 0.196 | 0.159 | 0.771 | 0.135 |
| | GB | 0.828 | 0.130 | 0.447 | 0.201 | 0.778 | 0.149 |
| RandomUnder | LR | 0.735 | 0.131 | 0.784 | 0.224 | 0.833 | 0.187 |
| | RF | 0.705 | 0.119 | 0.804 | 0.209 | 0.826 | 0.159 |
| | XGBoost | 0.715 | 0.121 | 0.774 | 0.209 | 0.804 | 0.147 |
| | GB | 0.704 | 0.120 | 0.794 | 0.208 | 0.820 | 0.183 |
| TomekLinks | LR | 0.951 | 0.000 | 0.000 | 0.000 | 0.838 | 0.186 |
| | RF | 0.950 | 0.067 | 0.005 | 0.009 | 0.808 | 0.176 |
| | XGBoost | 0.941 | 0.246 | 0.090 | 0.130 | 0.815 | 0.173 |
| | GB | 0.950 | 0.390 | 0.040 | 0.072 | 0.840 | 0.213 |
| ENN | LR | 0.946 | 0.192 | 0.030 | 0.051 | 0.840 | 0.192 |
| | RF | 0.938 | 0.158 | 0.050 | 0.074 | 0.830 | 0.182 |
| | XGBoost | 0.920 | 0.204 | 0.206 | 0.204 | 0.822 | 0.171 |
| | GB | 0.934 | 0.257 | 0.136 | 0.172 | 0.846 | 0.202 |
| NearMiss1 | LR | 0.437 | 0.051 | 0.553 | 0.092 | 0.503 | 0.066 |
| | RF | 0.196 | 0.041 | 0.688 | 0.077 | 0.432 | 0.054 |
| | XGBoost | 0.206 | 0.042 | 0.699 | 0.079 | 0.440 | 0.050 |
| | GB | 0.183 | 0.043 | 0.739 | 0.081 | 0.384 | 0.039 |
| NearMiss2 | LR | 0.663 | 0.103 | 0.768 | 0.182 | 0.775 | 0.139 |
| | RF | 0.143 | 0.049 | 0.905 | 0.093 | 0.664 | 0.098 |
| | XGBoost | 0.104 | 0.047 | 0.905 | 0.089 | 0.685 | 0.131 |
| | GB | 0.112 | 0.047 | 0.900 | 0.090 | 0.498 | 0.073 |
| NearMiss3 | LR | 0.707 | 0.101 | 0.633 | 0.174 | 0.750 | 0.142 |
| | RF | 0.655 | 0.084 | 0.617 | 0.147 | 0.670 | 0.114 |
| | XGBoost | 0.654 | 0.084 | 0.613 | 0.148 | 0.676 | 0.120 |
| | GB | 0.673 | 0.088 | 0.608 | 0.153 | 0.710 | 0.128 |

Table 2: Results of Machine Learning Models under Different Resampling Methods for Stroke Prediction

APPENDIX B

| Method | Model | True-Negative | False-Positive | False-Negative | True-Positive |
|--------------------|---------|---------------|----------------|----------------|---------------|
| None | LR | 972 | 0 | 50 | 0 |
| | RF | 970 | 2 | 50 | 0 |
| | XGBoost | 960 | 12 | 44 | 6 |
| | GB | 968 | 4 | 49 | 1 |
| SMOTE | LR | 767 | 205 | 14 | 36 |
| | RF | 906 | 66 | 41 | 9 |
| | XGBoost | 912 | 60 | 43 | 7 |
| | GB | 816 | 156 | 27 | 23 |
| RandomUnder | LR | 719 | 253 | 9 | 41 |
| | RF | 695 | 277 | 10 | 40 |
| | XGBoost | 685 | 287 | 14 | 36 |
| | GB | 657 | 315 | 11 | 39 |
| TomekLinks | LR | 971 | 1 | 49 | 1 |
| | RF | 968 | 4 | 50 | 0 |
| | XGBoost | 961 | 11 | 43 | 7 |
| | GB | 967 | 5 | 48 | 2 |
| ENN | LR | 964 | 8 | 45 | 5 |
| | RF | 955 | 17 | 38 | 12 |
| | XGBoost | 935 | 37 | 36 | 14 |
| | GB | 952 | 20 | 42 | 8 |
| NearMiss1 | LR | 340 | 632 | 22 | 28 |
| | RF | 131 | 841 | 11 | 39 |
| | XGBoost | 124 | 848 | 5 | 45 |
| | GB | 130 | 842 | 10 | 40 |
| NearMiss2 | LR | 652 | 320 | 11 | 39 |
| | RF | 91 | 881 | 3 | 47 |
| | XGBoost | 61 | 911 | 1 | 49 |
| | GB | 64 | 908 | 4 | 46 |
| NearMiss3 | LR | 656 | 316 | 21 | 29 |
| | RF | 619 | 353 | 22 | 28 |
| | XGBoost | 648 | 324 | 25 | 25 |
| | GB | 669 | 303 | 21 | 29 |

Table 3: Confusion Matrix Results of Machine Learning Models under Different Resampling Methods for Stroke Prediction

APPENDIX C

| Model + Method | LR + SMOTE | LR + RU | LR + TL | LR + ENN | LR + NM1 | LR + NM2 | LR + NM3 |
|-----------------------|--------------------|-----------------|-----------------|------------------|------------------|------------------|------------------|
| LR + None | 0.00* | 0.00* | 1.00 | 0.76 | 0.00* | 0.00* | 0.00* |
| LR + SMOTE | | 0.02* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| LR + RU | | | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| LR + TL | | | | 0.76 | 0.00* | 0.00* | 0.00* |
| LR + ENN | | | | | 0.00* | 0.00* | 0.00* |
| LR + NM1 | | | | | | 0.00* | 0.00* |
| LR + NM2 | | | | | | | 0.78 |
| Model + Method | RF + SMOTE | RF + RU | RF + TL | RF + ENN | RF + NM1 | RF + NM2 | RF + NM3 |
| RF + None | 0.00* | 0.00* | 0.84 | 0.77 | 0.00* | 0.00* | 0.00* |
| RF+ SMOTE | | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| RF + RU | | | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| RF + TL | | | | 0.92 | 0.00* | 0.00* | 0.00* |
| RF + ENN | | | | | 0.00* | 0.00* | 0.00* |
| RF + NM1 | | | | | | 0.048* | 0.00* |
| RF + NM2 | | | | | | | 0.00* |
| Model + Method | XGB + SMOTE | XGB + RU | XGB + TL | XGB + ENN | XGB + NM1 | XGB + NM2 | XGB + NM3 |
| XGB + None | 0.00* | 0.00* | 0.84 | 0.12 | 0.00* | 0.00* | 0.00* |
| XGB + SMOTE | | 0.00* | 0.00* | 0.02* | 0.00* | 0.00* | 0.00* |
| XGB + RU | | | 0.00* | 0.00* | 0.00* | 0.00* | 0.02* |
| XGB + TL | | | | 0.08 | 0.00* | 0.00* | 0.00* |
| XGB + ENN | | | | | 0.00* | 0.00* | 0.00* |
| XGB + NM1 | | | | | | 0.00* | 0.00* |
| XGB + NM2 | | | | | | | 0.00* |
| Model + Method | GB + SMOTE | GB + RU | GB + TL | GB + ENN | GB + NM1 | GB + NM2 | GB + NM3 |
| GB + None | 0.00* | 0.00* | 1.00 | 0.39 | 0.00* | 0.00* | 0.00* |
| GB + SMOTE | | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| GB + RU | | | 0.00* | 0.00* | 0.00* | 0.00* | 0.92 |
| GB + TL | | | | 0.39 | 0.00* | 0.00* | 0.00* |
| GB + ENN | | | | | 0.00* | 0.00* | 0.00* |
| GB + NM1 | | | | | | 0.00* | 0.00* |
| GB + NM2 | | | | | | | 0.00* |

Table 4(a): P-Value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

t-test: Two-sample Assuming Equal Variances

p-value* < 0.05: Significant difference between pair of Model + Method

| Model + Method | LR + None | LR + SMOTE | LR + RU | LR + TL | LR + ENN | LR + NM1 | LR + NM2 | LR + NM3 |
|----------------|--------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|
| RF + None | 0.84 | 0.00* | 0.00* | 0.84 | 0.92 | 0.00* | 0.00* | 0.00* |
| RF+ SMOTE | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| RF + RU | 0.00* | 0.00* | 0.21 | 0.00* | 0.00* | 0.00* | 0.03* | 0.02* |
| RF + TL | 0.69 | 0.00* | 0.00* | 0.69 | 0.92 | 0.00* | 0.00* | 0.00* |
| RF + ENN | 0.62 | 0.00* | 0.00* | 0.62 | 0.84 | 0.00* | 0.00* | 0.00* |
| RF + NM1 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| RF + NM2 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| RF + NM3 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.04* | 0.08 |
| XGB + None | 0.55 | 0.00* | 0.00* | 0.55 | 0.77 | 0.00* | 0.00* | 0.00* |
| XGB + SMOTE | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| XGB + RU | 0.00* | 0.00* | 0.05 | 0.00* | 0.00* | 0.00* | 0.15 | 0.09 |
| XGB + TL | 0.69 | 0.00* | 0.00* | 0.69 | 0.92 | 0.00* | 0.00* | 0.00* |
| XGB + ENN | 0.03* | 0.00* | 0.00* | 0.03* | 0.07 | 0.00* | 0.00* | 0.00* |
| XGB + NM1 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| XGB + NM2 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| XGB + NM3 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.40 | 0.57 |
| GB + None | 0.76 | 0.00* | 0.00* | 0.76 | 1.00 | 0.00* | 0.00* | 0.00* |
| GB + SMOTE | 0.00* | 0.045* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| GB + RU | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.81 | 0.60 |
| GB + TL | 0.76 | 0.00* | 0.00* | 0.76 | 1.00 | 0.00* | 0.00* | 0.00* |
| GB + ENN | 0.24 | 0.00* | 0.00* | 0.24 | 0.39 | 0.00* | 0.00* | 0.00* |
| GB + NM1 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| GB + NM2 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* |
| GB + NM3 | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.00* | 0.74 | 0.54 |

Table 4(b): P-Value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

t-test: Two-sample Assuming Equal Variances

p-value* < 0.05: Significant difference between pair of Model + Method

APPENDIX D

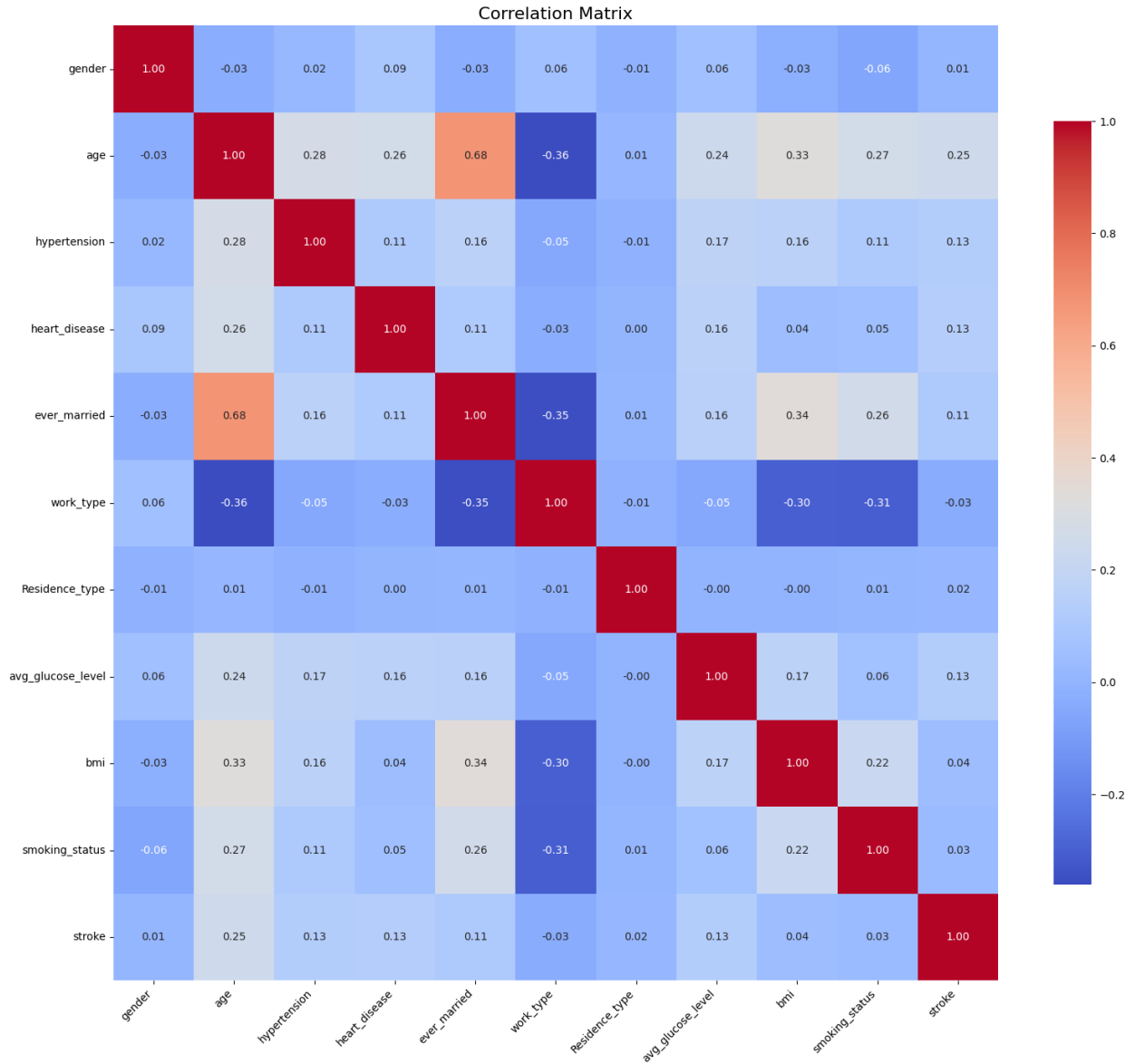


Figure 4: Correlation Matrix

APPENDIX E

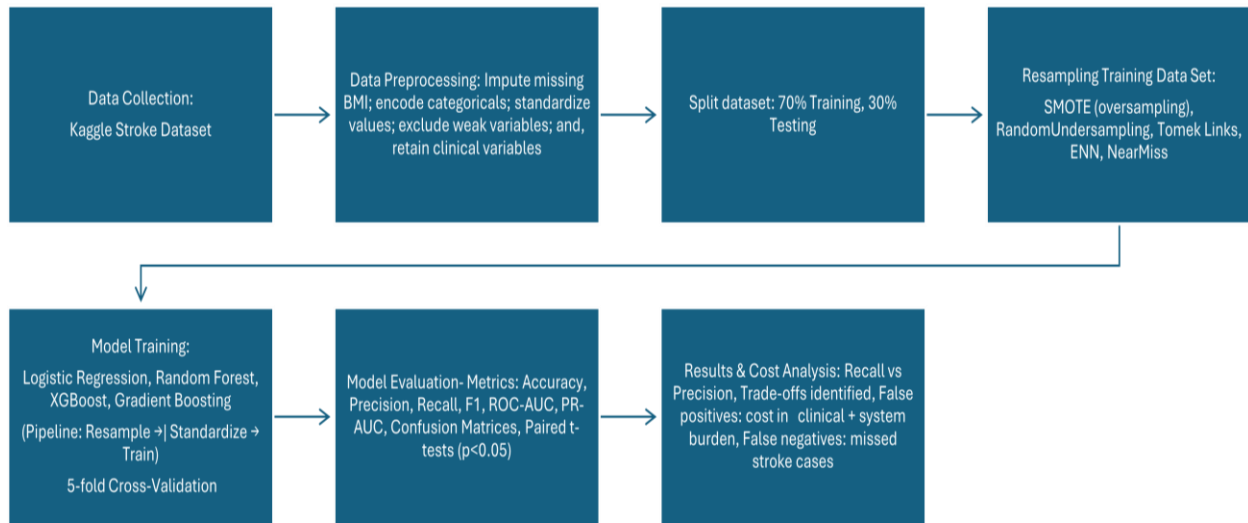
| Model + Method | LR + SMOTE | LR + RU | LR + TL | LR + ENN | LR + NM1 | LR + NM2 | LR + NM3 |
|-----------------------|--------------------|-----------------|-----------------|------------------|------------------|------------------|------------------|
| LR + None | 11.40 | 13.61 | 0.00 | 0.30 | 35.88 | 17.05 | 17.35 |
| LR + SMOTE | | 2.24 | -11.40 | -11.13 | 21.54 | 5.63 | 5.91 |
| LR + RU | | | -13.61 | -13.34 | 18.89 | 3.37 | 3.65 |
| LR + TL | | | | 0.30 | 35.88 | 17.05 | 17.35 |
| LR + ENN | | | | | 35.54 | 16.78 | 17.08 |
| LR + NM1 | | | | | | -15.06 | -14.75 |
| LR + NM2 | | | | | | | 0.28 |
| Model + Method | RF + SMOTE | RF + RU | RF + TL | RF + ENN | RF + NM1 | RF + NM2 | RF + NM3 |
| RF + None | 4.56 | 14.69 | 0.20 | 0.30 | 57.85 | 64.02 | 19.06 |
| RF+ SMOTE | | 10.35 | -4.37 | -4.27 | 48.32 | 52.94 | 14.67 |
| RF + RU | | | -14.51 | -14.42 | 30.27 | 33.06 | 4.18 |
| RF + TL | | | | 0.10 | 57.43 | 63.53 | 18.89 |
| RF + ENN | | | | | 57.23 | 63.29 | 18.80 |
| RF + NM1 | | | | | | 1.98 | -24.49 |
| RF + NM2 | | | | | | | -26.93 |
| Model + Method | XGB + SMOTE | XGB + RU | XGB + TL | XGB + ENN | XGB + NM1 | XGB + NM2 | XGB + NM3 |
| XGB + None | 3.89 | 15.03 | -0.20 | 1.55 | 57.20 | 69.61 | 17.42 |
| XGB + SMOTE | | 11.33 | -4.08 | -2.37 | 49.04 | 58.54 | 13.69 |
| XGB + RU | | | -15.21 | -13.62 | 29.35 | 34.66 | 2.28 |
| XGB + TL | | | | 1.74 | 57.06 | 70.18 | 17.59 |
| XGB + ENN | | | | | 53.95 | 65.10 | 15.99 |
| XGB + NM1 | | | | | | 3.81 | -26.16 |
| XGB + NM2 | | | | | | | -31.07 |
| Model + Method | GBC + SMOTE | GBC + RU | GBC + TL | GBC + ENN | GBC + NM1 | GBC + NM2 | GBC + NM3 |
| GB + None | 9.18 | 16.54 | 0.00 | 0.86 | 57.64 | 70.48 | 16.44 |
| GB + SMOTE | | 7.41 | -9.18 | -8.38 | 39.13 | 46.23 | 7.31 |
| GB + RU | | | -16.54 | -15.76 | 27.57 | 32.73 | -0.09 |
| GB + TL | | | | 0.86 | 57.64 | 70.48 | 16.44 |
| GB + ENN | | | | | 55.84 | 67.93 | 15.66 |
| GB + NM1 | | | | | | 3.87 | -27.70 |
| GB + NM2 | | | | | | | -32.88 |

Table 5(a): t-stat value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

| Model + Method | LR + None | LR + SMOTE | LR + RU | LR + TL | LR + ENN | LR + NM1 | LR + NM2 | LR + NM3 |
|-----------------------|------------------|-------------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|
| RF + None | -0.20 | 11.22 | 13.43 | -0.20 | 0.10 | 35.65 | 16.87 | 17.17 |
| RF+ SMOTE | -4.76 | 6.84 | 9.09 | -4.76 | -4.47 | 30.04 | 12.52 | 12.82 |
| RF + RU | -14.86 | -3.49 | -1.25 | -14.86 | -14.60 | 17.45 | 2.12 | 2.40 |
| RF + TL | -0.40 | 11.04 | 13.26 | -0.40 | -0.10 | 35.42 | 16.70 | 16.99 |
| RF + ENN | -0.50 | 10.95 | 13.17 | -0.50 | -0.20 | 35.31 | 16.61 | 16.91 |
| RF + NM1 | -58.27 | -35.72 | -32.14 | -58.27 | -57.64 | -10.19 | -27.24 | -26.85 |
| RF + NM2 | -64.52 | -38.93 | -35.07 | -64.52 | -63.78 | -12.20 | -29.84 | -29.42 |
| RF + NM3 | -19.24 | -7.71 | -5.43 | -19.24 | -18.98 | 12.82 | -2.05 | -1.76 |
| XGB + None | -0.60 | 10.86 | 13.08 | -0.60 | -0.30 | 35.19 | 16.52 | 16.82 |
| XGB + SMOTE | -4.47 | 7.13 | 9.37 | -4.47 | -4.18 | 30.40 | 12.81 | 13.11 |
| XGB + RU | -15.56 | -4.18 | -1.93 | -15.56 | -15.30 | 16.67 | 1.44 | 1.72 |
| XGB + TL | -0.40 | 11.04 | 13.26 | -0.40 | -0.10 | 35.42 | 16.70 | 16.99 |
| XGB + ENN | -2.14 | 9.42 | 11.66 | -2.14 | -1.84 | 33.35 | 15.10 | 15.40 |
| XGB + NM1 | -58.44 | -35.81 | -32.23 | -58.44 | -57.81 | -10.25 | -27.31 | -26.92 |
| XGB + NM2 | -71.37 | -42.14 | -37.96 | -71.37 | -70.48 | -14.12 | -32.36 | -31.93 |
| XGB + NM3 | -17.94 | -6.48 | -4.22 | -17.94 | -17.68 | 14.13 | -0.84 | -0.56 |
| GB + None | -0.30 | 11.13 | 13.34 | -0.30 | 0.00 | 35.54 | 16.78 | 17.08 |
| GB + SMOTE | -9.45 | 2.00 | 4.25 | -9.45 | -9.18 | 23.97 | 7.65 | 7.94 |
| GB + RU | -16.80 | -5.39 | -3.13 | -16.80 | -16.54 | 15.33 | 0.24 | 0.52 |
| GB + TL | -0.30 | 11.13 | 13.34 | -0.30 | 0.00 | 35.54 | 16.78 | 17.08 |
| GB + ENN | -1.17 | 10.34 | 12.57 | -1.17 | -0.86 | 34.53 | 16.01 | 16.31 |
| GB + NM1 | -58.27 | -35.72 | -32.14 | -58.27 | -57.64 | -10.19 | -27.24 | -26.85 |
| GB + NM2 | -71.37 | -42.14 | -37.96 | -71.37 | -70.48 | -14.12 | -32.36 | -31.93 |
| GB + NM3 | -16.70 | -5.29 | -3.04 | -16.70 | -16.44 | 15.43 | 0.33 | 0.61 |

Table 5(b): T-stat value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

APPENDIX F: Process Model for Stroke Prediction Analysis



The workflow includes data collection, preprocessing, resampling, model training with 5-fold stratified cross-validation, and multi-metric evaluation. Results are analyzed to highlight the trade-offs between precision and recall, including the practical costs of false positives and false negatives.

What Agile Skills Are Needed for IT Jobs?

Stephanie Totty
stephanie.totty@mtsu.edu
Middle Tennessee State University
Murfreesboro, TN 37132

Carol Clark
Carol.Clark@mtsu.edu
Middle Tennessee State University
Murfreesboro, TN 37132

Amy Harris
amy.harris@mtsu.edu
Middle Tennessee State University
Murfreesboro, TN 37132

Abstract

Careers in information technology (IT) require a myriad of skills and training. One important area of requirements is agile skills. Agile skills are the expertise that IT professionals use in an agile development environment. Examples of agile skills include sprint planning, user story creation, product backlog maintenance, and Scrum. It is assumed that regardless of IT area, the need for agile skills is essential in today's business environment. Given the importance of agile skills as part of the IS landscape and, thus, their importance to IS academic community, this research seeks to provide a deeper understanding of current industry needs for agile job skills. This is accomplished through the analysis of over 775,000 job postings referencing agile job titles and skills. Specifically, it seeks to answer the following questions: (1) What are the most common job titles associated with agile skills? (2) What skills are most associated with those job titles? and (3) How do the common and specialized skill requirements vary across job titles? Data was collected from job postings in the U.S. from July 2023 to June 2024. This study provides insights to information systems educators by 1) helping them guide their students during the job placement process, specifically when students express interest in agile, and 2) guiding curriculum design.

Keywords: Agile, Job Titles, Job Skills, IT Skills, Soft Skills, Technical Skills

Recommended Citation: Totty, S., Clark, C., Harris, A., (2026). What Agile Skills Are Needed for IT Jobs? *Journal of Information Systems Applied Research and Analytics*, v19(n2) pp 71-81. DOI# <https://doi.org/10.62273/FRGW2108>

What Agile Skills Are Needed for IT Jobs

Stephanie Totty, Carol Clark and Amy Harris

1. INTRODUCTION

Careers in information technology (IT) require a myriad of skills and training. These requirements are continually changing and should be monitored by those preparing students to enter the IT job field, those helping IT professionals to stay current, and IT professionals seeking to maintain credentials and/or successfully progress in their IT career.

One important area of requirements is agile job skills (e.g., Flynn, 2023; Sharp & Lang, 2018). It is assumed that regardless of IT area, the need for agile methodology skills (hereafter called agile skills for brevity) is essential in today's business environment. Agile skills are the expertise that IT professionals use in an agile development environment. Examples of agile skills include sprint planning, user story creation, product backlog maintenance, and Scrum. The approaches for software projects vary across companies. Many companies use waterfall, agile, and hybrid project approaches. The waterfall method is known as the long-standing traditional method. Agile methods originated for software projects in 2001 (Highsmith, 2001). When companies use hybrid project approaches, employees must be skilled in both waterfall and agile. So, the need for agile job skills is relevant even if a company is not developing IT projects using agile exclusively.

One such area where agile skills are important is in business analysis. Business analysis, often conducted by business analysts, includes problem definition, coordination of business value expectations, and solution development. According to the professional organization International Institute of Business Analysis (IIBA; 2024), 74% of the respondents said that they practice agile methodologies in their current role. IIBA also offers the Agile Analysis Certification that "meets the rising demand for collaboration between analyst and agile communities, enhancing project delivery" (IIBA, 2024, p. 14), another indicator of the importance of agile skills in business analyst jobs.

Given the importance of agile skills as part of the IS landscape and, thus, their importance to

the IS academic community, this research seeks to provide a deeper understanding of current industry needs for agile job skills. This is accomplished through the analysis of over 775,000 job postings referencing agile job titles and skills. Specifically, it seeks to answer the following questions:

- (1) What are the most common job titles associated with agile skills?
- (2) What skills are most associated with those job titles?
- (3) How do the common and specialized skill requirements vary across job titles?

The answers to these questions will benefit information systems (IS) educators. This study provides insights into job requirements for students who are interested in an agile work environment. It helps IS educators direct their students during the job placement process. Additionally, the results can guide IS curriculum development relating to agile topics.

2. LITERATURE REVIEW

Agile is a software development methodology that focuses on

- "Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan" (Beck et al., 2001, para. 2)

Agile came about because many software developers sought an alternative to "documentation driven, heavyweight software development processes" (Highsmith, 2001, para. 1). One key difference between agile and traditional approaches is that "planning efforts with the agile approaches are done more often and in an iterative manner" (Fernandez & Fernandez, 2008, p. 13). Another focus of the agile approach is its collaborative nature. "In agile environments, business analysis fosters collaboration, enables change, and ensures the software being developed effectively addresses business needs and delivers value to stakeholders." (IIBA, 2024, p. 22).

Over the years, agile has become a major software development methodology. Seventy-one percent of companies use agile exclusively or in combination with other methodologies (Flynn, 2023). Along with the popularity and growth in the agile development methodology, companies seek employees with agile skills.

Prior research has investigated several aspects of agile skills. According to Gartner, several skills—include core skills (Scrum, Kanban, metrics, user stories, customer focus, test-first, and continuous learning), value-added skills (collaborative development, ownership and collaboration, agile architecture), and specialized/emerging skills (agile database management and scaling agile)—are essential for agile application development teams (Rimol, 2022). When answering what skills are key to working with agile frameworks, one study reported that a higher skill level is needed for collaboration, effective communication, and teamwork (Cornide-Reyes et al., 2021). Further, the study reported that technical skills are required to work with agile frameworks, but technical skills are not required to “perform well on an agile team” (Cornide-Reyes et al., 2021, p. 84731).

Individuals benefit from having agile skills and working in agile environments. IT professionals with agile skills receive an estimated 22.6% higher salaries on average (Dattero et al., 2016). Furthermore, people in organizations involved in agile processes have higher job satisfaction (Huck-Fries et al., 2023).

Prior studies analyzing job postings have identified agile skills as needed for various IT career paths (e.g., Booker et al, 2024; Chumwatana & Hpone, 2025; Gunklach et al., 2025; Joshua et al., 2023; Siswipraptini et al., 2023). However, there is a dearth of research studies focused expressly on the job titles most associated with agile skills and identifying the most in-demand skills associated with those titles. While prior research on agile in the education context exists, much of the research focuses on *how* agile is taught (e.g., Castro Lopes & Fernandes, 2024; Marnewick, 2023; Sharp et al., 2020; Woods & Hulshult, 2024) rather than *what* should be taught about agile. Thus, the research presented here offers an important extension to existing research, specifically as it relates to IS education and curriculum development in an ever-changing industry environment.

3. METHODS

Since job postings are primarily comprised of the skills an ideal hire would possess, they provide a valuable data source for identifying current talent requirements. While they do not serve as a leading indicator of future in-demand skills, they do indicate current demands and are a useful tool for identifying near-term needs. We investigate our research questions using job postings data from Lightcast (Lightcast, 2025a). Lightcast regularly scrapes and stores job postings data from vetted online job posting sources. Because Lightcast scrapes job postings repeatedly from the same sources and because jobs are often posted on multiple sites, Lightcast deduplicates the data by making efforts to ensure 1) a posting is not collected again from the same source and 2) a posting listed on more than one site is only included from the original site based on fields such as company name, location, job title, and text similarity of the job posting (Harris & Clark, 2023). Lightcast also removes outliers and bad data from the dataset, such as “Postings where the employee must invest their own money” (Lightcast, 2025b, Data Curation section) and “Pyramid schemes/MLM postings” (Lightcast, 2025b, Data Curation section).

| Skill Type | Skill Type Definition |
|---------------------------|--|
| Specialized Skills | Skills that are primarily required within a subset of occupations or equip one to perform a specific task (e.g. “NumPy” or “Hotel Management”). Also known as technical skills or hard skills. |
| Common Skills | Skills that are prevalent across many different occupations and industries, including both personal attributes and learned skills. (e.g. “Communication” or “Microsoft Excel”). Also known as soft skills, human skills, and competencies. |
| Software Skills | Any software tool or programming component used to help with a job (e.g. Python, Workday, AutoCAD, Microsoft Excel, React.js, Accounting Software, and 3D Modeling Software would all be considered “Software Skills”). |

Table 1: Skill type names and definitions (Lightcast, 2025c).

After cleaning the data, Lightcast programmatically codes the text of each job posting into several fields—including each of the

skills listed in the job posting—that can be used for more detailed analysis (Lightcast, 2025b). For this coding, Lightcast disambiguates homonyms using a minimum threshold of complimentary words (Harris & Clark, 2023). The skills are mapped to Lightcast’s skill taxonomy that categorizes identified skills as specialized skills, common skills, software skills, and certifications. These skill categories are defined in Table 1.

| Skill | Skill Type |
|-----------------------------------|-------------------|
| Oracle Agile | Specialized Skill |
| Agile Auditing | Specialized Skill |
| SAFe Agile | Specialized Skill |
| Agile Methodology | Specialized Skill |
| Agile Leadership | Specialized Skill |
| Agile Coaching | Specialized Skill |
| Agile Certification | Certification |
| Agile Projects | Specialized Skill |
| Agile Management | Specialized Skill |
| Agile Modeling | Specialized Skill |
| Agile Testing | Specialized Skill |
| Agile Product Management | Specialized Skill |
| Agile Software Development | Specialized Skill |
| Agile Unified Process | Specialized Skill |
| Agile Product Development | Specialized Skill |
| Agile Project Management | Specialized Skill |
| Disciplined Agile Delivery | Specialized Skill |
| Scaled Agile Framework | Specialized Skill |
| IIBA Agile Analysis Certification | Certification |
| PMI Agile Certified Practitioner | Certification |
| Large Scale Agile Development | Specialized Skill |
| Agile Model Driven Development | Specialized Skill |

Table 2: Agile skills included in our search.

Using Lightcast’s web application, we ran the Job Postings Analytics report for postings marketed to the United States between July 2023 and June 2024. We searched Lightcast’s skills taxonomy using the keyword “agile”. After reviewing the skills that resulted from the search, we further refined our search to exclude skills not directly related to the agile methodology. Specifically, we removed the following skills from our search: “Agility”, “Change Agility”, “Learning Agility”, “Mental Agility”, “Agilent ChemStation”, and “Agilent VEE (Domain-Specific Programming Language)”. Table 2 shows the remaining skills included in our results.

Using the agile job postings that resulted from the search, we grouped the postings by job title

to identify the job titles most commonly seeking agile skills. We then further investigated the top 10 job titles by job posting volume, focusing on which specialized and common skills were most frequently mentioned for each title.

4.RESULTS

Our search resulted in 776,321 unique job postings referencing at least one of the agile skills and certifications and targeting the US between July 2023 and June 2024.

RQ1: What are the most common job titles associated with agile skills?

| Rank | Job Title | Unique Postings |
|------|-------------------------------|-----------------|
| 1 | Software Engineers | 24,775 |
| 2 | Project Managers | 12,299 |
| 3 | Business Analysts | 10,972 |
| 4 | Scrum Masters | 8,508 |
| 5 | Product Managers | 7,464 |
| 6 | Software Developers | 6,976 |
| 7 | Product Owners | 6,703 |
| 8 | Data Engineers | 6,555 |
| 9 | Java Developers | 5,851 |
| 10 | Full Stack Developers | 5,771 |
| 11 | DevOps Engineers | 5,486 |
| 12 | IT Project Managers | 5,312 |
| 13 | Program Managers | 4,978 |
| 14 | Principal Software Engineers | 4,886 |
| 15 | Full Stack Software Engineers | 4,705 |
| 16 | Systems Engineers | 4,510 |
| 17 | Business Systems Analysts | 4,449 |
| 18 | .NET Developers | 4,127 |
| 19 | Solutions Architects | 4,017 |
| 20 | Lead Software Engineers | 3,534 |
| 21 | Software Engineering Managers | 3,442 |
| 22 | Technical Project Managers | 3,290 |
| 23 | Salesforce Developers | 3,008 |
| 24 | Full Stack Engineers | 2,908 |
| 25 | Full Stack Java Developers | 2,766 |

Table 3: Top job titles for job postings referencing agile skills by posting volume.

Table 3 lists the top job titles referencing agile skills by posting volume. These job titles accounted for 27.18% of all job postings and 20.26% of job postings referencing agile skills. The job title “Software Engineers” accounted for

11.74% (24,775) of job postings referencing agile skills. Of the top 25 job titles, none of the job titles contained "agile" in the job title. Notably, 15 job titles (42.32% of agile postings) in the top 25 job titles include "engineer" or "developer". Using the agile job postings that resulted from the search, we grouped the postings by job title to identify the job titles most commonly seeking agile skills. We then further investigated the top 10 job titles by job posting volume, focusing on which specialized and common skills were most frequently mentioned for each title.

RQ2: What skills are most associated with those job titles?

Table 4 (see Appendix) lists the specialized skills most frequently appearing in job postings for the ten most common job titles ranked by frequency of appearance. Due, in part, to the method used, "Agile Methodology" was the first or second most frequently mentioned specialized skill for each of the top 10 job titles, being mentioned in more than 85% of job postings for each title. Other skills with high frequencies across the top 10 job titles include "Computer Science" (eight titles), "Scrum" (eight titles), and various programming languages (five titles; "Java", "JavaScript", "Python", "C#", or "SQL"). Forty-eight distinct specialized skills made the top ten for at least one job title. Thirty-two of those skills were unique to one of the top 10 job titles.

There are interesting findings in the rankings for Project Managers. The top 5 are "Project Management", "Agile Methodology", "Scrum (Software Development)", "Waterfall Methodology", and "Systems Development Life Cycle". Project managers have been directly associated with traditional development for decades. So, the specialized skills of "Waterfall Methodology" and "Systems Development Life Cycle" are not surprising. "Agile Methodology" and "Scrum (Software Development)" run counter to the association with traditional development. The widespread use of hybrid development approaches (e.g., Gemino et al., 2021) may explain these findings.

"Scrum (Software Development)" is a specialized skill that ranked in the top ten job postings for 8 out of 10 job titles. Scrum is the most used agile development approach. According to Digital.ai (2022), 87% of organizations reported using Scrum methodology in 2022, up from 56% in 2020. This skill, unsurprisingly, ranks 1 for

Scrum Masters and Product Owners, as these are specific roles in Scrum.

The "User Story" skill is ranked in the top 10 for only 4 job titles: Business Analysts, Scrum Masters, Product Managers, and Product Owners. It did not rank in the top ten for the other job titles including Software Developers. This could indicate 1) a lack of focus on Scrum or 2) that "Scrum (Software Development)" and "Agile Methodology" are used as umbrella terms instead of indicating specific user story skills in those job postings.

Table 5 (see Appendix) displays the top common skills for each job title. As expected, common skills were more consistent across job titles, with 27 distinct common skills compared to the 48 distinct specialized skills. "Communication" was the top ranked common skill across all job titles. This indicates the continuing need for soft skills across job titles. Both "Problem Solving" and "Management" were ranked in the top 10 for all 10 of the job titles. "Planning", "Writing", "Leadership", and "Innovation" ranked in most of the job titles.

RQ3: How do the common and specialized skill requirements vary across job titles?

Table 6 (See Appendix) lists the specialized and common skills combined. Every job title had a mix of skills to make the top 10 skills overall, but the mix of skill types differed across the job titles. Data Engineers appear to have a great need for specialized skills because all the top 10 skills were specialized skills. Software Engineers, Software Developers, Java Developers, and Full Stack Developers had nine specialized skills in their top 10. Project Managers and Product Owners are the job titles with the highest number of common skills in their overall top 10.

5. DISCUSSION AND CONCLUSION

Information systems educators prepare their students for the IT workforce. This study provides insights to IS educators by 1) helping them guide their students during the job placement process, specifically when students express interest in agile, and 2) guiding curriculum design. This finding suggests that a wide variety of jobs require agile skills, and the word "agile" may not be in the job title.

In this study, we found the most common job titles associated with agile skills span across different types of jobs ranging from project management-related job titles (e.g., Project Managers, Scrum Masters) to

developer/engineer job titles (e.g., Full Stack Developers, Software Engineers). We also identified which skills are most associated with those job titles and found that skills such as "Agile Methodology" and "Communication" ranked in the top 10 skills for all the top 10 job titles. This suggests that, like most jobs, jobs requiring agile skills require communication skills. However, we found a mix of top skills across the top 10 job titles, indicating that agile skills can be beneficial in jobs with different mixes of skills.

This research offers several practical applications. First, students learning about agile may want to work in a job that uses agile skills. We found many instances where a job title doesn't have "agile" in it, but the job requires agile skills. This indicates that students should be attentive in reading the full job description when applying for jobs, and that students need to broaden their search parameters to find hidden attributes in job descriptions. This research, indicating job titles connected to those skills, helps students know what jobs are relevant in their job search. This information can help instructors and advisors guide students to those jobs.

Second, our data indicated that Waterfall-like skills were sought even though we were searching for agile. This finding indicates the use of a hybrid approach among employers. In a related study, more widespread use of the hybrid approach was found than expected (Gemino et al., 2021). About fifty percent of their sample used the hybrid approach, which the authors suggest "represents a maturing of the project management discipline" (Gemino et al., 2021, p. 170). Because agile approaches are somewhat newer, students may want to focus exclusively on developing agile related skills. It's imperative that faculty introduce the reality of hybrid development.

On the surface, using the best of both waterfall and agile seems advantageous. However, hybrid software development requires skills from both approaches. Further, incorporating the hybrid approach creates complexity due to the prevailing organizational culture and the current skills of IT professionals. Curriculum and professional development should include the hybrid approach to prepare students and current IT professionals for this organizational environment. This means going beyond teaching waterfall and agile development separately to incorporating the combination of both when developing software solutions for businesses.

Third, this information could guide academic programs to develop curriculum that includes agile content. For example, several of the job titles were related to software engineering. This may indicate that programming-related disciplines could benefit from including agile skills development.

As with all research, this research has limitations. One limitation of the study was that we only used U.S. job postings. However, the results may be different in other countries. The skills required for agile workers vary between the U.S. and Germany (Aldenhoven et al., 2021). In line with this difference between countries, it is possible that the most common job titles associated with agile skills and the skills associated with those job titles may vary in other countries. Future research could look at international job postings for more generalizability or job postings in countries other than the U.S. for comparison.

This study is also limited by our measures and data. We assume that job postings directly indicate the skill requirements for jobs, in line with prior research (Harris & Clark, 2023). However, job postings are not always a direct reflection of the jobs. For example, job posting web sites may limit the length of the text of the job postings. If the job posting is not limited by the job posting web site, they are limited by the attention spans of the applicants. As mentioned in the methods section, job postings also serve as a lagging indicator of skill needs and are, thus, not predictive of what skills might be required in the future.

Furthermore, we assume that the job posting is both written by the poster and received by potential applicants with no outside motivations and biases. However, job postings could be written to scare away potential applicants (Clearpoint, 2016). Alternatively, job postings could be written to persuade potential applicants to apply. For example, software developers who use agile methodology more tend to have more favorable perceptions about their jobs than software developers who use agile methodology less (Sun & Schmidt, 2018). Employers may use these more positive perceptions to attract more applicants by including agile in the job postings. Furthermore, employees generally perceive agile transformation—moving from traditional software development approaches to an agile software development approach—as a positive change for the organization, but agile transformation comes with pains including things like resistance (Nerur et al., 2005). Including

agile skills in a job posting could signal to potential applicants that the agile transformation has already occurred. Employees will then reap the benefits of the agile environment without dealing with the struggles of the process of changing.

Future research can address the limitation related to job postings being a true representation of the job requirements by gathering data using other methodologies. For example, researchers could conduct interviews with hiring managers. This approach would avoid the issues noted above that are related specifically to job postings data.

Our study is also limited by the timeframe of the data. We pulled job postings between July 2023 and June 2024. This year of data provides some valuable insights into jobs that require agile skills. However, future research could investigate the change over time in jobs that require agile skills. This longitudinal study would provide additional insights for IS research with regards to how the IT job market has changed.

6. REFERENCES

- Aldenhoven, C. M., Brosch, D. K., Prommegger, B., & Krcmar, H. (2021). Differences in Skill Requirements for Agile Workers in the German and the US IT Industries. In F. Ahlemann, R. Schütte, & S. Stieglitz (Eds.), *Innovation through Information Systems. Lecture Notes in Information Systems and Organisation* (Vol. 47, pp. 582-595). Springer, Cham.
https://doi.org/10.1007/978-3-030-86797-3_38
- Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., & Jeffries, R. (2001). *Manifesto for Agile Software Development*.
<http://agilemanifesto.org/>
- Booker, Q.E., Rebman Jr., C.M., Wimmer, H., Levkoff, S.B., Powell, L., Breese, J.L., (2024). Data Analytics Position Description Analysis: Skills Review and Implications for Data Analytics Curricula. *Information Systems Education Journal*, 22(3), pp.76-87.
<https://doi.org/10.62273/SSQL7979>
- Castro Lopes, F.; Fernandes, S. The Use of Gamification for Learning SCRUM: Findings from a Case Study with Information Systems Students. *Trends High. Educ.* 2024, 3, 235–246.
<https://doi.org/10.3390/higheredu3020014>
- Chumwatana T., & Hpone, A. K. K. (2025). Bridging the IT skill gap with industry demands: An AI-driven text mining approach to job market trends using large language model. *Journal of Theoretical and Applied Information Technology*, 103(6), 2270-2282.
<https://www.jatit.org/volumes/Vol103No6/12Vol103No6.pdf>
- Clearpoint. (2016, June 9). How a Job Description Can Scare Away Candidates.
<https://clearpointco.com/blog/2016/06/09/job-description-scare-away-candidates-employment-agencies-houston-tx/>
- Cornide-Reyes, H., Riquelme, F., Noel, R., Villarroel, R., Cechinel, C., & Letelier, P. (2021). Key Skills to Work with Agile Frameworks in Software Engineering: Chilean Perspectives. *IEEE Access*, 9, 84724-84738.
<https://doi.org/10.1109/ACCESS.2021.3087717>
- Dattero, R., Galup, S. D., Kan, A., & Quan, J. (2016). It Pays to be Agile. *Journal of Computer Information Systems*, 57(3), 252-257.
<https://doi.org/10.1080/14659891.2016.1184025>
- Digital.ai. (2022). State of Agile Report.
<https://digital.ai/resource-center/analyst-reports/16th-state-of-agile-report/>
- Fernandez, D. J., & Fernandez, J. D. (2008). Agile Project Management — Agilism Versus Traditional Approaches. *Journal of Computer Information Systems*, 49(2), 10-17.
<https://doi.org/10.1080/08874417.2009.11646044>
- Flynn, J. (2022, November 27). 16 Amazing Agile Statistics [2023]: What Companies Use Agile Methodology. Zippia.
<https://www.zippia.com/advice/agile-statistics/>
- Gemino, A., Reich, B. H., & Serrador, P. M. (2021). Agile, Traditional, and Hybrid Approaches to Project Success: Is Hybrid a Poor Second Choice? *Project Management Journal*, 52(2), 161–175.
<https://doi.org/10.1177/8756972820973082>

- Gunklach, J., Nadj, M., Michalczyk, S. et al. Beyond the Unicorn? Job Roles in Data Science. *Business Information Systems Engineering* (2025). <https://doi.org/10.1007/s12599-025-00954-2>
- Harris, A., & Clark, C. (2023). Understanding the Diffusion of Business Analysis Responsibilities. *Journal of Computer Information Systems*, 64(2), 278-288. <https://doi.org/10.1080/08874417.2023.2193964>
- Highsmith, J. (2001). History: The Agile Manifesto. <https://agilemanifesto.org/history.html>
- Huck-Fries, V., Nothaft, F., Wiesche, M., & Krcmar, H. (2023). Job Satisfaction in Agile Information Systems Development: A Stakeholder Perspective. *Information and Software Technology*, 163, 107289. <https://doi.org/10.1016/j.infsof.2023.107289>
- International Institute of Business Analysis (IIBA). (2024). 2024 Global State of Business Analysis Report, Part 2: Certification and Career Development. <https://www.iiba.org/career-resources/the-global-state-of-business-analysis/the-global-state-of-business-analysis-reports-and-surveys/the-global-state-of-business-analysis-2023/#part2>
- Joshua, E., Zarefard, M., & Marsden, N. (2023, November). Investigating Skill Requirements and Gender Bias in Job Openings for Human Computer Interaction Professionals Across the USA, Australia, Germany, India, and South Africa. In *2023 9th International HCI and UX Conference in Indonesia (CHIuXiD)* (pp. 71-76). IEEE. <https://doi.org/10.1109/CHIuXiD59550.2023.10452730>
- Lightcast. (2025a). Lightcast Data: Basic Overview. <https://kb.lightcast.io/en/articles/6957498-lightcast-data-basic-overview>
- Lightcast. (2025b). Job Posting Analytics (JPA) Methodology. <https://kb.lightcast.io/en/articles/6957446-job-posting-analytics-jpa-methodology>
- Lightcast. (2025c). Skills. <https://kb.lightcast.io/en/articles/7934140-skills>
- Marnewick, C. (2023). Student experiences of project-based learning in agile project management education. *Project Leadership and Society*, 4, 100096. <https://doi.org/10.1016/j.plas.2023.100096>
- Nerur, S., Mahapatra, R., & Mangalaraj, G. (2005). Challenges of Migrating to Agile Methodologies. *Communications of the ACM*, 48(5), 72-78. <https://doi.org/10.1145/1060710.1060712>
- Rimol, M. (2022, March 3). 12 Essential Skills for Agile Developers. Gartner. <https://www.gartner.com/en/articles/12-essential-skills-for-agile-developers>
- Sharp, J. H., & Lang, G. (2018). Agile in Teaching and Learning: Conceptual Framework and Research Agenda. *Journal of Information Systems Education*, 29(2), 45-52. <https://aisel.aisnet.org/jise/vol29/iss2/1>
- Sun, W., & Schmidt, C. (2018). Practitioners' Agile-Methodology Use and Job Perceptions. *IEEE Software*, 35(2), 52-61. <https://doi.org/10.1109/MS.2018.1661333>
- Siswipraptini, P. C., Warnars, H. L. H. S., Ramadhan, A., & Budiharto, W. (2023). Information technology job profile using average-linkage hierarchical clustering analysis. *IEEE Access*, 11, 94647-94663. <https://doi.org/10.1109/ACCESS.2023.3311203>
- Woods, D., Hulshult, A., (2024). The Agile Student Practice Project: Simulating an Agile Project in the Classroom for a Real-World Experience. *Information Systems Education Journal*, 22(2), pp.70-81. <https://doi.org/10.62273/VAPJ1256>

Appendix

| Skill Rank | Software Engineers | Project Managers | Business Analysts | Scrum Masters | Product Managers | Software Developers | Product Owners | Data Engineers | Java Developers | Full Stack Developers |
|------------|--|--|---------------------------------------|---|---------------------------------------|--|---------------------------------------|--|--|--|
| 1 | Agile Methodology (91.31%) | Project Management (99.68%) | Agile Methodology (92.70%) | Scrum (Software Development) (100.00%) | Product Management (98.22%) | Agile Methodology (92.50%) | Scrum (Software Development) (99.82%) | Data Engineering (98.84%) | Java (Programming Language) (99.88%) | Full Stack Development (99.45%) |
| 2 | Software Engineering (82.53%) | Agile Methodology (87.10%) | Business Requirements (51.04%) | Agile Methodology (98.88%) | Agile Methodology (87.63%) | Computer Science (57.80%) | Agile Methodology (94.06%) | Agile Methodology (93.99%) | Agile Methodology (93.27%) | Agile Methodology (95.11%) |
| 3 | Software Development (64.41%) | Scrum (Software Development) (37.10%) | Business Process (45.73%) | Sprint Retrospectives (51.20%) | New Product Development (43.60%) | JavaScript (Programming Language) (52.12%) | User Story (64.17%) | SQL (Programming Language) (77.82%) | Spring Framework (52.32%) | JavaScript (Programming Language) (64.15%) |
| 4 | Computer Science (58.89%) | Waterfall Methodology (32.04%) | User Story (45.64%) | Sprint Planning (47.98%) | Product Roadmaps (41.77%) | Software Development (51.25%) | Product Management (53.07%) | Python (Programming Language) (73.96%) | Spring Boot (48.78%) | Angular (Web Framework) (55.28%) |
| 5 | Java (Programming Language) (40.73%) | Systems Development Life Cycle (26.96%) | Scrum (Software Development) (44.16%) | JIRA (45.72%) | Scrum (Software Development) (40.85%) | SQL (Programming Language) (40.41%) | New Product Development (31.60%) | Extract Transform Load (ETL) (64.30%) | Computer Science (44.45%) | Application Programming Interface (API) (48.80%) |
| 6 | Amazon Web Services (36.19%) | JIRA (24.71%) | Business Analysis (41.68%) | Continuous Improvement Process (39.79%) | Product Strategy (33.67%) | C# (Programming Language) (39.23%) | Product Roadmaps (30.96%) | Data Warehousing (54.71%) | SQL (Programming Language) (42.08%) | React.js (Javascript Library) (46.72%) |
| 7 | JavaScript (Programming Language) (35.41%) | Milestones (Project Management) (24.42%) | Project Management (41.52%) | Project Management (39.77%) | User Story (31.89%) | Java (Programming Language) (38.95%) | Project Management (30.48%) | Computer Science (53.90%) | Microservices (39.46%) | Java (Programming Language) (46.44%) |
| 8 | Python (Programming Language) (33.98%) | Project Planning (24.21%) | JIRA (35.33%) | Kanban Principles (26.76%) | Marketing (30.09%) | Angular (Web Framework) (32.37%) | JIRA (27.29%) | Data Pipelines (48.62%) | RESTful API (38.61%) | Amazon Web Services (44.81%) |
| 9 | Scrum (Software Development) (31.86%) | Project Scoping (21.94%) | Computer Science (26.08%) | User Story (26.55%) | Computer Science (29.38%) | Scrum (Software Development) (30.73%) | Computer Science (26.33%) | Amazon Web Services (46.01%) | Scrum (Software Development) (35.40%) | Computer Science (42.64%) |
| 10 | SQL (Programming Language) (31.62%) | Microsoft Project (21.14%) | Workflow Management (24.64%) | Software Development (25.28%) | Project Management (27.33%) | Git (Version Control System) (29.37%) | User Experience (UX) (24.54%) | Data Modeling (43.60%) | JavaScript (Programming Language) (33.65%) | SQL (Programming Language) (40.81%) |

Table 4: Top ranking specialized skills by frequency of appearance in agile job postings.

| Skill Rank | Software Engineers | Project Managers | Business Analysts | Scrum Masters | Product Managers | Software Developers | Product Owners | Data Engineers | Java Developers | Full Stack Developers |
|------------|--|-----------------------------|-------------------------------|---------------------------------|-----------------------------|--|---------------------------|--|--|--|
| 1 | Communication (38.53%) | Communication (67.10%) | Communication (57.84%) | Communication (61.82%) | Communication (57.64%) | Communication (33.89%) | Communication (60.57%) | Communication (40.98%) | Communication (36.80%) | Communication (48.12%) |
| 2 | Problem Solving (28.75%) | Management (56.99%) | Problem Solving (38.29%) | Planning (46.91%) | Leadership (41.97%) | Troubleshooting (Problem Solving) (24.53%) | Leadership (42.53%) | Problem Solving (28.47%) | Problem Solving (26.87%) | Problem Solving (29.73%) |
| 3 | Troubleshooting (Problem Solving) (23.98%) | Leadership (50.40%) | Management (33.86%) | Leadership (41.49%) | Prioritization (34.31%) | Problem Solving (23.28%) | Planning (38.59%) | Management (26.86%) | Troubleshooting (Problem Solving) (21.21%) | Troubleshooting (Problem Solving) (28.52%) |
| 4 | Management (22.63%) | Planning (42.69%) | Writing (23.61%) | Coaching (38.47%) | Customer Service (34.11%) | Management (20.04%) | Prioritization (38.45%) | Operations (18.83%) | Management (13.21%) | Management (20.00%) |
| 5 | Innovation (19.89%) | Problem Solving (30.84%) | Planning (20.99%) | Management (37.34%) | Problem Solving (32.06%) | Information Technology (15.19%) | Management (36.10%) | Troubleshooting (Problem Solving) (18.70%) | Information Technology (10.12%) | Operations (14.73%) |
| 6 | Planning (16.90%) | Coordinating (27.09%) | Detail Oriented (20.18%) | Problem Solving (32.65%) | Management (30.48%) | Writing (14.64%) | Problem Solving (28.91%) | Innovation (18.00%) | Writing (9.72%) | Innovation (14.09%) |
| 7 | Operations (15.93%) | Timelines (26.57%) | Microsoft Excel (19.60%) | Prioritization (26.65%) | Influencing Skills (29.31%) | Planning (13.50%) | Customer Service (25.91%) | Leadership (15.94%) | Self-Motivation (9.02%) | Writing (13.39%) |
| 8 | Writing (15.62%) | Presentations (20.78%) | Leadership (18.20%) | Decision Making (24.01%) | Innovation (27.69%) | Innovation (12.92%) | Writing (23.87%) | Mentorship (15.82%) | Innovation (8.99%) | Information Technology (11.16%) |
| 9 | Leadership (14.82%) | Program Management (19.80%) | Presentations (17.66%) | Mentorship (22.56%) | Planning (27.49%) | Detail Oriented (12.13%) | Innovation (20.71%) | Customer Service (12.74%) | Planning (7.83%) | Leadership (10.67%) |
| 10 | Mathematics (13.35%) | Writing (19.80%) | Microsoft PowerPoint (16.39%) | Information Technology (20.92%) | Sales (26.10%) | Operations (11.93%) | Research (19.87%) | Writing (12.57%) | Operations (7.67%) | Detail Oriented (9.77%) |

Table 5: Top ranking common skills by frequency of appearance in agile job postings.

| Skill Rank | Software Engineers | Project Managers | Business Analysts | Scrum Masters | Product Managers | Software Developers | Product Owners | Data Engineers | Java Developers | Full Stack Developers |
|------------|--|---------------------------------------|---------------------------------------|---|---------------------------------------|--|---------------------------------------|--|---------------------------------------|--|
| 1 | Agile Methodology (91.31%) | Project Management (99.68%) | Agile Methodology (92.70%) | Scrum (Software Development) (100.00%) | Product Management (98.22%) | Agile Methodology (92.50%) | Scrum (Software Development) (99.82%) | Data Engineering (98.84%) | Java (Programming Language) (99.88%) | Full Stack Development (99.45%) |
| 2 | Software Engineering (82.53%) | Agile Methodology (87.10%) | Communication (57.84%) | Agile Methodology (98.88%) | Agile Methodology (87.63%) | Computer Science (57.80%) | Agile Methodology (94.06%) | Agile Methodology (93.99%) | Agile Methodology (93.27%) | Agile Methodology (95.11%) |
| 3 | Software Development (64.41%) | Communication (67.10%) | Business Requirements (51.04%) | Communication (61.82%) | Communication (57.64%) | JavaScript (Programming Language) (52.12%) | User Story (64.17%) | SQL (Programming Language) (77.82%) | Spring Framework (52.32%) | JavaScript (Programming Language) (64.15%) |
| 4 | Computer Science (58.89%) | Management (56.99%) | Business Process (45.73%) | Sprint Retrospectives (51.20%) | New Product Development (43.60%) | Software Development (51.25%) | Communication (60.57%) | Python (Programming Language) (73.96%) | Spring Boot (48.78%) | Angular (Web Framework) (55.28%) |
| 5 | Java (Programming Language) (40.73%) | Leadership (50.40%) | User Story (45.64%) | Sprint Planning (47.98%) | Leadership (41.97%) | SQL (Programming Language) (40.41%) | Product Management (53.07%) | Extract Transform Load (ETL) (64.30%) | Computer Science (44.45%) | Application Programming Interface (API) (48.80%) |
| 6 | Communication (38.53%) | Planning (42.69%) | Scrum (Software Development) (44.16%) | Planning (46.91%) | Product Roadmaps (41.77%) | C# (Programming Language) (39.23%) | Leadership (42.53%) | Data Warehousing (54.71%) | SQL (Programming Language) (42.08%) | Communication (48.12%) |
| 7 | Amazon Web Services (36.19%) | Scrum (Software Development) (37.10%) | Business Analysis (41.68%) | JIRA (45.72%) | Scrum (Software Development) (40.85%) | Java (Programming Language) (38.95%) | Planning (38.59%) | Computer Science (53.90%) | Microservices (39.46%) | React.js (Javascript Library) (46.72%) |
| 8 | JavaScript (Programming Language) (35.41%) | Waterfall Methodology (32.04%) | Project Management (41.52%) | Leadership (41.49%) | Prioritization (34.31%) | Communication (33.89%) | Prioritization (38.45%) | Data Pipelines (48.62%) | RESTful API (38.61%) | Java (Programming Language) (46.44%) |
| 9 | Python (Programming Language) (33.98%) | Problem Solving (30.84%) | Problem Solving (38.29%) | Continuous Improvement Process (39.79%) | Customer Service (34.11%) | Angular (Web Framework) (32.37%) | Management (36.10%) | Amazon Web Services (46.01%) | Communication (36.80%) | Amazon Web Services (44.81%) |
| 10 | Scrum (Software Development) (31.86%) | Coordinating (27.09%) | JIRA (35.33%) | Project Management (39.77%) | Product Strategy (33.67%) | Scrum (Software Development) (30.73%) | New Product Development (31.60%) | Data Modeling (43.60%) | Scrum (Software Development) (35.40%) | Computer Science (42.64%) |

Note: Specialized skills are shaded gray.

Table 6: Top ranking combined skills (specialized and common) by frequency of appearance in agile job postings.